

UNIVERSIDAD DE GRANADA
DEPARTAMENTO DE GENÉTICA
COMPUTATIONAL GENOMICS AND BIOINFORMATICS



**REGIONES GENÓMICAS IMPLICADAS EN LA
METILACIÓN DIFERENCIAL DEL ADN**

Memoria de Tesis Doctoral

Guillermo Barturen Briñas

Granada 2014

Editor: Editorial de la Universidad de Granada
Autor: Guillermo Barturen Briñas
D.L.: GR 2140-2014
ISBN: 978-84-9083-159-5

Diseño de cubierta

© Eduard Marimon Beceril

emarimon@gmail.com

Los Drs. José L. Oliver Jiménez y Michael Hackenberg, como directores de la Tesis Doctoral que presenta el Licenciado Guillermo Barturen Briñas,

CERTIFICAN

que los trabajos desarrollados e incluidos en la presente memoria: “Regiones genómicas implicadas en la metilación diferencial del ADN”, son aptos para ser presentados y aspirar al Grado de Doctor en Biología por la Universidad de Granada.

Fdo.: Dr. José L. Oliver Jiménez

Fdo.: Dr. Michael Hackenberg

El doctorando Guillermo Barturen Briñas y los directores de la tesis Dr. José L. Oliver Jiménez y Dr. Michael Hackenberg garantizamos, al firmar esta Tesis Doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la Tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada, __ de abril de 2014

Los Directores de Tesis

Fdo.: Dr. José L. Oliver Jiménez

Fdo.: Dr. Michael Hackenberg

El Doctorando

Fdo.: Guillermo Barturen Briñas

Durante el desarrollo de esta Tesis Doctoral, el doctorando Guillermo Barturen Briñas ha disfrutado de una Beca AE del Programa de Ayudas para Formación del Personal Investigador del Departamento de Educación, Universidades e Investigación del Gobierno Vasco.

Este trabajo estuvo financiado por los proyectos de investigación BIO2008-01353 y BIO2010-20219 concedidos por el Gobierno Español, así como por los proyectos P06-FQM1858 y P07-FQM3163 concedidos por la Junta de Andalucía.

La investigación incluida en esta Tesis Doctoral se realizó en el Departamento de Genética de la Universidad de Granada.

A los que me han enseñado a amar la ciencia,
y a los que me han soportado en el proceso.

"There's real poetry in the real world.

Science is the poetry of reality."

Richard Dawkins

AGRADECIMIENTOS

Aunque a veces no lo parezca, todas las cosas tienen su final y, a pesar de que por miedo o inseguridad he postergado consciente o inconscientemente la redacción de esta memoria, aquí me encuentro redactando las primeras líneas del que será el último capítulo de esta Tesis Doctoral.

No recuerdo cuándo ni por qué decidí dedicarme a la investigación, pero actualmente no puedo imaginarme dedicándome a otra cosa. Esta pasión y la posibilidad de dedicarme a ella se la debo en gran medida a José L. Oliver por la oportunidad que me brindó; supongo que esta, es una deuda de esas que nunca se saldan. No existen palabras para agradecer a José L. Oliver y a Michael Hackenberg las innumerables horas dedicadas en mi formación y su inagotable paciencia durante estos años; sin ellos, esta Tesis no hubiera sido posible. Gracias Pepe. Gracias Mic.

Agradezco también a los colaboradores con los que he tenido la suerte de trabajar, ya que todos han contribuido de alguna manera a esta Tesis. En especial, agradecer a Steffi su contagiosa capacidad de trabajo, y a Ángel por mantener vivos los servidores contra los que tantas veces he atentado.

AGRADECIMIENTOS

Durante estos años me he sentido orgulloso de pertenecer al Departamento de Genética de la Universidad de Granada, donde las cosas se hacen bien, porque no parece que exista otra manera de hacerlas. En especial, mi agradecimiento a Esther Viseras por convertir la burocracia en mero trámite y por sus inestimables consejos para mejorar mi capacidad docente.

Como una Tesis no sólo se compone de trabajo y esfuerzo, gracias a todos los amigos con los que he compartido estos años en Granada, porque vosotros sois Granada y en Granada dejé de ser lo que era para empezar a ser lo soy. En especial a Loli, porque más que una amiga ha sido una madre. Y a Edu por no permitirme envolver la Tesis en una chapuza que no merecía ser llamada cubierta.

La deuda contraída con vosotros es otra de las que nunca se saldan. Por vuestro amor incondicional, por enseñarme a vivir y por confiar en mis decisiones. Gracias papá. Gracias mamá.

Los versos más sinceros son los que se escriben sin palabras y lo míos son todos para ti. Si sólo pudiera dedicar esta Tesis a una persona, esta Tesis sería para ti. Por tu cariño, por tu apoyo y por tu comprensión. Gracias Katu.

ABREVIATURAS Y GLOSARIO

Adaptador: Secuencia de ADN añadida a los extremos de los fragmentos a secuenciar en los protocolos de **secuenciación masiva**. Los adaptadores permiten anclar los fragmentos a secuenciar a una superficie sólida y definen el sitio donde comenzará la secuenciación.

Bisulfito: El **bisulfito** sódico es un compuesto químico utilizado en genética molecular para determinar el **estado de metilación** de las citosinas.

BSC: Secuencia reversa complementaria de la **secuencia de referencia** donde todas sus citosinas han sido sustituidas por timinas.

BSCRC: Secuencia reversa complementaria de **BSC**.

BSW: **Secuencia de referencia** donde todas sus citosinas han sido sustituidas por timinas.

BSWRC: **Secuencia** reversa complementaria de **BSW**.

BS-Seq: Protocolo de **secuenciación masiva** basado en el tratamiento previo del ADN con **bisulfito**. Incluye dos procesos de amplificación mediante *PCR*, resultando en secuencias que pueden proceder tanto de **BSW** y **BSC**, como de sus reversas complementarias (**BSWRC** y **BSCRC**).

Cluster: Conjunto de elementos que aparecen agrupados.

Cobertura: Cuando se refiere a la **secuenciación masiva**, fracción de posiciones con datos de resecuenciación. También se suele utilizar para referirse al promedio de veces que ha sido secuenciada cada posición.

Contenido en GC (G+C): fracción de guanina y citosina en una secuencia de ADN.

Contexto de metilación: Secuencia de ADN que contiene citosinas y por lo tanto puede presentar **metilación**. Habitualmente, los **contextos de metilación** son CpG, CpHpG y CpHpH (donde H hace referencia a cualquier nucleótido que no sea una citosina).

Curva ROC (Característica Operativa del Receptor): Representación gráfica de la **sensibilidad** frente a **1-especificidad**.

Densidad de CpGs: Fracción de dinucleótidos CpG en una secuencia de ADN, calculada como el número observado de CpGs dividido por la longitud total de la secuencia. Se suele multiplicar por 1000, para expresarla como número de CpGs por kilobase.

Desambiguación (de multilecturas): Selección de uno de los posibles alineamientos encontrados para una **lectura**.

Desaminación: Reacción química caracterizada por la ruptura del enlace de un grupo amino. El **bisulfito** induce esta reacción en citosinas no metiladas, permitiendo determinar *a posteriori* su **estado de metilación**.

DMIs: Islas CpG que presentan **metilación diferencial** al menos entre un par de tejidos.

DMIs-M: Islas CpG que se encuentran **metiladas** en la mayor parte de los tejidos y no **metiladas** en unos pocos.

DMIs-U: Islas CpG que se encuentran no **metiladas** en la mayor parte de los tejidos y **metiladas** en unos pocos.

DMRs (Regiones Diferencialmente Metiladas): Regiones del genoma caracterizadas por presentar diferencias de **metilación** entre tejidos o condiciones.

DNMTs (Metiltransferasas del ADN): Enzimas responsables de la adición de grupos metilo a los nucleótidos del ADN.

Ensamblado: Genoma del que se conoce su secuencia nucleotídica. Generalmente, el término es utilizado para referirse a los conocidos como **ensamblados o secuencias de referencia**, secuencias consenso de todo el genoma a partir de varios individuos de una misma especie. Por ejemplo: hg19, versión 19 del **ensamblado** consenso del genoma humano; panTro4, versión 4 de *Pan troglodytes*; rheMac3, versión 3 de *Macaca mulatta*; mm10, versión 10 de *Mus musculus*; tair10, versión 10 de *Arabidopsis thaliana* o itag2.4, versión 2.4 de *Solanum lycopersicum*.

Especificidad (Sp): Es la capacidad de un método para descartar correctamente una condición determinada.

Estado de metilación: clasificación discreta de los **niveles de metilación**.

FDR (“False Discovery Rate”): Método estadístico utilizado en análisis de hipótesis para corregir comparaciones múltiples; se utiliza para

controlar la proporción esperada de hipótesis nulas rechazadas incorrectamente.

FN (“False negatives”): Número de casos descartados incorrectamente.

FP (“False positives”): Número de casos identificados incorrectamente.

Genes domésticos o constitutivos (“Housekeeping genes”): Genes expresados en todos o la mayor parte de los tipos celulares, cuyos transcritos y/o proteínas son imprescindibles para la correcta función celular.

Genes tejido-específicos (“Tissue-specific genes”): Genes expresados en un número reducido de tipos celulares, cuyos transcritos y/o proteínas desempeñan funciones específicas.

Índice de enriquecimiento: Mide la asociación entre dos elementos genómicos. Este índice se calcula como la fracción del primero de los elementos dentro y fuera del segundo.

Isla CpG (CGI): Región del genoma que presenta una elevada **densidad de CpGs** en comparación con el resto del genoma.

k-mero (Palabra de ADN): secuencia de ADN con una longitud k y una composición determinada.

Lecturas (“Reads”): Secuencias, generalmente cortas, producidas por los métodos de **secuenciación**.

Lecturas basadas en código de color: Lecturas generadas por los secuenciadores de *SOLiD Applied Biosystems*, donde cada posición,

salvo la primera, está compuesta por un número (0-3) que representa la relación entre el nucleótido secuenciado en dicha posición y el anterior a este. Para determinar esta relación cada nucleótido se secuencia dos veces.

Lecturas basadas en secuencia: Tipo de **lecturas** generadas por la mayoría de los secuenciadores, donde cada posición muestra el nucleótido secuenciado.

Librería de lecturas: Conjunto de fragmentos de ADN junto con sus **adaptadores** que serán secuenciados mediante técnicas de **secuenciación masiva** para una muestra.

Mapear o Alinear: Localización de las **lecturas** en la **secuencia de referencia**.

Matriz de puntuación: Sistema de puntuación utilizado para inferir el alineamiento más probable.

MDRs (Regiones Determinantes de la Metilación): Regiones cortas de ADN con una alta **densidad de CpGs** y asociadas a **sitios de unión a factores de transcripción**. Autónomamente en *cis* mantienen hipometilados los promotores en células madre y pueden proteger de la **metilación de novo** durante la diferenciación.

MethylC-Seq: Protocolo de **secuenciación masiva** basado en el tratamiento previo del ADN con **bisulfito**. Incluye un solo proceso de amplificación mediante *PCR*, resultando en secuencias que proceden de **BSW** y **BSC**.

Metilación del ADN: Proceso bioquímico, mediado por **ADN metiltransferasas**, por el que un grupo metilo se une covalentemente a los nucleótidos del ADN.

Metilación diferencial: Cambio en los **niveles o estados de metilación** en regiones o posiciones concretas entre diferentes tipos celulares o condiciones.

Metilcitosina: Citosina que presenta un grupo metilo unido covalentemente al carbono 5 de su anillo aromático.

Metiloma (o Mapa de metilación): Conjunto de **niveles o estados de metilación** en el genoma completo de una muestra.

Metiloma de alta resolución: Conjunto de **niveles de metilación** de las citosinas individuales en el genoma completo de una muestra.

MIs: **Islas CpG** constitutivamente **metiladas** en todos los tejidos.

Multilecturas: **Lecturas** provenientes de experimentos de **secuenciación masiva** que **alinean** en varias localizaciones genómicas.

Navegador genómico: Herramienta informática que permite visualizar y navegar gráficamente por el genoma de una especie, además de representar múltiples anotaciones conjuntamente.

Nivel de metilación: Valor numérico de **metilación** en **contextos** o citosinas individuales. Varía entre 0 y 1 (no **metilado** y **metilado**).

Paralelización o Multiprocesado (“Multi-threading”): Habilidad de un programa para ejecutarse en varios hilos (*threads*) y realizar así múltiples tareas de manera simultánea.

Perl: Lenguaje de programación de alto nivel creado en 1987 por Larry Wall. Es especialmente adecuado para procesar textos o secuencias.

Pista (“Track”): Anotación genómica representada en un **navegador genómico**.

Profundidad (“Depth”): En **secuenciación masiva**, hace referencia al número de veces que ha sido re-secuenciada una posición concreta.

R13 (Región promotora): Región en torno al inicio de la transcripción (*TSS*), definida como 1,500 pb aguas arriba y 500 pb aguas abajo de este.

R8: Región en torno al sitio de finalización de la transcripción (*TES*), definida como 500 pb aguas arriba y 1,500 pb aguas abajo de este.

Ratio CpG [O/E]: Relación entre el número de CpGs observados y esperados en una región determinada. El número esperado se calcula a partir de las frecuencias observadas de citosinas y guaninas en dicha región.

Secuenciación direccional: Tipo de **secuenciación masiva** previo tratamiento del ADN con **bisulfito** realizado por el protocolo **MethylC-Seq**.

Secuenciación masiva (“High-Throughput Sequencing” o “Next-Generation Sequencing”): Técnica de secuenciación que permite la

secuenciación de múltiples fragmentos de ADN de manera simultánea. La **secuenciación masiva** se basa en la inmovilización de las muestras de ADN en un soporte sólido, reacciones cíclicas de secuenciación utilizando dispositivos automatizados y la detección de ciertos eventos moleculares.

Secuenciación no-direccional: Tipo de **secuenciación masiva** previo tratamiento del ADN con **bisulfito** realizado por el protocolo **BS-Seq**.

Secuencias *pair-end* (emparejadas): **Lecturas** secuenciadas a partir de ambos extremos de un fragmento de ADN.

Secuencias *single-end* (simples): **Lecturas** secuenciadas a partir de uno de los extremos de un fragmento de ADN.

Semilla (del alineamiento): Región del extremo 5' de las **lecturas** (parte con mayor calidad de secuenciación) que se alinea con el **genoma de referencia**.

Sensibilidad (*Sn*): Es la capacidad de un método para detectar correctamente una condición determinada.

SNPs (Polimorfismos de un solo nucleótido): Son variaciones en la secuencia de ADN de un solo nucleótido, cuya frecuencia en la población es mayor al 1%.

SNVs (Variaciones de un solo nucleótido): Son variaciones de un solo nucleótido frente al **genoma de referencia**, cuya frecuencia poblacional no se conoce.

Términos GO (“Gene Ontology”): Son términos organizados de manera jerárquica que definen las funciones, procesos o localizaciones de los productos génicos (<http://www.geneontology.org/>).

TFs (Factores de Transcripción): Son proteínas que regulan el proceso de transcripción del ADN.

TFBSs (Sitios de Unión a Factores de Transcripción): Regiones del ADN donde los **factores de transcripción** interaccionan con la secuencia nucleotídica.

TN (“True Negatives”): Número de casos descartados correctamente.

TP (“True Positives”): Número de casos identificados correctamente.

Uls: Islas CpG constitutivamente no **metiladas** en todos los tejidos.

Valor-p: En estadística, el valor- p es la probabilidad de encontrar un resultado al menos igual de extremo que el observado.

Valor de calidad PHRED (Q): Es la probabilidad de que el nucleótido obtenido esté correctamente secuenciado, expresada como $-10 * \log_{10}P$ (donde P es la probabilidad de error de la secuenciación).

Valor de predicción positiva (PPV): Es la proporción de situaciones correctamente identificadas como pertenecientes a una condición.

Ventana móvil (“Sliding-window”): En la predicción de **islas CpG** hace referencia a un segmento de longitud fija donde se analizan características de la secuencia de ADN, y se desplaza de manera secuencial por el genoma mediante pasos de una longitud fija.

RESUMEN

La metilación del ADN es la marca epigenética por excelencia, pudiendo intervenir tanto en la transcripción de los genes como en el mantenimiento de la estabilidad del genoma. Consiste en la unión covalente de un grupo metilo al carbono 5 de las citosinas, dando lugar a metilcitosinas, descritas por algunos autores como “la quinta base del ADN”.

En mamíferos, esta marca epigenética se encuentra sobre todo en los dinucleótidos CpG, que aparecen metilados en la mayor parte del genoma. La excepción son las islas CpG (*CG/s*), regiones abiertas de la cromatina, con una elevada densidad de CpGs y libres de metilación, lo que puede permitir la interacción con el ADN. Aproximadamente el 70% de los genes humanos presentan *CG/s*, entre los que figuran la totalidad de los genes domésticos y un 40% de los genes tejido específicos.

La metilación puede variar entre diferentes tejidos, individuos o tipos celulares. El mecanismo subyacente es la regulación de la interacción de diferentes proteínas con el ADN. El objetivo principal de esta Tesis es la identificación de regiones diferencialmente metiladas, que puedan servir como marcadores epigenéticos. Ello exige como requisito imprescindible disponer de mapas de metilación en genoma completo (metilomas) de alta calidad y procedentes de diferentes tejidos. Solo así se puede llevar

a cabo el estudio comparado de los mismos e identificar las regiones diferencialmente metiladas.

Por tanto, el primer paso fue el desarrollo de las herramientas bioinformáticas necesarias para la generación de metilomas de alta calidad a partir de datos de secuenciación masiva de ADN tratado con bisulfito: *NGSmethPipe* y *MethylExtract*. El primer programa permite preprocesar y alinear las lecturas cortas frente al genoma de referencia, mientras que el segundo infiere los niveles de metilación de citosinas individuales y las variaciones de un solo nucleótido (*SNVs*) a partir de dichos alineamientos. Ambos programas incorporan rigurosos controles de calidad, de forma que los metilomas generados para diferentes tejidos, especies o estados patológicos sean comparables. Asimismo, fue necesario el desarrollo de una base de datos relacional (*NGSmethDB*) para el almacenamiento, gestión y explotación de toda la información generada. Actualmente, *NGSmethDB* contiene metilomas para 6 especies y 114 tejidos y/o condiciones diferentes, para lo que ha sido necesario procesar 40 terabytes de lecturas.

Se sabe que la densidad de CpGs juega un papel crítico tanto en la unión de los factores de transcripción (*TFs*) como en la determinación del estado de metilación del ADN. Sin embargo, ni las herramientas existentes ni los estudios realizados en múltiples tejidos tienen en cuenta esta importante característica. Así pues, otro objetivo de esta Tesis ha sido la mejora del algoritmo *CpGcluster*, dando lugar a un nuevo algoritmo (*WordCluster*), capaz de identificar islas CpG (*CGIs*) estadísticamente significativas y con una alta densidad de CpGs. *WordCluster* presenta importantes ventajas con respecto a los métodos

clásicos de identificación de *CGIs*, destacando entre ellas la delimitación de dominios más cortos y homogéneos, pero estadísticamente significativos, y el que sus predicciones se asocian de manera más específica tanto con elementos reguladores, como con regiones conservadas del genoma.

El estudio estadístico de la metilación diferencial en las *CGIs* predichas por *WordCluster* ha permitido establecer que el uso combinado de la prueba exacta de Fisher y la binomial negativa es el mejor método para identificar *DMIs* (islas CpG diferencialmente metiladas). La mayoría de las *DMIs* pueden ser de dos tipos: *DMIs-M* (metiladas en la mayoría de los tejidos y no-metiladas en algún tejido) y *DMIs-U* (metiladas solamente en algún tejido y no-metiladas en el resto). El análisis funcional mediante el enriquecimiento en términos GO (*Gene Ontology*) muestra que las *DMIs-U* parecen estar implicadas en procesos de desarrollo y diferenciación, mientras que las *DMIs-M* se asocian con funciones tejido-específicas.

Por último, los estudios de enriquecimiento de las *DMIs* en elementos reguladores han revelado importantes diferencias con respecto a otras *DMRs* (Regiones diferencialmente metiladas), recientemente identificadas sin tener en cuenta la densidad de CpGs. Entre estas diferencias destacan el elevado enriquecimiento de las *DMIs* en promotores y exones y su empobrecimiento en intrones, así como la significativa menor proporción de *DMIs-M* asociadas con *TFBSs* cuando se las compara con las *DMRs*. Cabe destacar también que las *DMIs-U* solapan en mucho mayor grado con los *TFBSs* que las *DMIs-M*. Todas estas importantes características encontradas en las *DMIs* sugieren que

WordCluster puede ser el algoritmo adecuado para preseleccionar las regiones a analizar en futuros estudios de metilación diferencial.

ÍNDICE

| | |
|--|------------|
| ABREVIATURAS Y GLOSARIO | xv |
| RESUMEN | xxv |
| 1 INTRODUCCIÓN | 37 |
| 1.1 METILACIÓN DEL ADN | 39 |
| 1.2 ISLAS CpG | 41 |
| 1.2.1 Dinámica evolutiva de los dinucleótidos CpG | 42 |
| 1.3 METILACIÓN DIFERENCIAL | 43 |
| 2 OBJETIVOS | 49 |
| 3 PERFILES DE METILACIÓN DE ALTA CALIDAD | 53 |
| 3.1 MÉTODOS PARA DETERMINAR EL ESTADO DE METILACIÓN DEL ADN .. | 54 |
| 3.1.1 Metodologías previas al tratamiento con bisulfito | 55 |
| 3.1.2 Secuenciación masiva combinada con el tratamiento con bisulfito | 55 |
| 3.1.3 Herramientas bioinformáticas | 57 |
| 3.2 NGSmethPipe (DNA methylation Profiling from High-Throughput Sequencing Data) | 59 |

| | |
|---|------------|
| 3.2.1 Conjuntos de datos artificiales..... | 60 |
| 3.2.1.1 Simulación de lecturas cortas tratadas con bisulfito..... | 60 |
| 3.2.2 Preprocesado de los datos..... | 62 |
| 3.2.3 Alineamiento de lecturas cortas de ADN tratadas con bisulfito..... | 64 |
| 3.2.3.1 Alineamiento..... | 66 |
| 3.2.3.2 Selección de los alineamientos..... | 68 |
| 3.2.4 Mejora en la detección de adaptadores..... | 70 |
| 3.2.5 Efecto de la desambiguación de multilecturas..... | 71 |
| 3.2.6 Comparación con otros métodos..... | 73 |
| 3.2.6.1 Características..... | 74 |
| 3.2.6.2 Velocidad de procesado..... | 77 |
| 3.2.6.3 Eficiencia de los alineamientos..... | 78 |
| 3.3 MethylExtract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data..... | 81 |
| 3.4 MAPAS DE CALIDAD DE METIOMAS COMPLETOS..... | 96 |
| 4 MINERÍA DE DATOS Y VISUALIZACIÓN DE PERFILES DE METILACIÓN..... | 103 |
| 4.1 BASES DE DATOS PARA EL ANÁLISIS DE LA METILACIÓN DEL ADN..... | 104 |
| 4.2 NGSmethDB..... | 106 |

| | | |
|----------|---|------------|
| 4.2.1 | NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation | 108 |
| 4.2.2 | NGSmethDB: an updated genome resource for high-quality, single-cytosine resolution methylomes | 114 |
| 5 | PREDICCIÓN COMPUTACIONAL DE ISLAS CpG | 125 |
| 5.1 | IDENTIFICACIÓN DE CGIs BASADA EN LA SECUENCIA | 126 |
| 5.1.1 | Métodos de ventana | 127 |
| 5.1.2 | Métodos de adición | 128 |
| 5.1.3 | Métodos de agrupación de CpGs | 129 |
| 5.2 | WordCluster: detecting clusters of DNA words and Genomic elements | 130 |
| 5.3 | Prediction of CpG-island function: CpG clustering vs. sliding-window methods | 139 |
| 6 | ISLAS CpG DIFERENCIALMENTE METILADAS | 155 |
| 6.1 | MATERIAL Y MÉTODOS | 158 |
| 6.1.1 | Estados y niveles de metilación | 159 |
| 6.1.2 | Análisis estadísticos | 160 |
| 6.1.3 | Especificidad y sensibilidad | 162 |
| 6.1.4 | Clasificación de CGIs | 162 |
| 6.1.5 | Elementos genómicos utilizados para los análisis de enriquecimiento | 163 |
| 6.2 | IDENTIFICACIÓN DE ISLAS CpG DIFERENCIALMENTE METILADAS | 168 |

| | |
|---|------------|
| 6.2.1 Distribución de las diferencias de metilación..... | 170 |
| 6.2.2 Comparación entre métodos estadísticos | 172 |
| 6.2.3 Análisis basado en la binomial negativas combinado con la prueba exacta de Fisher | 174 |
| 6.3 CARACTERIZACIÓN DE LAS CLASES DE CGIs | 180 |
| 6.3.1 Características composicionales | 182 |
| 6.3.2 Enriquecimiento en elementos reguladores | 185 |
| 6.3.2.1 Regiones génicas | 186 |
| 6.3.2.2 Sitios de interacción con el ADN | 189 |
| 6.3.3 Conservación y variaciones de secuencia | 194 |
| 6.4 FUNCIONES REGULADAS POR LAS CGIs | 197 |
| 6.5 DISCUSIÓN | 202 |
| 6.6 MATERIAL SUPLEMENTARIO..... | 204 |
| 7 CONCLUSIONES..... | 215 |
| 8 CUESTIONES ABIERTAS..... | 221 |
| LISTADO DE FIGURAS..... | 225 |
| LISTADO DE TABLAS | 233 |
| REFERENCIAS..... | 237 |



INTRODUCCIÓN

La epigenética está siendo la rama de la genética más floreciente en los primeros años del siglo XXI. Si el siglo pasado fue considerado por muchos el siglo de la genética, el crecimiento exponencial de publicaciones científicas relacionadas con la epigenética (Figura 1.1) y los proyectos como *ENCODE* (Consortium 2004) o *ROADMAP Epigenomics* (Bernstein, Stamatoyannopoulos et al. 2010), destinados a analizar y comprender el “mapa epigenético”, hacen prever para los próximos años una revolución guiada por la epigenética, en el modo de entender y estudiar la regulación del genoma.

Actualmente, se manejan numerosas definiciones para el término epigenética, algunas procedentes de disciplinas tan dispares como la psicología (Gottlieb 1991) o la biología del desarrollo (Waddington 2012), que poco o nada tienen que ver con su definición genética. A grandes rasgos, el término epigenética hace referencia al estudio de cambios en la actividad génica que no están causados por variaciones en la secuencia de ADN. Sin embargo, las características de estos cambios



están aún por definir, ya que algunas de las propiedades consideradas intrínsecas a las marcas epigenéticas —heredables, autoperpetuables y reversibles (Bonasio, Tu et al. 2010)— no son comunes a todas ellas. Esta falta de acuerdo entre la definición y los cambios considerados epigenéticos, demuestra que a pesar del gran número de publicaciones y los grandes proyectos abordados en los últimos años, la epigenética sigue siendo en gran medida un área bastante desconocida. Podemos esperar, por tanto, que el número de publicaciones y proyectos en el área seguirán creciendo de manera exponencial.

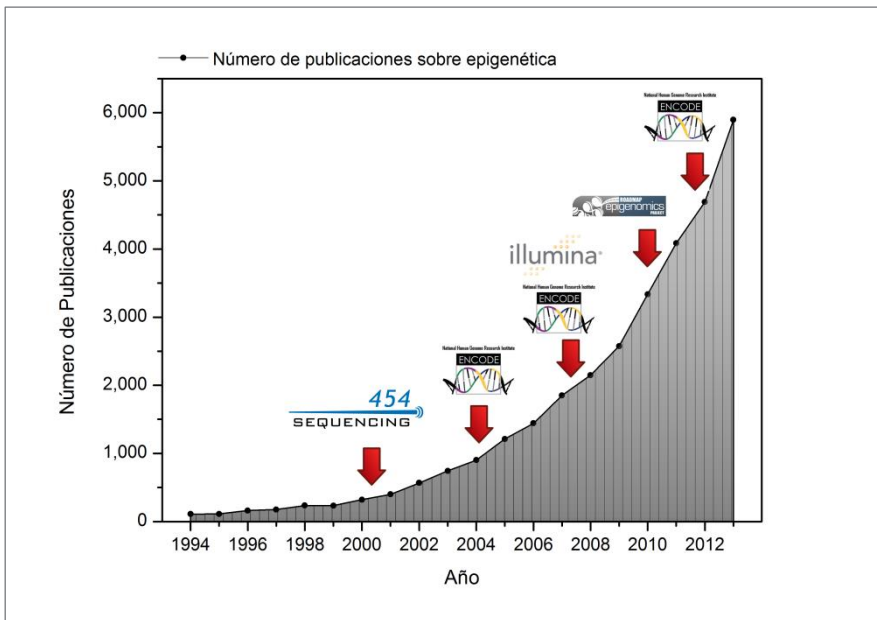


Figura 1.1. Número de publicaciones sobre epigenética entre los años 1994 y 2013. La gráfica representa el número de publicaciones que contienen los términos “Epigenetic” o “Epigenomic” incluidas en la base de datos PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>) durante los últimos 20 años (1994-2013). Además, se incluyen algunos de los hitos, tanto técnicos como científicos, más importantes en la investigación epigenética: la aparición del primer método de secuenciación masiva (“454 sequencing” en el año 2000), la fundación del consorcio ENCODE (Consortium 2004), la publicación de los resultados del proyecto piloto de ENCODE (Birney,

Stamatoyannopoulos et al. 2007), la incorporación de “*Illumina*” al mercado de la secuenciación masiva (2007), la fundación del consorcio “*ROADMAP Epigenomics*” (Bernstein, Stamatoyannopoulos et al. 2010) y la publicación en septiembre de 2012 de 29 artículos (<http://www.nature.com/encode/>) con los resultados obtenidos a partir del proyecto *ENCODE*.

1.1 METILACIÓN DEL ADN

La metilación del ADN es la marca epigénética por excelencia, ya que cumple todos los estándares establecidos: (i) puede modificar la actividad génica de múltiples maneras (como se verá a lo largo de esta Tesis), (ii) puede autopropetarse mediante metiltransferasas específicas de mantenimiento (Klose and Bird 2006), (iii) es una marca reversible, como se ha comprobado por ejemplo durante la diferenciación celular (Lister, Pelizzola et al. 2009, Laurent, Wong et al. 2010), y (iv) aunque la media es relativamente baja, algunas regiones concretas presentan evidencias de una elevada heredabilidad (Bell and Spector 2012).

Bioquímicamente, la metilación del ADN consiste en la unión covalente de un grupo metilo a los nucleótidos, mediada por metiltransferasas específicas del ADN (DNMTs). En eucariotas, esta modificación se encuentra restringida al carbono 5 de las citosinas, dando lugar a metilcitosinas que se han denominado “la quinta base del ADN” (Lister and Ecker 2009), debido a sus diferencias estructurales y funcionales con las citosinas no metiladas (Jones 2012).

La aparición de las técnicas de secuenciación masiva (“*High-Throughput Sequencing*” o “*Next-Generation Sequencing*”), en torno al año 2000 (Figura 1.1), revolucionó la secuenciación del genoma, y en

particular la investigación en epigenética. Su combinación con el tratamiento del ADN con bisulfito (Frommer, McDonald et al. 1992), permitió la obtención de mapas de metilación en genoma completo de una manera rápida y económica (véase [capítulo 3](#)). Estos avances tecnológicos han permitido el estudio de la metilación en genoma completo en múltiples tejidos, resultando en el descubrimiento de nuevos procesos regulados por la metilación, así como de características de la misma que se desconocían. Básicamente, la metilación interviene en la transcripción de los genes y mantiene la estabilidad del genoma. Sin embargo, estas funciones pueden regularse de diferentes maneras:

- Regulación de la transcripción:
 - Regulando la interacción del complejo de la ARN polimerasa con los promotores (Bell, Pai et al. 2011).
 - Interviniendo en mecanismos de *splicing* alternativo (Shukla, Kavak et al. 2011).
 - Regulando la unión de factores de transcripción a regiones potenciadoras (Hon, Rajagopal et al. 2013) y aisladoras (Wang, Maurano et al. 2012).
- Estabilidad genómica:
 - Silenciando elementos repetidos como retrotransposones (Yoder, Walsh et al. 1997).
 - Participando en la compensación de dosis de cromosomas sexuales (Sharp, Stathaki et al. 2011) y en la impronta de genes autosómicos (Li, Beard et al. 1993).

Observando los múltiples mecanismos reguladores en los que se encuentra implicada la metilación, no es de extrañar que se hayan

encontrado patrones aberrantes de metilación asociados a un gran número de enfermedades (Robertson 2005), entre las que destacan diferentes tipos de cánceres.

En mamíferos, esta marca epigenética se encuentra sobre todo en las citosinas de los dinucleótidos CpG (Lister, Pelizzola et al. 2009), que aparecen metiladas en la mayor parte del genoma. Sin embargo, existen ciertas regiones, llamadas islas CpG (*CGIs*), que se encuentran libres de esta marca epigenética y presentan una elevada densidad de CpGs. Esta distribución de la metilación a lo largo del genoma de mamíferos se ha denominado patrón global de metilación (Suzuki and Bird 2008).

1.2 ISLAS CpG

Las islas CpG (*CGIs*) son regiones del genoma con un alto contenido en G+C y una elevada frecuencia de dinucleótidos CpG en relación con el resto del genoma (Gardiner-Garden and Frommer 1987). Epigenéticamente, estas islas se han caracterizado por no presentar metilación y por ser regiones de cromatina abierta que pueden permitir la interacción con el ADN (Cooper, Taggart et al. 1983, Bird, Taggart et al. 1985). Aproximadamente el 70% de los promotores de genes anotados en el genoma humano presentan *CGIs* (Saxonov, Berg et al. 2006), entre los que figuran la totalidad de los genes domésticos y un 40% de los genes tejido específicos (Zhu, He et al. 2008).

1.2.1 Dinámica evolutiva de los dinucleótidos CpG

Las *CGs* son una excepción en el genoma de mamíferos, donde los CpGs se encuentran infrarrepresentados. En humanos, la frecuencia de guaninas y citosinas es aproximadamente 0.2, por lo que esperaríamos encontrar una frecuencia de $0.2 \times 0.2 = 0.04$ CpGs en el genoma. Sin embargo, la frecuencia observada es de 0.008 (Bird 1980), 5 veces menor que la esperada. La adición de grupos metilo a las citosinas del ADN aumenta su tasa de desaminación (Salser 1978), lo que conlleva un aumento de la mutabilidad de las metilcitosinas a timinas. En mamíferos, y en particular en humanos, casi la totalidad de las metilcitosinas presentes en el genoma se encuentran en contextos CpG (75.5% en la línea celular H1 y 99.98% en IMR90 (Lister, Pelizzola et al. 2009)), lo que explica la escasa proporción de CpGs presentes en sus genomas.

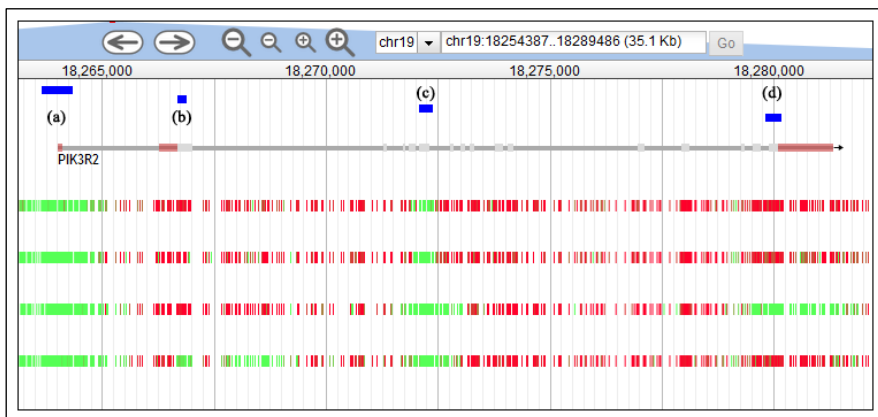


Figura 1.2. *CGs* predichas por *CpGcluster* en el gen *PIK3R2*. En la figura se muestra el gen *PIK3R2* (anotación de RefSeq (Pruitt, Tatusova et al. 2007)), las *CGs* predichas por *CpGcluster* (Hackenberg, Previti et al. 2006, Hackenberg, Carpena et al. 2011) y los niveles de metilación para los CpGs de 4 tejidos diferentes (Hackenberg, Barturen et al. 2011, Geisen, Barturen et al. 2014). En verde se representan los CpGs no metilados y en rojo los metilados. Las islas predichas en el gen *PIK3R2* ilustran diferentes tipos que pueden clasificarse en función de su estado de

metilación y/o de su localización: (a) y (c) son islas no metiladas en todos los tejidos, normalmente asociadas a las regiones promotoras de genes domésticos (a), pero también pueden encontrarse asociadas a otros elementos genómicos como exones (c); sin embargo, (b) y (d) son islas diferencialmente metiladas, reguladoras potenciales de la interacción de otras moléculas con el ADN.

Esta dinámica evolutiva no se observa en las *CG/s*, ya que generalmente no presentan metilación (islas (a) y (c) en la Figura 1.2). Sin embargo, se han encontrado *CG/s* con metilación diferencial (ver isla (b) y (d) en la Figura 1.2) y constitutivamente metiladas, cuya función parece ser la de regular las interacciones con el ADN de manera tejido específica (como factores de transcripción, intensificadores o aisladores) y estabilizar zonas estructurales del genoma, respectivamente (Dindot, Person et al. 2009, Doi, Park et al. 2009, Lister, Pelizzola et al. 2009).

1.3 METILACIÓN DIFERENCIAL

El término metilación diferencial hace referencia a variaciones sistemáticas en los niveles de metilación entre diferentes tejidos, individuos o tipos celulares (Figura 1.3). El mecanismo que subyace tras estas diferencias de metilación es la regulación de la interacción de diferentes moléculas con el ADN. Por ello, la identificación de las Regiones Diferencialmente Metiladas (*DMRs*) es de gran importancia a la hora de estudiar la regulación en el genoma. La secuenciación masiva en múltiples tejidos, previo tratamiento del ADN con bisulfito, ha permitido identificar un 21.8% de CpGs con metilación dinámica en el genoma humano (Ziller, Gu et al. 2013). En general, estas diferencias de metilación se han asociado tanto con funciones tejido-específicas (Ziller,

Gu et al. 2013), como con procesos de diferenciación y desarrollo celular (Lister, Pelizzola et al. 2009, Laurent, Wong et al. 2010).

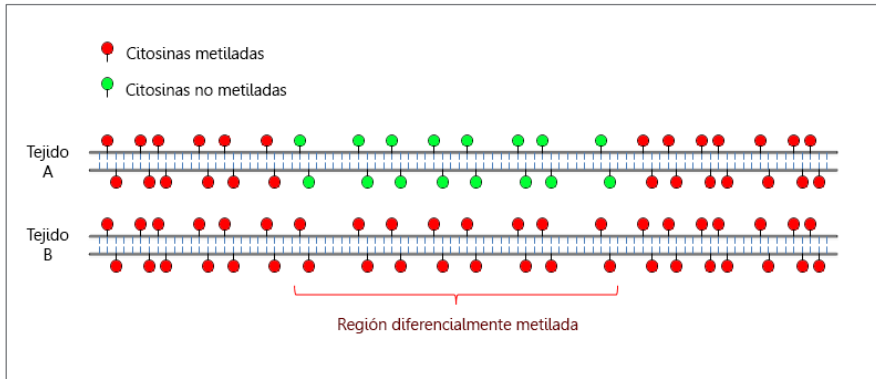


Figura 1.3. Representación esquemática de una región diferencialmente metilada. Los círculos representan citosinas en dinucleótidos CpG para dos muestras diferentes (Tejidos A y B); los círculos rojos simbolizan citosinas metiladas y los verdes no metiladas.

El estudio de la metilación diferencial se ha centrado hasta ahora en la propia metilación, buscando alguna función asociada a estas diferencias. Sin embargo, recientemente se ha demostrado la incapacidad de la metilación para inhibir la interacción de factores de transcripción (*TFs*) en sitios de unión con una baja densidad de CpGs, resultando en una desmetilación pasiva de dicha región (Stadler, Murr et al. 2011). Este descubrimiento podría suponer que multitud de regiones diferencialmente metiladas (aquellas que presentan una baja densidad de CpGs) a las que se les suponía alguna función, podrían ser simplemente un subproducto de la unión de factores de transcripción (Figura 1.4). Además, en las cercanías de los promotores se han identificado unas pequeñas regiones que determinan su metilación, denominadas Regiones Determinantes de la Metilación (*MDRs*). Estas regiones, además de presentar sitios de unión a factores de transcripción, suelen

tener una elevada densidad de CpGs (Lienert, Wirbelauer et al. 2011). Ambos hallazgos parecen indicar una gran importancia de la densidad de CpGs, tanto en la función como en la determinación del estado de la metilación.

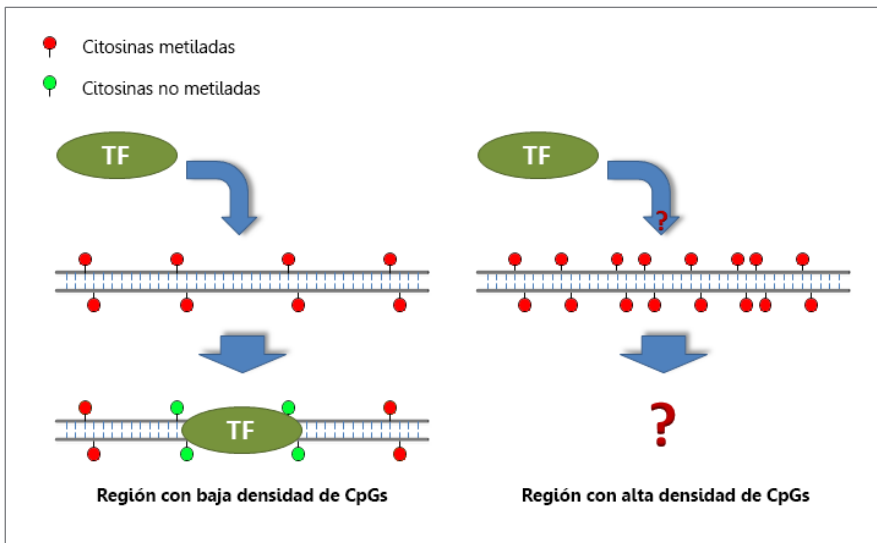


Figura 1.4. Representación esquemática de la interacción de factores de transcripción con el ADN. En la figura se muestran dos regiones hipermetiladas, una con baja densidad de CpGs (izquierda) y otra con alta densidad (derecha). Los círculos representan citosinas en dinucleótidos CpG (los rojos simbolizan citosinas metiladas y los verdes no metiladas). Por otro lado, el factor de transcripción (*TF*) con un sitio de unión a estas regiones se representa como una elipse verde.

Las regiones densas en CpGs del genoma se han asociado a la definición tradicional de islas CpG. Sin embargo, los algoritmos clásicos de predicción de islas son incapaces de detectar estas regiones, al no alcanzar el umbral de longitud prefijado. Un ejemplo de ello es que ni siquiera el 5% de las *DMRs* identificadas por Ziller y colaboradores se encuentran en las islas CpGs predichas por estos algoritmos (Ziller, Gu et al. 2013). Otros métodos, sin embargo, permiten identificar *clusters*

densos de CpGs sin estas limitaciones (véase [Capítulo 5](#)). Entre ellos destaca *CpGcluster* (Hackenberg, Previti et al. 2006), un algoritmo que identifica *clusters* de CpGs estadísticamente significativos, basándose exclusivamente en la composición de la secuencia. Algoritmos como este, que redefinen el concepto de isla CpG, pueden ser una buena opción para seleccionar aquellas regiones diferencialmente metiladas que presenten una densidad suficiente de CpGs como para ser funcionales y/o formar parte de *MDRs*.

OBJETIVOS

Los nuevos métodos de secuenciación masiva, tras el tratamiento del ADN con bisulfito, han permitido la obtención de metilomas de citosinas individuales para diferentes tejidos o condiciones experimentales. Los estudios de metilación diferencial realizados hasta la fecha han desvelado numerosas funciones previamente desconocidas de la metilación. Sin embargo, y aunque la densidad de grupos metilo parece tener una importancia capital, estos estudios no han tenido en cuenta la densidad de CpGs a la hora de intentar determinar la función de las Regiones Diferencialmente Metiladas (*DMRs*). Esta tesis doctoral tratará de identificar y caracterizar regiones diferencialmente metiladas con una elevada densidad de CpGs. Más concretamente, se investigarán las funciones y procesos celulares que pueden estar siendo regulados activamente por cambios en los patrones de metilación en humanos (Capítulo 6). Para ello, se abordarán los siguientes objetivos:

1. Desarrollo de la metodología y los programas bioinformáticos necesarios para procesar y cuantificar los niveles de metilación, a

partir de datos obtenidos mediante protocolos de secuenciación masiva. Este objetivo incluye tanto el alineamiento de lecturas cortas frente al genoma de referencia, obtenidas mediante los secuenciadores de nueva generación, como el postprocesado de estos alineamientos para obtener los niveles de metilación en citosinas individuales. Ambos pasos incluirán la implementación de rigurosos controles de calidad ([Capítulo 3](#)).

2. Desarrollo de herramientas bioinformáticas específicas para comparar la ingente cantidad de datos resultantes de los estudios genómicos en múltiples tejidos de una manera sencilla. Para ello se desarrollará una base de datos relacional que permita el almacenamiento y la comparación de los datos de metilación en citosinas individuales ([Capítulo 4](#)).
3. Optimización del algoritmo *CpGcluster*, con objeto de mejorar sus predicciones. Se ampliará el algoritmo para poder aplicarlo a cualquier contexto de secuencia o elemento genómico. Además, se desarrollará una interfaz web que facilite su uso ([Capítulo 5](#)).
4. Comparación del algoritmo mejorado con los métodos clásicos de identificación de islas CpGs. En particular, se comprobará su eficacia para caracterizar las regiones diferencialmente metiladas con elevada densidad de CpGs ([Capítulo 5](#)).

PERFILES DE METILACIÓN DE ALTA CALIDAD

La aparición de la secuenciación masiva y el tratamiento del ADN con bisulfito han permitido obtener metilomas de alta resolución (mapas genómicos con niveles de metilación para citosinas individuales). Estos experimentos generan millones de lecturas cortas que deben ser procesadas adecuadamente *in-silico* para obtener resultados fidedignos, lo que supone un reto importante a nivel bioinformático. En este capítulo se presentan y describen dos programas desarrollados en el marco de esta Tesis Doctoral: *NGSmethPipe* y *MethylExtract*. El primero permite preprocesar y alinear las lecturas cortas frente al genoma de referencia, mientras que el segundo infiere los niveles de metilación de citosinas individuales y las variaciones de un solo nucleótido (*SNVs*) a partir de dichos alineamientos. Estas herramientas incluyen el mayor número de controles de calidad entre los programas disponibles actualmente. Si se tiene en cuenta además su facilidad de uso y su capacidad para adaptarse a los requerimientos computacionales de los usuarios, ambas herramientas constituyen conjuntamente un protocolo muy potente y fiable para procesar los datos de experimentos de

secuenciación masiva de ADN tratado con bisulfito. Al final de este capítulo, se mostrarán resultados para conjuntos de datos procedentes de diversas publicaciones, procesados de manera uniforme mediante este protocolo.

Ambos programas y sus tutoriales se encuentran disponibles libremente en: <http://bioinfo2.ugr.es/NGSmethPipe/> y <http://bioinfo2.ugr.es/MethylExtract/>.

3.1 MÉTODOS PARA DETERMINAR EL ESTADO DE METILACIÓN DEL ADN

La metilación del ADN consiste en la adición de un grupo metilo a los nucleótidos del ADN. En eucariotas, esta modificación ocurre exclusivamente en las citosinas y suele localizarse en contextos de secuencia CpG y/o CpHpG (donde H: C, A o T). Esta modificación de la citosina ha sido ampliamente estudiada durante la última década, desarrollándose para ello numerosas técnicas que permiten detectarla. La metilación del ADN se elimina durante la *PCR* (Reacción en Cadena de la Polimerasa) y no puede detectarse mediante hibridación, ya que el grupo metilo se localiza en el surco mayor del ADN. Estos inconvenientes han lastrado esta área de la genómica, a diferencia de otras áreas, como la cuantificación de transcriptomas, que no presentan estas limitaciones. Todas las técnicas desarrolladas durante los últimos años para su detección se basan en algún pretratamiento del ADN antes de la hibridación, amplificación o secuenciación. Los principales pre-tratamientos son: la digestión por endonucleasas sensibles al grupo

metilo, la inmunoprecipitación sensible al grupo metilo y la conversión por bisulfito; revisadas por Laird en 2010 (Laird 2010).

3.1.1 Metodologías previas al tratamiento con bisulfito

Las metodologías previas al uso del bisulfito combinado con técnicas de secuenciación masiva presentan multitud de inconvenientes (Lister and Ecker 2009, Laird 2010):

- Las técnicas basadas en la digestión por endonucleasas seguidas de hibridación o electroforesis en geles bidimensionales sólo analizan una pequeña fracción del genoma (dependiendo su resolución de los sitios de restricción) y no pueden precisar el contexto donde se localiza la metilcitosina.
- En cuanto a las técnicas basadas en enriquecimiento por afinidad seguidas de hibridación (como *MeDIP*), a pesar de ser métodos rápidos y eficientes para determinar la metilación a lo largo del genoma, tampoco devuelven información de citosinas individuales. Presentan además un sesgo hacia regiones ricas en grupos metilo y poca sensibilidad en regiones con baja densidad de CpGs.

3.1.2 Secuenciación masiva combinada con el tratamiento de bisulfito

Aunque el tratamiento del ADN con bisulfito de sodio, para la obtención de niveles de metilación en citosinas individuales, se conoce desde el inicio de la década de los 90 (Frommer, McDonald et al. 1992), la obtención de metilomas completos de alta resolución no ha sido posible

hasta la aparición de los métodos de secuenciación masiva, como los desarrollados por *Illumina*, *Roche 454* o *SOLiD* (*Applied Biosystems*), por mencionar los más conocidos (Shendure and Ji 2008). Existen 2 protocolos basados en secuenciación masiva que no requieren de pretratamientos de digestión o enriquecimiento: *MethylC-Seq* (Lister, O'Malley et al. 2008) y *BS-Seq* (Cokus, Feng et al. 2008), de los que existe una revisión detallada (Lister and Ecker 2009). Ambos se basan en el tratamiento del ADN desnaturalizado con bisulfito de sodio, que provoca la desaminación de las citosinas no metiladas, preservando la integridad de las metilcitosinas. Tras la secuenciación del ADN tratado con bisulfito, el estado de metilación puede inferirse directamente a partir de las lecturas alineadas frente a un genoma de referencia: una citosina no modificada indicará la existencia de metilación en esa posición, mientras que una timina (resultado de la amplificación por *PCR* tras la desaminación) supondrá la existencia de una citosina no metilada.

En algunos estudios, se han tratado con bisulfito las regiones seleccionadas mediante inmunoprecipitación o enzimas de restricción y se han secuenciado, obteniéndose de esta manera la información de secuencia y evitando los problemas derivados del uso de *microarrays*, pero siguen presentando los inconvenientes heredados tanto de la inmunoprecipitación como del uso de endonucleasas. Por lo tanto, salvo en casos concretos (como estudios centrados en islas CpG, por ejemplo), los protocolos de referencia para el estudio de la metilación de genomas completos, son aquellos basados en la ruptura aleatoria del genoma (sonicación por ejemplo), selección de fragmentos por tamaño, tratamiento con bisulfito y secuenciación masiva.

Actualmente, el tratamiento con bisulfito seguido de secuenciación masiva se utiliza en numerosos proyectos para obtener metilomas en diferentes especies. La búsqueda de los términos “*Bisulfite*” y “*Methylation profiling by high throughput sequencing*” en la base de datos pública *GEO* (Barrett, Troup et al. 2009, Barrett, Wilhite et al. 2013) devuelve 178 experimentos para multitud de especies, como: *Homo sapiens*, *Mus musculus* o *Arabidopsis thaliana*, entre otros.

3.1.3 Herramientas bioinformáticas

En los experimentos de secuenciación masiva, la obtención de datos fiables y que cubran la mayor parte del genoma, requiere la resecuenciación a coberturas en torno a 15x (es decir, cada posición debe ser resecuenciada 15 veces por término medio), lo que obviamente supone un reto bioinformático importante a la hora de analizar los resultados. A pesar de la mejora sustancial que ha supuesto esta nueva metodología, para obtener resultados óptimos deben tenerse en cuenta las limitaciones y fuentes de error que presenta, revisadas recientemente (Robinson, Statham et al. 2010, Krueger, Kreck et al. 2012).

Habitualmente, el procesado de los datos consta de tres pasos: el preprocesado de las lecturas, el alineamiento de éstas y la inferencia de los niveles de metilación de citosinas individuales. Durante los últimos años, han ido apareciendo multitud de programas que alinean las lecturas generadas por los secuenciadores en formato *FastQ* (Cock, Fields et al. 2010) y analizan dichos alineamientos (generalmente en formato *SAM/BAM* (Li, Handsaker et al. 2009)) como: *BSMAP/RRBSMAP* (Xi and Li 2009), *RMAP* (Smith, Chung et al. 2009), *mrsFAST* (Hach, Hormozdiari

et al. 2010), *SOCS-B* (Ondov, Cochran et al. 2010), *BS Seeker* (Chen, Cokus et al. 2010), *MethylCoder* (Pedersen, Hsieh et al. 2011), *Bismark* (Krueger and Andrews 2011), *BRAT-BW* (Harris, Ponts et al. 2012) y *Bis-SNP* (Liu, Siegmund et al. 2012).

Entre estos programas figuran dos desarrollados en esta Tesis: *NGSmethPipe* (Hackenberg, Barturen et al. 2012), que preprocesa y alinea las lecturas y *MethylExtract* (Barturen, Rueda et al. 2013), que infiere los niveles de metilación. A lo largo de este capítulo, se describirán las principales fuentes de error y los métodos implementados en *NGSmethPipe* y *MethylExtract* para superarlas. Además, se compararán ambos programas con otros métodos, tratando de demostrar que *NGSmethPipe* y *MethylExtract* son los métodos más completos en cuanto a funciones y controles de calidad.

- 3.2 NGSmethPipe (adaptado y extendido a partir de Hackenberg, M, G Barturen, JL Oliver. 2012. DNA methylation Profiling from High-Throughput Sequencing Data. DNA Methylation: InTech - Open Access Publisher, ISBN 979-953-307-453-4)
-

Dirección de publicación:

<http://www.intechopen.com/books/dna-methylation-from-genomics-to-technology/dna-methylation-profiling-from-high-throughput-sequencing-data>

Página web de la aplicación:

<http://bioinfo2.ugr.es/NGSmethPipe/>

Breve descripción del programa:

NGSmethPipe es un programa para el preprocesado y alineamiento de lecturas cortas de ADN tratado con bisulfito. El programa, implementado en Perl, se compone de 2 subprogramas: (i) preprocesado del genoma de referencia; este paso sólo debe realizarse una vez por genoma y (ii) alineamiento de las lecturas cortas. Ambos subprogramas permiten la paralelización y el control del consumo de memoria del proceso, para adaptarse a la infraestructura informática del usuario.

3.2.1 Conjuntos de datos artificiales

La comparación entre los diferentes métodos se ha realizado utilizando lecturas simuladas. La generación de datos artificiales se basa en reproducir *in-silico* los resultados del protocolo experimental (lecturas cortas tratadas con bisulfito) y los errores que pueden sesgar los resultados finales. El uso de datos artificiales tiene la ventaja de que se conoce exactamente tanto la localización de las lecturas en el genoma, como el genotipo y los niveles de metilación del conjunto de datos analizado. Si se usaran como referencia otros datos experimentales, tales como *microarrays*, obviamente los errores del conjunto de datos serían más realistas, pero no podría asegurarse que la localización o valores devueltos por estos métodos fuesen los reales. La simulación de lecturas cortas tratadas con bisulfito supone: extraer aleatoriamente secuencias del genoma de referencia, simular la variación de secuencia presente en la población, asignar los niveles de metilación a las citosinas, simular los fallos del tratamiento con bisulfito, simular la inclusión del adaptador en el extremo 3' de las lecturas y simular, por último, los errores provocados durante la secuenciación.

3.2.1.1 Simulación de lecturas cortas tratadas con bisulfito

DNemulator (Frith, Mori et al. 2012) permite simular lecturas cortas tratadas con bisulfito, incluyendo todos los pasos necesarios comentados previamente, salvo la inclusión del adaptador en el extremo 3' para lo que se desarrolló un programa propio.

El programa que simula la inclusión de adaptadores permite fijar la probabilidad de inclusión del adaptador y ajusta la distribución de longitudes a una distribución normal con media y desviación estándar seleccionadas por el usuario. En este estudio, se utilizó la secuencia del adaptador de *Illumina* incluido en protocolos de secuenciación experimentales *single-end* (http://support.illumina.com/downloads/illumina_adapter_sequences_letters.ilmn).

El conjunto de variantes que se utilizó en *DNemulator* fue *dbSNP135* (Sherry, Ward et al. 2001) y la tasa de error de conversión con bisulfito se fijó en un 1% (tasa de error experimental aproximada). Utilizando ambos programas, se simularon 2 conjuntos de lecturas en función de los parámetros de inclusión del adaptador:

- Para determinar la eficiencia del preprocesado detectando adaptadores (Figura 3.5) se simularon 1,000,000 de lecturas cortas con *DNemulator*, incluyéndose en todas ellas la longitud total del adaptador.
- Para comparar la fiabilidad de los diferentes métodos (Figura 3.6 y Figura 3.7) se simularon 2,000,000 de lecturas para 2 *contigs* diferentes del ensamblado humano hg19 (*GL000022.1* del cromosoma 2 y *GL000109.1* del cromosoma 12, ambos con un tamaño aproximado de 10 Mb). La probabilidad de inclusión del adaptador se fijó en un 46%, y las longitudes de inserción se ajustaron a una distribución normal con media 15 y desviación estándar 14. Estos valores se calcularon empíricamente tras la

detección de adaptadores en el conjunto de datos de Lister (Lister, Pelizzola et al. 2009).

Los errores de secuenciación se simularon a partir de la línea de calidad (conjunto de datos *h1* de Lister (Lister, Pelizzola et al. 2009)) asignada a cada lectura tras la inclusión de los adaptadores.

3.2.2 Preprocesado de los datos

Antes de alinear las lecturas frente al genoma de referencia deben tomarse ciertas precauciones para optimizar el alineamiento. Este preprocesado de las lecturas puede dividirse en: (i) la eliminación o manipulación de lecturas con baja calidad de secuenciación y (ii) la preparación de las lecturas para el alineamiento frente al genoma de referencia.

- (i) A pesar de las mejoras en los métodos de secuenciación, la calidad de las lecturas decae hacia el extremo 3', aumentando la probabilidad de que el nucleótido detectado no sea realmente el que está presente en la muestra. Lister y colaboradores (Lister, Pelizzola et al. 2009) propusieron recortar las lecturas antes del primer nucleótido con una baja calidad de secuenciación (valor de calidad *PHRED* ≤ 2 , es decir probabilidades de error de al menos el 50% (Cock, Fields et al. 2010)), descartando de esta manera la región de menor calidad de las lecturas (Figura 3.1, A). El valor de calidad *PHRED*, fue originalmente desarrollado en el programa *Phred* (Ewing and Green 1998, Ewing, Hillier et al.

1998) durante la secuenciación del genoma humano, y asigna automáticamente la probabilidad de error a cada nucleótido secuenciado a partir de las intensidades detectadas en cada posición.

- (ii) Otro paso que puede mejorar la precisión del alineamiento es la eliminación de la secuencia del adaptador (Figura 3.1, B). Cuando el fragmento de ADN a secuenciar es menor que el número de ciclos, se secuenciará parte del adaptador en 3'. Este adaptador debe ser detectado y eliminado, ya que dependiendo del algoritmo de alineamiento puede llevar a la pérdida de alineamientos o incluso al incremento de alineamientos erróneos (véase apartado 3.2.3).

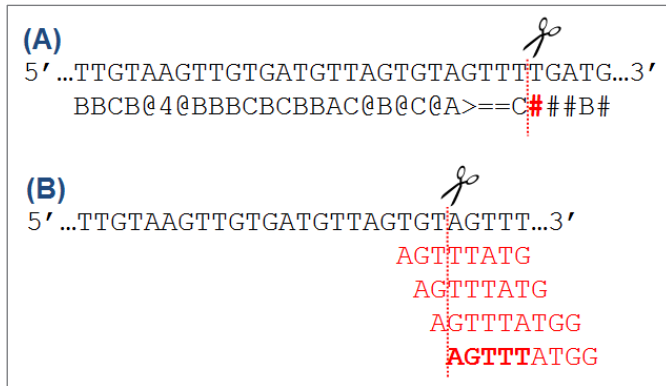


Figura 3.1. Preprocesado de las lecturas. La figura muestra los pasos necesarios para preparar las lecturas antes de ser alineadas. En primer lugar se elimina la región de menor calidad de las lecturas (A), recortando antes del primer nucleótido con una calidad de secuenciación menor de 2 (#) y posteriormente se busca de manera iterativa la secuencia del adaptador (B).

3.2.3 Alineamiento de lecturas cortas de ADN tratadas con bisulfito

El tratamiento con bisulfito del ADN (Frommer, McDonald et al. 1992), su secuenciación y su posterior comparación con un genoma de referencia, permite obtener el estado de metilación de citosinas individuales. El bisulfito de sodio provoca la desaminación espontánea de las citosinas no metiladas a uracilo, que durante la subsiguiente *PCR* serán amplificadas como timinas (Figura 3.2). De esta manera, comparando las citosinas de la secuencia sin tratar con las de la secuencia tratada con bisulfito, se deduce que las citosinas corresponden a posiciones metiladas y las timinas a las no metiladas.

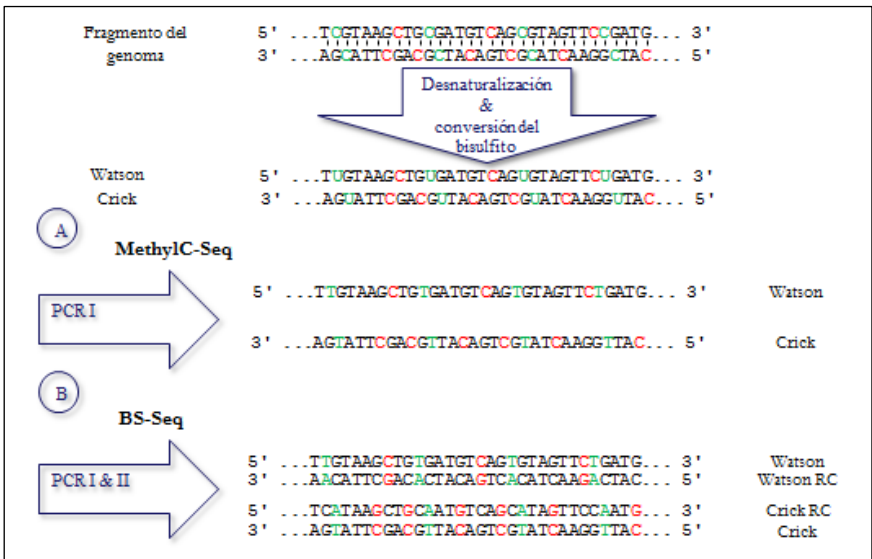


Figura 3.2. Tipos de secuenciación masiva con tratamiento de bisulfito: *MethyI-C-Seq* y *BS-Seq*. Tras la desnaturalización y el tratamiento con bisulfito del ADN se pierde la complementariedad de hebra, ya que las citosinas no metiladas se convierten en uracilos (coloreadas en verde). Durante la amplificación del ADN, los uracilos serán sustituidos por timinas. En la figura se muestran las lecturas resultantes de ambos protocolos: (A) El protocolo *MethyI-C-Seq* genera la librería de lecturas de manera direccional, donde encontraremos las secuencias

tratadas con bisulfito para la cadena Watson (*BSW*) y para la cadena Crick (*BSC*); (B) En el protocolo *BS-Seq* se realizan dos *PCRs* consecutivas, que resultan en la presencia tanto de las lecturas provenientes de *BSW* y *BSC* como de sus reversas complementarias (*BSWRC* y *BSCRC*).

En el caso de la secuenciación masiva, el cambio que permite determinar el estado de metilación también dificulta el alineamiento de las lecturas con el genoma de referencia (única manera de determinar su localización original). En los genomas de mamíferos, la metilación ocurre casi exclusivamente en el contexto CpG (4.8% de las citosinas del genoma humano), por lo que la mayoría de las citosinas (95.2%) se convertirán en timinas durante el tratamiento con bisulfito. Por lo tanto, este tratamiento reduce a un alfabeto de 3 letras (A,T,G) casi la totalidad de las lecturas, lo que provoca la reducción de la complejidad de las secuencias y complica el proceso de alineamiento. Otro problema específico de la secuenciación con bisulfito es que se duplica el tamaño del espacio de búsqueda de los alineamientos. Esto se debe a que el bisulfito sólo actúa sobre las citosinas, dejando intactas las guaninas de la hebra complementaria (Figura 3.2), lo que conlleva la pérdida de complementariedad de las cadenas Watson y Crick y por lo tanto, la aparición de dos referencias de búsqueda (*BSW* y *BSC*), donde deberá buscarse *in-silico* la localización original de las lecturas (Figura 3.3). Por último, y al igual que en protocolos sin tratamiento con bisulfito, la presencia de errores de secuenciación también complica el correcto alineamiento de las lecturas.

3.2.3.1 *Alineamiento*

Debido a las características específicas comentadas en el apartado previo, las lecturas tratadas con bisulfito no pueden ser alineadas de igual manera que las no tratadas, ya que las citosinas convertidas a timinas causarían desemparejamientos espurios frente a la referencia. En principio, debería ser posible alinear estas lecturas tratadas mediante alineadores convencionales, simplemente aumentando el número máximo de des-emparejamientos permitidos por el algoritmo. Sin embargo, esto conllevaría serias desventajas: (i) los alineamientos serían muy poco específicos, es decir, aumentaría el número de alineamientos erróneos y (ii) sería virtualmente imposible alinear y calcular los niveles de metilación para regiones con una gran densidad de CpGs (como las islas CpGs), ya que suelen ser regiones no metiladas donde la mayoría de las citosinas se convertirán a timina. Otra posibilidad podría ser generar para cada lectura todas las posibles combinaciones para las conversiones T/C, pero sería un proceso muy lento y computacionalmente muy costoso.

Descartada la aproximación basada en el aumento del número máximo de desemparejamientos, existen dos aproximaciones para alinear lecturas tratadas con bisulfito. En primer lugar, puede modificarse la matriz de puntuación de los alineamientos de tal forma que considere los desemparejamientos C/T (citosina en la referencia y timina en la lectura) o G/A (guanina en la referencia y adenina en la lectura) como emparejamientos; o también puede adaptarse la secuencia de referencia a un alfabeto de tres letras de tal manera que se ajuste a la reducida complejidad de las librerías tratadas con bisulfito (Figura 3.3).

Aparentemente, la primera aproximación será más precisa ya que maneja una mayor complejidad de secuencia en las lecturas y en la referencia. Sin embargo, este método genera un sesgo, alineando un mayor número de lecturas en regiones metiladas, lo que resulta en una sobreestimación de la metilación en el genoma (Krueger, Kreck et al. 2012).

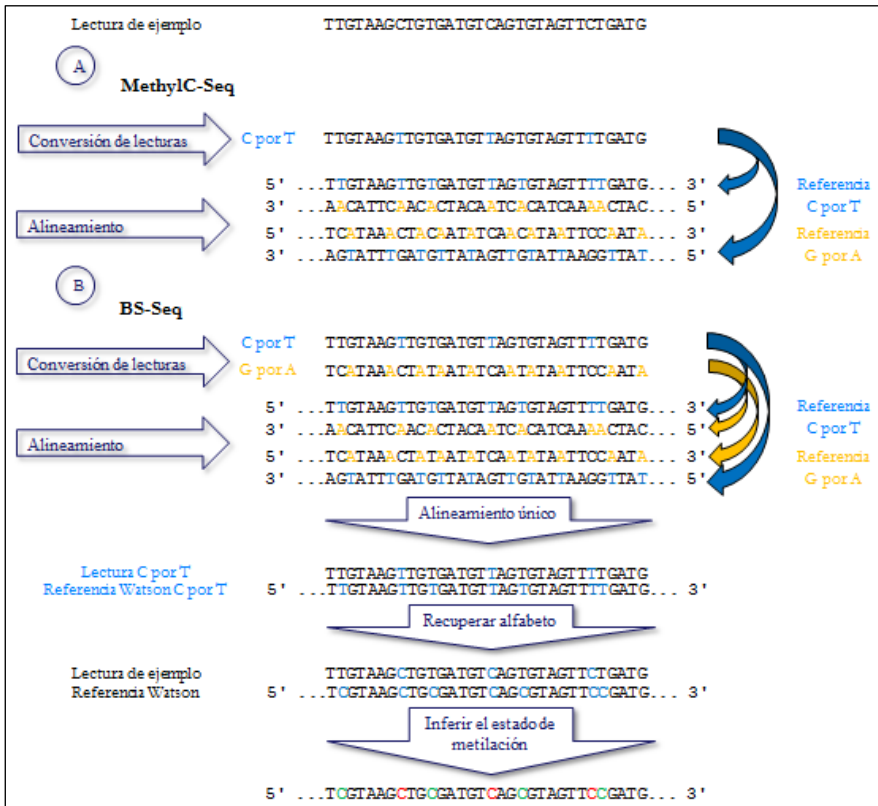


Figura 3.3. Alineamiento de lecturas tratadas con bisulfito basado en el modelo de referencia con tres letras. Para calcular los perfiles de metilación, en primer lugar las lecturas deben ser alineadas frente a un genoma de referencia. Una de las técnicas más utilizadas para lidiar con la reducción de complejidad causada por el bisulfito, es convertir tanto el genoma de referencia como las lecturas a un alfabeto de tres letras. En el protocolo *MethylC-Seq* (A), las lecturas vienen de las cadenas Watson (*BSW*) y Crick (*BSC*) convertidas por el bisulfito. Por lo tanto, todas las citosinas sin convertir en las lecturas se cambian por timinas y se las trata de alinear con la hebra directa de la referencia C>T y con la complementaria inversa de la referencia

G>A (flechas azules). Mientras que en el protocolo *BS-Seq* (B), las lecturas pueden alinearse con las dos hebras de ambas referencias (flechas azules y amarillas): referencia C>T (lecturas *BSW* y *BSWRC* en la **Figura 3.2**) y referencia G>A (lecturas *BSC* y *BSCRC* en la **Figura 3.2**). En caso de encontrar el alineamiento correcto, se revierten los cambios tanto en la referencia como en la lectura y se infieren los valores de metilación de manera directa: un desemparejamiento C/T indica una citosina no metilada (coloreadas en verde) y una citosina en ambas secuencias significa que la citosina se encuentra metilada (coloreadas en rojo). En el caso de las lecturas provenientes de las secuencias complementarias a las tratadas con bisulfito (sólo en el protocolo *BS-Seq*), el desemparejamiento G/A significaría la existencia de una citosina no metilada en la hebra complementaria.

3.2.3.2 Selección de los alineamientos

Una vez alineadas las lecturas, estas pueden clasificarse en tres tipos en función del resultado del alineamiento: (i) lecturas no alineadas, (ii) lecturas de alineamiento único y (iii) multilecturas (lecturas alineadas en múltiples localizaciones del genoma). Los alineamientos únicos se consideran correctos y serán incluidos en los siguientes pasos. En cuanto a las multilecturas, un gran número de protocolos de alineamiento las ignoran, resultando en una mayor sensibilidad que los métodos que las incluyen en los análisis subsiguientes. Sin embargo, esta estrategia limita los análisis a regiones únicas del genoma, descartando muchas familias multigénicas y la gran mayoría de los elementos repetidos, que pueden resultar de gran interés (Treangen and Salzberg 2012). Generalmente, las multilecturas pueden manejarse de dos maneras: incluyendo en el análisis todos los alineamientos posibles, o bien seleccionando el mejor alineamiento en función de algún criterio de calidad (desambiguación). La primera de las opciones supone aumentar de manera artificial la cobertura media de las lecturas sobre el genoma, y además introduce un gran número de alineamientos erróneos por cada lectura correctamente

alineada. La segunda, sin embargo, nos permitirá analizar esas regiones no únicas del genoma sin comprometer la calidad de los resultados finales (véase apartado 3.2.5).

En el caso de las lecturas tratadas con bisulfito, la desambiguación de multilecturas adquiere especial relevancia, ya que la reducción de la complejidad de secuencia y el aumento del espacio de búsqueda durante el alineamiento, aumentan la probabilidad de aparición de estas lecturas alineadas en múltiples posiciones. *NGSmethPipe*, implementa un método similar al utilizado por la herramienta *miRanalyzer* (Hackenberg, Sturm et al. 2009, Hackenberg, Rodriguez-Ezpeleta et al. 2011), que se resume en la Figura 3.4.

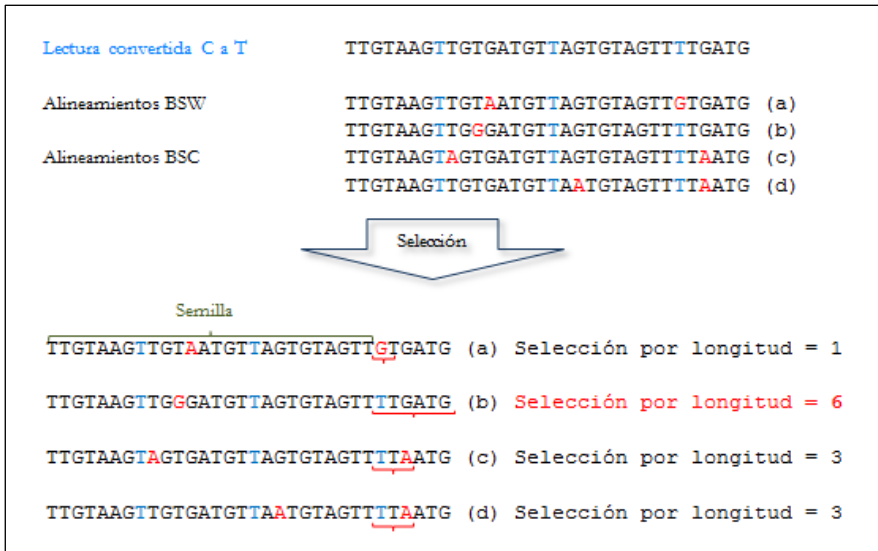


Figura 3.4. Desambiguación mediante extensión de la semilla. El alineamiento se realiza mediante *Bowtie*, usando una semilla por defecto de 26 pb y con 1 desparejamiento como máximo en la región semilla. En primer lugar, se preseleccionan los dos mejores alineamientos para cada secuencia de referencia (en este caso *BSW* y *BSC*), tomándose los alineamientos con menor número de desparejamientos dentro de la semilla y con menores valores de calidad en los desparejamientos a lo largo de la lectura. Entonces, se mide la longitud desde el final de la

semilla hasta el próximo desemparejamiento, seleccionando como mejor alineamiento aquel que presente la distancia más larga (el (b) en el ejemplo de la figura). Las citosinas convertidas se representan en azul, mientras que los desemparejamientos se encuentran coloreados en rojo, así como las distancias desde el final de la semilla al siguiente desemparejamiento.

3.2.4 Mejora en la detección de adaptadores

Como se comentó en el apartado 3.2.2, la eliminación de adaptadores es fundamental para alinear correctamente el mayor porcentaje posible de lecturas. Por lo tanto, su detección y posterior recorte debe realizarse de manera eficiente. *NGSmethPipe* implementa un método secuencial propuesto por Lister y colaboradores (Lister, Pelizzola et al. 2009), en el que las lecturas se limpian por calidad de secuenciación antes de buscar el adaptador. De esta manera se aumenta considerablemente la detección de adaptadores, como se muestra en la Figura 3.5.

La búsqueda del adaptador sin previo recorte por calidad (Figura 3.5, B), encuentra y recorta el adaptador en un 66.6% de las lecturas. Sin embargo, si se recorta previamente por calidad (Figura 3.5, A), los adaptadores se detectan en un 13.2% más de lecturas. Además, en un 22.2% de los casos, las lecturas se recortan por calidad pero no se localiza el adaptador, probablemente debido a que los adaptadores se encuentran dentro de la región recortada. En definitiva, usando la metodología implementada por *NGSmethPipe* se limpian el 99.8% de las lecturas, recortando el adaptador y eliminando aquellas posiciones con muy baja calidad de secuenciación que pueden influir en el alineamiento.

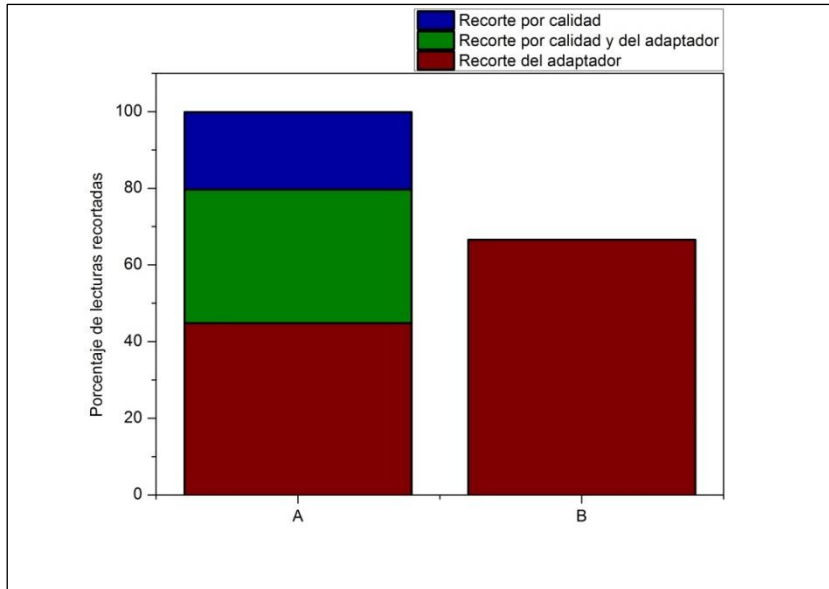


Figura 3.5. Eficiencia en el recorte de adaptadores. Se muestra el porcentaje de lecturas recortadas (conjunto de datos descritos en el apartado 3.2.1.1, donde todas las lecturas poseen el adaptador en 3') por calidad (azul), por calidad y adaptador (verde) o sólo adaptador (rojo). Se muestran dos metodologías diferentes, (A) *NGSmethPipe* y (B) búsqueda iterativa del adaptador sin previo recorte por calidad.

3.2.5 Efecto de la desambiguación de multilecturas

La desambiguación de multilecturas conduce a un aumento del número de lecturas correctamente alineadas, aunque como consecuencia, también aumenta el porcentaje de lecturas erróneamente mapeadas, como ya se comentó en el apartado 3.2.3.2 y se ilustra en la Figura 3.6. Este hecho se observa para todas las longitudes de semilla utilizadas (20-50 pbs). Tomando como referencia la longitud de la semilla por defecto (26 pbs), se han desambiguado correctamente unas 75,000 multilecturas, sin sobrepasar el 1% de lecturas erróneas. Esta comparación se repitió para el *contig GL000109.1* del cromosoma 12 de

hg19, encontrándose resultados semejantes (90,000 lecturas correctamente desambiguadas sin sobrepasar el 1% de error para los parámetros por defecto).

Más allá del número de lecturas que consigan mapearse correctamente, la importancia de utilizar el máximo número de multilecturas posibles, reside en la necesidad de obtener información de metilación en zonas del genoma con elementos repetidos. La función de la metilación de los elementos repetidos está ampliamente demostrada; por ejemplo, en regiones pericentroméricas contribuye al correcto alineamiento y a la segregación e integridad de los cromosomas durante la mitosis, así como a mantener inactivos los elementos transponibles, ya que si permaneciesen activos podrían resultar letales (Smith and Meissner 2013). Para demostrar la idoneidad del método, se han calculado los porcentajes de solapamiento de las multilecturas desambiguadas correctamente con elementos repetidos. Aproximadamente el 90% de las multilecturas desambiguadas solapan al menos en una base con elementos repetidos conocidos (Smit, Hubley et al. 1996-2010), mientras que el porcentaje de solapamiento de las lecturas únicas con estos elementos repetidos es del 42%, demostrando así la capacidad de la extensión de la semilla como método para alinear lecturas en secuencias repetidas del genoma.

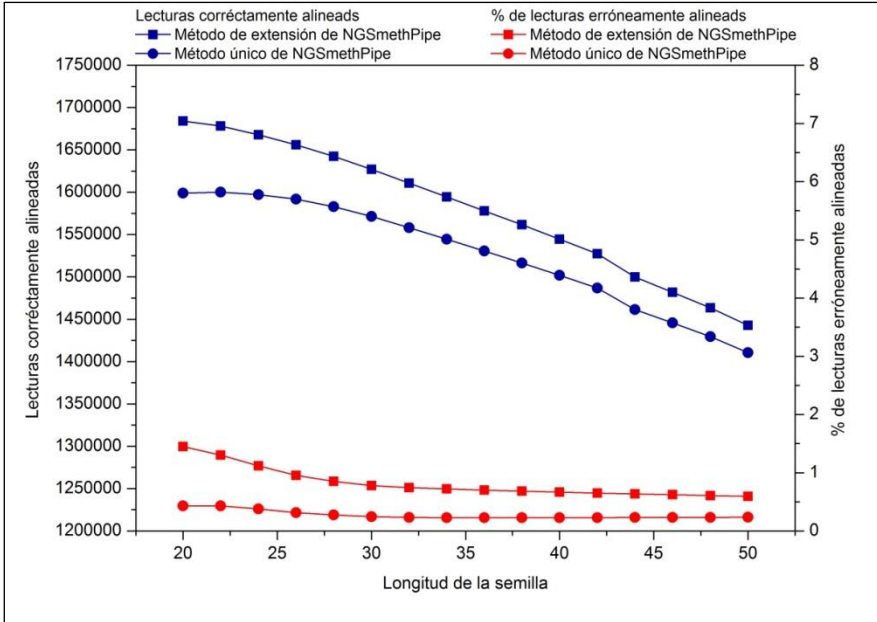


Figura 3.6. Comparación entre los alineamientos de lecturas únicas y la desambiguación de multilecturas mediante la extensión de la semilla. Se han simulado y alineado 2,000,000 de lecturas, tal y como se describe en el apartado 3.2.1.1, para el *contig GL000022.1* del cromosoma 2 de hg19. En azul se representan el número de alineamientos correctos (eje izquierdo de ordenadas) y en rojo el porcentaje de lecturas erróneamente alineadas (eje derecho de ordenadas). Estos valores se representan para diferentes longitudes de la semilla y dos metodologías diferentes: un método basado en la extensión de la semilla del alineamiento (cuadrados) y otro método que sólo selecciona lecturas con alineamientos únicos (círculos), ambos implementados por *NGSmethPipe*.

3.2.6 Comparación con otros métodos

Durante el desarrollo de esta Tesis, han ido apareciendo numerosas aplicaciones que abordan la problemática del alineamiento de lecturas tratadas con bisulfito: *RMAP* (Smith, Chung et al. 2009), *BSMAP* (Xi and Li 2009), *mrsFAST* (Hach, Hormozdiari et al. 2010), *SOCS-B* (Ondov, Cochran et al. 2010), *BS Seeker* (Chen, Cokus et al. 2010), *MethylCoder*

(Pedersen, Hsieh et al. 2011), *Bismark* (Krueger and Andrews 2011) y *BRAT-BW* (Harris, Ponts et al. 2012)). Además de estas aplicaciones, se han utilizado y propuesto numerosas metodologías (Cokus, Feng et al. 2008, Lister, O'Malley et al. 2008, Lister, Pelizzola et al. 2009, Bock, Tomazou et al. 2010, Gu, Bock et al. 2010, Harris, Wang et al. 2010). En los siguientes apartados se comparará *NGSmethPipe* con los programas actualmente disponibles.

3.2.6.1 Características

Actualmente, sólo un programa (*MethylCoder*) puede alinear los dos tipos de lecturas existentes: lecturas basadas en secuencias (*Illumina*, *Roche 454*) y lecturas basadas en códigos de color (*SOLiD*). Excepto *SOCS-B*, el resto de herramientas sólo pueden utilizarse para alinear lecturas basadas en secuencias. Debido a la manera en que se determinan experimentalmente las lecturas basadas en códigos de color, la detección de variaciones en la secuencia se realiza de manera muy eficiente, sin embargo la conversión de las lecturas a un alfabeto de tres letras para evitar el sesgo comentado en el apartado 3.2.3.1 es inviable (Krueger, Kreck et al. 2012). Por ello, es recomendable el uso de lecturas basadas en secuencias durante los estudios de metilación, de hecho la gran mayoría de estos estudios ya se basan en lecturas basadas en secuencias.

Otra diferencia importante, es la capacidad de los programas para procesar las diferentes librerías de lecturas basadas en secuencias procedentes de los protocolos experimentales (*MethylC-Seq* y *BS-Seq*): simples (*single-end*) y emparejadas (*pair-end*). Todos los programas

implementan el alineamiento de lecturas simples para el protocolo *MethylC-Seq*, sin embargo *mrsFAST*, *RMAP* y *MethylCoder* no implementan el protocolo no direccional (*BS-Seq*). En el caso de *MethylCoder*, en su página web se encuentra disponible una herramienta para convertir las lecturas no direccionales en direccionales. En cuanto a las lecturas emparejadas (*pair-end*) con secuenciación direccional, todos los métodos pueden alinearlas, salvo *RMAP*, *SOCS-B* y *BS Seeker*; mientras que las no direccionales sólo pueden procesarlas *Bismark* y *NGSmethPipe*.

En cuanto al método de alineamiento, ya se comentó la idoneidad del método de 3-letras utilizado por *mrsFAST*, *BS Seeker*, *MethylCoder*, *Bismark* y *NGSmethPipe*. Una vez alineadas las lecturas, también es muy importante la desambiguación de las multilecturas implementada por *Bismark* y *NGSmethPipe*. A su vez, también se comentó la importancia del preprocesado de las lecturas en alineadores que utilizan toda la secuencia de las lecturas (*SOCS-B* y *BS Seeker*) o desambiguan multilecturas a partir de alineamientos basados en semillas (*Bismark* y *NGSmethPipe*). De estos programas, sólo *NGSmethPipe* preprocesa las lecturas de manera eficiente, mientras que: *SOCS-B* recorta por calidad, *BS Seeker* recorta el adaptador y *Bismark* no incluye ningún tipo de preprocesamiento. Esta comparativa se resume en la Tabla 3.1.

| CARACTERÍSTICAS/PROGRAMAS | <i>mrsFAST</i> | <i>RMAP</i> | <i>SOCS-B</i> | <i>BS Seeker</i> | <i>BSMAP</i> | <i>BRAT-BW</i> | <i>MethylCoder</i> | <i>Bismark</i> | <i>NGSmethPipe</i> |
|---------------------------------------|----------------|-------------|---------------|------------------|--------------|----------------|---------------------|----------------|--------------------|
| Lecturas de entrada | BS | BS | BC | BS | BS | BS | BS/BC | BS | BS |
| Multiprocesadores | No | No | Si | Si* | Si | No | No | Si* | Si |
| Lenguaje | C | C++ | C++ | Python | C++ | C++ | Python/C | Perl | Perl |
| Recorte de lecturas | No | No | Si | No | Si | Si | No | No | Si |
| Recorte del adaptador | No | No | No | Si | Si | No | No | No | Si |
| Alineador | <i>mrsFAST</i> | <i>RMAP</i> | <i>SOCS</i> | <i>Bowtie</i> | <i>SOAP</i> | <i>BRAT</i> | <i>Bowtie/GSNAP</i> | <i>Bowtie</i> | <i>Bowtie</i> |
| Método | 3-letas | 4-letas | 4-letas | 3-letas | 4-letas | 2-letas | 3-letas/4-letas | 3-letas | 3-letas |
| Alineamiento basado en semilla | Si | Si | No | No | Si | Si | Si/Si | Si | Si |
| Q | No | Si | Si | No | No | No | Si/No | Si | Si |
| Simple (no direccional) | No | No | Si | Si | Si | Si | Si/Si | Si | Si |
| Emparejadas (direccional) | Si | No | No | No | Si | Si | Si/Si | Si | Si |
| Emparejadas (no direccional) | No | No | No | No | No | No | No | Si | Si |

Tabla 3.1. Características de los alineadores para lecturas tratadas con bisulfito. La fila “Lecturas de entrada” indica el tipo de lecturas que admite cada método: BS (basadas en secuencias) y BC (basadas en color). El asterisco (*) en la fila de multiprocesadores, significa que estos programas realizan el procesamiento múltiple sólo durante el alineamiento con *Bowtie*. La fila del “alineamiento basado en semilla”, especifica los métodos que realizan su alineamiento mediante el método de semilla. La fila “Q” indica que métodos utilizan la calidad de las secuencias durante el alineamiento. La fila “Simple (no direccional)” especifica los métodos que alinean lecturas simples para el protocolo *BS-Seq* (no direccional), ya que todos alinean lecturas singulares para el protocolo *MethylC-Seq* (direccional). Por último, las filas “Emparejadas” indican los métodos que alinean lecturas emparejadas provenientes de los protocolos *MethylC-Seq* (direccional) y *BS-Seq* (no direccional).

3.2.6.2 Velocidad de procesado

Dado que un experimento de resecuenciación puede llegar a producir fácilmente unos 3,000 millones de lecturas, claramente la velocidad de procesado es importante a la hora de elegir el programa de alineamiento. Los métodos basados en alfabetos de 3-letras que usan el alineador *Bowtie* (Langmead, Trapnell et al. 2009), parecen ser en general más rápidos que los programas que utilizan el método que modifica la matriz de puntuación del alineamiento (Frith, Mori et al. 2012). De los tres programas analizados en el estudio citado (*BS Seeker*, *MethylCoder*, *Bismark*), *BS Seeker* es el más rápido de los tres, probablemente debido a que la forma de alinear las lecturas es diferente a *MethylCoder* y *Bismark*.

Un aspecto importante para aumentar la velocidad de los alineamientos es la capacidad para procesar los datos en varios procesadores de manera simultánea (multiprocesado). *Bismark* y *BS Seeker* sólo utilizan el multiprocesado de *Bowtie* y además no permiten seleccionar el número de procesadores a utilizar (usando siempre un hilo por secuencia de referencia); mientras que *MethylCoder* no permite el uso de múltiples procesadores de manera simultánea. Sin embargo, *NGSmethPipe* paraleliza todo el proceso (desde el preprocesado hasta la desambiguación de multilecturas), y permite seleccionar tanto el número de hilos a utilizar como el número de lecturas procesadas por hilo. Esta característica permite adaptar la velocidad y el consumo de memoria de *NGSmethPipe* a las necesidades del usuario.

3.2.6.3 Eficiencia de los alineamientos

Aparte de la velocidad y las diferentes funciones de los programas disponibles, es de gran importancia la eficacia del método para alinear lecturas. En este apartado, se comparará *NGSmethPipe* con *Bismark*, que parece ser, por las comparaciones previas (Frith, Mori et al. 2012), el mejor método en términos de velocidad, funcionalidad y eficacia. Al igual que en el apartado 3.2.5 (Figura 3.6), la eficiencia del método se basa en la comparación entre lecturas correctamente alineadas y el porcentaje de lecturas erróneamente mapeadas (Figura 3.7). En general, *Bismark* alinea correctamente un mayor número de lecturas que *NGSmethPipe*. Sin embargo, salvo a longitudes largas de semilla, donde el número de lecturas alineadas correctamente es muy reducido, el porcentaje de lecturas erróneamente alineadas es mucho mayor en *Bismark* que en *NGSmethPipe*.

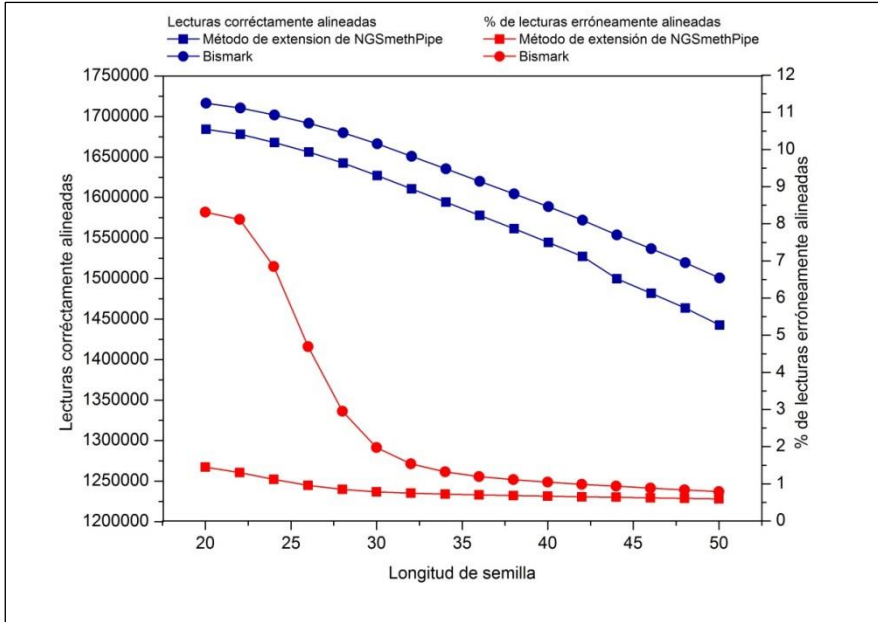


Figura 3.7. Comparación entre NGSmethPipe y Bismark. Se han simulado y alineado 2,000,000 de lecturas, tal y como se describe en el apartado 3.2.1.1, para el *contig GL000022.1* del cromosoma 2 de hg19. En azul se representan el número de alineamientos correctos (eje izquierdo de ordenadas) y en rojo el porcentaje de lecturas erróneamente alineadas (eje derecho de ordenadas). Estos valores se representan para diferentes longitudes de la semilla y dos programas diferentes NGSmethPipe (cuadrados) y Bismark (círculos). En ambos programas se han utilizado sus respectivos parámetros por defecto.

Analizando los resultados en detalle (Tabla 3.2), observamos que para igual longitud de semilla ($l=26$), Bismark alinea correctamente unas 35,000 lecturas más que NGSmethPipe, pero su porcentaje de error casi se quintuplica (de un 0.95% en NGSmethPipe a un 4.7%). Y cuando ambos métodos se igualan según el número de alineamientos correctos: Bismark ($l=32$) con 1,650,851 y NGSmethPipe ($l=26$) con 1,656,092 lecturas correctamente alineadas, Bismark (1.5%) sigue presentando un mayor porcentaje de error que NGSmethPipe. Al igual que en el apartado

3.2.5, la comparación se repitió para el *contig* *GL000109.1* de hg19, encontrándose resultados semejantes.

| PROGRAMAS (longitudes de semilla) | # alineamientos correctos | % alineamientos erróneos |
|-----------------------------------|---------------------------|--------------------------|
| <i>NGSmethPipe</i> (l=26) | 1,656,092 | 0.95 |
| <i>Bismark</i> (l=26) | 1,691,644 | 4.7 |
| <i>Bismark</i> (l=32) | 1,650,851 | 1.5 |

Tabla 3.2. Comparación entre *NGSmethPipe* y *Bismark*, para la misma longitud de semilla (l=26) y el mismo número de alineamientos correctos (aprox. 1.65e6).

- 3.3 Barturen G, Rueda A, Oliver JL et al. (2013) MethylExtract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000Research* 2013, 2:217.
-

Dirección de acceso PubMed:

<http://www.ncbi.nlm.nih.gov/pubmed/24627790.2>

Dirección de publicación:

<http://f1000research.com/articles/2-217/v2>

Página web de la aplicación:

<http://bioinfo2.ugr.es/MethylExtract/>

Breve descripción del programa:

MethylExtract permite inferir los niveles de metilación para citosinas individuales y las variaciones de un solo nucleótido a partir de lecturas tratadas con bisulfito, previamente alineadas con un genoma de referencia. Las características principales están implementadas en un solo programa escrito en Perl, totalmente paralelizable y con control de consumo de memoria, haciéndolo muy sencillo de utilizar y adaptable en función de la infraestructura informática del usuario. Además, *MethylExtract* incluye dos programas auxiliares para asignar estadísticamente la probabilidad del fallo en la conversión del bisulfito: (i) estimando la tasa de conversión del bisulfito a partir de un genoma completamente no metilado y (ii) calculando la probabilidad de que el nivel de metilación observado se encuentre fuera de un intervalo de confianza para el nivel real de metilación.



METHOD ARTICLE

REVISED **MethylExtract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data [v2; ref status: indexed, <http://f1000r.es/301>]**

Guillermo Barturen^{1,2}, Antonio Rueda^{1,2}, José L. Oliver^{1,2}, Michael Hackenberg^{1,2}

¹Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Granada, 18071, Spain

²Lab. de Bioinformática, Inst. de Biotecnología, Centro de Investigación Biomédica, Granada, 18016, Spain

v2 **First Published:** 15 Oct 2013, 2:217 (doi: 10.12688/f1000research.2-217.v1)
Latest Published: 21 Feb 2014, 2:217 (doi: 10.12688/f1000research.2-217.v2)

Abstract

Whole genome methylation profiling at a single cytosine resolution is now feasible due to the advent of high-throughput sequencing techniques together with bisulfite treatment of the DNA. To obtain the methylation value of each individual cytosine, the bisulfite-treated sequence reads are first aligned to a reference genome, and then the profiling of the methylation levels is done from the alignments. A huge effort has been made to quickly and correctly align the reads and many different algorithms and programs to do this have been created. However, the second step is just as crucial and non-trivial, but much less attention has been paid to the final inference of the methylation states. Important error sources do exist, such as sequencing errors, bisulfite failure, clonal reads, and single nucleotide variants.

We developed *MethylExtract*, a user friendly tool to: i) generate high quality, whole genome methylation maps and ii) detect sequence variation within the same sample preparation. The program is implemented into a single script and takes into account all major error sources. *MethylExtract* detects variation (SNVs – Single Nucleotide Variants) in a similar way to *VarScan*, a very sensitive method extensively used in SNV and genotype calling based on non-bisulfite-treated reads. The usefulness of *MethylExtract* is shown by means of extensive benchmarking based on artificial bisulfite-treated reads and a comparison to a recently published method, called *Bis-SNP*.

MethylExtract is able to detect SNVs within High-Throughput Sequencing experiments of bisulfite treated DNA at the same time as it generates high quality methylation maps. This simultaneous detection of DNA methylation and sequence variation is crucial for many downstream analyses, for example when deciphering the impact of SNVs on differential methylation. An exclusive feature of *MethylExtract*, in comparison with existing software, is the possibility to assess the bisulfite failure in a statistical way. The source code, tutorial and artificial bisulfite datasets are available at <http://bioinfo2.ugr.es/MethylExtract/> and <http://sourceforge.net/projects/methylextract/>, and also permanently accessible from [10.5281/zenodo.7144](https://zenodo.org/record/7144).

Article Status Summary**Referee Responses**

| Referees | 1 | 2 | 3 |
|---|------------|------------|------------|
| v1 published 15 Oct 2013 | report | report | report |
| v2 published 21 Feb 2014 REVISED | | | |

1 Michael Stadler, Friedrich-Miescher
Institute for Biomedical Research
Switzerland

2 Jörn Walter, University of Saarland
Germany

3 Felix Krueger, Babraham Institute UK

Latest Comments

Michael Hackenberg, University of Granada,
Spain
17 Feb 2014 (V1)

Corresponding authors: José L. Oliver (oliver@ugr.es), Michael Hackenberg (hackenberg@ugr.es)

How to cite this article: Barturen G, Rueda A, Oliver JL *et al.* (2014) MethylExtract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data [v2; ref status: indexed, <http://f1000r.es/301>] *F1000Research* 2014, 2:217 (doi: 10.12688/f1000research.2-217.v2)

Copyright: © 2014 Barturen G et al. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

Grant information: This work was supported by the Spanish Government [BIO2008-01353 to JLO and BIO2010-20219 to MH], and Basque country 'AE' grant (GB).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: No competing interests were disclosed.

First Published: 15 Oct 2013, 2:217 (doi: 10.12688/f1000research.2-217.v1)

First Indexed: 27 Jan 2014, 2:217 (doi: 10.12688/f1000research.2-217.v1)

REVISED Amendments from Version 1

This new version comprises several changes

- 1) The *MethylExtract* software was updated to version 1.5 including several important changes: i) compatibility to all Perl versions, ii) BAM files can be read directly (needs *samtools* installed), iii) several FLAG values can be given in order to suite for paired-end design
- 2) Several new benchmarking experiments were carried out like suggested by the referees: i) comparison of methylation profiling between *MethylExtract* and *Bis-SNP* using relaxed criteria (methylation values are considered as correct if they deviate only by 10% and 20% respectively from the real value), ii) analysis of artificial BS sequencing data for 5x and 35x coverage, iii) new runtime comparison which is based on the exactly same input files, iv) brief and descriptive comparison of results obtained from "real world" data (see comments to the referees)
- 3) The tutorial was completely revised including a new figure explaining how *MethylExtract* treats and indicates variant positions in the output files.

See referee reports

Introduction

DNA methylation at the cytosine carbon 5 position (5mC) is an important epigenetic mark in eukaryotic cells that is predominantly found in CpG or CpHpG (H = A,C,T) sequence contexts¹. Epigenetic modifications at the DNA level play important roles in embryonic development^{2,3}, transcription⁴, chromosome stability⁵, genomic imprinting⁶ and in the silencing of transposons in plants⁷. Furthermore, aberrant methylation is involved in the appearance of several disorders as cancer, immunodeficiency or centromere instability⁸. The methylation pattern along the genome sequence carries biologically relevant information. For example: methylated promoter regions are generally associated with silenced transcription and DNA methylation in the gene body of transcribed genes is often increased⁹. Given these findings, the generation of high quality whole genome methylation maps at a single cytosine resolution is an important step towards the understanding of how DNA methylation is involved in the regulation of gene expression or the generation of a pathologic phenotype. In addition, methylation maps may provide new insights into how the methylation patterns themselves are established.

Several high-throughput techniques have been developed able to generate whole genome methylation maps. In general, the techniques consist of a methylation-sensitive pre-treatment and a read-out step. The pre-treatments generally consist of digestion by methyl-sensitive endonucleases, methyl-sensitive immunoprecipitation or bisulfite conversion, while the read-out of the methylation information is done by hybridization, amplification or sequencing⁹. Recently, several promising techniques have been developed that link the bisulfite conversion with High-Throughput Sequencing (MethylC-Seq¹⁰, BS-Seq¹¹ or RRBS¹²). Briefly, the bisulfite treatment converts un-methylated cytosines into uracil (converted to thymine after PCR amplification) while leaving methylcytosines unconverted. After sequencing the bisulfite-treated genomic DNA, the methylation state can be recovered from the sequence alignments.

Therefore, the methylation profiling from High-Throughput Bisulfite Sequencing data can be divided into two steps: the alignment of the reads, and the read-out of the methylation levels from the alignment. The alignment of bisulfite-treated reads is highly non-trivial due to the reduced sequence complexity given that all cytosines except methylcytosines are converted to thymines. This challenge has been extensively addressed over the last years and several algorithms are available that either align the reads in a 3-letter space or adapt the alignment scoring matrix in order to account for the C/T conversions. Among these algorithms are *BSMAP*¹³, *Bismark*¹⁴, *MethylCoder*¹⁵, *NGSmethPipe*¹⁶, *BS Seeker*¹⁷, *Last*¹⁸ and *BRAT-BW*¹⁹. Note that some of these tools are not just alignment programs but can, in addition, perform the profiling of the methylation levels such as *Bismark* and *MethylCoder*. After alignment, the methylation states can be recovered: C/T mismatches indicate un-methylated cytosines while C/C matches reveal methylcytosines. However, several error sources—like sequencing errors, clonal reads, sequence variation, bisulfite failure and mis-alignments—can lead to a wrong inference of the methylation levels^{16,18,20}. For example, C→T or T→C (on converted cytosines) sequencing errors would be incorrectly interpreted as un-methylated or methylated respectively biasing the results towards lower or higher methylation levels. On the other side, bisulfite failures bias the methylation levels only to higher levels; un-methylated cytosines are not converted and therefore detected as methylcytosines. The existence of sequence variation is another important error source that was traditionally disregarded in the data analysis of whole genome bisulfite sequencing (WGBS) experiments. A C/T SNV would be interpreted as un-methylated cytosine. Given that over two thirds of all Single Nucleotide Polymorphisms (SNPs) occur in a CpG context, having two alleles: C/T or G/A²¹, sequence variation needs to be addressed as an important error source. A C/T SNV manifests on the complementary DNA strand as an adenine, while bisulfite deamination does not affect the guanine on the complementary strand (see Figure 1). This fact allows in principle to distinguish between sequence variation and bisulfite conversion and therefore to i) avoid wrong inference of the methylation state due to sequence variation and ii) detect sequence variation in the same sample preparation as the methylation levels. Profiling the methylation levels and the genotype of the sample from one experiment will be a very important step towards "putting the DNA back into methylation"²², as the impact and importance of certain DNA sequences on the methylation levels have been recently demonstrated²³. To our knowledge, the first program that performed a threshold-based detection of sequence variation in bisulfite sequencing experiments was *NGSmethPipe*¹⁶. This program detects sequence variation mainly to avoid wrong inference of the methylation level reporting those genome positions in the output. Only recently, the first state-of-the-art SNP calling algorithm based on the *Genome Analysis Toolkit* (*GATK*)²⁴ was implemented to detect both methylation levels and sequence variation at high precision in a single experiment (*Bis-SNP*).

Here we present *MethylExtract*, a multi-threaded tool for methylation profiling and sequence variation detection from alignments in standard BAM/SAM format²⁵. The tool is able to generate high quality methylation maps taking into account SNVs, putative bisulfite failures, reducing also the contribution of sequencing errors by means of the base quality PHRED score^{26,27}. In addition, it detects

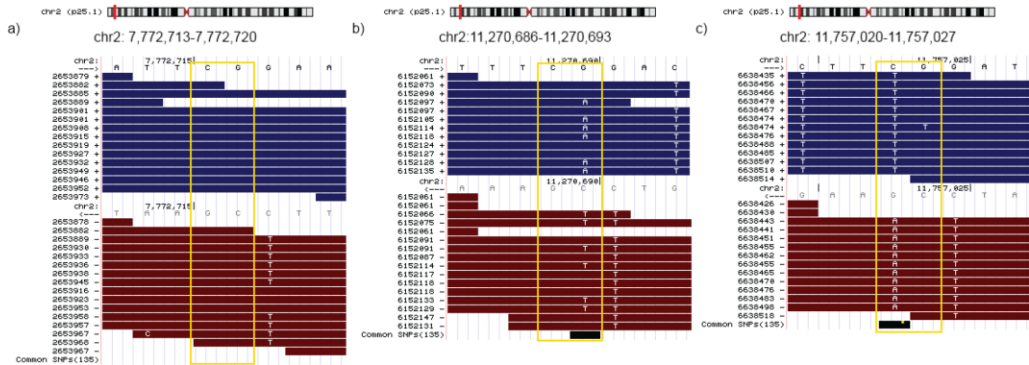


Figure 1. SNV detection in bisulfite converted reads. Sequence variation can be detected for a cytosine position analyzing the nucleotide frequency at the same position but on the complementary strand. Bisulfite conversion does not affect the guanine on the complementary strand, therefore the presence of any other base (H=A,C,T) might indicate the existence of an SNV. The figure illustrates three different situations: **(a)** a methylated cytosine in a CpG context without sequence variation (all reads that map to the position independently of the strand carry a cytosine in the corresponding position), **(b)** a heterozygous SNV (genotype C/T, SNV detected on the '+' strand) and **(c)** a homozygous SNV (genotype T/T, SNV detected on the '-' strand). The example in **b)** shows a heterozygous SNV; the 6 reads with A/G mismatch from a total of 11 reads mapping the position indicate a heterozygous variation. Furthermore, we can conclude that the cytosine allele is methylated (7 reads with C/C matches to the '-' strand). The case illustrated in part **c)**, shows 12 reads that show C/T mismatch ('+' strand in blue in the upper part). Without looking at the complementary strand, the inference would be a completely unmethylated cytosine. However, the 11 reads that map to the complementary strand show an A/G mismatch at the corresponding position (we would expect guanines in the case of bisulfite conversion). Note that on bisulfite treated datasets only G/A mapped on the '+' strand and C/T on the '-' strand (referred to the '+' strand) can be used for SNV calling purposes. The figure was generated using the UCSC Genome Browser⁴¹.

sequence variation based on *VarScan* methodology²⁸ reporting all detected SNVs in VCF format²⁹. Therefore, from a single sequencing experiment, *MethylExtract* obtains both the methylation levels and the sequence variation, which will increase the reliability of downstream analyses²³. We confirm its usefulness using extensive artificial BS data and a comparison to *Bis-SNP*. We show that while its SNV-calling performance is slightly less specific but more sensitive compared to *Bis-SNP*, *MethylExtract* performs better in methylation profiling, is easier to use and over twice as fast on a typical whole genome experiment.

Implementation Scope and workflow

MethylExtract is implemented in Perl and consists of one main script and two auxiliary scripts that are exclusively dedicated to the statistical assessment of the bisulfite error. In general, the program takes standard BAM/SAM file format as input (previously aligned reads) and performs methylation profiling and SNV calling taking into account several error sources like sequencing errors, clonal reads and bisulfite failures. *MethylExtract* writes two output files. First, the methylation information for each cytosine including the coordinates, sequence context (CG, CHG, CHH), number of methylcytosines, read coverage and mean base quality (PHRED) score. The second output file reports the sequence variation in standard VCF format²⁹.

Frequently, whole genome bisulfite experiments include the estimation of the bisulfite conversion rate through a completely unmethylated genome (lambda phage for example). If the bisulfite

conversion rate is known, statistical tests can be applied to infer whether an observed methylation level might be only due to failures of bisulfite conversion. The two auxiliary scripts allow i) estimating the bisulfite conversion rate by mapping the bisulfite-treated reads from the un-methylated genome only and ii) to apply a binomial statistics based test to infer the probability that the "real" methylation value lies within a given interval of the observed value.

Duplicated reads

The PCR step can lead to duplicated (clonal) reads, thus causing a bias in the read coverage. This bias might lead to incorrect inference at positions with allele-specific methylation (genetic imprinting), sequence variation, hemi-methylation, sequencing errors, bisulfite failure or those that are heterogeneous over the cell population. Frequently, the start coordinates of the alignments are used to eliminate duplicates like in *SAMtools*²⁵, adding a criterion to keep the best read among the duplicates. However, those approaches do not take into account that at a heterozygous locus two reads with the same start coordinate could represent two different alleles, thus not being clonal reads. The same applies for loci with genetic imprinting or hemi-methylation. To avoid the elimination of meaningful biological information, *MethylExtract* groups all reads that start at the same position in the genome and that have the same seed nucleotides with $Q \geq \text{'minQ'}$; and selects the read that has the highest number of bases with $Q \geq \text{'minQ'}$ (by default 'minQ' = 20) and the longest read in case of equal number of high quality positions. Furthermore, if there are multiple reads with the same selection values, only one will be selected in a random way. Two non-identical reads that align to exactly the same position in the chromosome can represent either

the existence of sequence variation or putative clonal reads with a sequencing error in at least one read (disregarding mis-alignments). To restrict the impact of sequencing errors we used only the seed region of the read, i.e. the region with the highest quality. The seed is defined as those nucleotides at the 5' end of the read (first 26 nt by default) that have a higher PHRED score than 'minQ'.

Note that the two types of methods, the ones that use only the coordinates and our method using the coordinates and the sequence, have advantages and disadvantages. If the sequence differences are considered, biological meaningful information like sequence variation, genetic imprinting or hemi-methylation is maintained; however, our approach will be vulnerable to sequencing errors and bisulfite errors. The default option is to not perform the detection of duplicated reads, and thus any of the publically available tools can be used optionally to remove clonal reads prior to run *MethylExtract*.

5' end trimming

The first nucleotides can be removed from the 5' end of the read (3 bp for the *MspI* restriction sites of non-directional reduced representation bisulfite sequencing (RRBS) protocol), as also implemented by *Bismark*¹⁴.

Eliminating reads with putatively high bisulfite conversion failure

The bisulfite conversion error probability of un-methylated cytosines is usually below 1% in modern protocols. However, even for such low values, some positions could be incorrectly profiled, i.e. some methylated cytosines are actually un-methylated. *MethylExtract* implements a method proposed by Lister *et al.*³⁰ to detect those reads with a high number of un-converted cytosines. By default, it eliminates reads with at least 90% of (presumably) unconverted cytosines in non-CpG contexts (Lister *et al.* used ≥ 3 methylated non-CpG cytosines). The default threshold is very conservative and only a rather small fraction of reads will be eliminated. Caution is needed if the user knows that the analyzed species (plants) or tissues (e.g. embryonic stem cells) contain an elevated number of DNA methylation in non-CpG contexts. In those cases, this step should be better skipped as otherwise a bias will be introduced into the analysis.

Controlling sequencing errors

Sequencing errors are another important cause of incorrect methylation profiling (and SNV calling). The contribution of the individual bases can be controlled by means of the assigned PHRED score (i.e. an upper limit of sequencing error contribution to the wrongly inferred methylation states). For example, when setting PHRED score ≥ 20 , thus accepting bases with a probability < 0.01 to be incorrectly called, the contribution of sequencing errors to the overall error would be less than 1%. By default, *MethylExtract* sets the minimum PHRED score to 20 ('minQ' parameter) which is then used for both methylation profiling and SNV calling (see below on the determination of the default values).

SNVs detection

SNVs are the most disregarded error source in the analysis of whole genome bisulfite sequencing data. Most tools would interpret a C

to T substitution as an un-methylated cytosine, although a certain number of them are actually SNVs, and therefore this inference would be wrong. A C/T SNV manifests on the complementary DNA strand as an adenine, while bisulfite deamination does not affect the guanine on the complementary strand³¹ (Figure 1). The SNVs detection algorithm implemented in *MethylExtract* is an adaptation of the widely used *varScan* algorithm²⁸. The main difference compared to SNV calling from non-bisulfite-treated DNA is the reduced amount of sequence information that can be used to detect sequence variation. The bisulfite treatment converts the un-methylated cytosines into thymines, and therefore, at cytosine positions nucleotides that might result from the bisulfite conversion cannot be used to detect sequence variation. For adenine and thymine, both strands can be used like in re-sequencing experiments. The algorithm works as follows: i) filter out positions that are covered by fewer reads than the minimum read depth ('minDepthSNV') – by default 'minDepthSNV' is set to 1, thus analyzing all positions that are covered by at least one read; ii) calculate the nucleotide frequencies including all base calls that pass the minimum PHRED score threshold ('minQ'); iii) discard nucleotides with frequencies below a given threshold ('varFraction'); iv) calculate a *p-value* for the variant positions (more than two nucleotides above 'varFraction') by means of Fisher's exact test, v) only those positions with a *p-value* below a given threshold are considered as SNVs ('maxPval'), and vi) the two nucleotides with the highest frequencies are determined as the putative genotype of the sample at this position. Detected sequence variation is reported in VCF output format, which can be used as input for SNP-annotation programs³² or *VCFtools*²⁹.

Statistical assessment of the bisulfite conversion error

Bisulfite conversion failure has been addressed using binomial statistics for the two possible outcomes; methylated and un-methylated³³. However, intermediate biologically meaningful states exist like allele specific methylation (with expected methylation levels of 0.5, if both homologous chromosomes have the same sequencing depth), or the reported partial methylation levels³⁰. Therefore, we developed a statistical test for the methylation levels and not for the methylation state previously proposed^{30,34}. To apply this test, the user needs to know the bisulfite conversion rate obtained in the experiment. This rate needs to be established using an un-methylated genome (lambda phage, chloroplast, etc). We supply two additional scripts to i) estimate the bisulfite conversion rate using the appropriate experimental data, and ii) associate a *p-value*, based on binomial statistics, to each of the extracted methylation levels, as well as a procedure to control the false discovery rate³⁵.

In order to calculate a *p-value* for a given methylation level, we first need to select an interval as we want to calculate the probability that the real methylation level lies within an interval of the observed methylation level. Once the interval is fixed, we can calculate the number of false methylcytosines that would not change the methylation level, e.g. the methylation level would stay within the error interval.

Once we have detected the maximum number of false methylcytosines that would maintain the methylation level within the error interval, we can calculate the *p-value* by means of the binomial distribution:

$$p\text{-value} = 1 - \sum_{k=0}^{fmc} \binom{mc}{k} p^k (1-p)^{mc-k}$$

being: p the bisulfite error rate, mc the number of observed methylcytosines at a given position and fmc the maximum number of allowed false methylcytosines. The p -value corresponds then to the probability to find more than fmc false methylcytosines at this position, e.g. the probability that the real methylation level lies outside the defined error interval.

To illustrate the method, let's assume that we have a position that is covered by 21 reads with 17 methylcytosines. In this situation, we would have a methylation level of 0.81. If we fix the error interval at 0.1, we could accept up to 2 false methylcytosines. For two false methylcytosines, the methylation level would be $(17-2)/21 = 0.714$ which lies within the error interval of $0.81-0.1 < 0.714$ while 3 false methylcytosines would lead to a methylation level of 0.67 which lies outside the tolerated error interval. Note that the coverage depth of the position (number of reads) does not appear in the equation, but it does to calculate the maximum number of false methylcytosines. In this way, a higher coverage will lead to a higher number of allowed false methylcytosines and therefore to smaller p -values. Finally, we implemented the Benjamini-Hochberg step-up procedure³⁵ to control for the false discovery rate in multiple testing. This step can be optionally activated by the user.

Results

General comparison to other available tools

MethylExtract is currently one of the programs with most implemented features related to quality control. Together with *Bis-SNP* it is the only program that detects sequence variation, both to avoid incorrect methylation profiling and to assess the genotype of the used sample. [Table 1](#) shows a comparison of the main features of all

programs that allow methylation profiling from aligned reads. Apart from the used method to call the sequence variation, another important difference between *MethylExtract* and *Bis-SNP* is the number of scripts involved to run a full analysis. *Bis-SNP* requires the execution of: i) 3 scripts to sort, add read group tags (required by *GATK* tools) and mark duplicates, ii) 4 scripts to realign the reads and recalibrate the base quality score, iii) 2 scripts to obtain and sort the SNVs and number of methylcytosines and iv) an additional script to calculate the methylation levels on a standard format. In summary, *Bis-SNP* needs 10 different scripts to process reads from bisulfite-treated experiments. On the other hand, *MethylExtract* unifies all analysis steps into a single program which makes it especially suitable for users without a bioinformatics background. Another feature that is currently unique to *MethylExtract* is the possibility to assess the bisulfite failure in a statistical way. In order to achieve this, *MethylExtract* provides an auxiliary script to estimate the bisulfite conversion rate, and a second script that calculates the probability that the observed methylation level lies outside the selected interval of the real methylation level due to bisulfite conversion failures.

Impact of SNVs on methylation levels

As mentioned above, sequence variants can lead to incorrect inference of methylation values. [Figure 2](#) illustrates the impact of C/T variation on the methylation values (C/(C+T) ratio) within CpGs contexts. Around 470,000 SNVs within CpG contexts (affecting to 2.08% of the CpG contexts on the genome) covered by at least 10 reads have been detected by *MethylExtract* in Lister's H1 dataset⁴⁰. [Figure 2](#) shows the methylation levels for non-variant positions (both alleles coincide with the reference) and for the variant sites, both in homozygosis and heterozygosis. The observed distribution of the methylation levels without variation has two maxima close to 0 and 1, which is similar to previous studies⁴⁰. However, for heterozygous positions detected by *MethylExtract*, the methylation levels present a local maximum at approximately 0.5 (the T allele

Table 1. Comparison of *MethylExtract* with different programs for methylation profiling and SNV calling.

| FEATURES//SOFTWARE | MethylCoder | BS_SEEKER | BRAT-BW | BSMAP/RRBSMAP | Bismark | Bis-SNP | MethylExtract |
|-----------------------------------|-------------|-----------|---------|---------------|---------|---------------|---------------|
| Input formats | * | * | * | Sam | Sam | Bam | Sam/Bam |
| 5' Trim | No | No | Yes | No | Yes | Yes | Yes |
| Bisulfite failure | No | No | No | No | No | No | Yes |
| Minimum depth | No | No | No | Yes | No | No | Yes |
| Base call errors | No | No | No | No | No | Yes | Yes |
| SNVs calling | No | No | No | No | No | Yes | Yes |
| Methylation output formats | * | * | * | * | *, bed | vcf, bed, wig | *, bed, wig |
| Variation output formats | - | - | - | - | - | vcf | vcf |

* Input formats: input formats used by each software. 5' trim: allows the trimming of the 5' end of the reads. Bisulfite failure: implementation of a step to discard reads where the bisulfite might have failed converting the un-methylated cytosines. Minimum depth: allows the user to discard positions with low coverage. Base call errors: discards positions that do not exceed a given minimum PHRED score value. SNVs calling: detects variation that can lead to wrong methylation levels or context estimation. Methylation output formats: available formats for the methylation results. Variation output formats: output formats for the sequence variation results. The asterisk (*) represents a non-standard input or output format, or the impossibility of extracting the methylation ratios from other alignment tools. The dash (-) represents the inexistence of SNV output format, because the software does not allow to detect them.

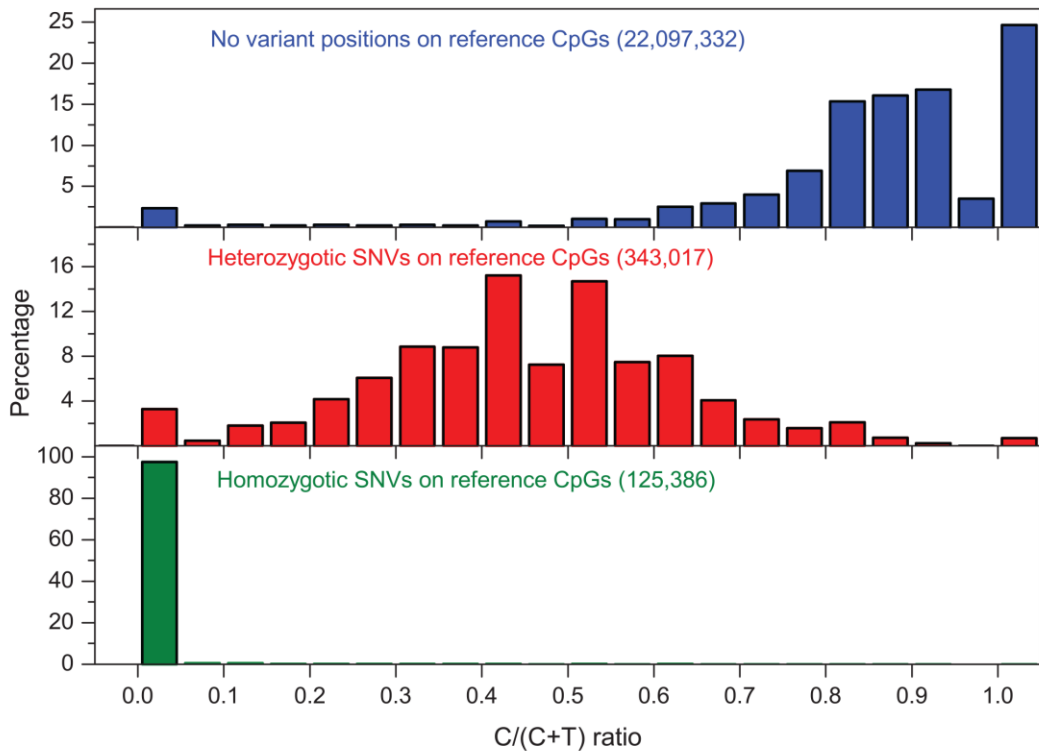


Figure 2. Distribution of $C/(C+T)$ ratios for cytosines within the CpG context in the H1 cell line. $C/(C+T)$ values for cytosines at non-variant and variant (homo- and heterozygotic) positions were shown. The minimum read coverage was set to 10 reads.

on one of the parental chromosomes biases the methylation levels to intermediate values, if the C allele is methylated) and a peak at 0 (if the C allele is un-methylated). Finally, for the homozygous positions where both chromosomes present the T allele, most of the methylation values are exactly 0. However, we know that no cytosine exist at those locus in the analyzed sample and therefore these values are incorrect and should be eliminated from the analysis. The incorrectly inferred methylation values for variant positions, both in homozygosis and heterozygosis, stress the need to detect and remove them from the analysis. For example, a CpG position with C→T SNV on both homologous chromosomes is eliminated by *MethylExtract*, as actually at this position no CpG exists in the sample. Furthermore, *MethylExtract* outputs the detected genotype of all profiled positions and therefore heterozygotic loci can be detected easily by the user and treated apart if wished.

Methylation profiling and SNV calling quality

MethylExtract implements several quality controls and is among the programs with most implemented features. Main features of *MethylExtract* are compared in Table 1 to a number of other, widely

used programs. The implementation was validated in several ways. *MethylExtract* takes aligned reads as input and therefore we first compared the methylation profiling quality achieved on artificial bisulfite data when using two different tools for aligning bisulfite-treated reads; *NGSmethPipe*¹⁶ and *Bismark*³⁶. Next, we quantify the correctly profiled methylation levels and SNVs as a function of the main quality parameters using *NGSmethPipe* as aligner. Finally the predictive power of *MethylExtract* to detect methylation levels and sequence variation was compared to *Bis-SNP*²⁴, both in terms of sensitivity and positive predictive value as it was proposed for datasets for which the number of true negatives tend to be much higher than false positives³⁷.

Generation of artificial BS data. For all further comparisons we will use artificial bisulfite data. The usage of this kind of data for benchmarking has the advantage that the true methylation levels and genotypes are known for each position, which is not true when using other experimental methods like microarrays as a golden standard. Artificial sequencing data has been used before in other studies assessing the SNV prediction quality of different algorithms³⁸.

To generate the artificial bisulfite data we used *DNemulator*¹⁸. We obtained two datasets from the human contig GL000022.1 (11.2Mb), one with all CpGs completely methylated, and the other one with all CpGs completely un-methylated. *DNemulator* allows also simulating the genotypes of a diploid genome by introducing the sequence variation from a set of confirmed SNPs (dbSNP135)¹⁹. Finally, we simulate a bisulfite conversion rate of 99%. The read quality scores are taken from real experimental data (Lister's H1 dataset²⁰). All together, we generated artificial bisulfite sequencing datasets at two different coverages; 15x and 20x which corresponds to the coverage usually achieved in whole genome bisulfite sequencing experiments.

MethylExtract with NGSmethPipe and Bismark input. *NGSmethPipe*¹⁶ is a tool to align bisulfite-treated reads which was developed by our group. It is based on the *Bowtie* aligner and uses a 3-letter alphabet to map the bisulfite-treated reads. The program implements a pre-processing to improve the mapping accuracy¹⁸ and an alignment seed extension in order to increase the number of mapped reads.

We launched both, *NGSmethPipe* and the well-established *Bismark* tool with default options to obtain the SAM/BAM input. Next we used *MethylExtract* on both input files to obtain the number of covered CpGs and the number of correctly recovered methylation values. Note that we know the correct methylation value for each position due to the use of artificial bisulfite data. A position is considered as correctly profiled, only if the obtained methylation value is identical to the real value. **Figure 3** shows the result of this comparison. It shows that the obtained CpG coverage and number of correctly profiled positions is nearly identical both as a function of read coverage (15x and 20x) and for the methylated and un-methylated input data. The only remarkable difference is that *NGSmethPipe* leads to a slightly higher CpG coverage at 20x for both data sets. Nevertheless, the main conclusion is that *MethylExtract* yields nearly identical results for input sets obtained from *NGSmethPipe* and *Bismark*.

Analysis of the MethylExtract quality parameters. Next, we aimed to assess the impact of certain quality parameters implemented in

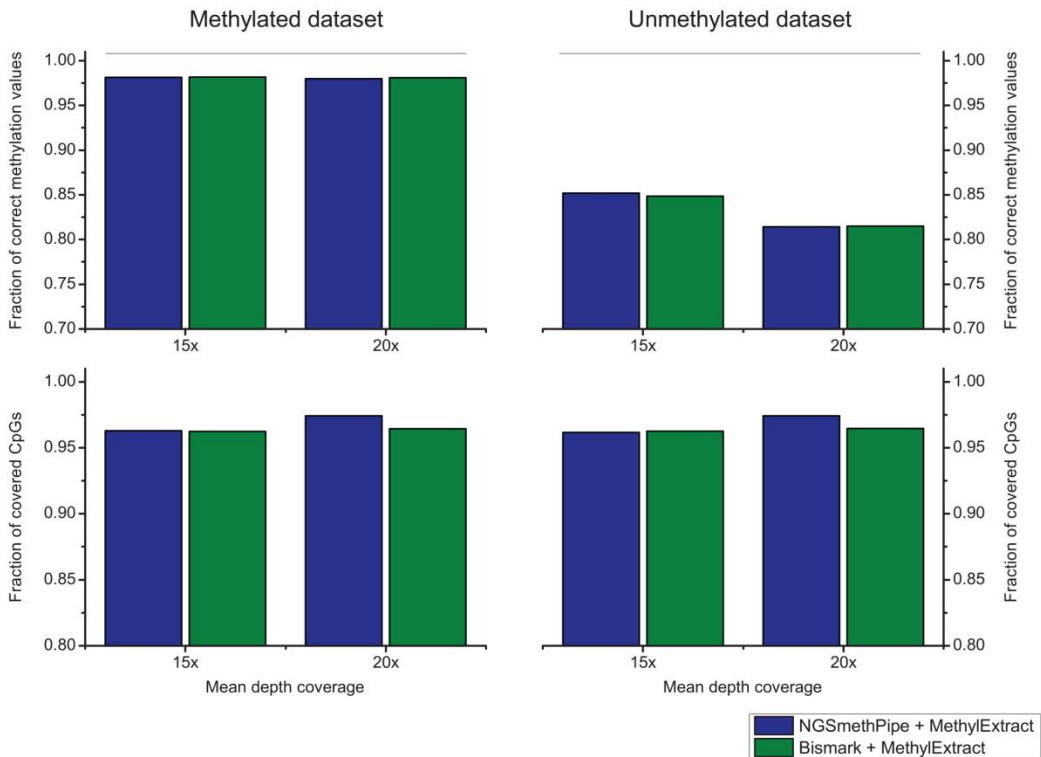


Figure 3. CpGs methylation profiling comparison for alignment methods. The results obtained from *MethylExtract* (correctly profiled methylation values and CpG coverage) using two bisulfite short read aligners, *NGSmethPipe* and *Bismark* are compared. The results are nearly independent of the used alignment algorithm.

MethylExtract on the methylation profiling and SNV calling capacity. To detect sequence variation, *MethylExtract* relies on two main parameters, i) the relative nucleotide frequencies ('varFraction') and ii) the corresponding *p*-value. The 'varFraction' parameter determines if a position shows putatively variation: the position is analyzed only if at least one nucleotide that differs from the reference sequence has relative frequencies higher than 'varFraction'. Only for these positions the corresponding *p*-value is calculated by means of a Fisher exact test. Figure 4 shows the impact of these parameters on the prediction sensitivity (Sn) and positive predictive value (PPV). Sequence variation is best detected by setting the 'varFraction' threshold close to 0.1 (yielding around 91% Sn and only 2% of false positives at a statistical significance of 0.05). If the 'varFraction' threshold is increased further, the probability to eliminate heterozygous loci increases steadily for positions with high bias in the read coverage between the two homologous chromosomes. If the *p*-value threshold is set to 0.01, a small increase in positive predictive value (PPV) is observed, but it causes a strong

decrease in sensitivity. Therefore, we determined a 'varFraction' of 0.1 and a *p*-value threshold of 0.05 as the best (default) parameters to detect sequence variations.

The minimum base quality ('minQ') and the coverage depth ('minDepthMeth' for the methylation profiling) thresholds might be also important parameters to control de quality of methylation profiling and SNV calling. To analyze the impact of the minimum PHRED score parameter ('minQ') we fix the minimum read coverage ('minDepthMeth') in 3, as suggested by Laurent *et al.*⁴⁰, 'varFraction' = 0.1 and 'maxPval' = 0.05 (default values derived above). Figure 5 shows the fraction of correctly profiled methylation values and the PPV for SNVs. It can be seen that the correctly profiled positions increase approximately 31% (from 68% to 99%) and the SNVs around 71% (27% – 98%), when the minimum PHRED score is increased from 0 (all base calls are accepted) to 30 (0.001 error probability). The major difference between the methylated and un-methylated datasets is observed for the profiling of the

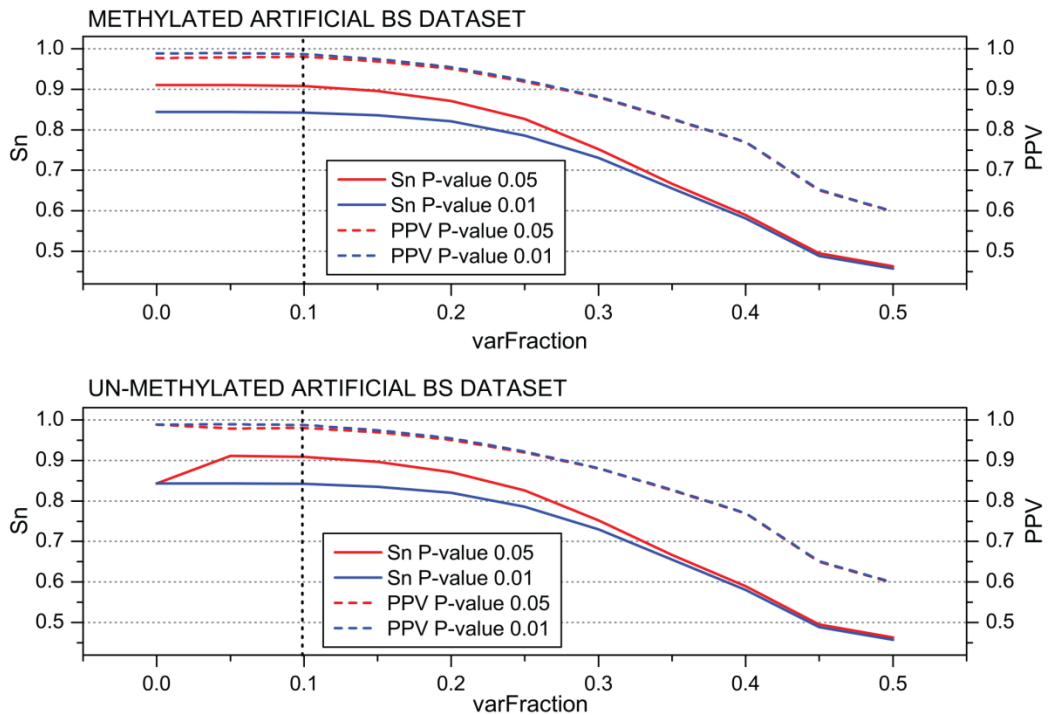


Figure 4. *MethylExtract* SNV calling as a function of the minimum relative nucleotide frequency ('varFraction'). The figures show the sensitivity (Sn) and the positive predictive value (PPV) for SNV detection using two different *p*-value thresholds. The graphs are based on the methylated (top) and un-methylated (bottom) artificial bisulfite datasets at a mean 20x read coverage.

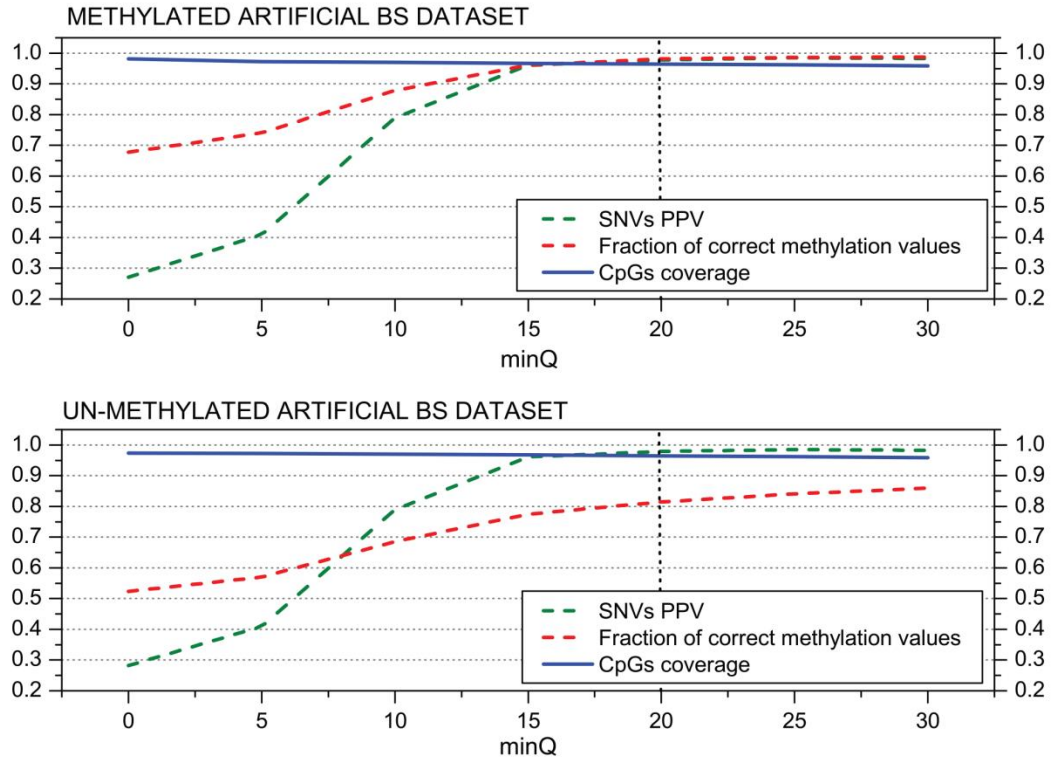


Figure 5. MethyExtract SNV calling and methylation profiling as a function of the base quality. Both graphs show the positive predictive value (PPV) for SNV calling and the fraction of correctly profiled CpG methylation values (methylation profiling) as a function of the minimum base quality (PHRED score parameter 'minQ'). The graphs are based on the methylated (top) and un-methylated (bottom) artificial bisulfite datasets at a mean 20x read coverage. Y-axis represents SNV PPV, Fraction of correct methylation values and CpG coverage. All of them vary between 0 to 1 therefore being represented together.

methylation level for which the percentage increases only from approximately 52% to 86%. The simulated bisulfite conversion failures will affect mainly un-methylated positions which can explain the observed differences. These results confirm that the 'minQ' threshold is critical to obtain high quality methylation profiling and genotyping results. The default value was set to 20 as higher values will lead to a coverage reduction compromising the SNV calling sensitivity.

Comparison with *Bis-SNP*

The comparison between *MethylExtract* and *Bis-SNP* needs to be based on identical alignment input files in BAM/SAM format. We obtained these files in a two-step process: First, we trim the input reads as it was done by Lister *et al.*³⁰ and second, we align

the bisulfite treated reads to the reference genome using *Bismark*³⁶ with default parameters. Note that we based this comparison on *Bismark*, as the realignment and recalibration steps implemented in *Bis-SNP* require the read mapping quality, which is currently not available in *NGSmethPipe*.

Both methods were used with default parameters. We first compared the detection of sequence variation (SNVs) in terms of Sn and PPV. *Figure 6*, shows that in general *Bis-SNP* is more specific (between 1.9% and 3.9% higher PPV), being *MethylExtract* more sensitive (between 1% and 3.1% higher Sn). This trend can be seen for both artificial bisulfite datasets as well as for both read coverages. However, when comparing the fraction of correctly recovered methylation values, drastic differences can be seen (*Figure 7*).

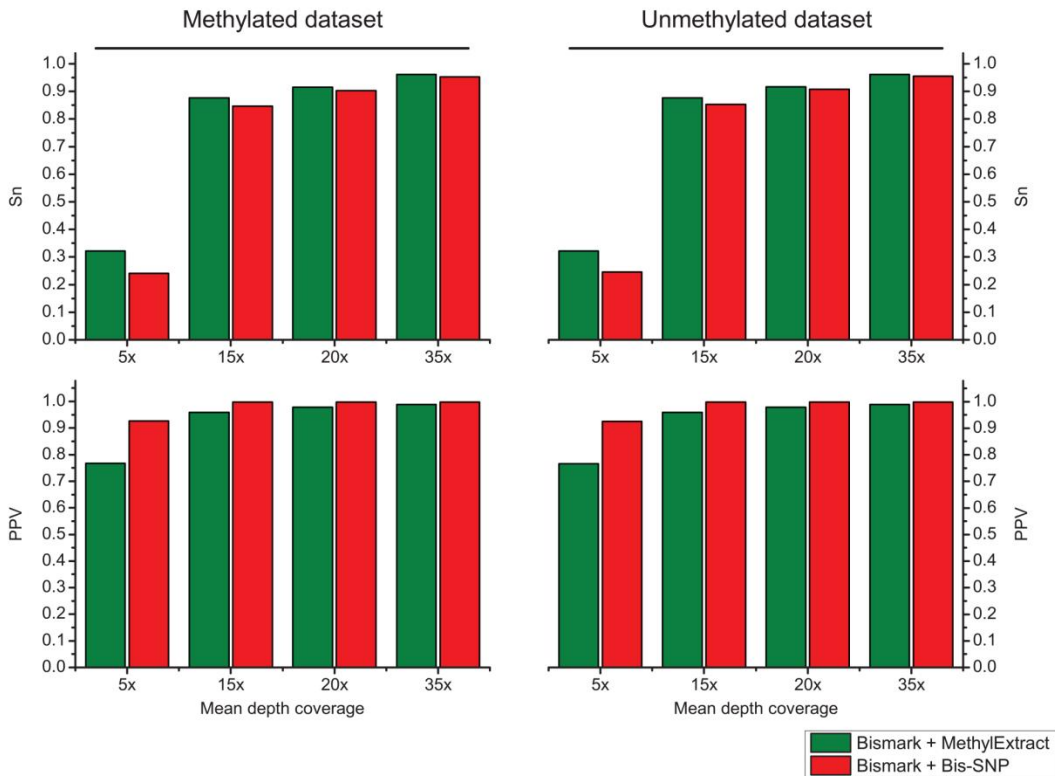


Figure 6. Comparison of SNV calling between *MethylExtract* and *Bis-SNP*. The top graph shows the sensitivity (Sn) and the bottom graph the specificity (PPV) obtained for the methylated and un-methylated artificial bisulfite datasets at two different mean coverages (5x, 15x, 20x and 35x).

Furthermore, when the criteria for correctly profiled methylation values are relaxed, *MethylExtract* still yields higher fractions than *Bis-SNP* (Supplementary Figure 1). While *Bis-SNP* yields a slightly higher number of covered positions (Fraction of covered CpGs), *MethylExtract* is more specific. In all four comparisons using the stringent criteria (no deviation from the real methylation values is allowed), *MethylExtract* yields over 20% more correctly profiled positions compared to *Bis-SNP*. One explanation for this difference might be the PHRED score quality threshold implemented in *MethylExtract*.

Runtime comparison to *Bis-SNP*

As mentioned before, only *Bis-SNP* and *MethylExtract* perform the detection of SNVs which constitutes an additional CPU demanding task. Therefore, we only compared these two programs in terms of CPU time using a reduced Lister's H1 dataset³⁰ on a 24 core Intel(R) Xeon(R) CPU X5650 2.67GHz machine. Available memory is crucial for both methods. In order to not bias the comparison, we limited the available memory to 15GB for both programs allowing up

to 15 threads. Both programs were tested using a 11GB BAM input file. After aligning with *Bismark*, we carried out the entire process for both tools (from the aligned reads to the methylation and SNV profiling). *MethylExtract* needed 6 hours 2 minutes to process the entire dataset including the sorting by coordinates and the removal of duplicated reads. *Bis-SNP* spend 2 hours 47 minutes sorting the file and removing putative clonal reads, 9 hours and 36 minutes realigning and recalibrating the reads, and 15 hours 54 minutes genotyping and retrieving the methylation levels. Therefore, it seems that *MethylExtract* is notably faster than *Bis-SNP* (approximately 4.5 times on this whole genome data set).

Conclusions

We present a user-friendly tool for methylation profiling and SNV calling in whole genome bisulfite sequencing experiments. *MethylExtract* takes standardized input formats (BAM/SAM) and writes out likewise broadly used file formats like WIG, BED and VCF. To show its usefulness, we compared it to *Bis-SNP*, a recently published method that is very similar in scope. Although *Bis-SNP* is

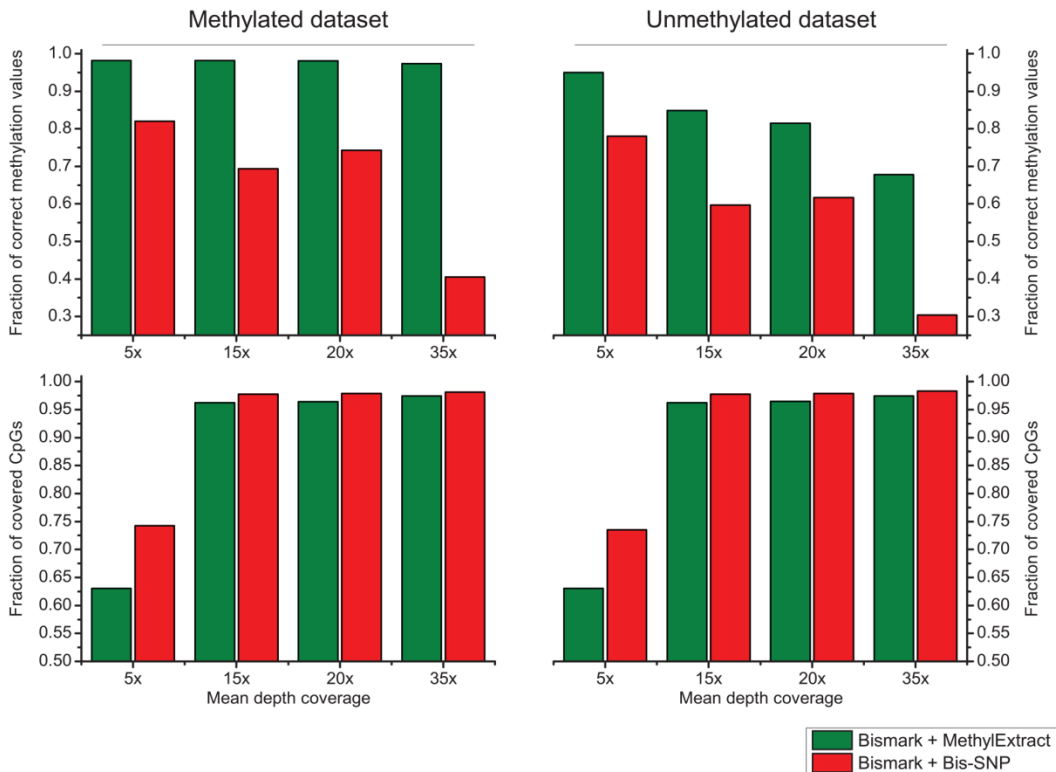


Figure 7. Comparison of CpG methylation values between *MethylExtract* and *Bis-SNP*. Both methods are compared in terms of fraction of correctly profiled CpG methylation values (top) and the fraction of recovered CpG positions (bottom).

more specific (less false positive predictions) in the detection of SNVs, *MethylExtract* is more sensitive (higher number of recovered SNVs). However, the main advantages of *MethylExtract* when compared to *Bis-SNP* seem to rely in the higher percentage of correctly profiled methylation values, as it reaches values over 20% higher compared to *Bis-SNP*. Other aspects that favor *MethylExtract* are its user-friendliness (everything is implemented into one script) and the run-time in comparison to *Bis-SNP* (over 4 times faster in a whole genome bisulfite sequencing experiment).

Availability and requirements

MethylExtract is freely available. The source code, the tutorial and artificial bisulfite datasets can be downloaded from the page <http://bioinfo2.ugr.es/MethylExtract/> and are also permanently accessible from [10.5281/zenodo.835142](https://doi.org/10.5281/zenodo.835142).

List of abbreviations used

5mC: DNA methylation at cytosine carbon 5 position; SNV: Single Nucleotide Variation; WGBS: whole genome bisulfite sequencing; SNP: Single Nucleotide Polymorphism; PPV: positive predictive value; PHRED score: the quality score to each base call assigned

by the program PHRED; SAM format: Sequence Alignment/Map format used for storing large nucleotide sequence alignments; BAM format: the compressed binary version of the SAM format.

Author contributions

GB wrote the code and carried out the experiments, AR helped with the benchmark experiments, and MH, JLO and GB designed the software and wrote the manuscript. All the authors critically read and approved the final version.

Competing interests

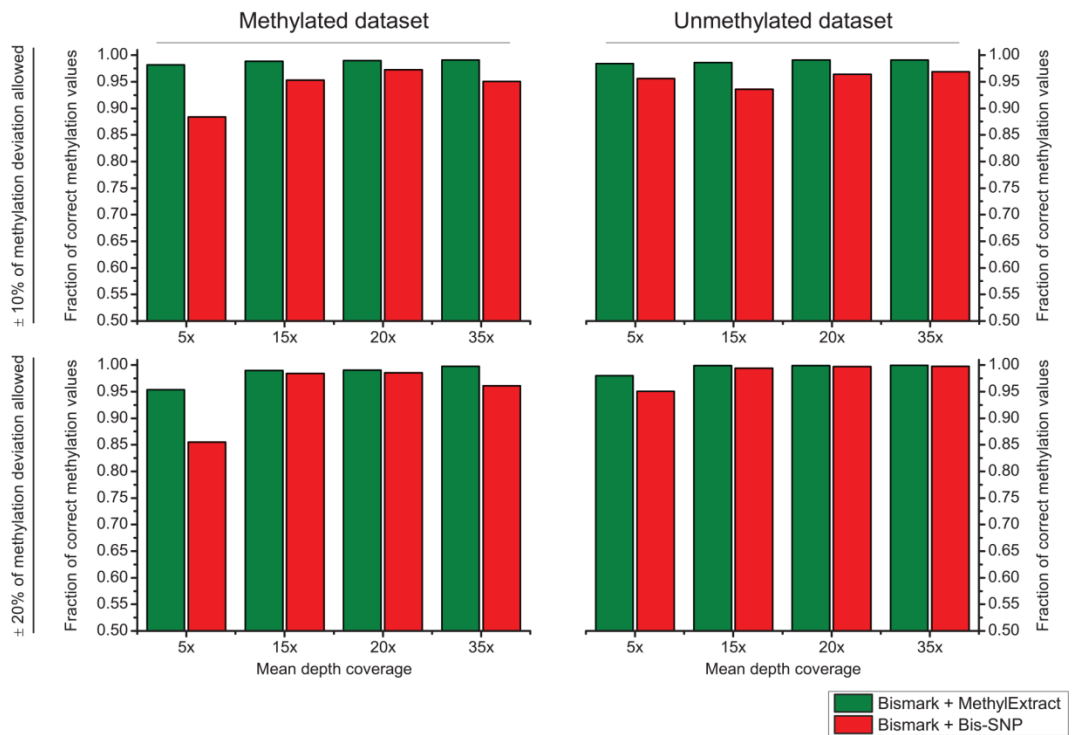
No competing interests were disclosed.

Grant information

This work was supported by the Spanish Government [BIO2008-01353 to JLO and BIO2010-20219 to MH], and Basque country 'AE' grant (GB).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material



Supplementary Figure 1. Methylation profiling comparison *MethylExtract* and *Bis-SNP* using relaxed criterion. Both methods are compared in terms of fraction of correctly profiled CpG methylation values. The upper part of the graph shows the result allowing up to 10% deviation from the real methylation values, while the lower part shows the outcome increasing this range to 20%. The analyses were done for unmethylated and methylated datasets at four different coverages (5x, 15x, 20x and 35x).

References

- Oliveira DC, Tomasz A, de Lencastre H: **The evolution of pandemic clones of methicillin-resistant *Staphylococcus aureus*: identification of two ancestral genetic backgrounds and the associated mec elements.** *Microb Drug Resist.* 2001; 7(4): 349–61.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Gu F, Doderer MS, Huang YW, et al.: **CMS: a web-based system for visualization and analysis of genome-wide methylation data of human cancers.** *PLoS One.* 2013; 8(4): e60980.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wasserkort R, Kalmár A, Valcz G, et al.: **Aberrant septin 9 DNA methylation in colorectal cancer is restricted to a single CpG island.** *BMC Cancer.* 2013; 13(1): 398.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Eden S, Cedar H: **Role of DNA methylation in the regulation of transcription.** *Curr Opin Genet Dev.* 1994; 4(2): 255–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Eden A, Gaudet F, Waghmare A, et al.: **Chromosomal instability and tumors promoted by DNA hypomethylation.** *Science.* 2003; 300(5618): 455.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Li E, Beard C, Jaenisch R: **Role for DNA methylation in genomic imprinting.** *Nature.* 1993; 366(6453): 362–5.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Kato M, Miura A, Bender J, et al.: **Role of CG and non-CG methylation in immobilization of transposons in *Arabidopsis*.** *Curr Biol.* 2003; 13(5): 421–6.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Jones PA: **Functions of DNA methylation: Islands, start sites, gene bodies and beyond.** *Nat Rev Genet.* 2012; 13(7): 484–92.
[PubMed Abstract](#) | [Publisher Full Text](#)

9. Laird PW: **Principles and challenges of genomewide DNA methylation analysis.** *Nat Rev Genet.* 2010; **11**(3): 191–203.
[PubMed Abstract](#) | [Publisher Full Text](#)
10. Lister R, O'Malley RC, Tonti-Filippini J, et al.: **Highly integrated single-base resolution maps of the epigenome in Arabidopsis.** *Cell.* 2008; **133**(3): 523–36.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Cokus SJ, Feng S, Zhang X, et al.: **Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning.** *Nature.* 2008; **452**(7184): 215–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Meissner A, Mikkelsen TS, Gu H, et al.: **Genome-scale DNA methylation maps of pluripotent and differentiated cells.** *Nature.* 2008; **454**(7205): 766–70.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. Lister R, Ecker JR: **Finding the fifth base: genome-wide sequencing of cytosine methylation.** *Genome Res.* 2009; **19**(6): 959–66.
[PubMed Abstract](#) | [Publisher Full Text](#)
14. Krueger F, Andrews SR: **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.** *Bioinformatics.* 2011; **27**(11): 1571–2.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Pedersen B, Hsieh TF, Ibarra C, et al.: **MethylCoder: software pipeline for bisulfite-treated sequences.** *Bioinformatics.* 2011; **27**(17): 2435–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Hackenberg M, Barturen G, Oliver JL: **DNA Methylation - From Genomics to Technology.** (ed. Tatarinova, T.) (In-Tech.), 2012.
[Publisher Full Text](#)
17. Chen PY, Cokus SJ, Pellegrini M: **BS Seeker: precise mapping for bisulfite sequencing.** *BMC Bioinformatics.* 2010; **11**: 203.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. Frith MC, Mori R, Asai K: **A mostly traditional approach improves alignment of bisulfite-converted DNA.** *Nucleic Acids Res.* 2012; **40**(13): e100.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
19. Harris EY, Ponts N, Le Roch KG, et al.: **BRAT-BW: efficient and accurate mapping of bisulfite-treated reads.** *Bioinformatics.* 2012; **28**(13): 1795–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
20. Krueger F, Kreck B, Franke A, et al.: **DNA methylome analysis using short bisulfite sequencing data.** *Nat Methods.* 2012; **9**(2): 145–51.
[PubMed Abstract](#) | [Publisher Full Text](#)
21. Tomso DJ, Bell DA: **Sequence context at human single nucleotide polymorphisms: overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG Islands.** *J Mol Biol.* 2003; **327**(2): 303–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
22. Bird A: **Putting the DNA back into DNA methylation.** *Nat Genet.* 2011; **43**(11): 1050–1.
[PubMed Abstract](#) | [Publisher Full Text](#)
23. Liener F, Wirbelauer C, Som I, et al.: **Identification of genetic elements that autonomously determine DNA methylation states.** *Nat Genet.* 2011; **43**(11): 1091–7.
[PubMed Abstract](#) | [Publisher Full Text](#)
24. Liu Y, Siegmund KD, Laird PW, et al.: **Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data.** *Genome Biol.* 2012; **13**(7): R61.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
25. Li H, Handsaker B, Wysoker A, et al.: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
26. Ewing B, Hillier L, Wendt MC, et al.: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res.* 1998; **8**(3): 175–85.
[PubMed Abstract](#) | [Publisher Full Text](#)
27. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res.* 1998; **8**(3): 186–94.
[PubMed Abstract](#) | [Publisher Full Text](#)
28. Koboldt DC, Chen K, Wylie T, et al.: **VarScan: variant detection in massively parallel sequencing of individual and pooled samples.** *Bioinformatics.* 2009; **25**(17): 2283–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. Danecek P, Auton A, Abecasis G, et al.: **The variant call format and VCFtools.** *Bioinformatics.* 2011; **27**(15): 2156–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
30. Lister R, Pelizzola M, Dowen RH, et al.: **Human DNA methylomes at base resolution show widespread epigenomic differences.** *Nature.* 2009; **462**(7271): 315–22.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
31. Weisenberger DJ, Campan M, Long TI, et al.: **Analysis of repetitive element DNA methylation by MethyLight.** *Nucleic Acids Res.* 2005; **33**(21): 6823–36.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
32. Cingolani P, Platts A, Wang le L, et al.: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; Iso-2; Iso-3.** *Fly (Austin).* 2012; **6**(2): 80–92.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
33. Bastone P, Bravo IG, Lochelt M: **Feline foamy virus-mediated marker gene transfer: identification of essential genetic elements and influence of truncated and chimeric proteins.** *Virology.* 2006; **348**(1): 190–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
34. Schultz MD, Schmitz RJ, Ecker JR: **'Leveling' the playing field for analyses of single-base resolution DNA methylomes.** *Trends Genet.* 2012; **28**(12): 583–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
35. Negre V, Grunau C: **The MethDB DAS server: adding an epigenetic information layer to the human genome.** *Epigenetics.* 2006; **1**(2): 101–5.
[PubMed Abstract](#) | [Publisher Full Text](#)
36. Krueger F, Andrews SR: **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.** *Bioinformatics.* 2011; **27**(11): 1571–2.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
37. Bursat M, Guigo R: **Evaluation of gene structure prediction programs.** *Genomics.* 1996; **34**(3): 353–67.
[PubMed Abstract](#) | [Publisher Full Text](#)
38. You N, Murillo G, Su X, et al.: **SNP calling using genotype model selection on high-throughput sequencing data.** *Bioinformatics.* 2012; **28**(5): 643–50.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
39. Sherry ST, Ward MH, Kholodov M, et al.: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res.* 2001; **29**(1): 308–11.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
40. Laurent L, Wong E, Li G, et al.: **Dynamic changes in the human methylome during differentiation.** *Genome Res.* 2010; **20**(3): 320–31.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
41. Karolchik D, Kuhn RM, Baertsch R, et al.: **The UCSC Genome Browser Database: 2008 update.** *Nucleic Acids Res.* 2008; **36**(Database issue): D773–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
42. Barturen G, Rueda A, Oliver JL, et al.: **MethylExtract release 1.5.** *Zenodo.* 2014.
[Data Source](#)

3.4 MAPAS DE CALIDAD DE METILOMAS COMPLETOS

Usando *NGSmethPipe* y *MethylExtract* se han procesado los conjuntos de datos tratados con bisulfito incluidos en la Tabla 3.3, obtenidos del repositorio público *GEO* (Barrett, Troup et al. 2009, Barrett, Wilhite et al. 2013). La utilización de un mismo protocolo para procesar estos datos experimentales permite comparar metilomas procedentes de diferentes estudios. Además, la metodología utilizada es con diferencia la que presenta un mayor número de controles de calidad, lo que asegura la obtención de metilomas muy precisos, al reducir los errores que pueden sesgar los resultados finales. En lugar de estimar la probabilidad de error para los niveles de metilación (protocolo adicional de *MethylExtract*), se ha fijado una profundidad mínima que asegure la fiabilidad de los datos, ya que no todos los experimentos incluyen un genoma no metilado con el que estimar la tasa de conversión del bisulfito. Aunque Laurent y colaboradores (Laurent, Wong et al. 2010) fijaron en 3 lecturas la profundidad mínima necesaria para reproducir en un 79% la metilación extraída por medio de *microarrays*, en esta comparación se ha fijado en 5, aumentando de esta manera la reproducibilidad de los resultados.

| Valores \ Muestras | <i>bcell</i> | <i>cd133hsc</i> | <i>fibro</i> | <i>h1</i> |
|--|---------------------|---------------------|---------------------|---------------------|
| Cobertura de CpGs en el genoma (%) | 81.75 | 76.21 | 78.32 | 86.35 |
| Profundidad media por CpG (SD) | 9.6 (5.4) | 8.6 (5.2) | 13.5 (10.4) | 20.7 (13.2) |
| Metilación media por CpG (SD) | 0.74 (0.32) | 0.79 (0.33) | 0.62 (0.35) | 0.8 (0.25) |
| % CpGs metilados (>= 0.8) | 62.09 | 72.19 | 45.46 | 74.08 |
| % CpGs no metilados (<=0.2) | 11.93 | 12.78 | 19.22 | 7.12 |
| Cobertura media de secuenciación (SD) | 10.7 (5.3) | 8.7 (5) | 10 (8.3) | 24 (11.4) |
| # Variaciones detectadas (%) | 4,673,543 (0.15) | 3,705,060 (0.12) | 2,145,487 (0.07) | 8,011,720 (0.26) |

| Valores \ Muestras | <i>hcc1954</i> | <i>hmec</i> | <i>hspc</i> | <i>imr90</i> |
|--|---------------------|---------------------|---------------------|----------------------|
| Cobertura de CpGs en el genoma (%) | 95.27 | 95.26 | 85.22 | 90.89 |
| Profundidad media por CpG (SD) | 24.8 (13.8) | 21.4 (8.6) | 11 (6.2) | 24.3 (14.1) |
| Metilación media por CpG (SD) | 0.61 (0.39) | 0.68 (0.35) | 0.77 (0.32) | 0.63 (0.34) |
| % CpGs metilados (>= 0.8) | 51.51 | 56.22 | 68.1 | 45.34 |
| % CpGs no metilados (<=0.2) | 26.84 | 17.53 | 12.39 | 17.79 |
| Cobertura media de secuenciación (SD) | 26.9 (12) | 21.3 (9) | 12.5 (6) | 27 (12.2) |
| # Variaciones detectadas (%) | 6,691,674 (0.22) | 6,896,094 (0.22) | 5,441,273 (0.18) | 10,719,970 (0.35) |

| Valores \ Muestras | <i>prefrontalcortex_hs1570</i> | <i>wa09</i> | <i>wa09fibro</i> |
|--|--------------------------------|---------------------|---------------------|
| Cobertura de CpGs en el genoma (%) | 80.03 | 82.07 | 79.75 |
| Profundidad media por CpG (SD) | 8.8 (4.7) | 16.6 (13.1) | 15 (12.3) |
| Metilación media por CpG (SD) | 0.8 (0.3) | 0.73 (0.3) | 0.7 (0.34) |
| % CpGs metilados (>= 0.8) | 73.03 | 59.52 | 58.44 |
| % CpGs no metilados (<=0.2) | 9.6 | 11.92 | 14.79 |
| Cobertura media de secuenciación (SD) | 8.3 (4.6) | 11.8 (10) | 11 (9.8) |
| # Variaciones detectadas (%) | 2,869,812 (0.09) | 2,346,422 (0.08) | 2,001,390 (0.06) |

Tabla 3.3. Resultados obtenidos con *NGSmethPipe* y *MethylExtract* para una serie de datos analizados en esta tesis. Los conjuntos de datos que se muestran provienen de diversas publicaciones: *bcell*, *cd133hsc* y *hspc* (Hodges, Molaro et al. 2011); *fibro*, *wa09fibro* y *wa09*

(Laurent, Wong et al. 2010); *h1* e *imr90* (Lister, Pelizzola et al. 2009); *hmec* y *hcc1954* (Hon, Hawkins et al. 2012); *prefrontalcortex_hs1570* (Zeng, Konopka et al. 2012). Todos ellos presentan valores de metilación para al menos el 75% de sus CpGs, con una profundidad mínima de 5 lecturas. Para cada conjunto de datos se muestra la media y desviación estándar (SD) para la cobertura de secuenciación, tanto de CpGs como de cualquier posición del genoma, y para los niveles de metilación en CpGs. También se muestra el porcentaje de CpGs metilados (niveles de metilación ≥ 0.8) y no metilados (niveles de metilación ≤ 0.2), así como el número de variaciones detectadas frente a la referencia, con su porcentaje sobre el total de posiciones en el genoma entre paréntesis. Para evitar confusiones, se han mantenido los nombres originales asignados por los respectivos autores.

En la Tabla 3.3 se muestra un resumen de los resultados obtenidos para los mejores conjuntos de datos procesados (cobertura mínima del 75% de los CpGs y con una profundidad de secuenciación de al menos 5 lecturas). A pesar de la alta cobertura genómica seleccionada, la profundidad de secuenciación media por posición es muy variable (8.6 a 24.8). También la media de metilación en contextos CpG presenta cierta variación, pudiendo distinguirse las líneas celulares de células madre (no diferenciadas: *h1* y *wa09*) y los tipos celulares frescos (diferenciados: *bcell*, *cd133hsc*, *hspc* y *prefrontalcortex_hs1570*) de las líneas celulares de fibroblastos (diferenciadas: *fibro*, *hmec* e *imr90*), por presentar estas últimas una media de metilación generalmente inferior. Estas diferencias también pueden apreciarse al representar las distintas muestras en función de sus proporciones de CpGs metilados y no metilados (Figura 3.8). En cuanto a las variaciones de secuencia detectadas frente a la referencia, el número de variaciones encontradas para las muestras de menor profundidad media (tipos celulares frescos) se encuentran dentro del rango de variaciones por individuo descritas por el proyecto 1000 genomas, mientras que el porcentaje de variaciones para muestras con mayor profundidad media se mantiene dentro del orden de magnitud

descrito (Abecasis, Altshuler et al. 2010). Estas muestras de mayor profundidad (líneas celulares) presentan en algunos casos, como *h1* o *imr90*, casi 3 veces más variaciones que los tejidos frescos, lo que posiblemente se deba al gran número de mutaciones espontáneas que pueden acumular estas líneas celulares (Maitra, Arking et al. 2005). Otra posible causa es la mayor profundidad empleada en el análisis de estas muestras, ya que ello aumentaría la significación estadística de las variaciones existentes.

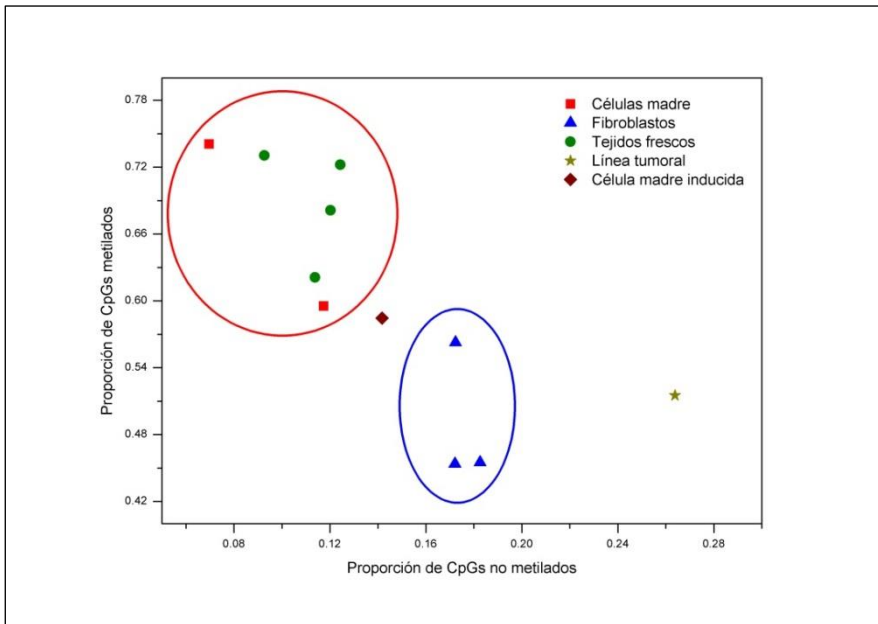


Figura 3.8. Representación de las proporciones de CpGs metilados y no metilados para cada muestra. Se representan las células madre (cuadrados rojos) y los tejidos frescos (círculos verdes) formando un subconjunto rodeado por un círculo rojo, a su vez las líneas celulares diferenciadas (triángulos azules) se encuentran rodeados por un círculo azul. También se incluyen en la figura una línea de células madre en la que se ha inducido la diferenciación (*wa09fibro*, rombo marrón) y una línea tumoral (*hcc1954*, estrella amarilla).

En la Figura 3.8 pueden observarse claramente las diferencias de células madre y tejidos frescos (círculo rojo) frente a líneas celulares diferenciadas (círculo azul), las primeras presentan mayores proporciones de CpGs metilados y menores proporciones de no metilados comparadas con los fibroblastos provenientes de líneas celulares. Además, se observa cómo las células madre en las que se ha inducido su diferenciación (rombo marrón) presentan características intermedias a ambos subconjuntos, tal y como se describe en la publicación original (Laurent, Wong et al. 2010), mientras que la línea tumoral (estrella amarilla) muestra un mayor grado de hipometilación, también previamente descrito (Hon, Hawkins et al. 2012).

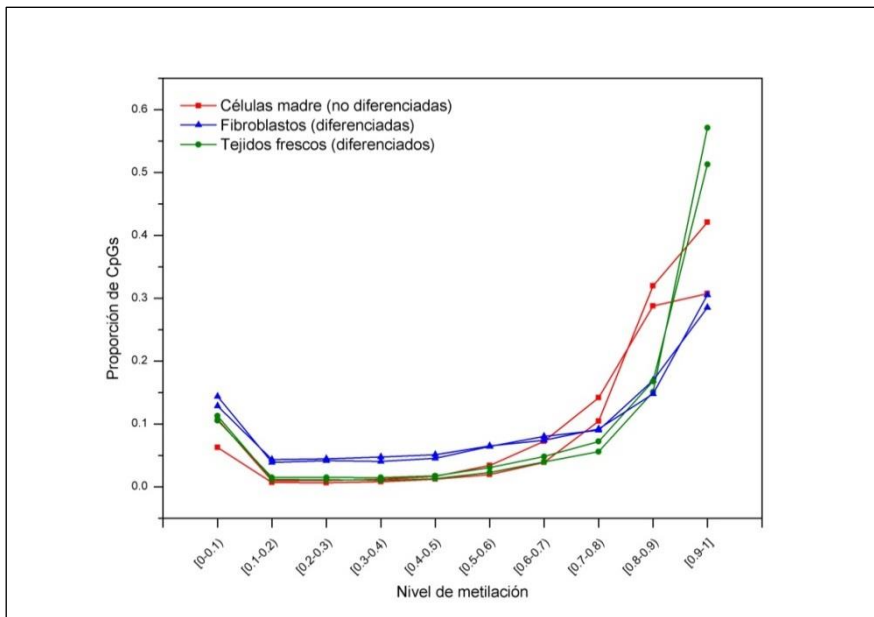


Figura 3.9. Distribuciones de los niveles de metilación en CpGs para diferentes tejidos. Se representan las proporciones de CpGs para 10 grupos de niveles de metilación en 6 tejidos, 2 para cada clase de las previamente descritas: tejidos frescos (*cd133hsc* y *hspc*) en verde, líneas celulares de fibroblastos (*fibro* e *imr90*) en azul y líneas celulares de células madre (*h1* y *wa09*) en rojo.

En cuanto a las distribuciones de los niveles de metilación en contextos CpG (Figura 3.9), todas las muestras presentan una distribución bimodal con los máximos en los extremos de la distribución (CpGs totalmente metilados y totalmente no metilados). La diferencia más significativa se encuentra en los valores intermedios de metilación, donde los tejidos diferenciados provenientes de líneas celulares presentan una mayor proporción con respecto a las células madre y a los tejidos frescos. Las diferencias entre líneas celulares diferenciadas y no diferenciadas, tanto en este aspecto como en las diferentes proporciones de CpGs metilados y no metilados, ya habían sido descritas por Lister y colaboradores (Lister, Pelizzola et al. 2009). Sin embargo, la relación encontrada entre tejidos frescos y células madre, y las diferencias de estos tejidos frescos con líneas celulares diferenciadas no se habían descrito previamente. Se observan diferencias considerables entre ambos conjuntos de muestras diferenciadas (tejidos frescos y líneas celulares), tanto en los niveles extremos de la distribución, como en los valores intermedios. Estas diferencias se explicarían por una desmetilación generalizada de los contextos CpG, debida probablemente a la inmortalización de estos tejidos, un proceso semejante al que se ha descrito en tejidos tumorales (Hon, Hawkins et al. 2012).

MINERÍA DE DATOS Y VISUALIZACIÓN DE PERFILES DE METILACIÓN

La gran cantidad de datos resultantes de los experimentos de secuenciación, ha hecho imprescindible el desarrollo de herramientas que posibiliten la comparación eficaz de la información obtenida. Durante la secuenciación del primer ensamblado del genoma humano (Lander, Linton et al. 2001, Venter, Adams et al. 2001), se desarrollaron en paralelo herramientas y bases de datos que posibilitaron la comparación y visualización de la información genómica resultante, publicándose las más utilizadas y difundidas un año después del primer ensamblado: “UCSC Genome Browser” (Kent, Sugnet et al. 2002) y “Ensembl Genome Database” (Hubbard, Barker et al. 2002). Tras la irrupción de la epigenómica, con proyectos como “ENCODE” (Consortium 2004) o “Roadmap Epigenomics” (Bernstein, Stamatoyannopoulos et al. 2010), que aportan datos sobre sitios de unión a factores de transcripción, modificaciones de histonas, transcripción de ARNs, conformación de la

cromatina o metilación del ADN, estas bases de datos se han convertido en herramientas imprescindibles para el manejo de la información contenida en el ADN.

En este capítulo se revisan las principales bases de datos para el análisis de perfiles de metilación del ADN, y se presenta *NGSmethDB* (Hackenberg, Barturen et al. 2011, Geisen, Barturen et al. 2014), una base de datos desarrollada durante esta Tesis Doctoral. *NGSmethDB* pretende ser la base de datos de referencia para el almacenamiento, análisis y estudio comparado de metilomas completos de alta resolución. Esta base de datos contiene metilomas completos procesados de manera uniforme mediante los protocolos descritos en el capítulo anterior: *NGSmethPipe* (Hackenberg, Barturen et al. 2012) y *MethylExtract* (Barturen, Rueda et al. 2013), lo que permite el estudio comparado de metilomas de diferentes tejidos, especies o estados patológicos.

4.1 BASES DE DATOS PARA EL ANÁLISIS DE LA METILACIÓN DEL ADN

La metilación del ADN es una de las modificaciones epigenéticas más estudiadas durante los últimos años, lo que ha dado lugar a la aparición de numerosas bases de datos especializadas: *MethDB* (Grunau, Renault et al. 2001, Amoreira, Hindermann et al. 2003, Negre and Grunau 2006), *MethyCancer* (He, Chang et al. 2008), *DiseaseMeth* (Lv, Liu et al. 2012), *MethylomeDB* (Xin, Chanrion et al. 2012), *PEpiD* (Shi, Hu et al. 2013), *Cancer Methylome System* (Gu, Doderer et al. 2013) o *NGSmethDB* (Hackenberg, Barturen et al. 2011, Geisen, Barturen et al. 2014), esta

última desarrollada en esta Tesis Doctoral. A su vez, el proyecto *ENCODE* mantiene una estrecha colaboración con el buscador genómico de la *UCSC*, almacenando los resultados en su base de datos. Por su parte el proyecto *Roadmap Epigenomics* permite acceder a un buscador basado en el de la *UCSC* con todos sus resultados integrados. En general, todas estas bases de datos permiten descargar, comparar y visualizar los niveles de metilación de diferentes regiones del genoma. La excepción es *MethDB*, que fue concebida originalmente como un repositorio para datos de metilación y lleva sin ser actualizada desde 2009 (por lo que no contiene metilomas de alta resolución). Las diferencias más importantes entre estas bases de datos residen en el tipo de datos que contienen y el enfoque específico de cada una:

- *MethyCancer* agrupa datos de metilación procedentes de diversas fuentes, permitiendo la comparación entre tejidos sanos y muestras cancerosas, pero lleva sin ser actualizada desde 2008 y no incluye metilomas de alta resolución.
- *DiseaseMeth* permite la comparación de datos de metilación procedentes de diversos experimentos, centrándose en el estudio de numerosas enfermedades.
- *MethylomeDB* y *PEpiD* son bases de datos tejido-específicas orientadas al estudio de ciertas patologías en determinados tejidos (depresión y esquizofrenia en el cerebro y cáncer de próstata, respectivamente).
- *Cancer Methylome System* incluye datos de metilación basados en *microarrays* para una amplia variedad de muestras tumorales, permitiendo localizar regiones diferencialmente metiladas.

Actualmente, el protocolo de referencia para el estudio de la metilación se basa en los datos de secuenciación masiva, previo tratamiento del ADN con bisulfito (Frommer, McDonald et al. 1992). Ninguna de las bases de datos mencionadas más arriba permite la comparación y visualización de metilomas completos de alta resolución. Esto se debe a la gran diversidad de programas y opciones que utilizan para el preprocesado de los datos, el alineamiento de las lecturas o la inferencia de los niveles de metilación, por lo que los resultados son muy heterogéneos y difíciles de comparar. Por el contrario, la base de datos *NGSmethDB* almacena metilomas completos de alta resolución obtenidos bajo condiciones uniformes, es decir procesados mediante los mismos programas y las mismas opciones. Esto permite el estudio comparado de dichos metilomas entre diferentes tejidos, especies o condiciones patológicas.

4.2 *NGSmethDB*

NGSmethDB surge como una base de datos flexible y sencilla de utilizar. Permite la comparación y visualización de los niveles de metilación para diferentes contextos (CpG, CpHpG y CpHpH) y en cualquier región del genoma. Además, unifica el protocolo bioinformático que procesa los datos de secuenciación masiva, permitiendo el estudio de metilomas procedentes de diferentes estudios. Las funciones de *NGSmethDB* pueden resumirse de la siguiente forma:

- Navegador genómico integrado para la visualización y comparación de metilomas de alta resolución (basado en

JBrowse (Skinner, Uzilov et al. 2009, Westesson, Skinner et al. 2013)).

- Enlaces para la visualización de metilomas en el navegador genómico de la *UCSC* (Kent, Sugnet et al. 2002, Meyer, Zweig et al. 2013), lo que posibilita la comparación con multitud de anotaciones y modificaciones epigenéticas (utilizando la función “*tracks hub*” proporcionada por el propio buscador (Karolchik, Hinrichs et al. 2009, Meyer, Zweig et al. 2013).
- Detección de citosinas con metilación diferencial en los tejidos de la base de datos (dentro de diferentes contextos de metilación: CpG, CpHpG o incluso CpHpH en el caso de plantas).
- Análisis de la metilación de diferentes regiones génicas para múltiples tejidos.
- Análisis de la metilación en listas de coordenadas genómicas (cualquier región del genoma) para múltiples tejidos.

Actualmente, la base de datos contiene metilomas para 6 especies (ensamblados: hg19, panTro4, rheMac3, mm10, tair10 e itag2.4) y 114 tejidos y/o condiciones diferentes que pueden consultarse en <http://bioinfo2.ugr.es/NGSmethDB/>.

- 4.2.1 Hackenberg, M, G Barturen, JL Oliver. 2011. NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data. *Nucleic Acids Res* 39:D75-79.
-

Dirección de acceso PubMed:

<http://www.ncbi.nlm.nih.gov/pubmed/20965971>

Dirección de publicación:

http://nar.oxfordjournals.org/content/39/suppl_1/D75.long

Página web de la base de datos:

<http://bioinfo2.ugr.es/NGSmethDB/> (*Versión actualizada*)

Breve descripción:

NGSmethDB es una base de datos orientada al almacenamiento, visualización y comparación de niveles de metilación de citosinas individuales. La base de datos contiene mapas de metilación para diferentes contextos de secuencia en múltiples especies, tejidos y/o condiciones. Las funciones disponibles son:

- Un visualizador genómico basado en *GBrowse* (Stein, Mungall et al. 2002), que permite comparar diferentes pistas de metilación con otras anotaciones para un mismo ensamblado.
- Un conjunto de herramientas de minería de datos que permiten obtener contextos no metilados o diferencialmente metilados entre tejidos. También se puede obtener el estado de metilación de promotores o de regiones genómicas definidas por el usuario.

NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data

Michael Hackenberg*, Guillermo Barturen and José L. Oliver*

Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, 18071-Granada and Lab. de Bioinformática, Inst. de Biotecnología, Centro de Investigación Biomédica, 18100-Granada, Spain

Received August 14, 2010; Revised September 28, 2010; Accepted September 30, 2010

ABSTRACT

Next-generation sequencing (NGS) together with bisulphite conversion allows the generation of whole genome methylation maps at single-cytosine resolution. This allows studying the absence of methylation in a particular genome region over a range of tissues, the differential tissue methylation or the changes occurring along pathological conditions. However, no database exists fully addressing such requirements. We propose here NGSmethDB (<http://bioinfo2.ugr.es/NGSmethDB/gbrowse/>) for the storage and retrieval of methylation data derived from NGS. Two cytosine methylation contexts (CpG and CAG/CTG) are considered. Through a browser interface coupled to a MySQL backend and several data mining tools, the user can search for methylation states in a set of tissues, retrieve methylation values for a set of tissues in a given chromosomal region, or display the methylation of promoters among different tissues. NGSmethDB is currently populated with human, mouse and *Arabidopsis* data, but other methylomes will be incorporated through an automatic pipeline as soon as new data become available. Dump downloads for three coverage levels (1, 5 or 10 reads) are available. NGSmethDB will be useful for experimental researchers, as well as for bioinformaticians, who might use the data as input for further research.

INTRODUCTION

DNA methylation is a common epigenetic mark that can be found in eukaryotes exclusively at cytosine residues (5^{mc}C). This modification has important roles in

embryonic development, as shown by early lethality in mice that lack DNA methyltransferases (DNMTs), the inactivation of the X chromosome in female cells or the establishment and maintenance of allele-specific expression of imprinted genes (1–3). Numerous studies over the past decades suggest that cytosine DNA methylation functions to maintain the repressed chromatin state and therefore stably silence promoter activity (4). In animal genomes, the predominantly methylated sequence context is the dinucleotide CpG, while non-CpG methylation exists in plants that is targeted to transposable elements by a mechanism that depends upon small interfering RNAs (5). Recently, methylation at sequence contexts CHH and CHG has been detected in human undifferentiated cells (6).

Many different techniques have been developed for DNA methylation profiling (7,8). The detection methods can be divided into a methylation-dependent pretreatment and an analytical step. The first step is necessary as 5^{mc}C is not readily distinguished from unmethylated cytosine by hybridization-based methods and PCR amplification erases the DNA methylation information. Basically, three different pretreatments can be distinguished: enzyme digestion, affinity enrichment (immunoprecipitation) and sodium bisulphite conversion. The information on the DNA methylation is finally read out by a gel-based, array-based or sequencing-based analysis. Virtually, all combinations of these two steps exist. Depending on the specific combination used, we can distinguish between 'single cytosine' and 'region wide' profiling of methylation states. The region wide methods detect normally the methylation states of known CpG islands or unmethylated fragments using either enzyme digestion or immunoprecipitation. There are several drawbacks with these methods. Apart from the errors introduced by the methylation-dependent pretreatment, only 'mean values' of the regions can be detected. Although for many experiments it might be sufficient to get information whether a

*To whom correspondence should be addressed. Tel: +34 958243261; Fax: +34 958244073; Email: mlhack@gmail.com
Correspondence may also be addressed to José L. Oliver. Tel: +34 958243261; Fax: +34 958244073; Email: oliver@ugr.es

© The Author(s) 2010. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.5>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

given region is methylated or unmethylated, for others it will be not. For example, recently it has been shown that many CpG islands show internal fluctuations that can be resolved by means of single-cytosine resolution analysis (9,10). Furthermore, single-cytosine resolution data can be critical to resolve the methylation states, and the possible functionality, of very small islands (islets) or even orphan CpG dinucleotides (10,11).

However, to completely exploit the full potential of single-base resolution whole genome methylation maps, a specifically designed database is needed. Given the lack of single base data in the past, current databases are only focused either on specific regions and/or on pathologic situations (12,13).

In the next years, however, whole genome methylation data will become available for many new tissues, pathological conditions and species and it will be of critical importance to store and unify this information in an adequate way. We therefore propose here NGSmethDB, a database for single-cytosine resolution methylation data. The database uses a web interface based on GBrowse (14) and coupled to a MySQL backend, which allows to visualize the methylation data in a genomic context together with many other annotations, as well as full data downloads. In addition, a set of powerful data mining tools are also implemented, so the user can filter, analyze and retrieve data in many different ways. For example, the user can search for unmethylated or differentially methylated cytosines in a selected set of tissues, or display and analyze the promoter methylation of RefSeq genes. Finally, the database extends the commonly used focus on CpG dinucleotides to the recently discovered non-CpG targets for DNA methylation in undifferentiated tissues (6).

FEATURES AND SCOPE

The NGSmethDB database can be divided into two parts. First, the content can be visualized, together with many other common annotations, by means of a web interface based on GBrowse (14) coupled to a MySQL backend; and second, several user-friendly data mining tools are provided so the average user can generate its own data sets easily. Currently, the database holds information on three species (human, mouse and *Arabidopsis*) and 52 different tissues (21 unique tissues). Furthermore, two different methylation contexts are considered, CpG and CWG, but other non-CpG contexts, as CAH or CHH, will be soon available. Currently, the database holds methylation data of 696 599 217 cytosines for human (hg18), 69 459 481 cytosines for mouse (mm8) and 16 321 229 cytosines for *Arabidopsis* (TAIR8). A detailed and updated database statistical table is maintained on-line: <http://bioinfo2.ugr.es/gbrowse2/StatGraphs/datasourcesrpt.php>. A summary of the publications where the data were generated from is also maintained and updated on-line: <http://bioinfo2.ugr.es/gbrowse2/DataSource/datasourcesrpt.php?start=1>.

We encourage data submissions of new methylation data in order to populate and maintain updated NGSmethDB.

For most data, the methylation information for the cytosines is directly available for the three mentioned genome assemblies. In these cases, we populate the database with these processed data. For other cases, we used the LiftOver tool (15) to convert the coordinates from other assemblies, or developed scripts to process the raw data (like *fastaq* files) in order to obtain the methylation information for all covered cytosines. All methylation values for both CpG and CWG contexts are calculated taking into account both strands. The assigned methylation value is therefore a weighted mean between the context in the direct and reverse strands. Which means that it is the sum of reads that indicate methylation (cytosine not converted to uracil/thymine) mapped to the specific position in the '+' strand and those mapped to the '-' strand, divided by the total number of reads mapped to the position regardless of the strand.

Genomic browser interface

The GBrowse genome viewer (14) connected to a MySQL backend is used to set up a web browser interface for NGSmethDB. Features of the browser include the ability to scroll and zoom through arbitrary regions of a genome, to enter a region of the genome by searching for a landmark or performing a full text search of features, as well as the ability to enable and disable feature tracks and change their relative order and appearance. The user can also upload private annotations to view them in the context of the existing ones at the NGSmethDB web site.

Apart from the methylation data, the following related annotations are currently available on the NGSmethDB browser: (i) CpGcluster CpG islands (16); (ii) Takai-Jones CpG islands (17); (iii) RefSeq genes (18); (iv) HMR conserved TFBSs (19); (v) CisRED regulatory elements (20); and (vi) the chromosome sequence (hg18, mm8 and TAIR8 genome assemblies) and G + C content.

The methylation information of a given context is represented by the coordinate of the cytosine on the direct strand. To display the methylation values of the cytosines we use a color gradient from white (methylation value = 0, unmethylated in all reads) to red (methylation value = 1, methylated in all reads). To demonstrate the usefulness of the web interface, we analyzed the promoter region of the gene *TIAM1* (Figure 1). It can be seen that this promoter is differentially methylated among the different tissues.

Data mining tools

Currently, five different ways are implemented to retrieve raw data from the database. For all five possibilities, two different sequence contexts and three coverage levels exist. We detected not just the methylation values of CpG dinucleotides but also for the cytosines in a CWG (CAG or CTG) context. The methylation value at a given position (cytosine) is calculated as explained before taking both strands into consideration. We stored three different coverage levels in the database: cytosines covered by at least 1, 5 and 10 reads.

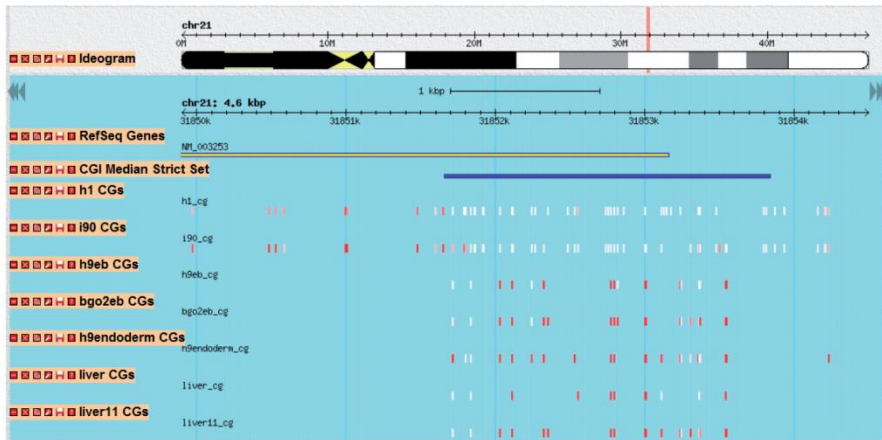


Figure 1. Visualization of methylation states of CpG dinucleotides in different tissues in the NGSmethDB genome browser. The promoter region of the gene *TIAM1* (NM_003253) is shown. The different methylation values are displayed by means of a color gradient from white (unmethylated in all reads) toward red (methylated in all reads).

Dump download

This option shows first an overview of current database content, including a short description of the tissue, the genome coverage in %, a link to PubMed, and raw data files for #read ≥ 1 , #read ≥ 5 and #read ≥ 10 coverage. The files show the chromosome, chromosome-start and chromosome-end coordinates, the sequence methylation context (either CpG or CWG), the number of reads and the cytosine methylation ratio.

Retrieve unmethylated contexts

This tool can be used to retrieve all unmethylated cytosines in a given set of tissues. The user has to select the sequence context (CG or CWG), the read coverage, the threshold for unmethylation (often a threshold of 0.2 is used, i.e. all cytosines with values ≤ 0.2 are considered to be unmethylated) and the tissues. The tool will detect all cytosine contexts showing lower methylation ratios than the chosen threshold in all selected tissues. The provided output file holds the chromosome, chromosome start- and end-coordinates and the methylation values in all selected tissues. Note that this tool can be also used to retrieve all CpGs which are present in every single analyzed tissue by setting the threshold to one. In doing so, cytosines with methylation data in all tissues will be reported regardless of its methylation state, i.e. cytosines that are not covered by at least the number of chosen coverage threshold (1, 5 or 10) in any of the analyzed tissues will not be reported in the output.

Retrieve differentially methylated contexts

By means of this tool all differentially methylated cytosine contexts can be determined in a given set of tissues. All parameters of the 'Retrieve unmethylated contexts' (see above) are available here, plus one additional

parameter: the threshold for the methylation value which defines whether a cytosine is considered to be methylated (often a threshold of 0.8 is used, i.e. all cytosines with higher values than ≥ 0.8 are considered to be methylated). We define a cytosine as differentially methylated if it is unmethylated in at least one tissue and methylated in at least one other tissue. The tool reports those differentially methylated cytosine contexts that are either methylated or unmethylated in all analyzed tissues, i.e. those contexts that show intermediate methylation in only one tissue will not be reported.

Get methylation states of promoter regions

This tool allows depicting the methylation states of all cytosine contexts within the promoter region of RefSeq genes. We define the promoter region as beginning 1.5kb upstream of the Transcription Start Site (TSS) and ending 500bp downstream of the TSS. The user needs to provide a valid RefSeq name (NM_*) or a unique TAIR gene id (ATxGxxxxx) and the desired coverage. The output is displayed by default as an overview table that summarizes the fluctuation along the promoter as well as over the different tissues. A detailed table can also be generated (Figure 2).

Retrieve methylation data for chromosome region

All methylation values for a selected set of tissues can be retrieved for a given chromosomal region, once the user provides the start and end chromosome coordinates.

CONCLUSIONS

Over the next years, methylation data for a growing number of tissues, cell types, pathological conditions and diverse species will all be available. In most of the original

| | | | | | | | | | |
|------|----------|----|---|---|--|---|-------|-------|-------|
| 1293 | 31853369 | 3 | 3 | 0 | fibro hesc headfibro | | 0.004 | 0.000 | 0.013 |
| 1304 | 31853358 | 3 | 3 | 0 | fibro hesc headfibro | | 0.000 | 0.000 | 0.000 |
| 1313 | 31853349 | 6 | 3 | 2 | fibro hesc headfibro | h9endoderm h9esc | 0.405 | 0.000 | 1.000 |
| 1320 | 31853342 | 11 | 3 | 8 | fibro hesc headfibro | bgo2esc h9afp h9eb h9endoderm h9esc h9noafp hct116 liver24 | 0.728 | 0.000 | 1.000 |
| 1330 | 31853332 | 12 | 3 | 9 | fibro hesc headfibro | bgo2eb bgo2esc h9afp h9eb h9endoderm h9esc h9noafp hct116 liver24 | 0.751 | 0.000 | 1.000 |
| 1338 | 31853324 | 3 | 3 | 0 | fibro hesc headfibro | | 0.004 | 0.000 | 0.013 |
| 1340 | 31853322 | 12 | 3 | 9 | fibro hesc headfibro | bgo2eb bgo2esc h9afp h9eb h9endoderm h9esc h9noafp hct116 liver24 | 0.752 | 0.000 | 1.000 |
| 1345 | 31853317 | 3 | 3 | 0 | fibro hesc headfibro | | 0.000 | 0.000 | 0.014 |
| 1352 | 31853310 | 3 | 3 | 0 | fibro hesc headfibro | | 0.000 | 0.000 | 0.000 |
| 1358 | 31853304 | 3 | 3 | 0 | fibro hesc headfibro | | 0.009 | 0.000 | 0.028 |
| 1364 | 31853298 | 6 | 3 | 3 | fibro hesc headfibro | h9afp h9noafp hct116 | 0.512 | 0.000 | 1.000 |
| 1384 | 31853278 | 3 | 3 | 0 | fibro hesc headfibro | | 0.000 | 0.000 | 0.000 |
| 1398 | 31853264 | 8 | 8 | 0 | bgo2eb fibro h9afp h9noafp hct116 hesc headfibro liver11 | bgo2eb h9afp h9noafp hct116 liver11 | 0.003 | 0.000 | 0.028 |
| 1402 | 31853260 | 8 | 3 | 5 | fibro hesc headfibro | | 0.627 | 0.000 | 1.000 |
| 1408 | 31853254 | 8 | 8 | 0 | bgo2eb fibro h9afp h9noafp hct116 hesc headfibro liver11 | | 0.000 | 0.000 | 0.000 |
| 1412 | 31853250 | 4 | 3 | 1 | fibro hesc headfibro | h9esc | 0.253 | 0.000 | 1.000 |
| 1414 | 31853248 | 9 | 3 | 6 | fibro hesc headfibro | bgo2eb h9afp h9esc h9noafp hct116 liver11 | 0.670 | 0.000 | 1.000 |
| 1418 | 31853244 | 10 | 3 | 7 | fibro hesc headfibro | bgo2eb h9afp h9eb h9esc h9noafp hct116 liver24 | 0.697 | 0.000 | 1.000 |
| 1429 | 31853233 | 4 | 3 | 1 | fibro hesc headfibro | h9endoderm | 0.253 | 0.000 | 1.000 |
| 1432 | 31853230 | 4 | 3 | 1 | fibro hesc headfibro | h9endoderm | 0.253 | 0.000 | 1.000 |
| 1435 | 31853227 | 4 | 3 | 1 | fibro hesc headfibro | h9endoderm | 0.250 | 0.000 | 1.000 |
| 1438 | 31853224 | 11 | 3 | 8 | fibro hesc headfibro | bgo2eb h9afp h9eb h9esc h9noafp hct116 liver11 liver24 | 0.730 | 0.000 | 1.000 |
| 1441 | 31853221 | 11 | 3 | 8 | fibro hesc headfibro | bgo2eb h9afp h9eb h9esc h9noafp hct116 liver11 liver24 | 0.728 | 0.000 | 1.000 |
| 1445 | 31853217 | 4 | 3 | 1 | fibro hesc headfibro | h9endoderm | 0.261 | 0.000 | 1.000 |
| 1447 | 31853215 | 3 | 3 | 0 | fibro hesc headfibro | | 0.008 | 0.000 | 0.012 |
| 1463 | 31853199 | 3 | 3 | 0 | fibro hesc headfibro | | 0.009 | 0.000 | 0.027 |
| 1465 | 31853197 | 3 | 3 | 0 | fibro hesc headfibro | | 0.000 | 0.000 | 0.000 |

Figure 2. Detailed analysis of the promoter region of the gene *TIAM1* (NM_003253) by means of the NGSmethDB data mining tools. The table shows the following columns: relative coordinate towards the point TSS-1.5 kb, the chromosomal coordinate of the cytosine, the number of tissues for which methylation data exists, the number of tissues where the cytosine were found to be unmethylated, the number of tissues where the cytosine were found to be methylated, the tissue names where the cytosine is methylated and unmethylated, respectively, the mean methylation value among all tissues, the minimum and maximum methylation values over all tissues. By means of the color code, green for unmethylation (value ≤ 0.2) and red for methylation (value ≥ 0.8), the situation can be rapidly analyzed. For example, if both minimum and maximum values are green for one cytosine position, this means that this cytosine is unmethylated in all analyzed tissues. On the other hand, if the minimum value is green and the maximum value is red, this indicates differential methylation over the different tissues for the given cytosine.

publications, the authors focus on concrete questions and scarcely the whole potential of the data can be exploited. To get more out of these data, a joint analysis with data from other tissues and/or species is needed. To carry out such analysis, data must be first stored in an appropriate way in a database. We propose here NGSmethDB, a new database with a very broad scope to facilitate the analysis of methylation data from different sources. Heterogeneous methylation data can be either simultaneously visualized through a powerful web interface or selectively downloaded by means of the provided data mining tools that allow the user to design new experiments and retrieve exactly the adequate data for them. Thus, we are confident that the database will be of great usefulness both for experimental and bioinformatics researchers.

FUNDING

The Spanish Government grant (BIO2008-01353 to J.L.O.); ‘Juan de la Cierva’ (to M.H.); Basque Country ‘Programa de formación de investigadores’ grant (to G.B.). Funding for open access charge: The Spanish Government grant (BIO2008-01353 to J.L.O.).

Conflict of interest statement. None declared.

REFERENCES

- Chen,T. and Li.E. (2004) Structure and function of eukaryotic DNA methyltransferases. *Curr. Top. Dev. Biol.*, **60**, 55–89.
- Karpf,A.R. and Matsui,S. (2005) Genetic disruption of cytosine DNA methyltransferase enzymes induces chromosomal instability in human cancer cells. *Cancer Res.*, **65**, 8635–8639.
- Dodge,J.E., Okano,M., Dick,F., Tsujimoto,N., Chen,T., Wang,S., Ueda,Y., Dyson,N. and Li.E. (2005) Inactivation of Dnmt3b in mouse embryonic fibroblasts results in DNA hypomethylation, chromosomal instability, and spontaneous immortalization. *J. Biol. Chem.*, **280**, 17986–17991.
- Bird,A.P. and Wolffe,A.P. (1999) Methylation-induced repression: belts, braces, and chromatin. *Cell*, **99**, 451–454.
- Chan,S.W., Henderson,I.R. and Jacobsen,S.E. (2005) Gardening the genome: DNA methylation in Arabidopsis thaliana. *Nat. Rev. Genet.*, **6**, 351–360.
- Lister,R., Pelizzola,M., Downe,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.
- Laird,P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
- Beck,S. and Rakyen,V.K. (2008) The methylome: approaches for global DNA methylation profiling. *Trends Genet.*, **24**, 231–237.
- Hodges,E., Smith,A.D., Kendall,J., Xuan,Z., Ravi,K., Rooks,M., Zhang,M.Q., Ye,K., Bhattacharjee,A., Brizuela,L. et al. (2009) High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Res.*, **19**, 1593–1605.
- Hackenberg,M., Barturen,G., Carpena,P., Luque-Escamilla,P.L., Previti,C. and Oliver,J.L. (2010) Prediction of CpG-island

- function: CpG clustering vs. sliding-window methods. *BMC Genomics*, **11**, 327.
11. Wong, N.C., Wong, L.H., Quach, J.M., Canham, P., Craig, J.M., Song, J.Z., Clark, S.J. and Choo, K.H. (2006) Permissive transcriptional activity at the centromere through pockets of DNA hypomethylation. *PLoS Genet.*, **2**, e17.
 12. Ongenaert, M., Van Neste, L., De Meyer, T., Menschaert, G., Bekaert, S. and Van Criekinge, W. (2008) PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res.*, **36**, D842–D846.
 13. Amoreira, C., Hindermann, W. and Grunau, C. (2003) An improved version of the DNA Methylation database (MethDB). *Nucleic Acids Res.*, **31**, 75–77.
 14. Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
 15. LiftOver: <http://genome.ucsc.edu/cgi-bin/hgLiftOver> (May 2010, date last accessed).
 16. Hackenberg, M., Previti, C., Luque-Escamilla, P.L., Carpena, P., Martínez-Aroza, J. and Oliver, J.L. (2006) CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics*, **7**, 446.
 17. 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.
 18. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
 19. Weirauch, M. and Raney, B. (2007) HMR Conserved transcription factor binding sites. *UCSC Genome Browser*. <http://genome.ucsc.edu/cgi-bin/hgGateway> (May 2010, date last accessed).
 20. Robertson, G., Bilenky, M., Lin, K., He, A., Yuen, W., Dagpinar, M., Varhol, R., Teague, K., Griffith, O.L., Zhang, X. *et al.* (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res.*, **34**, D68–D73.

- 4.2.2 Geisen S, Barturen G, Alganza AM, Hackenberg M, Oliver JL 2014. NGSmethDB: an updated genome resource for high quality, single-cytosine resolution methylomes. *Nucleic Acids Res* 42: D53-59.
-

Dirección de acceso PubMed:

<http://www.ncbi.nlm.nih.gov/pubmed/24271385>

Dirección de publicación:

<http://nar.oxfordjournals.org/content/42/D1/D53.long>

Página web de la base de datos:

<http://bioinfo2.ugr.es/NGSmethDB/>

Breve descripción:

Actualización de la base de datos *NGSmethDB*, donde además de aumentar el número de especies y tejidos, se ha incluido el contexto de secuencia CpHpH para plantas. También se ha mejorado considerablemente la interfaz web y se han incluido algunas herramientas nuevas que facilitan la interacción con los datos almacenados:

- El navegador genómico se ha sustituido por *JBrowse* (Skinner, Uzilov et al. 2009), más apropiado para datos procedentes de secuenciación masiva.
- Las pistas de metilación se han incluido en el buscador genómico de la *UCSC*, donde se encuentran disponibles múltiples anotaciones.

- El estado de metilación puede obtenerse para otras regiones génicas, además de los promotores.
- Las herramientas de minería de datos devuelven resultados para múltiples tejidos a la vez, permitiendo identificar regiones diferencialmente metiladas.
- Se ha optimizado el proceso de actualización de la base de datos, lo que permitirá aumentar el número de especies y tejidos.

NGSmethDB: an updated genome resource for high quality, single-cytosine resolution methylomes

Stefanie Geisen¹, Guillermo Barturen¹, Ángel M. Alganza¹, Michael Hackenberg^{1,2,*} and José L. Oliver^{1,2,*}

¹Facultad de Ciencias, Departamento de Genética, Universidad de Granada, 18071-Granada, Spain and

²Laboratorio de Bioinformática, Instituto de Biotecnología, Centro de Investigación Biomédica, 18100-Granada, Spain

Received September 13, 2013; Revised November 3, 2013; Accepted November 4, 2013

ABSTRACT

The updated release of 'NGSmethDB' (<http://bioinfo2.ugr.es/NGSmethDB>) is a repository for single-base whole-genome methylome maps for the best-assembled eukaryotic genomes. Short-read data sets from NGS bisulfite-sequencing projects of cell lines, fresh and pathological tissues are first pre-processed and aligned to the corresponding reference genome, and then the cytosine methylation levels are profiled. One major improvement is the application of a unique bioinformatics protocol to all data sets, thereby assuring the comparability of all values with each other. We implemented stringent quality controls to minimize important error sources, such as sequencing errors, bisulfite failures, clonal reads or single nucleotide variants (SNVs). This leads to reliable and high-quality methylomes, all obtained under uniform settings. Another significant improvement is the detection in parallel of SNVs, which might be crucial for many downstream analyses (e.g. SNVs and differential-methylation relationships). A next-generation methylation browser allows fast and smooth scrolling and zooming, thus speeding data download/upload, at the same time requiring fewer server resources. Several data mining tools allow the comparison/retrieval of methylation levels in different tissues or genome regions. NGSmethDB methylomes are also available as native tracks through a UCSC hub, which allows comparison with a wide range of third-party annotations, in particular phenotype or disease annotations.

INTRODUCTION

DNA methylation is an epigenome mark involved in key biological processes (1–3), such as embryonic development, transcription, genomic imprinting, learning, memory or age-related cognitive decline (4–7). DNA methylation plays an important role in the origin and function of CpG islands (CGIs). Aberrant methylation (mostly hypermethylation) of CGIs has been implicated in the appearance of several disorders, such as cancer, immunodeficiency or centromere instability (8–14).

Many different techniques are available for DNA methylation profiling (15,16). Region-wide methods detect the methylation states of known CGIs or unmethylated fragments using either enzyme digestion or immunoprecipitation, but frequently only 'mean values' of the corresponding regions can be derived from these methods. The advent of next-generation sequencing (NGS), together with bisulfite conversion of DNA, allows the generation of whole genome methylation maps at single-cytosine resolution (17–19). This provides an opportunity for studying important biological phenomena, such as the absence of methylation in a particular genome region over a range of tissues, the differential tissue methylation or the changes occurring along pathological conditions.

Several methylation databases centered in gene loci (20–23), tissues (24,25) or diseases (26–28) have been compiled. However, a wide variety of methodologies to pre-process the data, aligning the reads or inferring the methylation states has been used in compiling these databases, thus leading to methylomes obtained with very different methods or parameter sets to be included into the same database, which can bias downstream analyses. Additional problems are the regional resolution or the partial coverage of only some specific genome regions,

*To whom correspondence should be addressed. Tel: +34 958243261; Fax: +34244073; Email: oliver@ugr.es
Correspondence may also be addressed to Michael Hackenberg. Tel: +34 958249695; Fax: +34244073; Email: hackenberg@ugr.es

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

© The Author(s) 2013. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

which makes it difficult to use these data for comparative analyses. However, the single-base whole-genome methylomes stored in the new version of the 'NGSmethDB' database are all obtained using the same set of programs/scripts, and derived under the same settings and quality controls, thus allowing consistent comparative analyses of whole-genome methylomes.

NGSmethDB CONTENT

Publicly available short-read data sets from NGS bisulfite-sequencing projects for different cell lines, fresh tissues and pathological tissues were downloaded mainly from NCBI GEO (29). An updated list of the data sets used for each genome, with detailed information on the source cell-line or tissue, is maintained online (<http://bioinfo2.ugr.es/NGSmethDB/database.php>).

To date, the database includes 87 methylome maps generated for CpG and CpHpG (H = A,C,T) sequence contexts in five different species for the most recent genome assembly: *Homo sapiens* (hg19), *Pan troglodytes* (panTro4), *Macaca mulatta* (rheMac3), *Mus musculus* (mm10) and *Arabidopsis thaliana* (tair10). The number of available methylomes by species was also increased: *Homo sapiens* (17), *Pan troglodytes* (5), *Macaca mulatta* (6), *Mus musculus* (30) and *Arabidopsis thaliana* (18). We restructured the database allowing the easy incorporation of novel species and/or methylomes, which ensures that the database will be always well-curated and maintained.

EPIGENOME-WIDE METHYLOME MAPS

A flow diagram delineating the implementation and main features of NGSmethDB is shown in Figure 1. Short-read data sets were pre-processed and aligned to the corresponding reference genome using 'NGSmethPipe' (31), and then profiling the methylation levels by means of 'MethylExtract' (32).

Alignment of short-reads

NGSmethPipe (<http://bioinfo2.ugr.es/NGSmethPipe/>) implements several pre-processing steps to improve the alignment quality, like the trimming prior to the adapter detection. It uses 'Bowtie' (33) as an external aligner applied on a three-letter alphabet. To map a higher number of reads without compromising the mapping quality, NGSmethPipe uses a 'seed extension' method applied to the Bowtie alignments, similar to that used in 'miRanalyzer' (34,35). Short-read alignment per se is a highly parameterized process. Adding the NGSmethPipe-specific parameters results in obtaining a notable parameter space. Relaxed parameters will lead to a higher coverage (i.e. many cytosines can be profiled), but a higher number of incorrect alignments can also be expected. On the contrary, strict parameters might lead to a lower coverage, thereby discarding a considerable amount of valuable information. For the presented database, we carried out a careful study to measure alignment accuracy as a function of the seed length and number of mismatches to obtain the

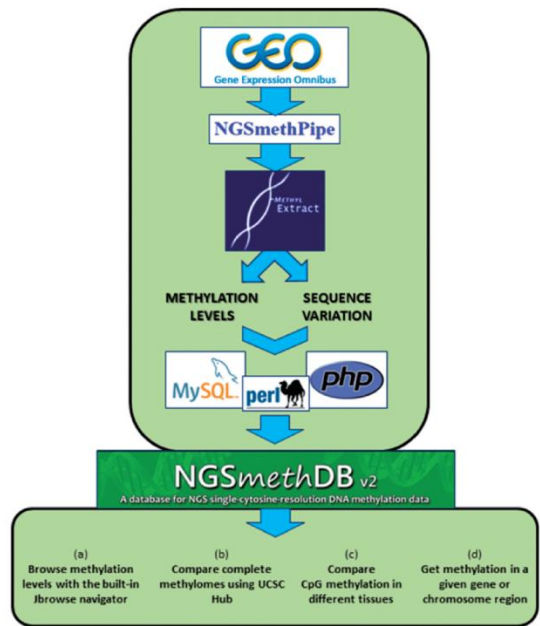


Figure 1. Flow diagram showing the implemented steps and main features of NGSmethDB.

best parameter set. NGSmethPipe now uses these settings as default options (see the 'Quick start' section in <http://bioinfo2.ugr.es/NGSmethPipe/Manual.html> for a complete list of defaults).

Methylation profiling

For the methylation profiling carried out by Methyl-Extract (<http://bioinfo2.ugr.es/MethylExtract/>), we implemented a number of stringent quality controls, carefully chosen to minimize important error sources [see (32) for a complete description]:

- (1) A first potential error source in methylation profiling is the bisulfite conversion failure. In modern protocols, usually <1% of all unmethylated cytosines fail to be converted by bisulfite treatment. Thus, some positions are incorrectly profiled, i.e. some inferred methylcytosines are actually unmethylated. To cope with this error, we first implemented (as an option) a method proposed by Lister (17) to detect reads with a high number of unconverted cytosines: if this option is activated, the reads with at least 90% of unconverted cytosines in non-CpG contexts were eliminated. Second, when a non-methylated genome is available (e.g. the chloroplast genome for *Arabidopsis* data sets), MethylExtract can associate a *P*-value, based on binomial statistics, and a false discovery rate to the extracted methylation levels [see (32) for details]. For the sake of uniformity, and given the lack of non-methylated genomes for all

the included species, we do not use this feature in populating NGSmethDB. However, when using the data mining tools, the user can choose the minimum coverage required for a cytosine methylation context. In addition, the methylation browser shows all the individual methylation values.

- (2) Other potential sources for incorrect methylation profiling are sequencing errors. We used the assigned Phred score (36) to limit the contribution of incorrectly sequenced bases. By setting $Q \geq 20$, we are only accepting bases with a $P < 0.01$ to be incorrectly called.
- (3) In methylation profiling, SNVs are probably the most disregarded error source. Over two-thirds of all SNPs occur in a CpG context, having two alleles: C/T or G/A (37). Most other tools would interpret a C>T substitution as an unmethylated cytosine, although a certain number of them are actually SNVs, and therefore the inference would be wrong. A C/T SNV manifests on the complementary DNA strand as an adenine, while bisulfite deamination does not affect the guanine on the complementary strand (38). We take advantage of this observation to detect putative SNVs by means of a threshold method based on VarScan, thus avoiding subsequent erroneous methylation profiling.
- (4) Duplicated (clonal) reads provoked by the polymerase chain reaction step adds another layer of potential errors in methylation profiling. MethylExtract implements an option to delete duplicated reads without eliminating meaningful biological information. In populating NGSmethDB, we used this option of MethylExtract.
- (5) Lastly, when needed, we carried out 5' end trimming of reads. As implemented in 'Bismark' (39), the first N nucleotides are removed from the 5' end of the read (3nt in case of the MspI restriction sites of the reduced representation bisulfite sequencing protocol).

Methylome maps

The resulting high-quality methylomes, obtained under uniform settings as indicated earlier in the text, were stored in a 'MySQL' database back-end, which is used to serve visualization, data mining and database dumps. Methylation maps for minimum coverages of 1, 3, 5 or 10 reads (<http://bioinfo2.ugr.es/NGSmethDB/database.php>) were generated. We used 'Perl' scripts to automate data parsing and database management.

An outstanding feature of MethylExtract is the calling of SNVs from the same sequence library of bisulfite-treated DNA used to infer methylation states. Therefore, besides methylation tracks, SNV tracks were also generated for each sample and made available for download or visualization through the methylation browser.

THE METHYLATION BROWSER

The user interface was improved by replacing 'Gbrowse' with 'Jbrowse' (40,41), resulting in a methylation browser

with a fast and smooth scrolling and zooming mechanism (Figure 2). This speeds data download and upload, and requires light server resources.

Users can include their own data in 'bigWig', 'VCF', 'gff' or 'bed' formats (<https://genome.ucsc.edu/FAQ/FAQformat.html>), thus comparing their data directly with the NGSmethDB methylomes. User data sets are not uploaded to the server, but instead opened directly via the Java interface. This ensures a quick and stable data integration without compromising the server stability and response time.

RefSeq (30) gene names were indexed, thus making them searchable via the browser interface. In addition, NGSmethDB includes many other annotation tracks (CpGislands, promoters, SNPs, repeats, isochores, phastCons) that can be viewed and compared with the methylation maps.

A detailed manual (<http://bioinfo2.ugr.es/NGSmethDB/manual.php>) guides the user through the different steps to quickly browse the Web site and download NGSmethDB methylation maps. Furthermore, a general and context-dependent help about searching, moving, zooming and showing/hiding tracks with JBrowse has been interactively integrated in the proper methylation browser window.

A UCSC TRACK HUB FOR NGSmethDB METHYLomes

We also made NGSmethDB methylation maps directly available through a UCSC track hub, a web-accessible directory of genomic data that can be viewed on the UCSC genome browser (<http://genome.ucsc.edu/goldenPath/help/hgTrackHubHelp.html>). Therefore, high-quality NGSmethDB methylomes can be visualized and tuned on the UCSC genome browser as native tracks. This allows the comparison with a wide range of third-party annotations, in particular phenotype and disease associations, or the ENCODE annotation tracks.

DATA MINING TOOLS

Similar to the first version of NGSmethDB, the user interface was based on the practical appeals of epigenome-wide analysis: namely, the possibility to (i) obtain methylation values for particular chromosomal regions or tissues, (ii) analyze promoter methylation for a set of tissues and (iii) compare methylation patterns across a set of different tissues. To this end, three different database mining tools were developed to allow the user to filter, compare, analyze and download the methylation data in different species, tissues, developmental stages or diseases:

- (1) Comparison of cytosine methylation levels in different tissues. The user can select the sequence context (CG or CHG) and the methylation states for comparison: methylated versus unmethylated, methylated versus intermediate, unmethylated versus intermediate or all of them.

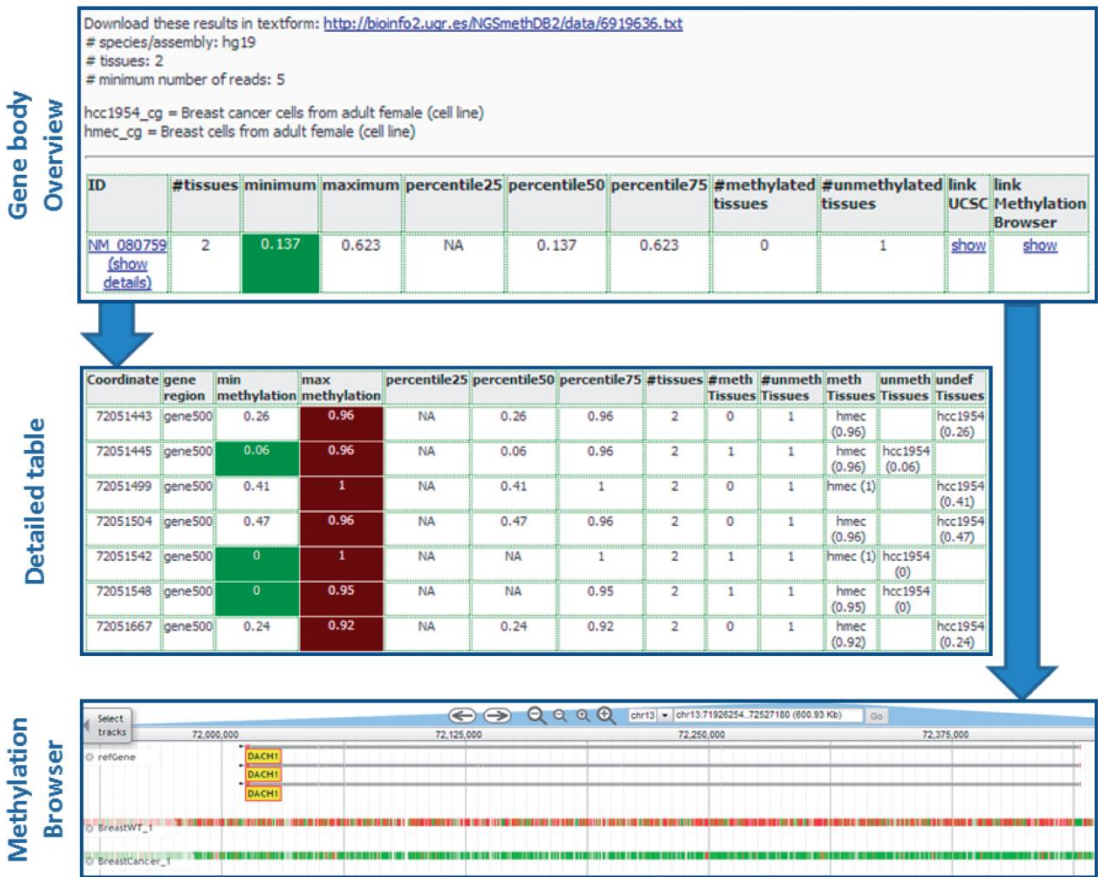


Figure 2. Gene hypomethylation in the *DACH1* tumor suppressor gene. The figure shows the average CpG methylation in the gene body (Gene body Overview), the methylation levels at single cytosines (detailed table) and its visualization in the methylation browser for normal (*hmec*) and cancer (*hcc1954*) breast cell lines. Average and single-base CpG methylation levels can be downloaded for further analysis. Short-read samples GSM721195 HMEC-methylC-Seq and GSM721194 HCC1954-methylC-Seq (42), downloaded from GEO (29), were used to generate the corresponding methylome maps.

- (2) The methylation states of different gene regions, including gene body, promoters, 3' ends, exons and introns, can be retrieved/downloaded.
- (3) Methylation data for single cytosines within a given chromosome region can be retrieved/downloaded; a detailed table is provided with direct links to our methylation browser and the UCSC genome browser.

New features in this version of the database are the possibility to supply a customized set of regions in bed format (<https://genome.ucsc.edu/FAQ/FAQformat.html>) to obtain the methylation levels or a gene list to retrieve the data in a given gene region. Depending on the amount of requested data (mainly, the number of tissues), some of these tools might take several hours to process the requested data. To overcome this limitation, we

implemented PHP sessions (<http://php.net/manual/en/ref.session.php>), thus offering the user the possibility to submit >1 job at a time. An ID is assigned to each submitted job. Running jobs are shown under the header 'running', providing the possibility to also cancel the jobs. Once finished, a long life link becomes available, allowing the user to retrieve the results within 30 days. If there are >5 jobs running from the same user, the next job gets queued and will be executed automatically as soon as the previous job has finished.

WORKING EXAMPLES

As a first example, the hypomethylation of the *DACH1* tumor suppressor gene (42) was analyzed by means of NGSmethDB. Human *DACH1* on chromosome 13

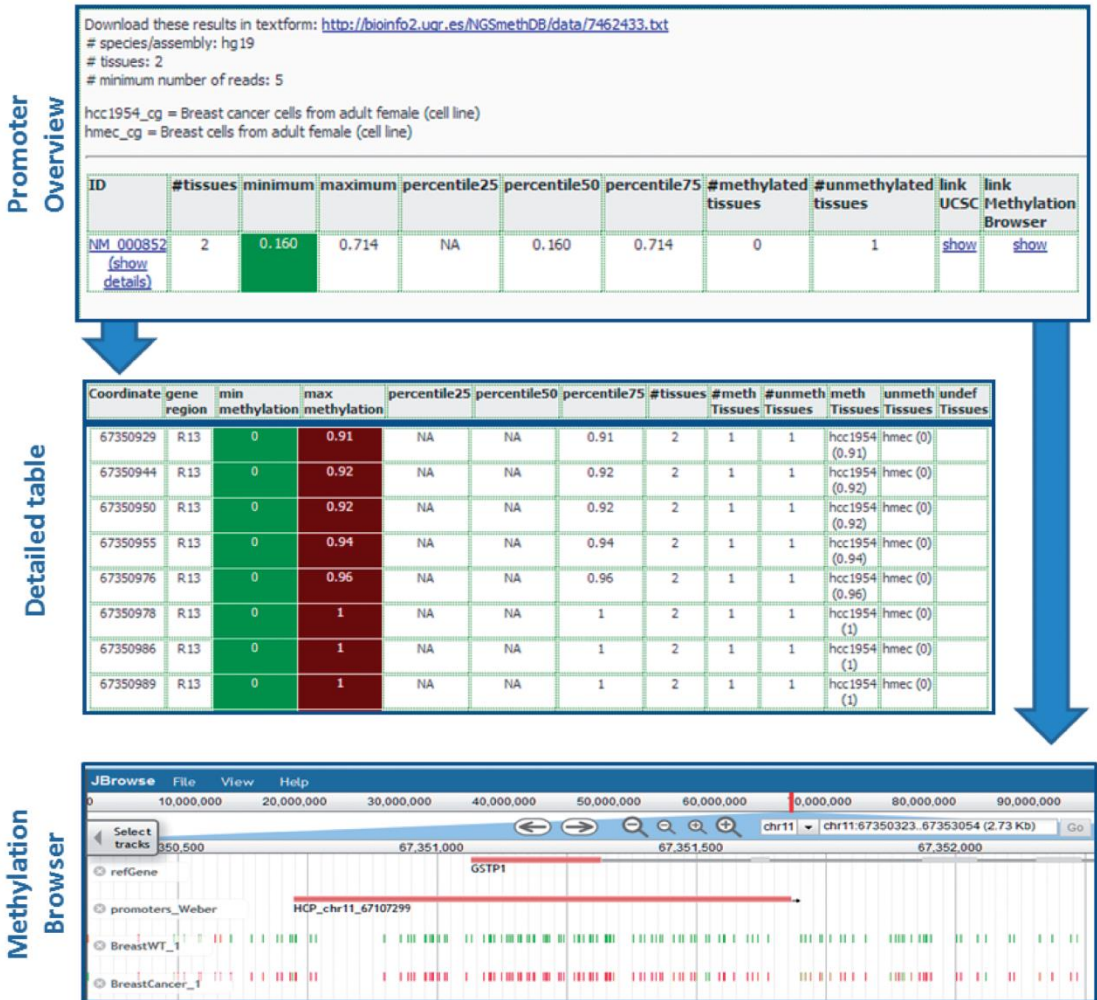


Figure 3. *GSTP1* hypermethylation in breast cancer. *GSTP1* codes for the glutathione S-transferase Pi-1. The screenshot of the NGSmethDB methylation browser (bottom) corresponds to positions 67349906–67356735 of the human chromosome 11. The promoter region, as defined in ref. (44), and the NGSmethDB methylation maps for normal (*hmec*) and cancer (*hcc1954*) breast cell lines are shown. The healthy breast promoter appears as unmethylated (green vertical bars), whereas the breast cancer tissue is heavily methylated (red vertical bars). Some rows of the detailed methylation table at single cytosines with coverage of at least five reads are shown (middle).

encodes a chromatin-associated protein that associates with other DNA-binding transcription factors to regulate gene expression and cell fate determination during development. Figure 2 shows the results when analyzing the gene body methylation of this gene for normal (*hmec*) and cancer (*hcc1954*) breast cell lines. NGSmethDB first shows a summary statistics of the methylation levels across the used set of tissues (Figure 2, top), also providing links to a table with detailed methylation levels at single cytosine resolution (Figure 2, middle) and its visualization in the methylation browser (Figure 2, bottom). A global gene

hypomethylation in breast cancer, as compared with healthy tissue, can be clearly appreciated.

A second example shows the analysis of the hypermethylation of the *GSTP1* promoter in cancer. This gene codes for the glutathione S-transferase Pi-1, an enzyme involved in cellular detoxification of xenobiotics and carcinogens, being a promising biomarker for cancer diagnosis and prognosis (43). The methylation map of the promoter region in normal and cancer breast tissue provided by NGSmethDB is shown in Figure 3 (bottom). A detailed table with methylation values at individual CpGs is shown in Figure 3 (middle).

NGSmethDB analysis clearly shows the hypermethylation of this promoter region in breast cancer.

Lastly, NGSmethDB methylomes have been used to compile 'CpGislandEVO' (45), a specialized genome platform for the comparative evolutionary genomics of CGIs. Both databases may be useful for studies relating DNA methylation and the evolutionary rates of different genome elements (46).

CONCLUSIONS

NGSmethDB provides high-resolution epigenome-wide methylome maps for a collection of the best-assembled eukaryotic genomes. All methylome maps stored in the database were obtained under uniform conditions, i.e. using strictly the same bioinformatics protocol for all raw data sets including the same parameter settings and the same stringent quality controls. SNV variants, obtained jointly with methylation values, have also been provided as accompanying tracks, which may facilitate to analyze the relation between DNA methylation and sequence variation. To widen comparative studies, the NGSmethDB methylome maps are connected to a UCSC track hub, thus allowing the comparison to third-part phenotype or disease annotation tracks.

ACKNOWLEDGEMENT

Beta testing of the database by Cristina Gómez Martín and Ernesto Aparicio Puerta is acknowledged.

FUNDING

Spanish Government [BIO2008-01353 to J.L.O. and BIO2010-20219 to M.H.], and Basque country 'AE' grant (to G.B.). Funding for open access charge: Department of Genetics, University of Granada, Spain.

Conflict of interest statement. None declared.

REFERENCES

- Bird, A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.
- Bird, A.P. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
- Ziller, M.J., Gu, H., Muller, F., Donaghey, J., Tsai, L.T., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E. *et al.* (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature*, **500**, 477–481.
- Schubeler, D. (2012) Molecular biology. Epigenetic islands in a genetic ocean. *Science*, **338**, 756–757.
- Oliveira, A.M., Hemstedt, T.J. and Bading, H. (2012) Rescue of aging-associated decline in Dnmt3a2 expression restores cognitive abilities. *Nat. Neurosci.*, **15**, 1111–1113.
- Zovkic, I.B., Guzman-Karlsson, M.C. and Sweatt, J.D. (2013) Epigenetic regulation of memory formation and maintenance. *Learn Mem.*, **20**, 61–74.
- Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D. *et al.* (2013) Global epigenomic reconfiguration during mammalian brain development. *Science*, **341**, 1237905.
- Baylin, S.B., Esteller, M., Rountree, M.R., Bachman, K.E., Schubeler, K. and Herman, J.G. (2001) Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Hum. Mol. Genet.*, **10**, 687–692.
- De Smet, C., Lurquin, C., Lethe, B., Martelange, V. and Boon, T. (1999) DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter. *Mol. Cell. Biol.*, **19**, 7327–7335.
- Esteller, M., Corn, P.G., Baylin, S.B. and Herman, J.G. (2001) A gene hypermethylation profile of human cancer. *Cancer Res.*, **61**, 3225–3229.
- Issa, J.P. (2004) CpG island methylator phenotype in cancer. *Nat. Rev. Cancer*, **4**, 988–993.
- Riazalhosseini, Y. and Hoheisel, J.D. (2008) Do we use the appropriate controls for the identification of informative methylation markers for early cancer detection? *Genome Biol.*, **9**, 405.
- Krebs, A.R. and Schubeler, D. (2012) Tracking the evolution of cancer methylomes. *Nat. Genet.*, **44**, 1173–1174.
- Wasserkort, R., Kalmár, A., Valez, G., Spisak, S., Krispin, M., Toth, K., Tulassay, Z., Sledziewski, A.Z. and Molnar, B. (2013) Aberrant septin 9 DNA methylation in colorectal cancer is restricted to a single CpG island. *BMC Cancer*, **13**, 398.
- Beck, S. and Rakan, V.K. (2008) The methylome: approaches for global DNA methylation profiling. *Trends Genet.*, **24**, 231–237.
- Laird, P.W. (2010) Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
- 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.
- Laurent, L., Wong, E., Li, G., Huynh, T., Tsigros, 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.
- Bock, C., Kiskinis, E., Verstaep, G., Gu, H., Boulting, G., Smith, Z.D., Ziller, M., Croft, G.F., Amoroso, M.W., Oakley, D.H. *et al.* (2011) Reference maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell*, **144**, 439–452.
- Amoreira, C., Hindermann, W. and Grunau, C. (2003) An improved version of the DNA Methylation database (MethDB). *Nucleic Acids Res.*, **31**, 75–77.
- Grunau, C., Renault, E., Rosenthal, A. and Roizes, G. (2001) MethDB—a public database for DNA methylation data. *Nucleic Acids Res.*, **29**, 270–274.
- Negre, V. and Grunau, C. (2006) The MethDB DAS server: adding an epigenetic information layer to the human genome. *Epigenetics*, **1**, 101–105.
- Ongenaert, M., Van Neste, L., De Meyer, T., Menschaert, G., Bekaert, S. and Van Criekinge, W. (2008) PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res.*, **36**, D842–D846.
- Xin, Y., Chanrion, B., O'Donnell, A.H., Milekic, M., Costa, R., Ge, Y. and Haghghi, F.G. (2012) MethylomeDB: a database of DNA methylation profiles of the brain. *Nucleic Acids Res.*, **40**, D1245–D1249.
- Shi, J., Hu, J., Zhou, Q., Du, Y. and Jiang, C. (2013) PEPiD: a prostate epigenetic database in mammals. *PLoS One*, **8**, e64289.
- Lv, J., Liu, H., Su, J., Wu, X., Li, B., Xiao, X., Wang, F., Wu, Q. and Zhang, Y. (2012) DiseaseMeth: a human disease methylation database. *Nucleic Acids Res.*, **40**, D1030–D1035.
- He, X., Chang, S., Zhang, J., Zhao, Q., Xiang, H., Kusonmano, K., Yang, L., Sun, Z.S., Yang, H. and Wang, J. (2008) MethyCancer: the database of human DNA methylation and cancer. *Nucleic Acids Res.*, **36**, D836–D841.
- Gu, F., Doderer, M.S., Huang, Y.W., Roa, J.C., Goodfellow, P.J., Kizer, E.L., Huang, T.H. and Chen, Y. (2013) CMS: a web-based system for visualization and analysis of genome-wide methylation data of human cancers. *PLoS One*, **8**, e60980.
- Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
- Pruitt, K.D., Tatusova, T., Brown, G.R. and Maglott, D.R. (2012) NCBI reference sequences (RefSeq): current status, new features

- and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
31. Hackenberg, M.H., Barturen, G. and Oliver, J.L. (2012) In: Tatarinova, T. and Kerton, O. (eds), *DNA Methylation - From Genomics to Technology*. In-Tech, Rijeka, Croatia, p. 27.
 32. Barturen, G., Rueda, A., Oliver, J.L. and Hackenberg, M. (2013) MethylExtract: high-quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *Fl1000Research*, **2**, 217–232.
 33. 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.
 34. Hackenberg, M., Rodriguez-Ezpeleta, N. and Aransay, A.M. (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.*, **39**, W132–W138.
 35. Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J.M. and Aransay, A.M. (2009) miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.*, **37**, W68–W76.
 36. Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
 37. Tomso, D.J. and Bell, D.A. (2003) Sequence context at human single nucleotide polymorphisms: overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands. *J. Mol. Biol.*, **327**, 303–308.
 38. Weisenberger, D.J., Campan, M., Long, T.I., Kim, M., Woods, C., Fiala, E., Ehrlich, M. and Laird, P.W. (2005) Analysis of repetitive element DNA methylation by MethyLight. *Nucleic Acids Res.*, **33**, 6823–6836.
 39. Krueger, F. and Andrews, S.R. (2011) Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, **27**, 1571–1572.
 40. Skinner, M.E. and Holmes, I.H. (2010) Setting up the JBrowse genome browser. *Curr. Protoc. Bioinformatics*, **Chapter 9**, Unit 9.13.
 41. Skinner, M.E., Uzilov, A.V., Stein, L.D., Mungall, C.J. and Holmes, I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
 42. Hon, G.C., Hawkins, R.D., Caballero, O.L., Lo, C., Lister, R., Pelizzola, M., Valsesia, A., Ye, Z., Kuan, S., Edsall, L.E. et al. (2012) Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.*, **22**, 246–258.
 43. Heyn, H. and Esteller, M. (2012) DNA methylation profiling in the clinic: applications and challenges. *Nat. Rev. Genet.*, **13**, 679–692.
 44. Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M. and Schubeler, D. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.*, **39**, 457–466.
 45. Barturen, G., Geisen, S., Dios, F., Hamberg, E.J.M., Hackenberg, M. and Oliver, J.L. (2013) CpGislandEVO: a database and genome browser for comparative evolutionary genomics of CpG islands. *Biomed. Res. Int.*, **2013**, 1–6.
 46. Chuang, T.J., Chen, F.C. and Chen, Y.Z. (2012) Position-dependent correlations between DNA methylation and the evolutionary rates of mammalian coding exons. *Proc. Natl Acad. Sci. USA*, **109**, 15841–15846.

PREDICCIÓN COMPUTACIONAL DE ISLAS CpG

Las islas CpG (*CG/s*) son pequeñas regiones con una composición excepcional con respecto al resto del genoma y suelen encontrarse libres de metilación. Generalmente, las *CG/s* se han caracterizado de manera computacional mediante límites arbitrarios basados en sus características composicionales. Sin embargo, la aparición de nuevos algoritmos computacionales en los últimos años ha puesto en duda la definición clásica que atribuían los algoritmos convencionales a estas *CG/s*.

En este capítulo, se resumen los tipos de algoritmos existentes para la predicción de *CG/s* y se incluye la actualización y mejora del algoritmo *CpGcluster*. Además, se desarrolla una exhaustiva comparación de los métodos clásicos con *CpGcluster*, que pretende determinar el método más apropiado para definir y caracterizar estas regiones tan singulares del genoma.

5.1 IDENTIFICACIÓN DE CG/s BASADA EN LA SECUENCIA

Inicialmente, las CG/s se identificaron de manera experimental en ratón mediante la digestión del ADN con la enzima HpaII, enzima de restricción sensible a las metilcitosinas (Cooper, Taggart et al. 1983, Bird, Taggart et al. 1985). El análisis en ratón y humano de estos productos de digestión, determinó la existencia de unas 26,000 regiones no metiladas compuestas por agrupaciones de CpGs (Antequera and Bird 1993). A partir de estos resultados, las CG/s se caracterizaron como regiones del ADN con una longitud mínima de 200 pbs, un contenido de guaninas y citosinas de al menos el 50% y una proporción de CpGs [observados/esperados] ($CpG [o/e]$) mayor de 0.6 (Gardiner-Garden and Frommer 1987). La excepcional composición de las CG/s y la aparición del primer ensamblado del genoma humano (Lander, Linton et al. 2001, Venter, Adams et al. 2001), permitió predecir su localización en el genoma a partir de métodos *in-silico*. La escasa variación interindividual en humanos, entre un 0.1 y un 0.4% (Sachidanandam, Weissman et al. 2001, Jorde and Wooding 2004, Abecasis, Auton et al. 2012), asegura la obtención de una predicción muy fiable, a partir de la secuencia de referencia. Basándose en esta premisa, han ido apareciendo múltiples algoritmos para predecir la localización de las CG/s, que pueden agruparse en métodos basados en ventanas móviles, métodos de adición y métodos basados en la agrupación (*clustering*) de CpGs.

5.1.1 Métodos de ventana

Los métodos clásicos para detectar *CGIs* se basan en ventanas móviles (“*Sliding Windows Methods*”) de una longitud determinada, donde se analizan características composicionales como la proporción de guaninas y citosinas (*G+C*) y/o la proporción de *CpGs* observados/esperados (*CpG [o/e]*). Estos métodos identifican las *CGIs* en función de umbrales inferiores arbitrarios de estas características, considerando *CGIs* las regiones, que presentando una longitud mínima, cumplan los criterios composicionales establecidos. Los principales métodos de ventana son: *Takai & Jones* (Takai and Jones 2002) y *CpGProD* (Ponger and Mouchiroud 2002).

Estos métodos de ventana, a pesar de ser ampliamente utilizados, presentan varios inconvenientes importantes:

- Los cambios en los parámetros resultan en predicciones dispares.
- El gran número de parámetros configurables (5 en el caso de *Takai & Jones*) dificulta la realización de un estudio exhaustivo que determine los valores óptimos.
- La necesidad de adaptar estos parámetros arbitrarios en función de la especie en estudio, imposibilita la comparación evolutiva de las *CGIs*.
- Son métodos matemáticamente incompletos debido a la imposibilidad de detectar todas las secuencias que cumplan sus criterios composicionales (Hsieh, Chen et al. 2009).

5.1.2 Métodos de adición

Los métodos de adición (“*Running Sum Methods*”) surgen como una alternativa a los métodos de ventana descritos previamente. Estos métodos se basan en la localización de secuencias de ADN con una mayor frecuencia de dinucleótidos CpG que las regiones colindantes. Los más conocidos de este grupo son: *CpGreport* (Rice, Longden et al. 2000, Olson 2002) y el algoritmo de predicción utilizado por la UCSC (Kent, Sugnet et al. 2002, Meyer, Zweig et al. 2013).

Estos algoritmos suponen una mejora considerable con respecto a los métodos de ventana, tanto a nivel biológico como matemático:

- Las *CGIs* predichas empiezan y terminan por dinucleótidos CpG.
- *CpGreport* permite detectar *CGIs* por debajo de los umbrales clásicos de longitud, que originalmente se consideraban artefactos del método pero cuya función se ha demostrado recientemente (Wong, Wong et al. 2006).
- Identifican las *CGIs* en base a la composición del genoma, aunque de manera local, reduciendo el número de parámetros a tener en cuenta.

Sin embargo y a pesar de estas mejoras, estos métodos mantienen la necesidad de elegir parámetros arbitrarios para identificar las *CGIs*.

5.1.3 Métodos de agrupación de CpGs

Los métodos de agrupación o “*clustering*” son los algoritmos más avanzados en la detección de *CGIs*. Los principales algoritmos de agrupación son: *CpGcluster* (Hackenberg, Previti et al. 2006), basado en las distancias entre CpGs; *cgClusters* (Glass, Thompson et al. 2007), que identifica regiones con alta densidad de CpGs y el método *CGI_HMM* (Wu, Caffo et al. 2010), basado en modelos ocultos de Markov. El primero de estos métodos fue *CpGcluster*, que identifica las agrupaciones de CpGs (*CGIs*) basándose en las distancias entre estos dinucleótidos.

Los 3 métodos de agrupación son muy completos matemáticamente hablando, y basan sus predicciones en las características composicionales presentes en la secuencia a analizar, lo que permite comparar las anotaciones obtenidas entre diversas especies y/o ensamblados. Sin embargo, todos ellos presentan ciertos inconvenientes:

- Tanto *cgClusters*, como *CGI_HMM* requieren de un estudio previo para ajustar los parámetros necesarios, que pueden modificar drásticamente las anotaciones finales. A su vez, *CpGcluster* requiere verificar la bondad de la mediana como aproximación si se analizan otras especies.
- Además, *CGI_HMM* necesita que el genoma en estudio tenga anotaciones de elementos repetidos Alu, ya que descarta estos elementos al inicio del proceso.

- 5.2 Hackenberg, M, P Carpena, P Bernaola-Galvan, G Barturen, AM Alganza, JL Oliver. 2011. WordCluster: detecting clusters of DNA words and genomic elements. *Algorithms Mol Biol* 6:2.
-

Dirección de acceso PubMed:

<http://www.ncbi.nlm.nih.gov/pubmed/21261981>

Dirección de publicación:

<http://www.almob.org/content/6/1/2>

Interfaz web:

<http://bioinfo2.ugr.es/wordCluster/wordCluster.php>

Breve descripción:

En esta Tesis se ha mejorado *CpGcluster* y se ha ampliado el algoritmo para cualquier *k-mero* (“palabra” de ADN) o elemento que pueda representarse en el genoma mediante coordenadas cromosómicas. Debido a la generalización del método para cualquier “palabra”, la actualización se ha llamado *WordCluster* (Hackenberg, Carpena et al. 2011), y las mejoras implementadas han sido:

- El algoritmo calcula la intersección entre las distribuciones teórica y experimental de distancias, con lo que ya no es necesario utilizar la mediana como aproximación. Esto permite usar el algoritmo para cualquier especie sin necesidad de comprobar previamente que la aproximación es válida.
- El cálculo de la intersección puede realizarse para cromosomas individuales o para todo el genoma.

- El cálculo de distancias se realiza estrictamente dentro de los *contigs*, es decir no se permite la existencia de Ns entre dos palabras consecutivas.
- Para facilitar su uso, el algoritmo se ha implementado junto con una interfaz web.
- Los resultados incluyen un resumen con la estadística básica para cada cromosoma del genoma y para las agrupaciones que solapan con diferentes regiones génicas.
- Se incluye también un análisis funcional de enriquecimiento en términos GO (*Gene Ontology*) para los genes con los que solapan las agrupaciones detectadas.

SOFTWARE ARTICLE

Open Access

WordCluster: detecting clusters of DNA words and genomic elements

Michael Hackenberg^{1*}, Pedro Carpena^{2,3}, Pedro Bernaola-Galván², Guillermo Barturen¹, Ángel M Alganza¹, José L Oliver^{1*}

Abstract

Background: Many k -mers (or DNA words) and genomic elements are known to be spatially clustered in the genome. Well established examples are the genes, TFBSs, CpG dinucleotides, microRNA genes and ultra-conserved non-coding regions. Currently, no algorithm exists to find these clusters in a statistically comprehensible way. The detection of clustering often relies on densities and sliding-window approaches or arbitrarily chosen distance thresholds.

Results: We introduce here an algorithm to detect clusters of DNA words (k -mers), or any other genomic element, based on the distance between consecutive copies and an assigned statistical significance. We implemented the method into a web server connected to a MySQL backend, which also determines the co-localization with gene annotations. We demonstrate the usefulness of this approach by detecting the clusters of CAG/CTG (cytosine contexts that can be methylated in undifferentiated cells), showing that the degree of methylation vary drastically between inside and outside of the clusters. As another example, we used *WordCluster* to search for statistically significant clusters of olfactory receptor (OR) genes in the human genome.

Conclusions: *WordCluster* seems to predict biological meaningful clusters of DNA words (k -mers) and genomic entities. The implementation of the method into a web server is available at <http://bioinfo2.ugr.es/wordCluster/wordCluster.php> including additional features like the detection of co-localization with gene regions or the annotation enrichment tool for functional analysis of overlapped genes.

Background

Genome entities as diverse as genes [1], CpG dinucleotides [2], transcription factor binding sites (TFBSs) [3] or ultra-conserved non-coding regions [4] usually form clusters along the chromosome sequence. Such spatial clustering often translates into genome structures with a clear functional and/or evolutionary meaning: gene clusters encoding the same or similar products and originated through gene duplication events, CpG islands, cis-regulatory modules, etc. Thus, the spatial clustering of functional genome elements (in general, words or k -mers) would somewhat remember the situation in

literary texts, where keywords show a strong clustering, whereas common words are randomly distributed [5].

Despite its potential importance, no algorithm exists to detect the clustering of DNA words in a rigorous way. Most current methods are based on densities and sliding-window approaches or arbitrary distances. For example, the Galaxy work suite ([6], <http://main.g2.bx.psu.edu/>) implements an algorithm which lets the user decide to fix the maximum distance between two entities and the minimum number of entities in the cluster. Recently, we developed an algorithm to detect clusters of CpG dinucleotides in DNA sequences based on the distance between neighboring CpGs, then assigning a statistical significance [7]. Now, we generalize the method to any k -mer or any arbitrary combination of them, as well as to any other genome entity defined by its chromosome coordinates.

* Correspondence: mlhack@gmail.com; oliver@ugr.es

¹Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva s/n, 18071-Granada & Lab. de Bioinformática, Centro de Investigación Biomédica, PTS, Avda. del Conocimiento s/n, 18100-Granada, Spain

Full list of author information is available at the end of the article

Implementation

The *WordCluster* algorithm allows the detection of clusters for DNA words (k -mers) and genomic elements (genes, transposons, SINES, TFBSs, etc.). The algorithm is based on the distances between the entities and an assigned p -value.

The algorithm

The algorithm is basically the same for k -mers and genomic elements except for the detection of the coordinates and the way the success probabilities are calculated. Briefly the algorithm performs the following steps:

1. Detection of all k -mer copies in the chromosomes, storing its coordinates (this step is unique to the detection of k -mer clusters as the genomic elements already come defined by its coordinates). The copies are detected in a non-overlapping way, i.e. once a copy is found the search is resumed at the end of the word, thus preventing the detection of overlapping copies.
2. Calculation of the distances between consecutive copies. The distance is defined as: "start coordinate of the downstream copy" minus "end coordinate of the upstream copy". This implies that the minimum distance is 1 when the two entities are located directly next to each other.
3. Detection of the clusters, defined as those chromosomal regions where all distances are equal or below a given maximum distance. A cluster is defined by its start and end coordinates and the number of k -mers or genomic elements it contain.
4. Calculation of the statistical significance for each cluster by means of the negative binomial distribution. A p -value threshold is then used to filter out those clusters which are not statistically significant.

A main difference to the originally described algorithm is the way N -runs in the DNA sequence (ambiguous sequence sites occupied by any nucleotide) are treated. While the original *CpGcluster* method allows up to 10 N s between two consecutive CpGs, *WordCluster* detects the DNA words and the distances strictly within the contigs, i.e. not a single N is allowed to lie between two copies.

Statistical significance

From now on, we will have to use the word k -mer in different contexts. Therefore, to avoid confusion we define as "target k -mer(s)" the k -mer(s) which are being analysed, i.e. those for which the clusters are going to be detected. On the contrary, "no-target k -mer(s)" are all the remaining k -mer(s). We use k -mer in a generic way, referring to all DNA words of length k .

The statistical significance is calculated as the cumulative density function of the negative binomial distribution:

$$P_{N,p}(n_f) = \binom{n_f + (n-1) - 1}{(n-1) - 1} \cdot p^{n-1} \cdot (1-p)^{n_f}$$

being n the number of target k -mers within the cluster, n_f the number of "failures", i.e. the number of no-target k -mers. For example, if we are detecting clusters of AGCT, all k -mers other than AGCT would be considered as failures. Finally, p is the success probability, i.e. the probability to find a target k -mer or genomic element within the DNA sequence. Note that in the above equation we use $(n-1)$ instead of n , as the first appearance of a target k -mer within the cluster is trivial (i.e. all the clusters start with a target k -mer). While the negative binomial distribution can be defined in the same way for k -mers and genomic elements, differences exist in the way the number of "failures" and the success probability are calculated.

For k -mers, the number of failures n_f is simply given by

$$n_f = L_c - n \cdot k$$

being L_c the length of the cluster, k the length of the target k -mer and n the number of non-overlapping target k -mers in the cluster. The number of failures is the number of no-target k -mers within the cluster. For example, given the target k -mer ATGC, the cluster ATGCATGC would give $n_f = 0$ while ATGCAATGC would give $n_f = 1$. Each k -mer can overlap with itself and other k -mers, but here we consider just non-overlapping occurrences. In such a case, the probabilities for k -mers are given by the following equation

$$p = \frac{N}{(L_s - k + 1) - N \cdot (k - 1)}$$

being N the number of non-overlapping occurrences of the target k -mers in the sequence, k the length of the k -mer and L_s the sequence length. The formula is simply the number of target k -mers in the sequence divided by the total number of k -mers in the sequence. As we do not consider overlapping instances, $N \cdot (k-1)$ was subtracted from the total number of k -mers $(L_s - k + 1)$, as those sequence positions are not considered, in order to take this effect into account.

For genomic elements, it is less clear how to define the number of failures. For example, one has a cluster with 5 elements which have mean length of 300 bp and 250 bp of distance on average between each other. The question is how many "no-elements" contain this

cluster, i.e. how many failures. We define the number of failures as

$$n_f = \text{ceiling}\left(\frac{L_{no}}{L_{mean}}\right)$$

being L_{no} the number of bases in the cluster not belonging to the genomic element and L_{mean} the mean length of the genomic element. Thus, this number is an approximation to the number of “no-elements” within the cluster. Finally, the success probability is then given as

$$p = \frac{N \cdot L_{mean}}{L_S}$$

being L_S the length of the sequence, L_{mean} the mean length of the genomic elements and N the number of genomic elements.

Distance models

The maximum distance is the main parameter of the algorithm determining the copies belonging to each cluster. We have shown previously [7] that, for most human chromosomes, the median of the observed distance distribution of CpGs lies near the intersection between the observed and the expected distance distribution. The intersection can be interpreted as the point separating the intra-cluster from the inter-cluster distances. In this new tool, we added two more distance models based on the direct detection of the mentioned intersection (one genome wide and the other for each chromosome separately). In this way, *WordCluster* implements a total of 4 different distance models:

1. Percentile distance: The distance corresponding to a given percentile of the observed distance distribution is calculated and used as the maximum distance threshold.
2. Chromosomal intersection: The distance corresponding to the intersection between the observed and the expected distributions is used as the maximum distance (see Figure 1).
3. Genome intersection: The distance distributions for all chromosomes are merged, then calculating the distance corresponding to the “genome intersection point”. If this distance model is chosen, the success probabilities (i.e. the probability to find the target k -mers in the chromosome) are not calculated for each chromosome separately (like in the two models above), but a genome wide success probability (probability to find the target k -mers) is calculated.
4. Fixed distance: the user can set the distance threshold.

Webserver

We implemented the described algorithm into a web server. The tool uses PHP for the interaction with the user, to access the core program (written in Java) and the MySQL database. Two types of input data can be supplied: 1) a group of k -mers and a genomic sequence to be scanned by the program (the user can upload his own sequence or choose one of the 24 genome assemblies stored in our database - see below); and 2) a file in BED format [8,9] with the coordinates of the genomic elements whose clustering properties should be analyzed. No mandatory input parameters exist, but the user can select between different distance models (the default is the chromosome intersection) and set the cut-off for the statistical significance (the default here is $p\text{-value} \leq 1E-5$).

The output generated by the web server depends on whether the user chooses a genome assembly from our database or supplies an anonymous sequence. The minimum output consists of the basic statistics of the clusters (base composition, entity composition and statistical significance) and the statistics by chromosome. Furthermore, for all species in the database, the co-localization of detected clusters with different gene regions (promoters, introns, etc.) is reported.

Finally, for some species (human, mouse, rat, cow, *C. elegans*, zebrafish and chicken) an enrichment/depletion analysis for the genes overlapped by the clusters is carried out using the Gene Ontology [10] and the Annotation-Modules database [11,12].

Database

Currently, the genomes of 24 genome assemblies are stored into our database. The following sequences were downloaded from the UCSC genome browser or the corresponding project homepages (plant genomes): Human (hg18, hg19), Mouse (mm8, mm9), Rat (rn4), Fruit fly (dm3), *Anopheles gambiae* (anogam1), Honey bee (apimel2), Cow (bosTau4), Dog (canFam2), *C. briggsae* (cb3), *C. elegans* (ce6), Sea squirt (ci2), Zebrafish (danrer5), Chicken (galgal3), Stickleback (gasacu1), Medaka (orylat2), Chimp (pantro2), *Rhesus macaque* (rhmac2), *S. cerevisiae* (saccer1), *Tetraodon* (tetnig1), *Arabidopsis thaliana* (tair8, tair9), and *Zea mays* (zm1). To determine the co-localization with genes, we used RefSeq genes whenever they were available [13], Ensembl genes otherwise [14].

Results and Discussion

To demonstrate the ability of our algorithm in finding biologically significant and relevant clusters in the genome, at the same time illustrating the different distance models, we carried out three analysis: 1) detection of

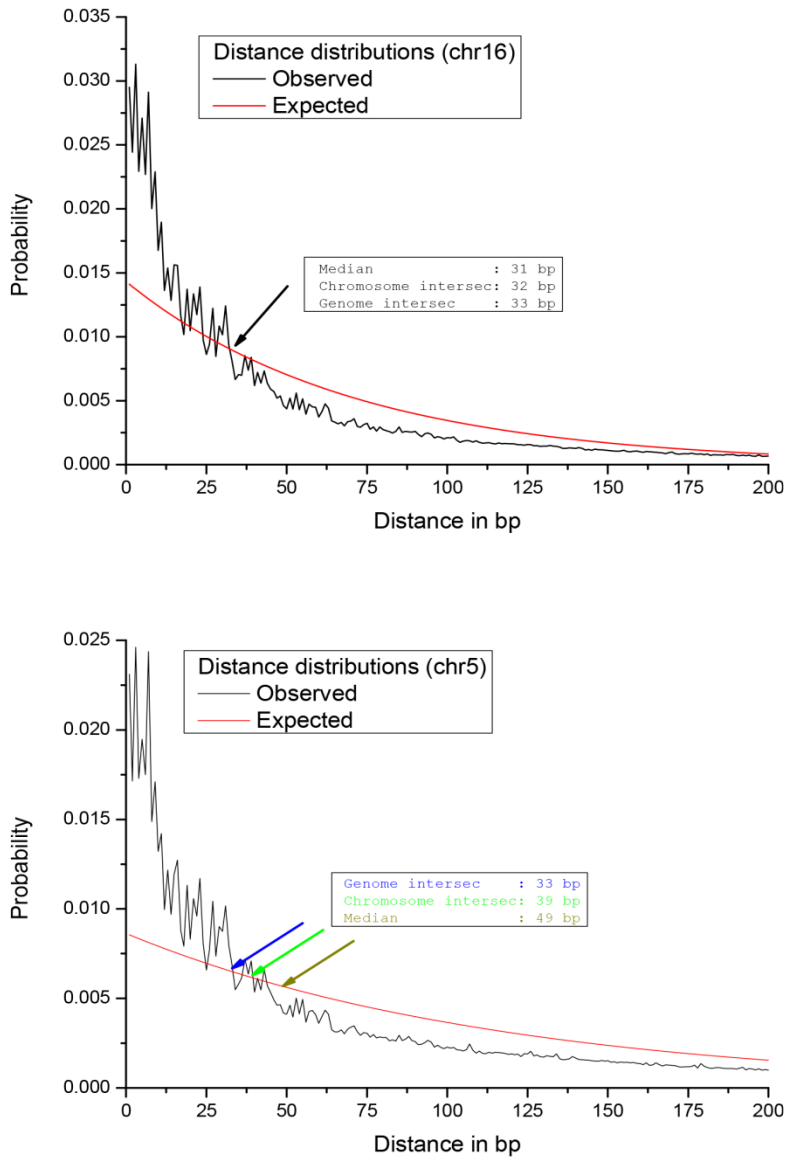


Figure 1 Distance distributions. Expected and observed distance distributions for human chromosomes 16 (above) and 5 (below). It can be seen that for chr16 the median, the chromosome intersection and the genome intersection are very close (within 1 bp), while for chromosome 5 notable differences exist (from 33 bp to 49 bp).

clusters of CpGs (CpG islands) using different distance models, 2) detection of clusters of the word CWG (where W = A, T) and 3) detection of clusters of olfactory receptor genes in the human chromosome 11.

Detection of CpG islands with different distance models

We choose this example as the detection of CpG islands was the reason to develop the algorithm from which *WordCluster* [7] was derived. In the original *CpGcluster* algorithm, we used the percentile of the observed distance distribution as distance model (apart from the fixed distance), suggesting the median as the default parameter. We did this since we observed that the intersection between the observed and expected distance distributions is often very close to the median of the observed distance distribution (see Figure 1). This intersection can be interpreted in the following way. When the observed curve lies above the expected, theoretical curve, it means that more CpGs exist at this distance than expected by chance. We can observe in Figure 1 that this is generally the case for short distances, thus indicating the clustering (overrepresentation of short distances) of CpG dinucleotides. The intersection defines the “reversal point”, i.e. at larger distances than this point, the CpG dinucleotides are not clustered any more. Therefore, it might be that the strict use of the intersection defines better clusters than the use of the median, which is a mere approximation to the intersection point. Furthermore, we observed that for some chromosomes the intersection and the median differ slightly. To clarify the impact of this change in the maximum distance, we predict CpG islands by means of the median (cpg50), the chromosome intersection (cpgISc) and the genome intersection (cpgISg), then assessing the prediction quality by some of the criteria previously described [7,15]. Table 1 shows that the mean length of both intersection models are clearly below the mean length of the original cpg50 islands. This can be explained as the intersection models produce on average shorter distance thresholds, which leads to fragmentation, shortening and disappearance of some cpg50 CpG islands. Consequently, the chromosome intersection

model (cpgISc) predicts fewer islands than the original cpg50 algorithm (3979). Nevertheless, the genome intersection (cpgISg) yields more predictions compared to cpg50 (5535). The latter observation can be explained as the predictions are done with a single, genome wide probability. The *p-value* assigned to each cluster depends on the success probability, and in G+C rich chromosomes the genome wide probability is much lower than the chromosome probability. This leads to smaller *p-values* in G+C rich chromosomes, so that more islands can pass the *p-value* threshold. For example, cpg50 predicts 2434 islands in chromosome 22 while cpgISg predicts 5197. Of course, in AT-rich chromosomes this effect is reverted but less pronounced (the difference between genome wide and chromosome probabilities are smaller in AT-rich compared to GC-rich chromosomes), and therefore a higher total number of islands are predicted.

Next, we analyzed the predictions under functional aspects. Table 2 shows the overlap of the predictions with RefSeq genes [13], Alu elements and phylogenetically conserved PhastCons elements [16]. The cpgISg predictions show the highest overlap with the promoter region (R13), and conserved PhastCons elements, simultaneously showing the lowest overlap with spurious Alu elements. This might indicate that cpgISg predictions are slightly better than the other two, the original cpg50 and cpgISc. However, 1) the differences seem to be rather small and 2) a more detailed analysis would be needed to resolve this question.

Independently of this open question, we can summarize: 1) the chromosome intersection seems to be a good replacement for the median and furthermore removes one input parameter from the method, as the intersection is a fixed statistical property of the chromosome; 2) the genome intersection may be used when the expected clusters are known to be not dependent on the chromosome. The CpG islands are probably not dependent on the chromosome, as the biological mechanisms forming and maintaining them are probably the same for all chromosomes. This may suggest the use of the genome intersection, which is confirmed by producing slightly better results than the other two tested distance models.

Table 1 *WordCluster* predictions of CpG clusters*

| Method | # | Length ± SD | GC ± SD | OE ± SD |
|--------|--------|---------------|------------|---------------|
| cpg50 | 198703 | 273.2 ± 246.4 | 63.8 ± 7.5 | 0.855 ± 0.265 |
| cpgISc | 194725 | 218.7 ± 200.1 | 65.6 ± 7.7 | 0.916 ± 0.273 |
| cpgISg | 204238 | 202.6 ± 183.8 | 66.3 ± 7.5 | 0.930 ± 0.274 |

*Basic statistic of CpG island predictions using three different distance models: cpgISg (genome intersection), cpg50 (Median) and cpgISc (chromosome intersection). The number of predicted islands, the length, the G+C content and the observed to expected ratios are shown. Note that the original cpg50 algorithm predicts 198702 islands, i.e. one less than *WordCluster* with the median model. This is due to the changes introduced regarding the N-runs (see main text).

Detection of CWG clusters

Besides the conventional CpG context, the CWG context has recently been shown to be a potential target for methylation [17]. *WordCluster* detects 84996 CAG/CTG clusters in the human genome (NCBI 36, hg18) significant at the 1E-5 level using the chromosome intersection (Table 1). We found a high number of statistically significant CWG clusters scattered along all human chromosomes, many of which are overlapping gene regions (Table 3). To check if the detected clusters

Table 2 Biological meaning of WordCluster predictions*

| Method | #islands | #TSS overlap | #R13 overlap | #Alu overlap | #PhastCons overlap |
|--------|----------|--------------|---------------|---------------|--------------------|
| cpg50 | 198703 | 12432 (6.3%) | 30660 (15.4%) | 80323 (40.4%) | 48787 (24.6%) |
| cpg1Sc | 194724 | 11926 (6.1%) | 34567 (17.8%) | 70144 (36.0%) | 48930 (25.1%) |
| cpg1Sg | 204238 | 12156 (6.0%) | 37616 (18.4%) | 70456 (34.5%) | 52335 (25.6%) |

*Comparison of three *WordCluster* predictions of CG clusters (CpG islands) using three different distance models: cpg1Sg (genome intersection), cpg50 (median) and cpg1Sc (chromosome intersection). The overlap with two gene regions (TSS and R13), Alu elements and phylogenetically conserved PhastCons elements have been measured and both absolute numbers and percentages are given.

Table 3 Clusters of CWG trinucleotides*

| N | 84996 |
|--|----------|
| Genome coverage (bp) | 15700789 |
| Average length (bp) | 184.7 |
| No. of clusters co-locating with gene regions: | |
| TSS | 272 |
| TSS ± 100 bp | 686 |
| 5'UTR | 4712 |
| Introns | 29326 |
| Exons | 1852 |
| 3'UTR | 1658 |

*Statistically significant clusters of CAG and CTG trinucleotides detected by *WordCluster* in the human genome (hg18). We used the "genome intersection" distance model and a *p*-value threshold of 1E-05.

Table 4 Clusters of OR genes in human chromosome 11*

| Cluster | chromStart | chromEnd | length | count | <i>p</i> -value |
|---------|------------|-----------|--------|-------|-----------------|
| 1 | 4345160 | 5178488 | 833329 | 53 | 1.60E-49 |
| 2 | 5269273 | 5559687 | 290415 | 21 | 6.80E-21 |
| 3 | 5697096 | 6177989 | 480894 | 28 | 2.70E-25 |
| 4 | 48194938 | 48344593 | 149656 | 9 | 2.50E-08 |
| 5 | 48398372 | 48505102 | 106731 | 9 | 1.70E-09 |
| 6 | 49876392 | 49960613 | 84222 | 7 | 2.60E-07 |
| 7 | 51250039 | 51384376 | 134338 | 11 | 1.60E-11 |
| 8 | 54842612 | 55380573 | 537962 | 32 | 4.00E-29 |
| 9 | 55427396 | 56344568 | 917173 | 66 | 6.30E-65 |
| 10 | 56495101 | 56580184 | 85084 | 7 | 2.90E-07 |
| 11 | 57555001 | 57964200 | 409200 | 22 | 3.20E-19 |
| 12 | 58833691 | 59056759 | 223069 | 12 | 1.30E-10 |
| 13 | 123181329 | 123481891 | 300563 | 16 | 5.40E-14 |

*Chromosome coordinates, length, number of OR genes and *p*-values for all statistically significant OR gene clusters in chromosome 11.

might be biologically meaningful, we compared the percentage of methylated words (CAG and CTG) inside and outside of the clusters. We observed that 26.7% of all CAG/CTG trinucleotides are methylated inside the clusters while 45.3% of them are methylated when located outside a cluster. It seems therefore, as occurs in CpG islands, that CAG/CTG clusters remain unmethylated with a much higher probability than the bulk DNA.

Detection of olfactory gene clusters

As a third example, we used *WordCluster* to search for significant clusters of olfactory receptor (OR) genes, the largest multigene family in multicellular organisms whose members are known to be clustered within vertebrate genomes [18,19]. Table 4 shows the basic statistics for the 13 clusters of OR genes detected by our algorithm in human chromosome 11. Figure 2 shows a comparative analysis of the clusters predicted by *WordCluster* to the clusters currently annotated in the CLIC/HORDE database [19] in a selected region of chromosome 11. Our algorithm predicts a higher number of clusters, being all of them statistically significant.

Conclusions

WordCluster generalizes the previous *CpGcluster* algorithm [7] to any word or genomic element in the genome, at the same time associating a statistical significance to the clusters found. It outperforms current methods relying on densities and sliding-window approaches or arbitrarily chosen distance thresholds. The implementation as a web server connected to a MySQL backend allows for co-localization studies with

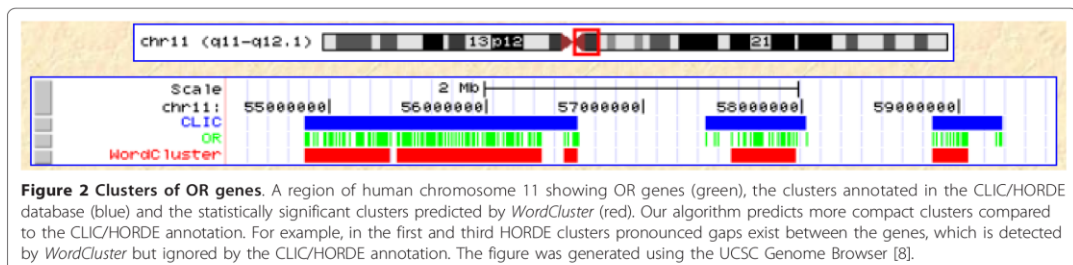


Figure 2 Clusters of OR genes. A region of human chromosome 11 showing OR genes (green), the clusters annotated in the CLIC/HORDE database (blue) and the statistically significant clusters predicted by *WordCluster* (red). Our algorithm predicts more compact clusters compared to the CLIC/HORDE annotation. For example, in the first and third HORDE clusters pronounced gaps exist between the genes, which is detected by *WordCluster* but ignored by the CLIC/HORDE annotation. The figure was generated using the UCSC Genome Browser [8].

different gene regions, as well as for genome wide enrichment/depletion analysis of functional terms (GO).

Availability and requirements

The *WordCluster* webserver (<http://bioinfo2.ugr.es/wordCluster/wordCluster.php>) is freely available. No registering is needed but every access is logged. For large jobs, a long-life web link to the results is provided.

List of abbreviations used

k-mer: DNA word (oligonucleotide) with length *k*; SINES: Short interspersed nuclear elements; TSS: Transcription Start Site; TFBS: Transcription Factor Binding Site; R13: promoter region [TSS-1500 bp; TSS+500 bp].

Acknowledgements

The Spanish Government grants BIO2008-01353 to JLO, mobility PR2009-0285 to PC, Spanish Junta de Andalucía grants P07-FQM3163 to PC and P06-FQM1858 to PB are acknowledged. The Spanish 'Juan de la Cierva' grant to MH and Basque Country 'Programa de formación de investigadores del Departamento de Educación, Universidades e Investigación' grant to GB are also acknowledged.

Author details

¹Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva s/n, 18071-Granada & Lab. de Bioinformática, Centro de Investigación Biomédica, PTS, Avda. del Conocimiento s/n, 18100-Granada, Spain. ²Dpto. de Física Aplicada II, E.T.S.I. de Telecomunicación, Universidad de Málaga 29071-Málaga, Spain. ³Division of Sleep Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.

Authors' contributions

MH developed and implemented the algorithm and wrote the manuscript (with JLO), PC and PB carried out the theoretical analysis of word clustering and help with the interpretation of statistical results, GB and AMA retrieve and organize the genome and methylation databases, and JLO developed the algorithm and wrote the manuscript (with MH). All the authors critically read and approved the final version.

Competing interests

None declared

Received: 30 August 2010 Accepted: 24 January 2011

Published: 24 January 2011

References

1. Durand D, Sankoff D: Tests for gene clustering. *J Comput Biol* 2003, **10**:453-482.
2. Gardiner-Garden M, Frommer M: CpG islands in vertebrate genomes. *Journal of molecular biology* 1987, **196**:261-282.
3. Makeev VJ, Lifanov AP, Nazina AG, Papatsenko DA: Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic acids research* 2003, **31**:6016-6026.
4. Sandelin A, Bailey P, Bruce S, Engstrom PG, Klos JM, Wasserman WW, Ericson J, Lenhard B: Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* 2004, **5**:99.
5. Carpena P, Bernaola-Galván P, Hackenberg M, Coronado AV, Oliver JL: Level statistics of words: finding keywords in literary texts and DNA. *Phys Rev E* 2008, **79**:035102-035104.
6. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, et al: Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 2005, **15**:1451-1455.
7. Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Aroza J, Oliver JL: CpGcluster: A distance-based algorithm for CpG-island detection. *BMC Bioinformatics* 2006, **7**:446.

8. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, et al: The UCSC Genome Browser Database: 2008 update. *Nucleic acids research* 2008, **36**:D773-779.
9. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**:841-842.
10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 2000, **25**:25-29.
11. Hackenberg M, Matthiesen R: Annotation-Modules: a tool for finding significant combinations of multisource annotations for gene lists. *Bioinformatics (Oxford, England)* 2008, **24**:1386-1393.
12. Hackenberg M, Matthiesen R: Algorithms and methods for correlating experimental results with annotation databases. *Methods in molecular biology (Clifton, NJ)* 2009, **593**:315-340.
13. Pruitt KD, Tatusova T, Maglott DR: NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* 2007, **35**:D61-65.
14. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, et al: Ensembl 2009. *Nucleic acids research* 2009, **37**: D690-697.
15. Hackenberg M, Barturen G, Carpena P, Luque-Escamilla PL, Previti C, Oliver JL: Prediction of CpG-island function: CpG clustering vs. sliding-window methods. *BMC Genomics* 2010, **11**:327.
16. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005, **15**:1034-1050.
17. Lister R, Pelizzola M, Downen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al: Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009, **462**:315-322.
18. Aloni R, Olender T, Lancet D: Ancient genomic architecture for mammalian olfactory receptor clusters. *Genome biology* 2006, **7**:R88.
19. The HORDE Project [<http://genome.weizmann.ac.il/horde/>]. [<http://bioportal.weizmann.ac.il/HORDE>].

doi:10.1186/1748-7188-6-2

Cite this article as: Hackenberg *et al.*: *WordCluster*: detecting clusters of DNA words and genomic elements. *Algorithms for Molecular Biology* 2011 **6**:2.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



- 5.3 Hackenberg, M, G Barturen, P Carpena, PL Luque-Escamilla, C Previti, JL Oliver. 2010. Prediction of CpG-island function: CpG clustering vs. sliding-window methods. BMC Genomics 11:327.
-

Dirección de acceso PubMed:

<http://www.ncbi.nlm.nih.gov/pubmed/20500903>

Dirección de publicación:

<http://www.biomedcentral.com/1471-2164/11/327>

Breve descripción:

Debido al debate generado en torno a los algoritmos más apropiados para definir las *CG/s*, en esta Tesis Doctoral se ha realizado un estudio comparado de sus principales características:

- Se han comparado las características composicionales de las predicciones de *CG/s* entre los métodos clásicos de ventana y *CpGcluster*.
- Se ha analizado la asociación de las diferentes predicciones con ciertos elementos genómicos que puedan sugerir su función, como por ejemplo: promotores, sitios alternativos de la transcripción, sitios de unión del complejo Polycomb 2 o isletas CpG verificadas experimentalmente.
- Además, se han utilizado datos de metilación de citosinas individuales para determinar la capacidad de las diferentes predicciones para definir dominios de metilación homogéneos.

Prediction of CpG-island function: CpG clustering vs. sliding-window methods

Michael Hackenberg*^{1,2}, Guillermo Barturen^{1,2}, Pedro Carpena^{3,4}, Pedro L Luque-Escamilla⁵, Christopher Previti⁶ and José L Oliver*^{1,2}

Abstract

Background: Unmethylated stretches of CpG dinucleotides (CpG islands) are an outstanding property of mammal genomes. Conventionally, these regions are detected by sliding window approaches using %G + C, CpG observed/expected ratio and length thresholds as main parameters. Recently, clustering methods directly detect clusters of CpG dinucleotides as a statistical property of the genome sequence.

Results: We compare sliding-window to clustering (i.e. *CpGcluster*) predictions by applying new ways to detect putative functionality of CpG islands. Analyzing the co-localization with several genomic regions as a function of window size vs. statistical significance (*p-value*), *CpGcluster* shows a higher overlap with promoter regions and highly conserved elements, at the same time showing less overlap with *Alu* retrotransposons. The major difference in the prediction was found for short islands (CpG islets), often exclusively predicted by *CpGcluster*. Many of these islets seem to be functional, as they are unmethylated, highly conserved and/or located within the promoter region. Finally, we show that window-based islands can spuriously overlap several, differentially regulated promoters as well as different methylation domains, which might indicate a wrong merge of several CpG islands into a single, very long island. The shorter *CpGcluster* islands seem to be much more specific when concerning the overlap with alternative transcription start sites or the detection of homogenous methylation domains.

Conclusions: The main difference between sliding-window approaches and clustering methods is the length of the predicted islands. Short islands, often differentially methylated, are almost exclusively predicted by *CpGcluster*. This suggests that *CpGcluster* may be the algorithm of choice to explore the function of these short, but putatively functional CpG islands.

Background

The methylation of CpG dinucleotides is an important epigenetic modification of DNA, required in mammals for embryonic development, genomic imprinting and X-chromosome inactivation [1-3]. Around 80% of all CpG dinucleotides are methylated in mammal genomes. The exceptions are short stretches of CpG dinucleotides (CpG islands or CGIs), which are predominantly hypomethylated in healthy tissues [4,5]. CGIs are thought to be predominantly located in the promoter region of genes; around 70% of all genes have a CGI overlapping its pro-

motor region. Moreover, virtually all housekeeping genes are associated to CGIs, while only half of the tissue specific genes show such association [6]. Given its location in the promoters, CGIs may play important roles in the regulation of gene expression. An example is the aberrant methylation of CGIs observed in many cancer types [7-11]. Moreover, evidence exist that the differential or tissue specific methylation of CpG islands may be involved in the regulation of tissue specific genes [12].

Accurate prediction tools are therefore needed and a considerable effort has been carried out over the last decade to detect CGIs in mammal genomes. Many different algorithms have been proposed, most of them based on the criteria of Gardiner-Frommer [1]. These authors proposed in 1987 thresholds for the detection of CGIs: GC-content (50%), CpG observed/expected (O/E) ratio (0.6) and length (200 bp). Many of the published methods

* Correspondence: mlhack@gmail.com, oliver@ugr.es¹

Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva s/n, 18071, Granada, Spain

¹ Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva s/n, 18071, Granada, Spain

Full list of author information is available at the end of the article

simply readjust these thresholds. However, it has been shown that filtering criteria-based definitions of CpG islands are mathematically incomplete and non-operational, as the sliding window methods frequently fail to identify a large percentage of subsequences that meet the filtering criteria [13].

Recently, methods based on the clustering of CpGs along the genome sequence detect CGIs as a statistical property, thereby not relying on thresholds of GC-content, O/E ratio and length. The first algorithm published in this category was the *CpGcluster* method [14], which detects the CGIs by means of the distances between CpGs, then assigning a statistical significance to each cluster of CpG dinucleotides. Subsequently, *CpGcluster* was followed by other methods detecting CGIs by means of the CpG densities [15-18]. In the same way, many other features could also contribute to determine the boundary of individual CpG-islands, such as transcription factors and nucleosome location. The nucleosome code could be an important ingredient of future CGI models, although sequence features will probably remain as the principal component (see, for example, [19]). Epigenetic information may be also of help in detecting CGIs by making use of contextual information [20].

Given the conceptual differences between sliding window algorithms (SWA) using a high parameter space and those detecting CGIs as a statistical property of the CpG clustering in DNA sequences, disagreement exist on the way CGIs should be predicted. Recently, a comparison between islands detected by the window-based Takai-Jones (TJ) program [21] and those detected by *CpGcluster* was published [22]. The comparison evaluated mainly the co-localization of CGIs and known promoters and concludes an overall advantage for the TJ approach over *CpGcluster*.

We present here new ways to detect putative function of CGIs, emphasizing the basic difference between *CpGcluster* and SWA predictions: the statistical significance introduced by *CpGcluster* instead of the conventional length threshold. We show that the statistical significance assigned to each *CpGcluster* island is a key criterion to control the overlap with promoter regions, evolutionarily conserved elements and spurious *Alu* elements. Finally, we show that many short (<200 bp) islands (CpG islets) may be also functional, given its overlap with either promoter or evolutionary conserved regions and the absence of methylation in at least one tissue. As many of these islets are exclusively predicted by *CpGcluster*, this may be the algorithm of choice for experimental essays aimed to verify the function of these short islands.

Results and Discussion

The way sliding-window approaches and *CpGcluster* detect CGIs are conceptually different. While SWA

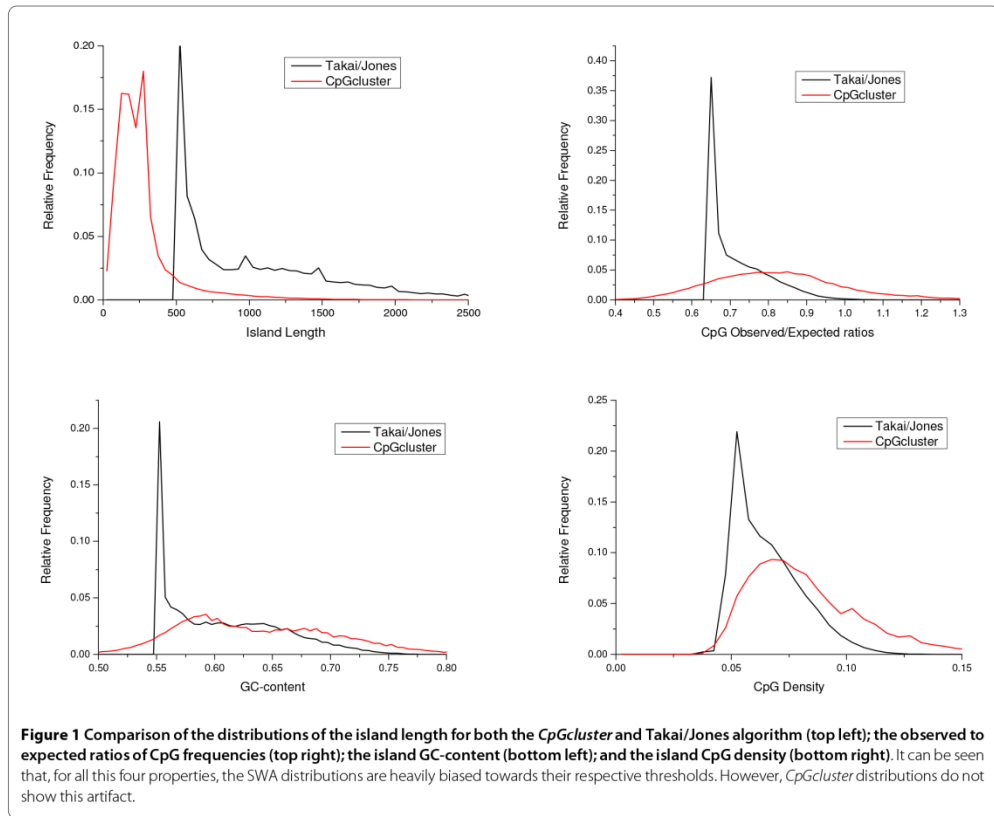
detect regions above the thresholds of G + C, O/E, min CpG and length, *CpGcluster* predicts statistically significant clusters of CpGs as CGIs. As a first consequence, the statistical properties of the predicted islands are different as well (Figure 1); e.g. in SWA approaches the distributions of important CGI properties like %G + C and O/E ratio are heavily biased towards the user thresholds.

Therefore, the first part of this work is basically aimed to clarify: 1) the differences between the length threshold used by SWA and the statistical significance used by *CpGcluster*; and 2) the consequences that the differences in the number of predicted islands and the mean length might have on the prediction quality.

Prediction quality has been assigned conventionally by the percentage of overlap with promoter regions and spurious *Alu* elements. In the original publication of *CpGcluster* [14] we added the overlap with evolutionarily conserved elements or PhastCons [23] as an indicator of putative functionality. Here, we add several new types of analysis to assess the prediction quality, namely the capability to distinguish between different methylation domains or different alternative Transcription Start Sites (TSSs).

CpG islands in the promoter region

Since CpG islands are preferentially located in the promoter region of genes, this fact has been extensively used to assess the quality of CGI predictions [24]. Recently, it has been claimed [22] that a higher percentage of TJ islands (35%) are located within the promoter when compared to *CpGcluster* islands (14.7%). In Table 1, we show a similar analysis as carried out in [22], but extending the comparison to other window based programs and different prediction sets for the *CpGcluster* algorithm. When considering *CpGcluster* islands with $p\text{-value} \leq 1E-5$ (the original relaxed set), the CGI fraction overlapping the promoter region is effectively smaller than for the other programs. However, note that the numbers of CGIs predicted by window-based methods are far below the number predicted by *CpGcluster*. To allow for an unbiased comparison, we obtained a second, strict set of *CpGcluster* islands simply by increasing the required statistical significance to $p\text{-value} \leq 1E-20$ (i.e. filtering out the less significant islands), then obtaining a total 25,454 CGIs. This number is within the range of recent estimates for the complete human somatic cell CGI complement [25]. The strict, more statistically significant set of *CpGcluster* islands shows now the highest overlap (52.4%) with the promoter region. This advantage looks even more important when considering that the genome coverage of our strict set (0.65%) was the lowest one. This indicates a high specificity of *CpGcluster*, which strongly supports our original claim that the $p\text{-value}$ is the most important



parameter to distinguish promoter CGIs from the rest of genome islands [14].

A comparison of length and p-value thresholds

The main quality parameter in SWA is the window size (CGI length threshold). Originally, the window size was set to 200 bp to assure that the detected regions surpass the G + C and O/E criterion not due to chance alone [1]. Subsequently, this threshold was increased to 500 bp in order to reduce the false positive rate by eliminating spurious *Alu* elements [21]. This criterion was replaced in *CpGcluster* by the statistical significance (*p-value*), a more robust and reliable way to distinguish true CGIs from stochastic noise, disregard island length [14]. Note that the *p-value* is not just a different expression for the island length. A non-linear relation exists between the *p-values* and the lengths of the predicted *CpGcluster* islands, as the *p-value* depends on both the island length and the island density (Figure 2).

To evaluate the discrimination power of *CpGcluster p-value* against window size, we generated a series of island-set predictions, each one containing the same number of islands, by appropriately varying the window size or the *p-value* thresholds. Next, we determined the overlap of the resulting islands with the promoter regions, PhastCons elements [23] and *Alu* repeats. The island sets selected by *p-value* clearly outperformed those selected by length: a higher percentage of *CpGcluster* islands overlap with promoters (Figure 3) and PhastCons elements (Figure 4) along the entire range of the two parameters, at the same time reducing the overlap with spurious *Alu* elements (Figure 5). Table 2 shows the correspondence between the number of predicted islands, *p-value* and window length.

The results in Figures 3, 4, 5 are straightforward in comparing the relative strengths of the two main parameters involved in CGI quality (length and *p-value*). The increased stringency in the conventional parameters used by the TJ program excluded contaminating *Alu* elements,

Table 1: Co-localization of CpG islands and the promoter region.

| Method | Number of predicted islands | Genome coverage (%) | Promoter overlap (R13) | |
|--------------------|-----------------------------|---------------------|------------------------|-------|
| | | | Number of islands | % |
| TJ | 37,323 | 1.43 | 14,034 | 37.60 |
| UCSC | 27,639 | 0.74 | 13,369 | 48.40 |
| CpGproD | 76,886 | 2.81 | 14,814 | 19.30 |
| <i>CpGcluster:</i> | | | | |
| relaxed set* | 198,702 | 1.90 | 30,660 | 15.43 |
| strict set** | 25,454 | 0.65 | 13,349 | 52.40 |

* p -value $\leq 1E-5$; ** p -value $\leq 1E-20$

but it also reduced the number of gene promoter associated islands, suggesting that bona fide CGIs were also being discarded [25]. However, raising the statistical significance (i.e. decreasing the p -value) of CpGcluster leads to an exponential increase in the overlap with promoters or PhastCons, simultaneously decreasing the overlap with Alu elements. CpGcluster algorithm is, therefore, a more rational and powerful way to increase CGI prediction quality. An additional advantage is that CpGcluster p -value would be particularly useful in comparative genomics of CGIs, making possible the comparison of CGIs with the same statistical significance, but obtained from different species, despite variations in G + C content or CpG density.

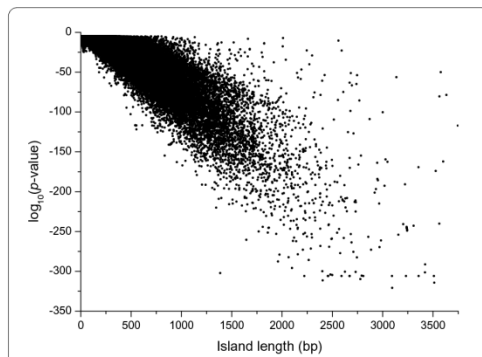


Figure 2 The length of CpGcluster islands vs. the logarithm of the assigned p -value. It can be seen that no linear correlation exists and that the relation between p -value and length is more complex, e.g. the p -value depends on both the island length and the island density.

Prediction of unmethylated regions

The most important criterion to assess putative functionality of a CpG island is the absence of methylation. Therefore, the comparison to experimentally verified, unmethylated regions is another important analysis type to establish prediction quality.

Recently, the methylation status of 697 hypermethylated and 6,987 hypomethylated promoter regions in WI38 primary lung fibroblast [26] have been used to compare the prediction quality of TJ and CpGcluster algorithms [22]. In this study, the prediction quality was measured in the following way: i) true positives (TP): hypomethylated promoters containing a predicted island, ii) false positives (FP): hypermethylated promoters containing a predicted island, iii) true negatives (TN): hypermethylated promoters not containing a predicted island, and iv) false negatives (FN): hypomethylated promoters not containing a predicted island.

However, in our opinion, there is an important pitfall in such an approach. It is known that the methylation state of a given region can change among different tissues; therefore, assigning a "false positive" label to a predicted island which has been shown to be methylated in a single tissue may be misleading, as the same prediction could be perfectly "true positive" if measured in a different tissue.

Fortunately, Weber *et al.* [26] also determined the methylation states in sperm. Analyzing fibroblast and sperm data together, we observed that 11,260 regions are unmethylated in both tissues but 1,550 are unmethylated in one tissue but methylated in the other one. This means that around 12% of the regions are differentially methylated; therefore, a substantial number of FPs were actually TPs. Given these data, in our opinion, without the knowl-

Table 2: Correspondence between the number of predicted islands, log (p-value) and window length.

| No. of predicted islands | log (p-value) | Window length |
|--------------------------|---------------|---------------|
| 193,856 | 5.06509 | 200 |
| 139,013 | 6.1864 | 250 |
| 109,907 | 7.19943 | 300 |
| 69,477 | 9.82744 | 350 |
| 52,687 | 11.85626 | 400 |
| 42,392 | 13.73824 | 450 |
| 37,293 | 14.96788 | 500 |
| 33,691 | 15.95388 | 550 |
| 30,881 | 16.8824 | 600 |
| 28,162 | 18.18919 | 650 |
| 26,192 | 19.45203 | 700 |

edge of the methylation state in a vast number of different tissues, the number of "false positive predictions" cannot be assessed in this way.

We therefore based our quality assessment on sensitivity, a measure not dependent on the false positive rate, as well as on the estimation of the lower bound for the posi-

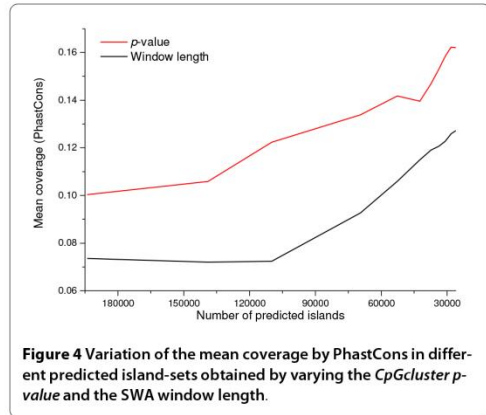


Figure 4 Variation of the mean coverage by PhastCons in different predicted island-sets obtained by varying the *CpGcluster* p-value and the SWA window length.

tive predictive value (PPV, see Data and Methods), a measure used in the gene prediction field under the name of specificity [27]. We used two different experimentally validated sets of unmethylated regions (see Data and Methods) to assess the quality of the 5 sets of predicted islands. Table 3 depicts the results when taking genome-wide, experimentally verified unmethylated CpG islands as reference (Bird's islands, [28]). The table shows that the *CpGcluster* relaxed set shows the highest sensitivity while the strict set shows the lowest one. When considering the lower boundary of the PPV (i.e. the method is at least as specific as this value), we observed the contrary pattern, the *CpGcluster* strict set now shows the highest PPV, while the relaxed set shows the lowest one. Table 4 seems to confirm this trend when using unmethylated regions which are mainly related to promoters [26]. These results indicate that *CpGcluster* is either the most sensitive or the most specific algorithm, depending on the applied p-

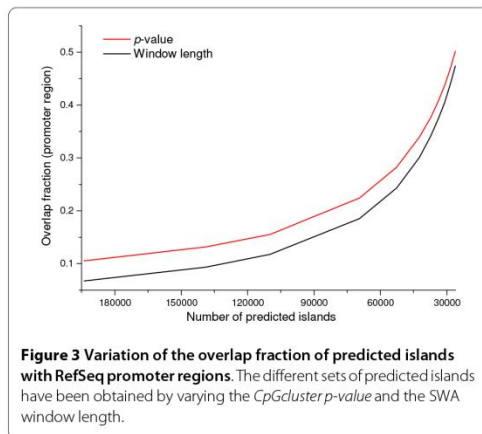


Figure 3 Variation of the overlap fraction of predicted islands with RefSeq promoter regions. The different sets of predicted islands have been obtained by varying the *CpGcluster* p-value and the SWA window length.

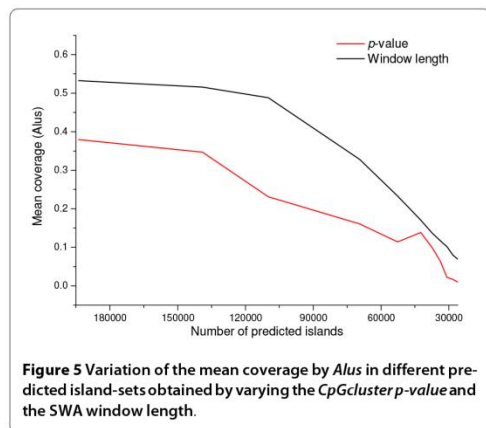


Figure 5 Variation of the mean coverage by *Alus* in different predicted island-sets obtained by varying the *CpGcluster* p-value and the SWA window length.

Table 3: Prediction of unmethylated regions (Bird's islands, N = 17,383).

| Method | Number of predicted islands | Number of islands overlapping a Bird's island | Number of Bird's islands 'touched' by the prediction | SN | PPV |
|---------------------|-----------------------------|---|--|-------|-------|
| TJ | 37,293 | 14,315 | 14,942 | 0.854 | 0.384 |
| UCSC | 27,639 | 13,858 | 14,256 | 0.816 | 0.501 |
| CpGproD | 76,886 | 14,250 | 15,346 | 0.875 | 0.185 |
| <i>CpGcluster</i> : | | | | | |
| relaxed set* | 198,702 | 29,235 | 15,497 | 0.939 | 0.147 |
| strict set** | 25,454 | 14,809 | 12,623 | 0.757 | 0.582 |

p*-value ≤ 1E-5; *p*-value ≤ 1E-20

value threshold. The finding for the relaxed set confirms the result reported by Han and Zhao [22]. Note, however, that *CpGcluster* strict set reaches the highest specificity but the lowest sensitivity. Interestingly, a recent study [29] also emphasizes that the *CpGcluster p*-value is a key attribute for distinguishing between constitutively methylated and unmethylated CGIs.

CpG islands in the domains bound by polycomb repressive complex 2

Functional clusters of CpGs are not limited to promoter regions, they are also found in other genomic locations. An example are the hyperconserved CpG domains largely overlapping the domains bound by polycomb repressive complex 2 (PRC2) [30], located far from the promoter

and playing an important role in transcriptional silencing during development. We determined the overlap of the CGIs predicted by different finders with the domains bound by PRC2. Table 5 shows that all the finders show high sensitivities and low PPVs in predicting these sites, being *CpGcluster* the algorithm obtaining the highest sensitivity (relaxed set).

Functional specificity vs. length of CpG islands

One of the most striking differences between SWA and the *CpGcluster* approach is the length of the predicted islands. SWA islands are on average much longer than *CpGcluster* islands (TJ = 1,094.9; UCSC = 764.5; CpG-ProD = 1,046.1; *CpGcluster* = 273.2 (relaxed set), or 727.5 (strict set)). Originally, CGIs were estimated to be on

Table 4: Prediction of unmethylated regions (Weber's regions, N = 13,277).

| Method | Number of predicted islands | Number of islands overlapping a Weber's region | Number of Weber's regions 'touched' by the prediction | SN | PPV |
|---------------------|-----------------------------|--|---|-------|-------|
| TJ | 37,293 | 10,179 | 9,965 | 0.755 | 0.273 |
| UCSC | 27,639 | 9,788 | 9,552 | 0.724 | 0.354 |
| CpGproD | 76,886 | 10,320 | 10,257 | 0.774 | 0.134 |
| <i>CpGcluster</i> : | | | | | |
| relaxed set* | 198,702 | 18,967 | 10,372 | 0.867 | 0.095 |
| strict set** | 25,454 | 9,633 | 8,378 | 0.663 | 0.378 |

p*-value ≤ 1E-5; *p*-value ≤ 1E-20

Table 5: Overlap of different CGIs with 3,465 domains bound by the polycomb repressive complex 2 (PRC2).

| Method | Number of predicted islands | Number of islands overlapping PRC2 domains | Number of PRC2 domains 'touched' by the prediction | SN | PPV |
|--------------------|-----------------------------|--|--|-------|-------|
| TJ | 37,293 | 3,523 | 3,033 | 0.891 | 0.094 |
| UCSC | 27,639 | 3,179 | 2,790 | 0.825 | 0.115 |
| CpGproD | 76,886 | 3,321 | 3,159 | 0.916 | 0.043 |
| <i>CpGcluster:</i> | | | | | |
| relaxed set* | 198,702 | 9,097 | 3,097 | 0.961 | 0.046 |
| strict set** | 25,454 | 3,424 | 2,372 | 0.758 | 0.135 |

*p-value ≤ 1E-5; **p-value ≤ 1E-20

average 1 kb long [1]. Frequently, more than one *CpGcluster* island can be found within the promoter region and furthermore, several *CpGcluster* islands are often embedded within one single conventional, SWA island. For instance, around 53% of all TJ islands host more than one *CpGcluster* island (Figure 6).

Given these facts, it might be that either conventional SWA predictions erroneously merge smaller islands into longer ones, or that *CpGcluster* erroneously fragments longer islands into many smaller ones. Next, we use alternative TSSs and single CpG resolution methylation data to shed light on these questions.

Alternative promoters

Frequently, *CpGcluster* predicts more than one island within the promoter region. It has been shown [22] that 37.8% of all RefSeq genes have more than one *CpGcluster* island, while only 3.2% have more than one TJ island. Following the premise "one promoter one CpG island", this observation was interpreted as a disadvantage of *CpGcluster* [22]. However, in recent years, new insights into the regulation of gene expression became available, showing among other things a frequent use of alternative TSSs. The existence of alternative TSSs opens the possibility that more than one island per gene might exist. Therefore, the high percentage of genes with more than one *CpGcluster* island might instead indicate a more specific relation of *CpGcluster* islands to alternative promoters or TSSs. To check this possibility, we used the DBTSS database [31]. Out of 15,194 RefSeq genes annotated in the latest DBTSS release, 7,895 (52%) have at least one alternative TSS. With such scenario, one might expect up to 52% of all promoters having more than one island in its promoter (one for each TSS). Given these numbers, the

reported 37.8% of genes with more than one *CpGcluster* island might look not so inadequate.

Conversely, this finding might indicate that the TJ algorithm artificially joins several functional islands into one single longer island. To further investigate this possibility, we estimated the number of islands simultaneously overlapping multiple TSSs annotated in the DBTSS database. Table 6 shows that the *CpGcluster* sets, both relaxed and strict, overlap a higher fraction of unique, and a lower fraction of multiple TSSs than the islands predicted by other programs, thus making *CpGcluster* predictions much more specific in overlapping individual TSSs.

Figure 7 shows a particular example of a bidirectional promoter region. The TSSs of the two genes, UFD1L and CDC45L, are overlapped by the same TJ or UCSC island, while *CpGcluster* predicts separate islands. This is inter-

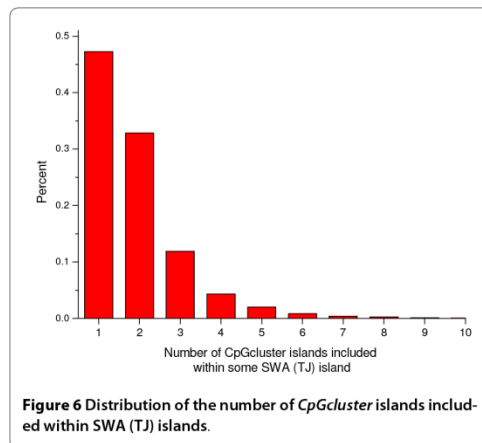


Figure 6 Distribution of the number of *CpGcluster* islands included within SWA (TJ) islands.

esting, as these two genes have very different expression breadths. Using the GeneAtlas2 expression data [32], we determined for UFD1L an expression breadth of 97.3% (expressed in 71 out of 73 healthy tissues), being therefore a housekeeping gene, while the CDC45L gene is expressed in just 15.1% (11 of 73) of all tissues. Given this differential gene expression pattern, a shared CpG island seems to be less specific than the scenario where each of the genes has its own island, as suggested by the prediction of *CpGcluster*.

In the human genome, there are a total of 166 bi-directional promoter pairs which share one long SWA CGI but two separated *CpGcluster* CGIs. The gene-pair shown in Figure 7 may be just an example of extreme differentiation in gene-expression: while the first member of the gene-pair is a housekeeping gene, the second one is a tissue-specific gene. However, one cannot reasonably expect that this may be the rule for all the bidirectional gene-pairs. In fact, after analyzing the expression profiles in a sample of 73 healthy tissues, only 16 (or 9.64%) gene-pairs show a completely divergent pattern of gene-expression (coexpression value ≤ 0.2 , see Methods), while 13 (or 7.83%) exhibit complete coexpression (coexpression value = 1). The remaining gene-pairs show intermediate values of coexpression.

On the other hand, by using single base resolution methylation data [33], we also analyzed methylation differences between the CGIs overlapping bi-directional promoters. We found that 10 (or 11.24%) of these island-pairs in H1 stem cells, and 15 (or 16.85%) in the IMR90 fetal lung fibroblasts, show significant differences (Mann-Whitney non-parametric test) in their methylation average ($p \leq 0.05$).

Heterogeneous methylation in long SWA islands

A functional CpG island should show a rather homogeneous methylation profile among the different CpGs and over the different tissues. For example, the existence of more than one methylation domain within a predicted island might indicate an erroneous merging of two small islands into a single longer island.

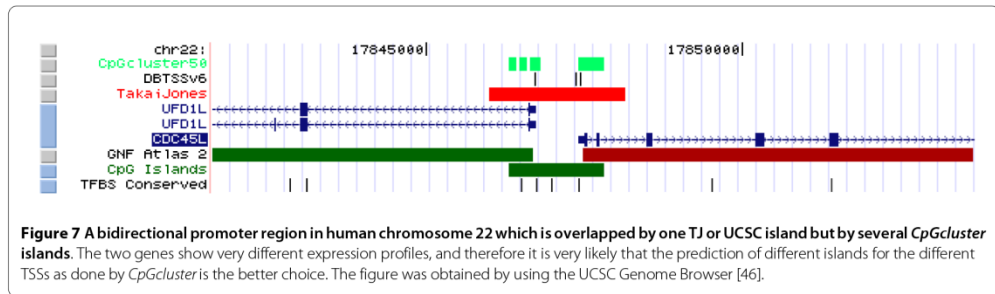
Here, we used single base resolution methylation data from different sources (see Data and Methods) to decide whether *CpGcluster* predicts too many short islands or SWA predict too many long islands. In doing so, we detect all TJ islands which harbor at least two *CpGcluster* islands. Next, we calculate the mean methylation for each *CpGcluster* island and the maximal difference in methylation over the different tissues. If many TJ islands exist with high methylation differences inside, this might indicate an erroneously joining of different methylation domains into a single island. Figure 8 shows a particular example from human chromosome 22. The region for which HEP data were available is just 317 bp long, showing a very pronounced change of the methylation values in embryonic liver cells. All SWA programs predict a very long island in this region, including completely the interesting region where the un-methylation/methylation border occurs. Only *CpGcluster* predicts precisely one CGI for each of the methylation domains.

Figure 9a shows the distribution of the maximum differences in the methylation of CpGs inside TJ islands for HEP data. It can be seen that very high differences occur, around 12% of all tested islands having higher differences than 30% in methylation. Methylation HEP data are available for only 5% of all tissues, and therefore the 12% of heterogeneous TJ islands merging several methylation domains might increase when data for more tissues

Table 6: Co-localization of CpG islands and alternative promoters.

| Method | Numbers of overlapping islands | | |
|----------------------------|--------------------------------|-----------------|----------------|
| | All the TSSs | Unique TSS | Multiple TSSs |
| TJ | 13,759 | 8,868 (64.45%) | 4,891 (35.55%) |
| UCSC | 11,826 | 8,143 (68.86%) | 5,518 (31.14%) |
| CpGproD | 15,319 | 9,801 (63.98%) | 5,518 (36.02%) |
| <i>CpGcluster</i> islands: | | | |
| relaxed set* | 15,095 | 12,034 (79.72%) | 3,061 (20.28%) |
| strict set** | 10,325 | 7,659(74.18%) | 2,666 (25.82%) |

* p -value $\leq 1E-5$; ** p -value $\leq 1E-20$



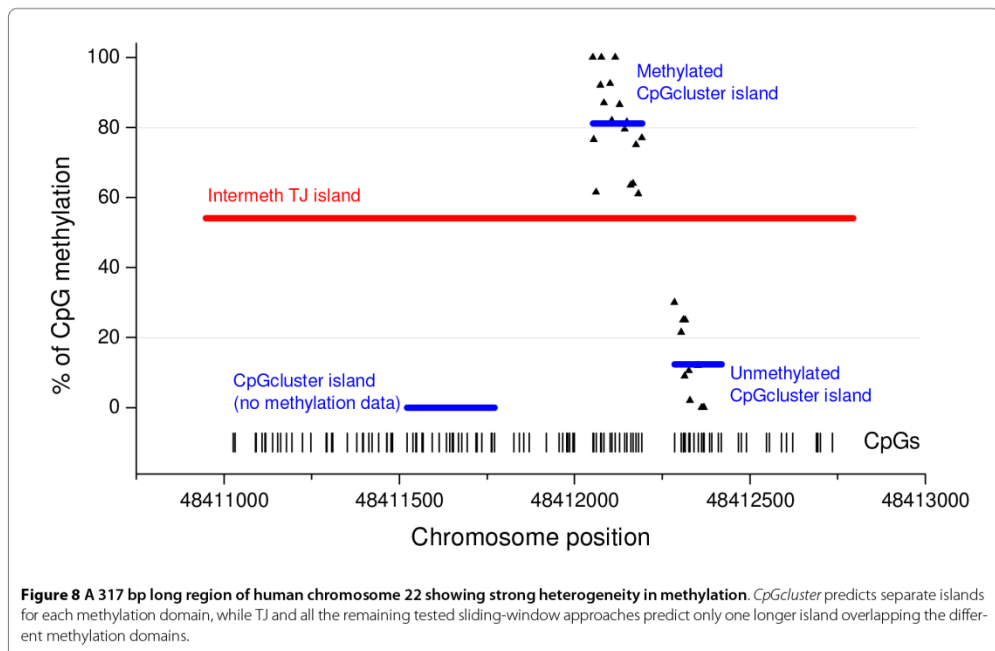
becomes available. A similar conclusion can be reached when methylation data for two human methylomes [33] were used (Figure 9b). Note that the complex methylation structure within CpG islands has been reported before within a different context, but also showing that many long CpG islands contain more than one methylation domain [34].

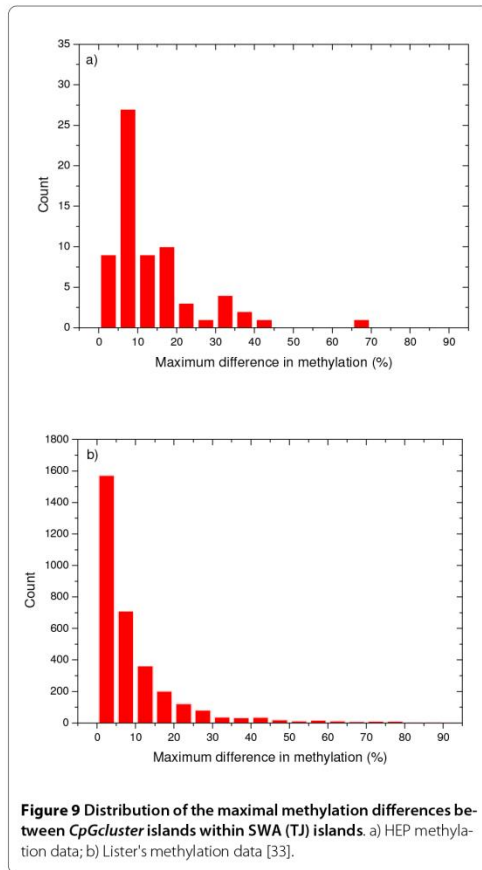
CpG-islets

CpG "islets", genomic regions not conventionally classified as CGIs because of their short length (<200 bp), but having a GC content and observed-to-expected CpG ratio characteristic of a CGI, have recently been identified in a 6.76-Mbp chromosomal region (10q25) containing a

neocentromere [35]. Some of these islets remain unmethylated, corresponding to sites of active transcription and/or boundaries that separate major chromatin sub-domains. This suggests that, as conventional islands, the islets can also participate in the maintenance of a particular genomic pattern of methylated/unmethylated CpGs, thus contributing to the differential regulation of gene expression [3-5].

Given their tiny size, islets remain undetected by SWA, conventional CGI finders [2,21,36-40], as all these programs share a length threshold above 200, or even 500 bp. Such length thresholds make conventional finders useless for the detection of CpG islets, since a relaxation of the length threshold will lead to a strong increment of false





positives. However, since *CpGcluster* [14] does not use any length threshold, it allows to identify short, but statistically significant CpG islets. A genome-wide search identifies a total of 88,137 CpG islets in the human genome with $p\text{-value} \leq 10E-5$. Table 7 shows that relatively high percentages of CpG islets overlap with different sets of promoters and evolutionarily conserved elements, thus suggesting a functional role for many of the predicted islets. Noteworthy, a high proportion of these overlapping islets are exclusively predicted by *CpGcluster*, but not by any of the remaining finders. This indicates that: 1) many of the small islands predicted by *CpGcluster* are not fragments of conventional islands, and 2) given the co-localization with functional regions, the islets might be indeed functional.

Using HEP data [41,42] and Lister et al. methylation levels of single cytosines [33], we also determined the number of unmethylated and differentially methylated

CpG 'islets' (Table 8). A high proportion of the sampled CpG islets were unmethylated or differentially methylated, thus again suggesting a functional role for CpG islets. This is a very important point, as differential methylation of islands/islets may be involved in the regulation of gene expression. Again, the proportion of these CpG islets exclusively predicted by *CpGcluster* is very high.

Conclusions

We systematically compared conventional SWA for detecting CGIs to a clustering method, namely the *CpGcluster* algorithm. We showed that both approaches perform very similar when predicting long, unmethylated regions or polycomb sites. However, we found three scenarios where the *CpGcluster* algorithm seems to have advantages. First, the statistical significance assigned to each *CpGcluster* island seems to be a better quality parameter than the window size of conventional finders, as it reduces more efficiently false positive predictions. Second, we have shown that *CpGcluster* islands co-localize in a more specific way to alternative TSSs and methylation domains. Third, we have shown that many of the small islands predicted by *CpGcluster* might be functional, given the overlap with conserved elements or promoter regions. Moreover, 30% of the differentially methylated islets are exclusively predicted by *CpGcluster*, which suggests this method as the option of choice for the experimental verification of islet functionality.

Methods

Sequence Data

We used human genome assembly NCBI 36.1 (hg18), downloaded from the UCSC genome browser <http://hgdownload.cse.ucsc.edu/downloads.html#human>.

Promoter data

To quantify the co-localization of the predictions with promoter regions and principal transcription start sites we used the RefSeq gene annotation [43]. We furthermore used the DBTSS database version 6.0 [44], as it annotates also alternative transcription start sites, as well as start sites which cannot be assigned to a known RefSeq transcript. From both, the RefSeq and DBTSS annotation, we extracted the coordinates of two regions; the transcription start site (TSS) and the promoter regions, defined as TSS-1500 bp to TSS+500 bp.

Genomic elements

We determined the overlap of CGIs with conserved elements (PhastCons) and spurious *Alu* elements. The evolutionarily conserved elements [23] and the *RepeatMasker* [45] annotation of repeated elements were downloaded from the UCSC table browser [46]. In general, we consider two measures to quantify the over-

Table 7: Overlap of CpG islets (N = 88,137) with different sets of promoters and evolutionarily conserved elements.

| Genome element | Number of overlapping CpG islets | Number of overlapping CpG islets exclusively predicted by CpG cluster |
|--------------------------------------|----------------------------------|---|
| Promoters from RefSeq database | 9,826 (11.15%) | 1,218 (12.40%) |
| TSSs from DBTSS database | 1,868 (2.12%) | 398 (21.31%) |
| Promoter regions from DBTSS database | 6,510 (7.39%) | 4,869 (74.79%) |
| PhastCons | 17,613 (19.98%) | 8,219 (46.66%) |

lap between CGIs and genomic elements. First, we define the mean coverage of a CGI prediction as the mean value of all coverage fractions. The coverage fraction can be calculated as the number of bases of an island corresponding to a given genomic element divided by the island length. Furthermore, we calculate the overlap fraction as the number of islands which overlap in at least one base with a given genomic element divided by the total number of predicted islands.

Island predictions

For SWA CGI finders, a CpG island was at least 200 bp long, which excluded the detection of any shorter tracts. To detect CpG-rich regions, disregarding its length, we used a recently published CpG island finder algorithm (*CpGcluster*, [14]) which does not rely on any length threshold but directly predicts statistically significant CpG clusters. Briefly, the *CpGcluster* algorithm can be divided into two steps. First, based on a distance threshold, the individual CpGs which are below this threshold are clustered along the DNA sequence. Second, by means

of the negative binomial distribution a *p-value* is assigned to each CpG cluster, which allows the prediction of highly significant clusters such as CpG islands.

We considered five computational predictions of CpG islands. For the *CpGcluster* algorithm [14], we generated two prediction sets by setting the assigned *p-value* to two different thresholds. We generated a relaxed set with *p-value* $\leq 1E-5$ and a strict set by setting the threshold to $1E-20$. We implemented the TJ algorithm, as explained in [21], by setting the thresholds to: length ≥ 500 bp, GC content $\geq 55\%$, $Obs_{CpG}/Exp_{CpG} \geq 0.65$ and $minCpG \geq 0.6 * L_{island}/16$ (to avoid "mathematical" islands). We generated the CpGprod prediction [38] running the program <http://pbil.univ-lyon1.fr/software/cpgprod.html> with default parameters. Finally, we downloaded the UCSC CpG island predictions from the UCSC table browser [46].

Gene coexpression analysis

We used the GeneAtlas2 expression data [32] to determine the co-expression of gene pairs sharing a bi-direc-

Table 8: Number of unmethylated and differentially methylated CpG 'islets'.

| Dataset | Methylation state* | Number of CpG islets | CpG 'islets' exclusively predicted by CpGcluster |
|--------------------------------------|---------------------------|----------------------|--|
| HEP (12 tissues)** | Unmethylated | 126 | 1 |
| | Differentially methylated | 26 | 8 |
| Lister et al. 2009 (2 cell lines)*** | Unmethylated | 4,460 | 1,472 |
| | Differentially methylated | 373 | 295 |

*Unmethylated: average methylation ≤ 0.2 ; differentially methylated: average methylation ≤ 0.2 in at least one tissue & average methylation ≥ 0.8 in at least one other tissue.

**The methylation state of 246 CpG 'islets' from chromosomes 6, 20 and 22 was determined by using 3,168 individual CpG sites (HEP project). We only included CpGs which have been detected in at least 2 clones or in at least 6 different tissues.

***We used the sequence reads obtained by MethylC-Seq for two human cell lines [33], H1 human embryonic stem cells and IMR90 fetal lung fibroblasts, to get the average methylation level of single cytosines at both DNA strands for these two methylomes. All islands need more than 50% of its CpGs covered. Only cytosines covered by at least 10 reads were counted.

tional promoter. The "coexpression value" for a couple of genes is the ratio of the number of tissues in which both genes are simultaneously expressed (signal levels > 200) or simultaneously not expressed (signal levels <= 200), and the number of healthy tissues with expression data.

Methylation data

Since the lack of methylation of a CpG island is a very good indicator of function [25], we used several different sources of experimental methylation data. Weber *et al.* [26] detected methylation states in two different tissues, fibroblast and sperm. We extracted 13,277 non-overlapping regions which are unmethylated in at least one of the two tissues (scaled 5mC log₂ ratio < 0.3). Next, we used 17,383 CpG island recently detected in blood cells by means of a new technique [28].

Finally, we assigned methylation states (unmethylated, methylated and differentially methylated) to our *CpG-cluster* predictions by means of the data from the HEP-human epigenome project [42]. The data comprises about 1.9 million CpG methylation values, obtained from the analysis of 2,524 amplicons across chromosomes 6, 20 and 22 in 43 samples (derived from 12 different tissues). We first calculated the mean methylation of each CpG dinucleotide over the different clones, then deleting all CpGs which have been detected in less than 2 clones or in less than 6 different tissues. Subsequently, the individual CpGs were labeled as methylated (mean methylation >= 80), intermediate methylated (80-20) and unmethylated (under 20) for each of the different tissues. Next, we define the methylation states of the CpGs over the different tissues in the following way: i) methylated CpG: methylated in more than 50% of tissues and never unmethylated, ii) unmethylated CpG: unmethylated in more than 50% of tissues and never methylated, iii) differentially methylated CpG: both, methylated and unmethylated in different tissues, the number of intermediate methylation states being smaller than 50%. Finally, we assign a methylation label to the CpG islands which have methylation data for more than 50% of its CpGs: i) methylated: more than 50% of the CpGs are methylated and no unmethylated CpG exist, ii) unmethylated: more than 50% of the CpGs are unmethylated and no methylated CpG exist, iii) differentially methylated: more than 50% of all CpGs need to be differentially methylated.

We also used the sequence reads obtained by MethylC-Seq for two human cell lines [33], H1 human embryonic stem cells and IMR90 fetal lung fibroblasts, to get the average methylation level of single cytosines at both DNA strands for these two methylomes. All islands need more than 50% of its CpGs covered. Only cytosines covered by at least 10 reads were counted.

Assessing prediction quality

When comparing the prediction of CpG islands to a gold standard (e.g. experimentally verified islands), we define:

- True Positives (TP): An island overlapping in at least 1 bp with the gold standard
- False Positives (FP): An island not overlapping with the gold standard
- False Negative (FN): An island in the gold standard that has not been predicted.

By means of these values, we then calculate the sensitivity and the Positive Predictive Value (also known as specificity in the gene prediction field [27]):

$$S_n = \frac{TP}{TP+FN}$$
$$PPV^{LB} = \frac{TP}{TP+FP}$$

Note that we consider all islands not overlapping with the gold standard as false positive predictions. However, no complete gold standard exists, and therefore an unknown number of these islands will be actually true positive predictions. This assumption does not affect the sensitivity, as FP does not occur in the equation, but it affects the PPV. Consequently, and since the PPV can only increase when some FPs turn out to be TPs, the value used in this work is the lower boundary PPV of the prediction, e.g. the worst case scenario when all islands which do not overlap with the gold standard are indeed false positives.

List of abbreviations

CGI: CpG island; CpG O/E ratio: Ratio between observed and expected CpG frequencies; CpG: dinucleotide CG; G + C content, %G + C: Molecular fraction of guanine and cytosine; PhastCons: Phylogenetic Conserved Elements; Sn: The sensitivity of the prediction; PPV: Positive Predictive Value of the prediction; SWA: Sliding-window approaches; TJ: Takai/Jones program or island; TSS: Transcription Start Site

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MH designed and performed the experiments and wrote the manuscript (with JLO), GB performed the search for alternative promoters and differential methylation, PC and PLE carried out the theoretical analysis of CpG clustering and help with the interpretation of statistical results, CP determined the overlap of the CGIs predicted by different finders with the domains bound by polycomb repressive complex, and JLO designed the experiments and wrote the manuscript (with MH). All the authors critically read and approved the final version.

Acknowledgements

We thank Andy Choo and Nicholas Wong, from the University of Melbourne (Parkville, Victoria, Australia), by sharing with us their data on CpG islets. We acknowledge the Spanish Government (Grant No. BIO2008-01353) and the Spanish Junta de Andalucía (Grant Nos. P06-FQM1858 and P07-FQM3163) financial support.

MH acknowledges financial support from the 'Juan de la Cierva' grant from the Spanish Government. GB acknowledges financial support from the 'Programa de formación de investigadores del Departamento de Educación, Universidades e Investigación' grant from the Basque Country Government.

Author Details

¹Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva s/n, 18071, Granada, Spain, ²Lab. de Bioinformática, Centro de Investigación Biomédica, PTS, Avda. del Conocimiento s/n, 18100, Granada, Spain, ³Dpto. de Física Aplicada II, E.T.S.I. de Telecomunicación, Universidad de Málaga 29071-Málaga, Spain, ⁴Division of Sleep Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA, ⁵Dpto. de Ingeniería Mecánica y Minería, EPS Jaén-Universidad de Jaén, Campus Las Lagunillas s/n A3-008, 23071-Jaén, Spain and ⁶Computational Biology Unit, Bergen Center for Computational Science & Sars Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgt 55, 5008 Bergen, Norway

Received: 26 January 2010 Accepted: 26 May 2010
Published: 26 May 2010

References

- Gardiner-Garden M, Frommer M: CpG islands in vertebrate genomes. *Journal of molecular biology* 1987, **196**(2):261-282.
- Larsen F, Gundersen G, Lopez R, Prydz H: CpG islands as gene markers in the human genome. *Genomics* 1992, **13**(4):1095-1107.
- Antequera F: Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci* 2003, **60**(8):1647-1658.
- Bird A: DNA methylation patterns and epigenetic memory. *Genes & development* 2002, **16**(1):6-21.
- Bird AP: CpG-rich islands and the function of DNA methylation. *Nature* 1986, **321**(6067):209-213.
- Zhu J, He F, Hu S, Yu J: On the nature of human housekeeping genes. *Trends Genet* 2008, **24**(10):481-484.
- Baylin SB, Esteller M, Rountree MR, Bachman KE, Schuebel K, Herman JG: Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Human molecular genetics* 2001, **10**(7):687-692.
- De Smet C, Lurquin C, Lethe B, Martelange V, Boon T: DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter. *Molecular and cellular biology* 1999, **19**(11):7327-7335.
- Esteller M, Corn PG, Baylin SB, Herman JG: A gene hypermethylation profile of human cancer. *Cancer research* 2001, **61**(8):3225-3229.
- Issa JP: CpG island methylator phenotype in cancer. *Nature reviews* 2004, **4**(12):988-993.
- Riazalhosseini Y, Hoheisel JD: Do we use the appropriate controls for the identification of informative methylation markers for early cancer detection? *Genome biology* 2008, **9**(11):405.
- Song F, Smith JF, Kimura MT, Morrow AD, Matsuyama T, Nagase H, Held WA: Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(9):3336-3341.
- Hsieh F, Chen SC, Pollard K: A Nearly Exhaustive Search for CpG Islands on Whole Chromosomes. In *The International Journal of Biostatistics Volume 5*. The Berkeley Electronic Press; 2009:1-24.
- Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Aroza J, Oliver JL: CpGcluster: A distance-based algorithm for CpG-island detection. *BMC Bioinformatics* 2006, **7**(1):446.
- Glass JL, Thompson RF, Khulan B, Figueroa ME, Olivier EN, Oakley EJ, Van Zant G, Bouhassira EE, Melnick A, Golden A, *et al*: CG dinucleotide clustering is a species-specific property of the genome. *Nucleic acids research* 2007, **35**(20):6798-6807.
- Su Juan Y, Asaithambi A, Liu Y: CpGIF: an algorithm for the identification of CpG islands. *Bioinformatics* 2008, **2**(8):335-338.
- Izarray RA, Wu H, Feinberg AP: A species-generalized probabilistic model-based definition of CpG islands. *Mamm Genome* 2009, **20**(9-10):674-680.
- Wu H, Caffo B, Jaffee HA, Izarray RA, Feinberg AP: Redefining CpG islands using hidden Markov models. *Biostatistics* 2010 in press.
- Hughes A, Rando OJ: Chromatin 'programming' by sequence - is there more to the nucleosome code than %GC? *J Biol* 2009, **8**(11):96.
- Bock C, Walter J, Paulsen M, Lengauer T: CpG island mapping by epigenome prediction. *PLoS Comput Biol* 2007, **3**(6):e110.
- Takai D, Jones PA: Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(6):3740-3745.
- Han L, Zhao Z: CpG islands or CpG clusters: how to identify functional GC-rich regions in a genome? *BMC Bioinformatics* 2009, **10**:65.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson S, Spiehl J, Hillier LW, Richards S, *et al*: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005, **15**(8):1034-1050.
- Saxonov S, Berg P, Brutlag DL: A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(5):1412-1417.
- Illingworth RS, Bird AP: CpG islands--a rough guide'. *FEBS letters* 2009, **583**(11):1713-1720.
- Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D: Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature genetics* 2007, **39**(4):457-466.
- Burset M, Guigo R: Evaluation of gene structure prediction programs. *Genomics* 1996, **34**(3):353-367.
- Illingworth R, Kerr A, Desousa D, Jorgensen H, Ellis P, Stalker J, Jackson D, Clee C, Plumb R, Rogers J, *et al*: A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS biology* 2008, **6**(1):e22.
- Previti C, Harari O, Zwir I, del Val C: Profile analysis and prediction of tissue-specific CpG island methylation classes. *BMC Bioinformatics* 2009, **10**:116.
- Tanay A, O'Donnell AH, Damelin M, Bestor TH: Hyperconserved CpG domains underlie Polycomb-binding sites. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(13):5521-5526.
- Yamashita R, Suzuki Y, Wakaguri H, Tsuritani K, Nakai K, Sugano S: DBTSS: DataBase of Human Transcription Start Sites, progress report 2006. *Nucleic acids research* 2006:D86-89.
- Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al*: A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(16):6062-6067.
- Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, *et al*: Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009, **462**(7271):315-322.
- Hodges E, Smith AD, Kendall J, Xuan Z, Ravi K, Rooks M, Zhang MQ, Ye K, Bhattacharjee A, Brizuela L, *et al*: High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Res* 2009, **19**(9):1593-1605.
- Wong NC, Wong LH, Quach JM, Canham P, Craig JM, Song JZ, Clark SJ, Choo KH: Permissive transcriptional activity at the centromere through pockets of DNA hypomethylation. *PLoS genetics* 2006, **2**(2):e17.
- Li W, Bernaola-Galván P, Haghghi F, Grosse L: Applications of recursive segmentation to the analysis of DNA sequences. *Computers & chemistry* 2002, **26**(5):491-510.
- Luque-Escamilla PL, Martínez-Aroza J, Oliver JL, Gómez-Lopera JF, Román-Roldán R: Compositional searching of CpG islands in the human genome. *Phys Rev E* 2005, **71**:6.
- Ponger L, Mouchiroud D: CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics (Oxford, England)* 2002, **18**(4):631-633.
- Takai D, Jones PA: The CpG island searcher: a new WWW resource. *silico biology* 2003, **3**(3):235-240.
- Wang Y, Leung FC: An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics (Oxford, England)* 2004, **20**(7):1170-1177.
- Human Epigenome Project [<http://www.epigenome.org/>]
- Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, *et al*: DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature genetics* 2006, **38**(12):1378-1385.

43. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic acids research* 2007:D61-65.
44. Wakaguri H, Yamashita R, Suzuki Y, Sugano S, Nakai K: **DBTSS: database of transcription start sites, progress report 2008.** *Nucleic acids research* 2008:D97-101.
45. **RepeatMasker** [<http://www.repeatmasker.org/>]
46. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, *et al.*: **The UCSC Genome Browser Database: 2008 update.** *Nucleic acids research* 2008:D773-779.

doi: 10.1186/1471-2164-11-327

Cite this article as: Hackenberg *et al.*, Prediction of CpG-island function: CpG clustering vs. sliding-window methods *BMC Genomics* 2010, **11**:327

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit





ISLAS CpG DIFERENCIALMENTE METILADAS

La metilación del ADN es uno de los procesos epigenéticos más estudiados durante los últimos años. Aunque inicialmente la mayoría de los estudios se centraban en la regulación del inicio de la transcripción mediante modificaciones en las regiones promotoras de los genes (Bell, Pai et al. 2011), los avances en los métodos de secuenciación masiva han permitido ampliar estos horizontes. Actualmente, se ha descrito su implicación en múltiples funciones asociadas a diversas regiones del genoma. La metilación en el cuerpo génico, en lugar de inhibir la transcripción, al igual que ocurre en los promotores, se ha demostrado que la estabiliza (Hellman and Chess 2007). A su vez, las variaciones en la metilación pueden intervenir indirectamente en el *splicing* alternativo (Shukla, Kavak et al. 2011), o incluso regular la actividad de potenciadores de la transcripción (Hon, Rajagopal et al. 2013). También es conocida la existencia de elevados niveles de metilación en torno a los elementos repetidos, como en el caso de los centrómeros (Moarefi and Chedin 2011), o de los elementos transponibles para mantener la estabilidad del genoma (Yoder, Walsh et al. 1997). Además de la

diversidad de funciones de la metilación, la cuantificación de la misma en diferentes tipos celulares ha permitido identificar la existencia de aproximadamente un 22% de CpGs autosómicos diferencialmente metilados (Ziller, Gu et al. 2013), lo que muestra un escenario mucho más dinámico de lo que originalmente se pensaba. En general, podemos afirmar que la mayoría de los procesos en los que interfiere la metilación se asocian a la interacción de diferentes moléculas con el ADN, la cual podría encontrarse regulada por estados variables de metilación.

Recientemente, se ha observado que la presencia de ciertas regiones denominadas *MDRs* (Regiones Determinantes de la Metilación), caracterizadas por su elevada densidad de CpGs y la presencia de sitios de unión a factores de transcripción, pueden establecer patrones de hipometilación en *cis* de manera tejido específica (Lienert, Wirbelauer et al. 2011). La elevada densidad de CpGs parece ser un factor determinante en la regulación de la unión de factores de transcripción, ya que la metilación en regiones con baja densidad de CpGs es incapaz de bloquear la unión de dichos factores (Stadler, Murr et al. 2011). Esto sugiere que aquellos sitios de unión donde deba regularse de manera indispensable la interacción del ADN con sus factores de transcripción, deberían presentar una alta densidad de CpGs.

Actualmente, ni las herramientas existentes (Zhang, Liu et al. 2011, Hansen, Langmead et al. 2012, Li, Garrett-Bakelman et al. 2013) para la identificación de Regiones Diferencialmente Metiladas (*DMRs*), ni los estudios realizados en múltiples tejidos (Hon, Rajagopal et al. 2013, Ziller, Gu et al. 2013), tienen en cuenta la densidad de CpGs presentes en la región. Además, como ya se comprobó (Hackenberg, Barturen et al.

2010) y se ha corroborado en un estudio reciente (Ziller, Gu et al. 2013), las diferencias de metilación suelen encontrarse en regiones fuera de la definición clásica de *CGIs* (Islas CpG), por lo que los algoritmos clásicos no son de ayuda a la hora de detectar regiones densas de CpGs que puedan presentar metilación diferencial. Sin embargo, otros algoritmos como *WordCluster* (Hackenberg, Carpena et al. 2011), un algoritmo mejorado a partir de *CpGcluster* (Hackenberg, Previti et al. 2006), que usa una aproximación diferente, permite identificar regiones con una alta densidad de CpGs que probablemente muestren metilación diferencial (Hackenberg, Barturen et al. 2010).

Aunque la caracterización de regiones diferencialmente metiladas en múltiples tejidos ha sido abordada recientemente (Ziller, Gu et al. 2013), el estudio de la metilación diferencial en *CGIs* sigue presentando un gran interés por varios motivos:

- Las *CGIs* son regiones evolutivamente conservadas, en la mayoría de los casos debido a una función crítica para el correcto funcionamiento del organismo (Cohen, Kenigsberg et al. 2011).
- La función de la metilación está íntimamente relacionada con la densidad de citosinas metiladas en la región (Stadler, Murr et al. 2011), por lo que la existencia de metilación diferencial en regiones con una baja densidad de CpGs no implica necesariamente que presenten alguna función.
- Y por último, y no menos importante, definir los contenedores de interés previa caracterización de su estado de metilación, permite identificar regiones diferencialmente metiladas que no variarán su

asociación con otros elementos al incluir nuevos tejidos en el estudio, ya que la longitud y la localización de las *CGIs* no dependen de los tejidos analizados.

De esta manera, la caracterización y análisis de la metilación diferencial en *CGIs* nos permitirá llegar a conclusiones más sólidas que tomando en consideración todas las regiones que puedan presentar metilación diferencial. A lo largo de este capítulo, definiremos y caracterizaremos un conjunto de *CGIs* diferencialmente metiladas basándonos en las predicciones de *WordCluster*. Además, analizaremos su asociación con elementos reguladores y su función, para determinar la posible influencia de la densidad de CpGs en los procesos regulados por la metilación diferencial.

6.1 MATERIAL Y MÉTODOS

Los tipos celulares utilizados han sido extraídos de la base de datos *NGSmethDB* (Hackenberg, Barturen et al. 2011, Geisen, Barturen et al. 2014). El conjunto de tejidos seleccionados incluye tanto líneas celulares como tejidos frescos, y ninguno de ellos proviene de procesos inducidos de diferenciación o de estados patológicos: células B (*bcell*), células madre hematopoyéticas CD133 + (*cd133hsc*), células madre hematopoyéticas (*hspc*) (Hodges, Molaro et al. 2011); fibroblastos epiteliales (*fibro*), células madre H9 (*wa09*) (Laurent, Wong et al. 2010); células madre (*h1*), fibroblastos de pulmón (*imr90*) (Lister, Pelizzola et al. 2009); células mamarias (*hmec*) (Hon, Hawkins et al. 2012); células mononucleares de sangre periférica (*pbmc*) (Li, Zhu et al. 2010); células

del córtex pre-frontal (*prefrontalcortexhs1570*) (Zeng, Konopka et al. 2012) y esperma (*spermdonor1*) (Molaro, Hodges et al. 2011). Las descripciones y la estadística básica de estos tipos celulares se encuentran en las pestañas de contenido y estadística de *NGSmethDB* (<http://bioinfo2.ugr.es/NGSmethDB/>). Los tejidos del conjunto seleccionado presentan 5 lecturas en al menos el 75% de los CpGs del genoma.

6.1.1 Estados y niveles de metilación

Los estados de metilación se han clasificado en 5 clases excluyentes (Tabla 6.1), que se usan tanto para los análisis estadísticos que requieren de clasificaciones discretas (a nivel de CpGs individuales), como para referirse a la metilación dentro de las *CGIs*.

| Etiqueta | Estado de metilación | Nivel de metilación |
|-----------------|-----------------------------|----------------------------|
| U | No metilado | $0 \geq x \leq 0.2$ |
| T | Parcialmente no metilado | $0.2 > x < 0.4$ |
| I | Metilación intermedia | $0.4 \geq x \leq 0.6$ |
| N | Parcialmente metilado | $0.6 > x < 0.8$ |
| M | Metilado | $0.8 \geq x \leq 1$ |

Tabla 6.1. Clasificación discreta de los niveles de metilación.

Los niveles de metilación para las regiones definidas como *CGIs* se han calculado ponderando cada nivel de metilación con su correspondiente profundidad de secuenciación (Schultz, Schmitz et al. 2012):

$$\text{Nivel de metilación} = \sum_{i=1}^n C_i / \sum_{i=1}^n (C_i + T_i) \quad [1]$$

donde n es el número de CpGs en la región, C hace referencia a las citosinas y T a las timinas.

6.1.2 Análisis estadísticos

En este apartado se describen los análisis estadísticos utilizados para la detección de islas CpG diferencialmente metiladas: binomial negativa, prueba exacta de Fisher, prueba de la t de Student y prueba de los rangos de signos de Wilcoxon.

Binomial negativa

Cuando se comparan dos tejidos, la probabilidad de que las diferencias de metilación encontradas en una CGI dada no se deban al azar se puede calcular por medio de la distribución binomial:

$$p = 1 - \sum_{dCpGi} \binom{CpGi}{dCpGi} \rho^{dCpGi} (1 - \rho)^{CpGi - dCpGi} \quad [2]$$

donde $dCpGi$ es el número de CpGs diferencialmente metilados en la isla, $CpGi$ el número de CpGs para ambos tejidos comparados y ρ es la probabilidad de encontrar CpGs diferencialmente metilados en todo el genoma, calculada como:

$$\rho = dCpG / CpG \quad [3]$$

siendo $dCpG$ el número de CpGs diferencialmente metilados y CpG el número de CpGs con datos para los tejidos comparados en todo el genoma. Los CpGs considerados diferencialmente metilados, son aquellos cuyos tejidos presentan estados de metilación M y U, o cualquiera de estos más el I.

Prueba exacta de Fisher

Para tomar en cuenta los niveles de metilación intermedios, se ha aplicado la prueba exacta de Fisher a una tabla 2x3, donde se comparan los CpGs con estados de metilación: M, U e I en cada par de tejidos. La prueba exacta de Fisher se ha abordado desde dos aproximaciones diferentes:

- Datos no emparejados: se incluyen todos los CpGs M, U o I con profundidad suficiente.
- Datos emparejados: sólo se incluyen los CpGs que presenten estados de metilación M, U o I y profundidad suficiente en ambos tejidos.

Prueba de la t de Student y prueba de los rangos de signos de Wilcoxon

Ambas pruebas estadísticas son análisis para distribuciones continuas, por lo que se incluyen los niveles de metilación para todos los CpGs con profundidad suficiente.

6.1.3 Especificidad y sensibilidad

La especificidad (Sp) y sensibilidad (Sn) representadas en la curva ROC (Figura 6.2) se han calculado para todos los pares de tejidos por separado y se definen como:

$$Sp = TN / (FP + TN) \quad [4]$$

$$Sn = TP / (TP + FN) \quad [5]$$

donde TN hace referencia a las CG/s no detectadas como diferencialmente metiladas con diferencias absolutas de metilación por debajo de 0.2, FP son CG/s con valores de metilación significativamente diferentes y con diferencias absolutas de metilación por debajo de 0.2, TP son CG/s con diferencias absolutas de metilación superiores a 0.2 detectadas como diferencialmente metiladas, y FN son CG/s por encima del umbral y sin valores de metilación significativamente diferentes.

El umbral que define las “verdaderas” diferencias de metilación se ha fijado en 0.2, ya que toda CGI con diferencias superiores a este valor, necesariamente presentará diferentes estados de metilación en el par de tejidos comparados (según se han definido en el apartado 6.1.1).

6.1.4 Clasificación de CG/s

Las CG/s del conjunto relajado de $CpGcluster$ (Hackenberg, Previti et al. 2006, Hackenberg, Carpena et al. 2011) se han clasificado en función de la metilación observada en todos los tejidos analizados (se han descartado todas aquellas CG/s que presentan datos de metilación para

menos de la mitad de los tejidos). Esta clasificación define 4 clases de *CG/s*:

- *DMIs* (islas diferencialmente metiladas); son *CG/s* con diferencias estadísticamente significativas de metilación para al menos un par de tejidos.
- *MIs* (islas constitutivamente metiladas); son *CG/s* sin diferencias significativas de metilación y que no presenten ningún tejido fuera de los estados de metilación M o N.
- *UIs* (islas constitutivamente no metiladas); son *CG/s* sin diferencias significativas de metilación y que no presenten ningún tejido fuera de los estados de metilación U o T.
- *NAs*, *CG/s* que no cumplen los requisitos establecidos para el resto de las clases.

6.1.5 Elementos genómicos utilizados para los análisis de enriquecimiento

En este apartado se describen los conjuntos de datos utilizados para los estudios de enriquecimiento de las clases de *CG/s* en elementos reguladores: regiones génicas, *TFBSs* (Sitios de unión a factores de transcripción), sitios de hipersensibilidad a la DNasa I, potenciadores, Aisladores, sitios conservados y *SNPs* (Polimorfismos de un solo nucleótido).

Regiones génicas

Las regiones génicas se han extraído a partir de las tablas de genes *refSeq* (Pruitt, Tatusova et al. 2007), seleccionando sólo aquellos genes que codifican proteínas. La región génica se ha definido como una zona que comprende el cuerpo génico de los genes (desde la posición de inicio de la transcripción hasta el final de la misma) y un entorno génico de 1,000 pb (500 pb aguas arriba (5') del inicio de la transcripción y 500 pb aguas abajo (3') del final de la transcripción). Los exones e intrones se definen tal y como aparecen en la base de datos citada, mientras que la región promotora (R13) comprende 1,500 pb aguas arriba y 500 pb aguas abajo del sitio de inicio de la transcripción, y la región del final de la transcripción (R8), 500 pb aguas arriba y 1,500 pb aguas abajo del final de la transcripción.

Debido a la continuidad existente entre las diferentes regiones génicas, una misma *CGI* podría solapar con más de una región. Para evitar esta redundancia, la clasificación de las *CGIs* en dichas regiones se ha organizado de manera jerárquica, por lo que las islas solapantes con una categoría no serán utilizadas en los análisis de las regiones subsiguientes, siguiendo el orden: R13, R8, exones e intrones.

Elementos reguladores

Los elementos reguladores utilizados se han tomado de las tablas disponibles (<http://genome.ucsc.edu/>) en el buscador de la UCSC (Meyer, Zweig et al. 2013, Rosenbloom, Sloan et al. 2013). Prácticamente, la totalidad de los elementos utilizados se han derivado a partir de datos de *ENCODE* (Consortium, Bernstein et al. 2012):

- Clusters de TFBSs uniformemente procesados (V3), que combinan los resultados para 91 tipos celulares y 189 factores de transcripción.
- Clusters de sitios de hipersensibilidad a la DNasa I (V2) para 125 tipos celulares.
- Clusters de sitios de unión a las subunidades de la ARN polimerasa II, extraídos a partir del conjunto de *clusters* de TFBSs.
- Clusters de sitios de unión a las subunidades de la ARN polimerasa III, extraídos a partir del conjunto de *clusters* de TFBSs.
- Potenciadores VISTA, conjunto de potenciadores determinados experimentalmente *in-vivo* (Visel, Minovitsky et al. 2007).
- Potenciadores potencialmente activos, se han definido como regiones del genoma con sitios de unión a factores de transcripción con actividad de unión específica a potenciadores distales de la transcripción de la ARN polimerasa II (GO:0003705, según la base de datos Gene Ontology (Ashburner, Ball et al. 2000)) y con las modificaciones de histonas asociadas a este tipo de elementos: mono-metilación de la lisina 4 en las histonas H3 (*H3K4me1*) y acetilación en la lisina 27 de las histonas H3 (*H3K27ac*). Esta definición se basa en estudios previos (Zentner, Tesar et al. 2011) y los datos se han obtenido del conjunto uniformemente procesado de ENCODE. Los tejidos seleccionados presentan datos tanto para las marcas de las histonas (sólo se han tomado aquellas regiones con

valores- $p \leq 1e-5$) como para los factores de transcripción; en la Tabla 6.2 se resume el conjunto de datos utilizados.

| Tipos celulares | Genes <i>GO:0003705</i> |
|-----------------|---|
| <i>GMI2878</i> | ATF2, BHLHE40, CEBPB, CREB1, FOXM1, MEF2A, MEF2C, NFATC1, SPI1, RELA, SRF, USF1, USF2 |
| <i>HMEC</i> | JUN |
| <i>K562</i> | BHLHE40, CEBPB, CREB1, JUN, MEF2A, SPI1, SRF, USF1, USF2 |
| <i>A549</i> | BHLHE40, CEBPB, CREB1, FOXA1, FOXA2, USF1 |
| <i>HeLAS3</i> | CEBPB, JUN, USF2 |
| <i>HepG2</i> | BHLHE40, CEBPB, CREB1, FOXA1, FOXA2, HNF4A, JUN, SRF, USF1, USF2 |
| <i>HCT-116</i> | CEBPB |

Tabla 6.2. Tipos celulares y *TFBSs* utilizados para definir el conjunto de potenciadores potencialmente activos.

- Aisladores, la única proteína en vertebrados descrita con función de unión a regiones aisladoras es el CTCF (Bell, West et al. 1999, Kim, Abdullaev et al. 2007). Además de su función como aislador fuera del cuerpo génico, recientemente se ha demostrado su implicación en la regulación del *splicing* alternativo del gen CD45, mediante cambios en la metilación del exón 5 de dicho gen (Shukla, Kavak et al. 2011). Por lo tanto, los *clusters* de sitios de unión a CTCF presentes en el conjunto de *ENCODE* descrito previamente, se han dividido en dos en función de su localización: sitios de unión localizados en regiones intergénicas y sitios de unión solapantes con exones.

Elementos conservados y variaciones de secuencia

La asociación de las *CGIs* con elementos conservados se ha analizado en dos conjuntos obtenidos mediante algoritmos diferentes: *PhastCons* para 46 vertebrados (Siepel, Bejerano et al. 2005, Pollard, Hubisz et al. 2010) y *GERP* para 35 mamíferos (Cooper, Stone et al. 2005, Davydov, Goode et al. 2010). Ambos métodos se describen en detalle en las tablas del buscador de la *UCSC* (<http://genome.ucsc.edu/>). También se ha incluido un conjunto de regiones en las que se ha predicho la existencia de un sesgo mutacional hacia GC, *phasBiasGC* (Duret and Galtier 2009, Hubisz, Pollard et al. 2011).

En cuanto a las variantes de secuencia, se han seleccionado variaciones de un solo nucleótido de la versión 138 de la base de datos *dSNP* (Sherry, Ward et al. 2001) para dos subconjuntos: i) polimorfismos comunes (*dbSNP common*), variaciones con una frecuencia poblacional mayor del 1% (*SNPs*) y ii) variaciones potencialmente asociadas al desarrollo de patologías (*dbSNP flagged*); estas últimas se encuentran asociadas a algún locus incluido en *LSDB* (Horaitis, Talbot et al. 2007) o en *OMIM* (Hamosh, Scott et al. 2005) y sus frecuencias poblacionales son inferiores al 1%.

6.2 IDENTIFICACIÓN DE ISLAS CpG DIFERENCIALMENTE METILADAS

En los últimos años se ha desarrollado una amplia variedad de métodos para identificar regiones diferencialmente metiladas entre tejidos, como *QDMR* (Zhang, Liu et al. 2011), *BSmooth* (Hansen, Langmead et al. 2012) o *eDMR* (Li, Garrett-Bakelman et al. 2013). Sin embargo, estos métodos no tienen en cuenta la densidad de CpGs, a pesar de la importancia que ésta puede tener para la función de la metilación, tal y como se comentó al inicio de este capítulo. Para tomar en cuenta la densidad de CpGs, antes de analizar las diferencias de metilación, deberían seleccionarse regiones con una elevada densidad de CpGs. Para este propósito, se han utilizado las islas predichas por *CpGcluster* (Hackenberg, Previti et al. 2006, Hackenberg, Carpena et al. 2011), que permite identificar *clusters* densos de CpGs altamente significativos.

Una vez definidos los contenedores donde se estudiará la metilación diferencial (las *CG/s*), se debe comprobar si existen diferencias de metilación entre los diferentes tejidos. Aunque el método más extendido para determinar diferencias de metilación en CpGs individuales es la prueba exacta de Fisher (Lister, Pelizzola et al. 2009), no existen estudios que permitan decidir cuál es el análisis estadístico más apropiado para determinar diferencias entre regiones previamente definidas. Sin embargo, existen numerosas aproximaciones estadísticas para determinar estas diferencias, que varían tanto en la prueba estadística a utilizar, como en la organización de los datos (emparejados o no, medir las diferencias utilizando los niveles de metilación continuos o

clasificarlos de manera discreta). Por lo tanto, en este apartado compararemos diferentes aproximaciones estadísticas para determinar cuál es la más apropiada a la hora de identificar diferencias en los niveles de metilación en regiones previamente definidas. Los métodos estadísticos a comparar serán: análisis basado en la binomial negativa, prueba exacta de Fisher, prueba exacta de Fisher para valores emparejados, prueba de la t de Student y prueba de los rangos de signos de Wilcoxon (véase apartado 6.1.2).

6.2.1 Distribución de las diferencias de metilación

En primer lugar, se comparan las distribuciones de las diferencias absolutas de metilación dentro de las *CGI*s identificadas como diferencialmente metiladas por los diferentes métodos estadísticos.

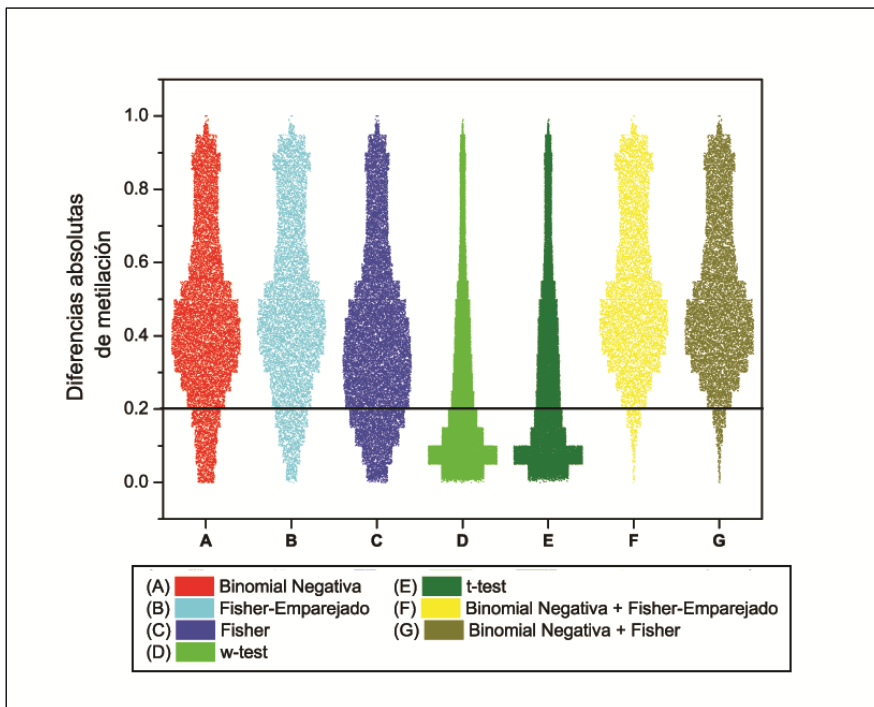


Figura 6.1. Diferencias absolutas de metilación para cada *CGI* identificada como diferencialmente metilada por cada método estadístico. Las diferencias de metilación se calcularon entre las líneas celulares *h1* e *imr90* (Lister, Pelizzola et al. 2009), incluyéndose en el análisis sólo aquellos contextos CpG con una profundidad mayor a 5. Los resultados mostrados en (F) y (G), se obtuvieron aplicando dos test estadísticos conjuntamente (el análisis basado en la binomial negativa junto con cada una de las variantes de la prueba exacta de Fisher). En estos análisis conjuntos, ambos valores-*p* deben mantenerse por debajo del umbral seleccionado (0.05). La línea horizontal es el límite superior de las diferencias que pueden caer dentro de una misma clase, ya que 0.2 es la amplitud de las clases descritas en Material y métodos (apartado 6.1.1).

Los diagramas de dispersión (Figura 6.1), indican que aquellos análisis que previamente agrupan los valores de metilación en clases discretas (prueba exacta de Fisher y análisis basado en la binomial negativa), identifican dos subgrupos biológicamente significativos en función de sus diferencias de metilación:

- Subgrupo en torno a 0.4: estas diferencias se corresponden con regiones donde mayoritariamente uno de los tejidos se encuentra no-metilado (0-0.2) o metilado (0.8-1), mientras que el otro tejido presenta valores intermedios de metilación (0.4-0.6). Estos dominios de metilación intermedia o parcial (*PMDs*), se han descrito en tejidos diferenciados (Lister, Pelizzola et al. 2009), transformándose en dominios metilados al inducir estas células a la pluripotencia (Lister, Pelizzola et al. 2011). A su vez, la presencia de *PMDs* se ha descrito en regiones potenciadoras (Stadler, Murr et al. 2011) y en promotores de genes improntados (Bell, Pai et al. 2011).
- Subgrupo en torno a 0.9: este subgrupo se corresponde con aquellas *CGIs* que varían su metilación entre las clases más extremas de la distribución, pasando de encontrarse metiladas (0.8-1) a no-metiladas (0-0.2), o viceversa. Estos cambios drásticos de metilación se han asociado con regiones promotoras y cuerpos génicos (Laurent, Wong et al. 2010), provocando variaciones en los niveles de expresión de dichos genes (Dindot, Person et al. 2009, Irizarry, Ladd-Acosta et al. 2009).

Por otro lado, los análisis que comparan niveles continuos de metilación (prueba de la *t* de Student y prueba de los rangos de signos de

Wilcoxon), con la significación estadística utilizada (valor- $p \leq 0.05$), no sólo no muestran la existencia de estos subgrupos funcionales, sino que además, identifican como diferencialmente metiladas un gran número de islas que no sobrepasan diferencias de metilación de 0.2 (es decir, dichas islas se encontrarán muy probablemente dentro de la misma clase de metilación). Si se aumenta la significación estadística (umbral de valor- $p \leq 1e-5$), comienzan a vislumbrarse los subgrupos descritos previamente; sin embargo, el número de *CG/s* con diferencias de metilación detectadas es entre 2 y 4 veces menor que en los análisis para datos discretos.

Por otra parte, se ha observado que al utilizar el análisis basado en la binomial negativa conjuntamente con la prueba exacta de Fisher (método combinado), se reduce el número de islas con metilación diferencial significativa que se encuentran por debajo del umbral crítico de diferencias de metilación (Figura 6.1, F y G).

6.2.2 Comparación entre métodos estadísticos

Según las distribuciones observadas en el apartado anterior, parece que los métodos estadísticos combinados son los que mejor definen los subgrupos descritos previamente, además de presentar un menor número de pares identificados como diferencialmente metilados con diferencias de metilación menores a 0.2. Sin embargo, utilizando las distribuciones de la Figura 6.1 es muy difícil diferenciar entre ambos métodos combinados. Por lo tanto, se han comparado los diferentes métodos mediante una curva *ROC* (Característica Operativa del Receptor) para todos los pares de tejidos seleccionados en el estudio (Figura 6.2).

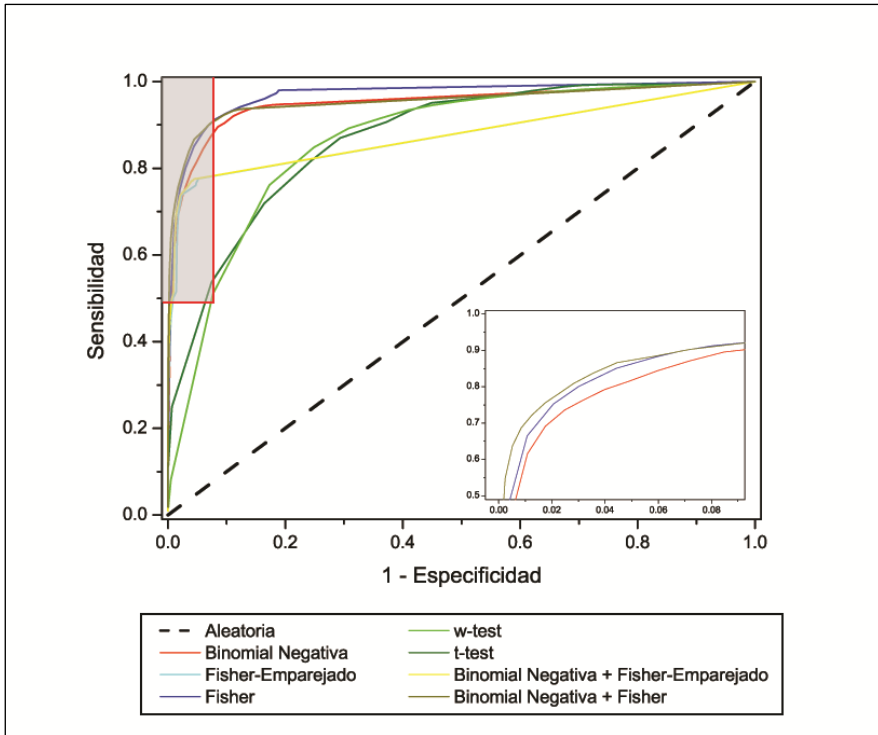


Figura 6.2. Curva ROC de los diferentes análisis estadísticos. Se representa la sensibilidad (S_n) frente a $1 -$ especificidad ($1 - S_p$) de la detección de diferencias de metilación para todos los pares de tejidos incluidos en el estudio (Apartado 6.1). Los valores- p utilizados en los análisis estadísticos varían entre $1e-10$ y 1. En la parte inferior derecha de la imagen se muestra una ampliación de la zona resaltada para los análisis que muestran mejores resultados (Binomial negativa, Fisher y Binomial negativa + Fisher). La línea negra discontinua muestra una predicción aleatoria.

A pesar de las aparentes diferencias observadas en la Figura 6.1 entre los métodos combinados e individuales, sobre todo en cuanto al número de pares diferencialmente metilados por debajo del umbral establecido, al compararlos en una curva ROC (Figura 6.2), los métodos con mejor rendimiento son: el método combinado de la Binomial negativa + Fisher, y ambos por separado. De entre estos métodos, aunque con pequeñas diferencias, el método combinado es el que mejores resultados

presenta en la región de interés de la gráfica (zona resaltada y ampliada, donde se sitúan los valores- p por debajo de 0.05), ya que a iguales valores de sensibilidad presenta mayor especificidad que cualquiera de los otros métodos. Por lo tanto, este será el método estadístico utilizado en adelante para determinar las *CGIs* con metilación diferencial.

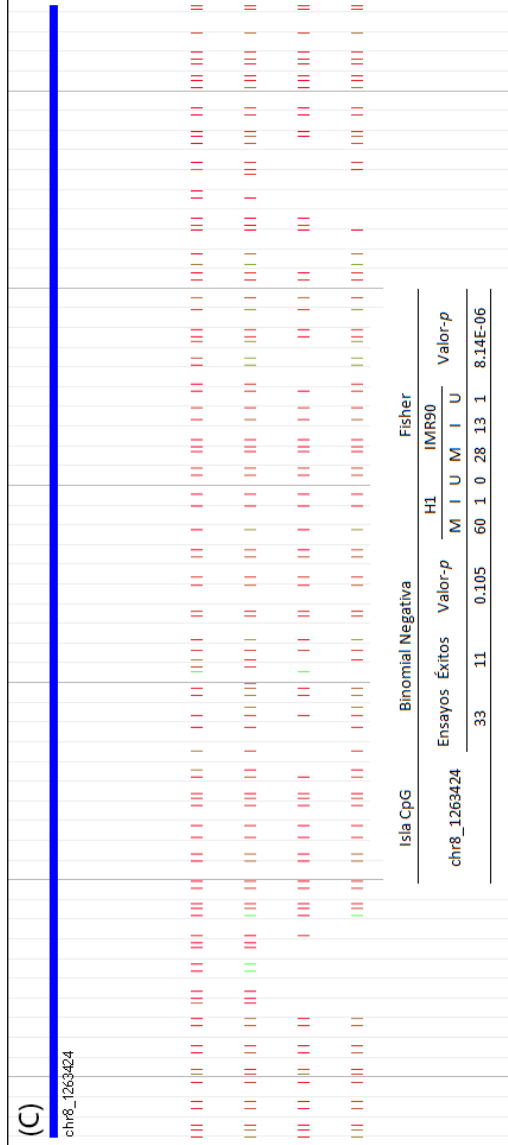
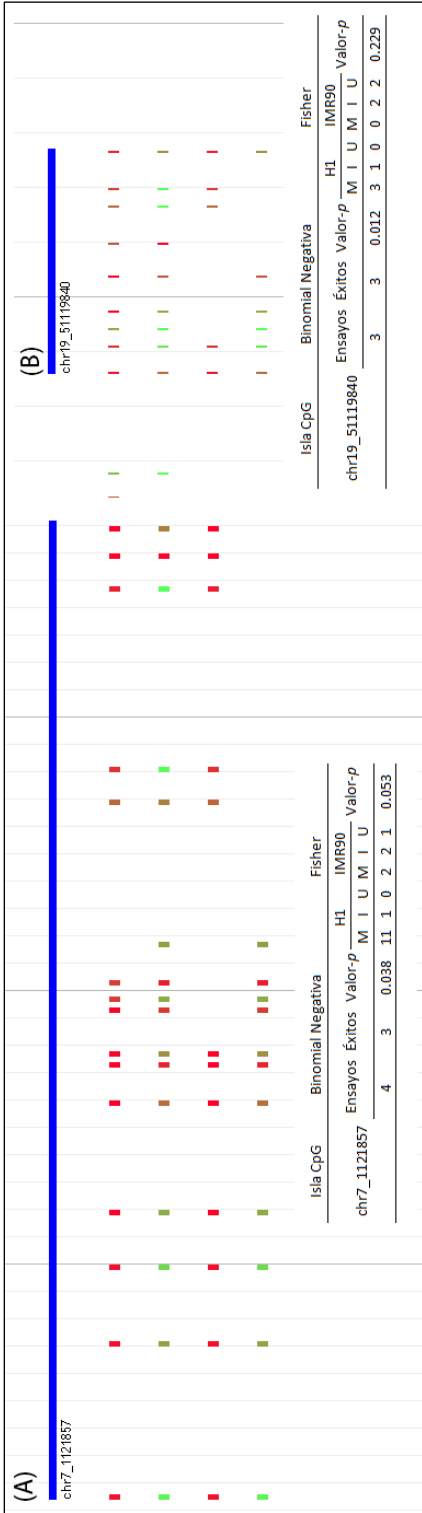
6.2.3 Análisis basado en la binomial negativa combinado con la prueba exacta de Fisher

Estadísticamente, identificar diferencias entre dos series numéricas usando secuencialmente dos análisis de manera conjunta, es algo atípico. Sin embargo, biológicamente tiene mucho sentido, ya que cada método captura características diferentes, que no podrían ser tomadas en cuenta por separado:

- Análisis basado en la binomial negativa: analiza aquellos CpGs con profundidad suficiente para ambos tejidos dentro de la *CGI*, en función de la probabilidad de encontrar diferencias de metilación entre dicho par de tejidos a lo largo de todo el genoma (véase apartado 6.1.2).
- Prueba exacta de Fisher: analiza el número de CpGs que caen en cada una de las 3 clases principales de metilación para cada uno de los tejidos comparados (véase apartado 6.1.2).

Por un lado, la binomial comprueba si el número de CpGs con diferencias observadas, teniendo en cuenta el total de CpGs en la región, se encuentran por azar en el par de tejidos analizados. Mientras que la prueba exacta de Fisher, analiza la existencia de diferencias globales de

metilación. En la Figura 6.3 puede observarse este comportamiento, donde las imágenes A y B presentan diferencias significativas según el análisis basado en la binomial negativa, y sin embargo no se aprecian diferencias significativas al observar la región en su conjunto, hecho que recogen los valores- p de la prueba exacta de Fisher. En el caso de las imágenes C y D, la prueba exacta de Fisher detecta diferencias significativas en estas regiones, debido a un sesgo en los CpGs con datos para ambos tejidos. Por su parte, el análisis de la binomial concluye que no existe un número suficiente de pares de CpGs diferencialmente metilados en la región para considerar las diferencias significativas, tal y como puede observarse visualmente.



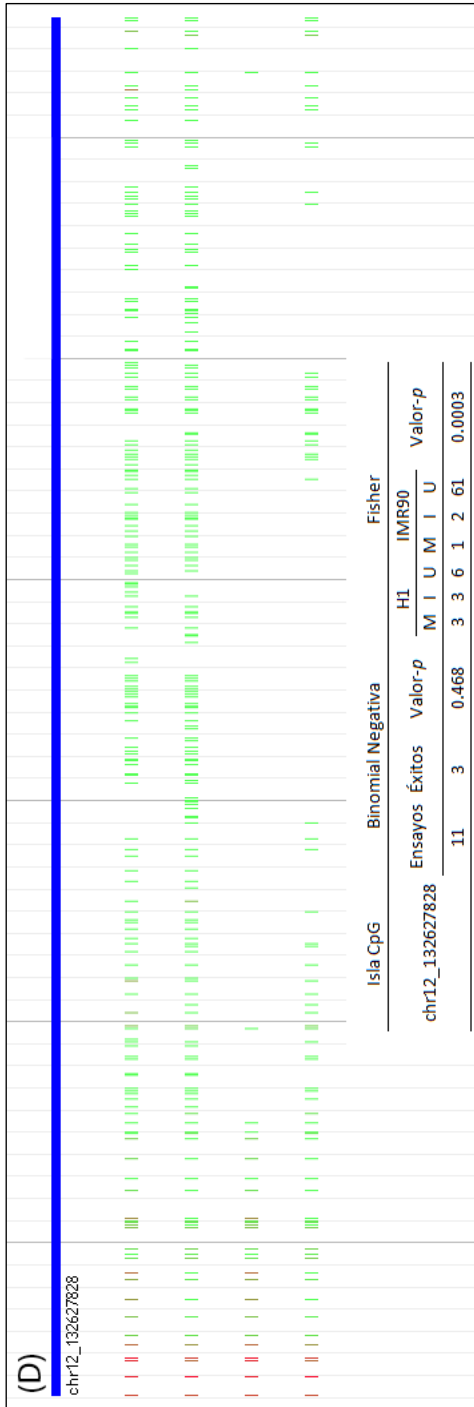


Figura 6.3. CGIs con metilación significativamente diferente en función del análisis estadístico utilizado. Las imágenes A y B muestran CGIs identificadas como diferencialmente metiladas por la binomial negativa y no por la prueba exacta de Fisher. Las imágenes C y D, muestran ejemplos del caso contrario. Las franjas azules representan las CGIs, las 2 primeras pistas muestran valores de metilación para todos los CpGs presentes en las islas, mientras que las 2 pistas inferiores representan CpGs con una profundidad mínima de 5 (profundidad mínima utilizada en los análisis estadísticos). Los tejidos incluidos en todos los casos son *h1* e *imr90* (ver apartado 6.1), donde *h1* es la primera pista representada e *imr90* la segunda en los pares descritos previamente. En las tablas bajo las imágenes se incluye el identificador de la isla (Isla CpG) y los parámetros utilizados en cada uno de los análisis estadísticos, así como el valor-*p* resultante en cada caso.

Al aplicar el análisis combinado en las *CGs* para los 66 pares de tejidos seleccionados, se han identificado un 6% de pares diferencialmente metilados. Además, la media de las diferencias absolutas de metilación observada en los pares con diferencias significativas en ambos análisis estadísticos es aproximadamente el doble que la de las comparaciones cuya significación difiere en función del análisis utilizado (Tabla 6.3).

| | Comparaciones | Comparaciones diferencialmente metiladas | Binomial Negativa (NS) | Fisher (NS) |
|--|-----------------|--|------------------------|-----------------|
| # | 8,818,209 | 547,169 | 176,814 | 130,654 |
| Diferencias de metilación ($X \pm SD$) | 0.09 \pm 0.16 | 0.59 \pm 0.25 | 0.29 \pm 0.2 | 0.23 \pm 0.16 |

Tabla 6.3. Diferencias de metilación por pares de tejidos. En la tabla se muestran: el número de comparaciones con datos suficientes para los 66 pares de tejidos analizados en las *CGs*, el número de comparaciones con un valor-*p* significativo para ambos análisis estadísticos y las comparaciones con diferencias significativas para uno de los análisis, pero no significativas para el otro. Además, para cada uno de estos casos se muestran las medias pesadas (X) y desviaciones estándar (SD) de las diferencias absolutas de metilación.

Se puede formular la hipótesis de que los tipos celulares semejantes o pertenecientes al mismo linaje celular podrían presentar perfiles de metilación similares. Una manera indirecta de comprobar la validez del método estadístico para identificar diferencias de metilación con sentido biológico, sería comprobar si los tipos celulares semejantes se agrupan en un árbol, cuando a modo de distancia se utiliza la fracción de *CGs* identificadas como diferencialmente metiladas entre los distintos pares de tejidos.

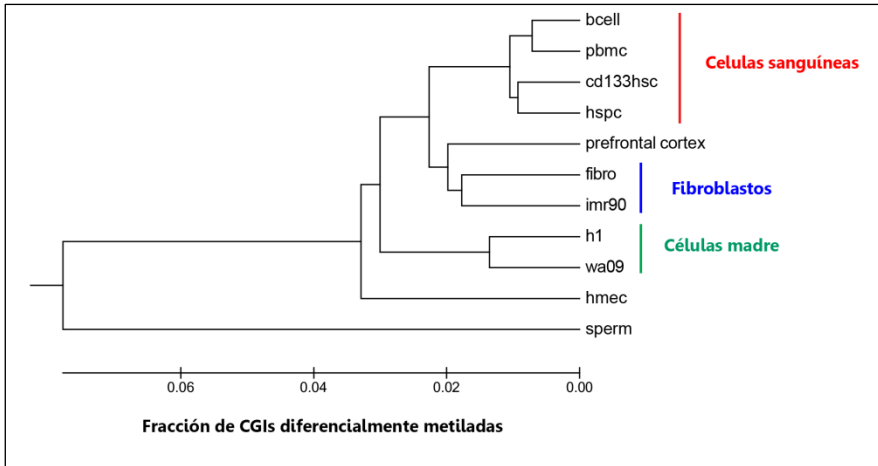


Figura 6.4. Árbol de distancias entre tipos celulares basado en la fracción de CGIs diferencialmente metiladas entre pares de tejidos. Se ha utilizado el programa *MEGA6* (Tamura, Stecher et al. 2013) y el método *UPGMA* (Nei and Kumar 2000).

Como podemos ver en la Figura 6.4, las diferencias de metilación agrupan los tipos celulares en 3 grupos: células madre (*h1* y *wa09*), células extraídas del torrente sanguíneo (*cd133hsc*, *pbmc* y *bcell*), todas ellas derivadas de células madre hematopoyéticas (*hspc*) y el resto de tejidos, entre los que se encuentran los fibroblastos (*imr90* y *fibro*). Por otro lado, el esperma aparece como el tipo celular con mayores diferencias de metilación con respecto al resto (véase apartado 6.4). Estos resultados revelan la asociación entre los perfiles de metilación de las CGIs y las semejanzas funcionales o de linaje de los diferentes tipos celulares, sugiriendo así que el método estadístico propuesto puede tener cierto sentido biológico.

6.3 CARACTERIZACIÓN DE LAS CLASES DE CGIs

Las CGIs incluidas en el conjunto relajado de *CpGcluster* (Hackenberg, Previti et al. 2006, Hackenberg, Carpena et al. 2011) se han clasificado en función de la metilación encontrada a lo largo de los diferentes tejidos (véase apartado 6.1.4): islas no metiladas (*UIs*), islas metiladas (*MIIs*) e islas diferencialmente metiladas (*DMIs*). Debido a que las muestras han sido tomadas de individuos de diferentes sexos, se han eliminado los cromosomas sexuales del estudio.

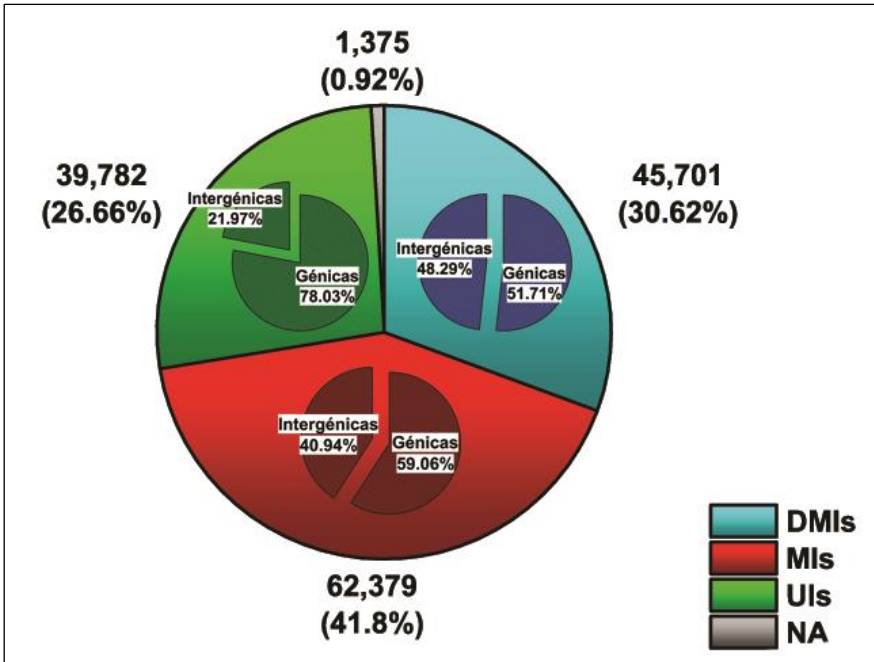


Figura 6.5. Clases de CGIs y porcentajes de solapamiento con los genes y con las regiones intergénicas. En la gráfica se muestra el número y porcentaje de islas CpG clasificadas (Apartado 6.1.4) como no metiladas (*UIs*), metiladas (*MIIs*), diferencialmente metiladas (*DMIs*) y no asignadas a ninguna de las clases (*NA*). En el interior de cada porción, otro gráfico circular muestra el solapamiento de cada clase con regiones génicas de *refSeq* (Pruitt, Tatusova et al.

2007), donde la región génica se ha definido como el cuerpo génico \pm 500 pb de entorno génico (Apartado 6.1.5).

A diferencia del conjunto estricto de *CpGcluster*, que presenta un 72% de *UIs* y solamente 5,517 *DMIs*, el conjunto relajado revela un total de 45,701 islas diferencialmente metiladas (Figura 6.5). Lo más destacado de esta clasificación, es el gran número de *CGIs* metiladas en todos los tejidos (42%) y su solapamiento con regiones génicas (casi de un 60%). Aunque pueda parecer contraintuitivo, la metilación de los cuerpos génicos se ha asociado con genes transcripcionalmente activos (Hellman and Chess 2007), lo que explicaría la presencia de un gran porcentaje de *CGIs* metiladas asociadas a las regiones génicas. Además, no es de extrañar el elevado porcentaje de solapamiento de *UIs* con las regiones génicas, ya que la totalidad de los genes domésticos presentan *CGIs* asociadas a sus promotores, debiendo encontrarse no metiladas para permitir su transcripción. En cuanto a las *DMIs* identificadas, solapan casi por igual con las regiones génicas e intergénicas. Esto podría deberse a las funciones relacionadas con la metilación diferencial, tanto fuera (asociada al control de la interacción de elementos reguladores con el ADN (Wiench, John et al. 2011, Hon, Rajagopal et al. 2013, Ziller, Gu et al. 2013)) como dentro de las regiones génicas (asociada al *splicing* alternativo (Maunakea, Nagarajan et al. 2010, Shukla, Kavak et al. 2011, Oberdoerffer 2012)).

6.3.1 Características composicionales

Aunque las *MI*s y *UI*s, presentan una metilación uniforme a largo de los tejidos, las *DMI*s pueden presentar una amplia variedad de combinaciones: desde estar metiladas en un tejido y en el resto no, a la situación inversa, donde todos los tejidos salvo uno se encontrarían metilados.

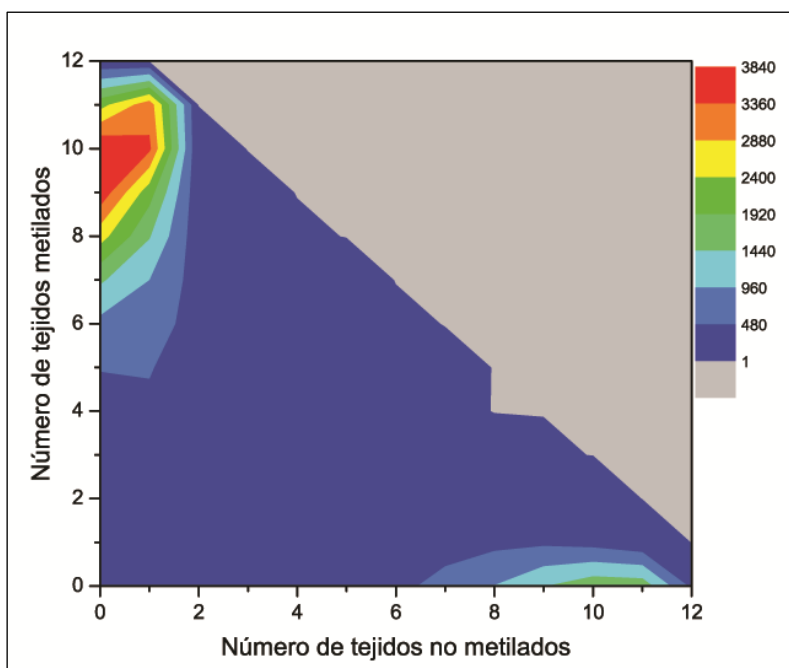


Figura 6.6. Número de tejidos metilados y no metilados en las *DMI*s. En el eje de abscisas se representan el número de tejidos no metilados y en el de ordenadas los metilados. Los colores simbolizan el número de *DMI*s, tal y como se muestra en la leyenda situada a la derecha de la figura.

En la Figura 6.6, se representa el número de *DMI*s según el número de tejidos metilados y no metilados. Se observa que el 66% de las *DMI*s pueden clasificarse en dos grupos muy diferentes en función del número

de tejidos metilados: a) *DMIs* principalmente metiladas (*DMIs* con al menos 8 tejidos metilados y como máximo 2 tejidos no metilados, *DMIs-M*), formando estas el grupo más numeroso (50% del total de *DMIs*); y b) *DMIs* principalmente no metiladas (*DMIs* con al menos 8 tejidos no metilados y como máximo 1 tejido metilado, *DMIs-U*). Estos resultados concuerdan en lo esencial con lo descrito en un estudio reciente sobre regiones diferencialmente metiladas, donde un elevado porcentaje de las *DMRs* se encuentran metiladas en la mayoría de los tejidos incluidos en el estudio (Ziller, Gu et al. 2013).

| | # | Longitud | Ratio [O/E] | GC | Densidad |
|---------------|--------|-----------------|-------------|-------------|-------------|
| UIs | 39,782 | 317.03 ± 272.69 | 0.95 ± 0.19 | 0.7 ± 0.07 | 0.11 ± 0.03 |
| DMIs | 45,701 | 210.5 ± 185.9 | 0.89 ± 0.23 | 0.66 ± 0.06 | 0.09 ± 0.03 |
| MIIs | 62,379 | 156.43 ± 91.66 | 0.92 ± 0.25 | 0.64 ± 0.07 | 0.09 ± 0.04 |
| DMIs-M | 22,925 | 192.45 ± 160.3 | 0.89 ± 0.23 | 0.65 ± 0.07 | 0.09 ± 0.03 |
| DMIs-U | 7,444 | 291.31 ± 264.9 | 0.9 ± 0.19 | 0.68 ± 0.06 | 0.10 ± 0.03 |

Tabla 6.4. Estadística básica de la composición de las clases de CGIs. En la tabla se muestran la media y la desviación estándar de las características composicionales de las clases de CGIs descritas (Apartado 6.1.4). Los valores que se muestran son: el número de CGIs de cada clase (#), su longitud, la proporción de CpGs observados/esperados (*Ratio [O/E]*), la fracción de G+C (GC) y la densidad de CpGs (medida como el número de CpGs dividido por la longitud de la isla).

En cuanto a sus características composicionales, las diferentes clases de CGIs presentan cierta variación entre ellas (Tabla 6.4). Generalmente, las *UIs* presentan los valores más elevados en todas las propiedades analizadas. La elevada longitud de estas CGIs puede deberse a su asociación con los promotores, donde se une específicamente el complejo de la ARN polimerasa (uno de los complejos de interacción con el ADN más grandes que se conocen), y por lo tanto,

debe existir una región libre de impedimentos estéricos lo suficientemente grande como para permitir su unión. Las diferencias en el resto de parámetros, probablemente se deban a la ausencia del sesgo mutacional causado por la metilación (Bird 1980).

Las *DMIs* presentan longitudes intermedias entre *UIs* y *MIIs*, lo que claramente se debe a la existencia de las dos clases de *DMIs* descritas previamente. Efectivamente, al determinar las longitudes de estas clases por separado, las *DMIs-U* presentan longitudes cercanas a las *UIs*, mientras que las *DMIs-M* presentan una mayor semejanza con las *MIIs*. Al igual que en el caso de las *UIs*, probablemente esta relación también se deba a la asociación diferencial de ambas clases de *DMIs* con las regiones promotoras (Figura 6.7). Como ocurre con la longitud, la densidad de CpGs también presenta una relación inversa con el número de tejidos metilados. Sin embargo, esta relación no se observa en la *ratio* $[O/E]$, donde las *DMIs* presentan *ratios* menores que el resto de clases.

Los resultados composicionales parecen indicar que a pesar de la existencia de otros mecanismos evolutivos que puedan explicar la existencia de *CGIs* en el genoma (Cohen, Kenigsberg et al. 2011), la ausencia del sesgo mutacional causado por la metilación de las citosinas (Bird 1980) parece ser el mecanismo que actúa con mayor fuerza sobre las *CGIs*. De esta manera, las *CGIs* con un mayor número de tejidos metilados tenderán a presentar tanto longitudes como densidades inferiores de CpGs, debido probablemente también a una mayor probabilidad de encontrarse metiladas en la línea germinal.

6.3.2 Enriquecimiento en elementos reguladores

En el apartado anterior se ha definido un conjunto de regiones con una alta densidad de CpGs, compuesto por diferentes clases en función de su estado de metilación en múltiples tejidos (*UIs*, *MIs* y *DMIs*). En este apartado estudiaremos su co-localización con elementos reguladores, con especial atención a las *DMIs*. Este estudio se ha realizado mediante el cálculo de índices de enriquecimiento (Hackenberg, Rueda et al. 2012) de las clases de *CGIs* en diferentes elementos genómicos. Estos índices expresan la fracción de *CGIs* que solapan con el elemento estudiado dividida por la fracción fuera del elemento.

6.3.2.1 Regiones génicas

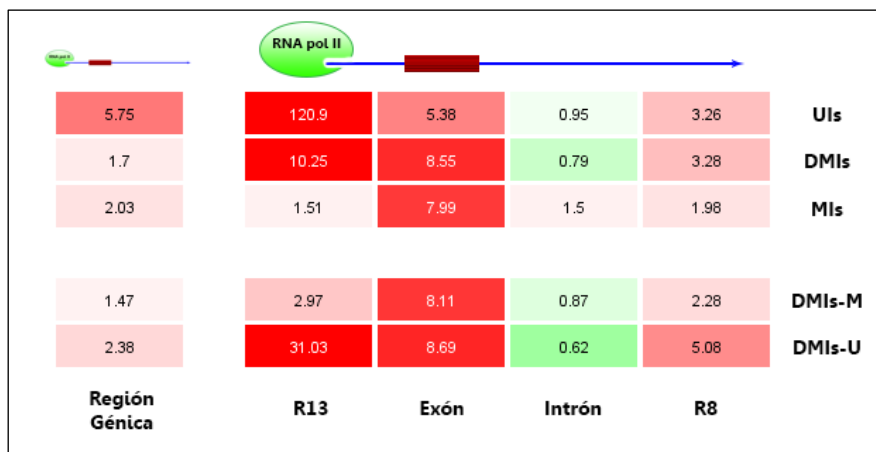


Figura 6.7. Enriquecimiento de CGIs en regiones génicas. En la tabla se muestran los índices de enriquecimiento de las clases de CGIs (filas) para diferentes regiones génicas (columnas). De izquierda a derecha: región génica, región promotora R13, exones, intrones y una región en torno al final de la transcripción etiquetada como R8 (estas regiones se definen en el apartado 6.1.5). La escala de color oscila entre 0 y 10, el color verde (0 a <1) simboliza empobrecimiento en el elemento estudiado y el color rojo (>1 a 10) enriquecimiento. Los colores próximos al blanco (1) serían casos de distribución aleatoria con respecto al elemento estudiado. Todos los valores de enriquecimientos/empobrecimiento presentan valores-*p* por debajo de 0.01.

En primer lugar, se ha estudiado el enriquecimiento de las CGIs en diferentes regiones génicas (definidas en el apartado 6.1.5). Como se observa en la Figura 6.7, las CGIs en general se asocian preferencialmente con las regiones promotoras y con los exones de los genes codificantes estudiados.

Regiones promotoras

Las UIs se encuentran altamente enriquecidas en las regiones promotoras, lo que concuerda con su papel en la iniciación de la transcripción de los genes asociados. A su vez, las DMIs también

presentan un alto enriquecimiento en los promotores, donde podrían estar relacionadas con la inhibición de la transcripción en los tejidos donde se encuentran hipermetiladas. Entre las dos clases de *DMIs* también existen diferencias de enriquecimiento, siendo las *DMIs-U* las que presentan un mayor enriquecimiento en estos elementos (inhibición en pocos tejidos), mientras que las *DMIs-M* mediarían en la regulación de unos pocos genes tejido-específicos. Las *MIs* presentan enriquecimientos inferiores al resto de clases, distribuyéndose casi de manera aleatoria fuera y dentro de los promotores. Probablemente, las *MIs* se asocien con genes tejido-específicos cuyos tejidos no han sido incluidos en este estudio, como veremos en el apartado 6.4.

Exones e intrones

El elevado enriquecimiento de *CGIs* en exones y su empobrecimiento en intrones, concuerda con el elevado contenido en GC de los exones con respecto a los intrones (Schwartz, Meshorer et al. 2009). Las *CGIs* asociadas principalmente a exones suelen presentar niveles de metilación elevados en la mayoría de los tejidos (*MIs* y *DMIs*). Este elevado enriquecimiento de *MIs* en exones no es sorprendente, ya que como se ha comentado, la presencia de regiones hipermetiladas en el cuerpo génico se asocia con altos niveles de transcripción (Hellman and Chess 2007); así pues, esta clase de *CGIs* podría cumplir una función importante estabilizando dicha transcripción. Por otro lado, recientemente se ha demostrado que la metilación diferencial en exones puede regular mecanismos de *splicing* alternativo (Shukla, Kavak et al. 2011), función en la que podrían estar interviniendo las *DMIs* presentes en estas regiones. También puede observarse un elevado enriquecimiento de

DMIs-U, e incluso de *UIs*, en exones. Estas *CGIs* podrían estar asociadas a inicios alternativos de la transcripción, tal y como se ha observado recientemente (Hackenberg, Barturen et al. 2010, Maunakea, Nagarajan et al. 2010).

Con respecto a los intrones, es interesante subrayar las diferencias de asociación entre las *DMIs* (empobrecidas) y las *DMRs* caracterizadas por Ziller y colaboradores, que presentan casi un 40% de solapamiento (Ziller, Gu et al. 2013). En el caso de los exones ocurre lo contrario, con los que solapan sólo un 5% de las *DMRs*, mientras que las *DMIs* se encuentran enriquecidas. Estas diferencias de asociación indican ciertas diferencias funcionales entre ambos conjuntos, además, como ya se ha comentado en este capítulo, teniendo en cuenta que la posible función de la metilación suele ir ligada a una elevada densidad de CpGs, muchas de las *DMRs* podrían no tener una función activa regulando la interacción de otras moléculas con el ADN. De hecho, por el momento sólo se ha demostrado de manera directa la función de diferencias de metilación en los exones (Shukla, Kavak et al. 2011), quedando aún por asignar una posible función a las diferencias en intrones.

Regiones de terminación de la transcripción

En cuanto a la región R8 (final de la transcripción), en los últimos años se ha demostrado la existencia de transcritos de ARNs anti-sentido en estas regiones (Jacquier 2009), que podrían estar reguladas por niveles diferenciales de metilación. Esto podría explicar el enriquecimiento (aunque no demasiado elevado) de *DMIs* en estas regiones.

6.3.2.2 Sitios de interacción con el ADN

Además de las regiones génicas, también se ha estudiado la asociación de las *CGIs* con otros elementos reguladores. En la Figura 6.8, se ha analizado su enriquecimiento en los *clusters* de sitios de unión a factores de transcripción (*TFBSs*) y sitios de hipersensibilidad a la DNasa1 definidos en el proyecto *ENCODE* (Consortium, Bernstein et al. 2012). Además, se han seleccionado algunos subconjuntos a partir de los datos de *ENCODE* donde la metilación diferencial puede jugar un papel importante. En general, las *UIs* y *DIMs-U* presentan elevados índices de enriquecimiento en todos los sitios de unión analizados.






| | Clusters TFBSs | | Clusters DNasa | | RNA pol II | | RNA pol III | | Potenciadores | | Aisladores | |
|---|----------------|-------|----------------|-------|------------|-------|-------------|-------|---------------|--------|------------|--|
| | ENCODE | | ENCODE | | ENCODE | | ENCODE | | VISTA | ENCODE | ENCODE | |
|  | 495.19 | 71.66 | 108 | 48.26 | 1.08 | 48.26 | 0.71 | 20.37 | 9.56 | 78.36 | Uis | |
|  | 7.07 | 9.17 | 8.85 | 5.05 | 8.85 | 5.05 | 1.69 | 3.48 | 4.72 | 18 | DMIs | |
|  | 0.92 | 1.44 | 1.61 | 0.0 | 1.61 | 0.0 | 0.31 | 0.26 | 0.59 | 2.82 | MIs | |
|  | 2.09 | 3.77 | 2.96 | 0.61 | 2.96 | 0.61 | 0.78 | 0.92 | 1.74 | 5.86 | DMIs-M | |
|  | 120.25 | 57.13 | 25.17 | 5.01 | 25.17 | 5.01 | 3.46 | 7.67 | 9.8 | 40.64 | DMIs-U | |

Figura 6.8. Enriquecimiento de CGIs en distintos elementos reguladores. En la tabla se muestran los índices de enriquecimiento de las clases de CGIs (filas) para diversos elementos reguladores (columnas). De izquierda a derecha: *clusters* de sitios de unión a factores de transcripción, *clusters* de sitios de hipersensibilidad a la DNasa, dos conjuntos de *clusters* de subunidades de la ARN polimerasa II y III, potenciadores de la base de datos VISTA (Visel, Minovitsky et al. 2007), potenciadores potencialmente activos (Zentner, Tesar et al. 2011), sitios de unión del factor de transcripción CTCF fuera de regiones génicas y sitios de unión del factor de transcripción CTCF solapantes con regiones exónicas. Los datos, salvo los potenciadores VISTA, se han extraído del proyecto ENCODE (Consortium, Bernstein et al. 2012). La escala de color oscila entre 0 y 10: el color verde (0 a <1) simboliza empobrecimiento de la clase de CGI en el elemento estudiado y el color rojo (>1 a 10) enriquecimiento. Los colores próximos al blanco (1) señalan casos de distribución aleatoria con respecto al elemento estudiado. Los valores de enriquecimiento/empobrecimiento con valores-*p* superiores a 0.01 se marcan con un asterisco. La descripción de los conjuntos de datos utilizados puede encontrarse en el apartado 6.1.5.

Clusters de ENCODE

Los resultados obtenidos para los *clusters* de sitios de unión a factores de transcripción y sitios de hipersensibilidad a la DNase muestran resultados semejantes. Ambos conjuntos son regiones donde se ha identificado alguna interacción con el ADN, es decir son regiones abiertas de la cromatina, y por lo tanto en los tejidos donde se produzca dicha interacción las regiones deberán encontrarse libres de metilación. En general, el enriquecimiento en estas de regiones de *UIs* y *DMIs-U* es muy superior al de *CGIs* con un mayor número de tejidos metilados (*DMIs-M*). Además, prácticamente el 100% de las *UIs* y un 97% de las *DMIs-U* solapan con alguno de estos elementos, mientras que sólo un 30% de las *DMIs-M* y un 9% de las *MIIs* lo hacen. Es destacable, que a diferencia del conjunto de *DMRs*, donde el 60% de los *TFBSs* analizados solapan con alguna región principalmente metilada (Ziller, Gu et al. 2013), las *DMIs-M* sólo cubren un 1% del total de *TFBSs* del conjunto de *ENCODE*.

Subunidades de ARN polimerasas

Para este análisis, se han seleccionado los *clusters* definidos por *ENCODE* para diferentes subunidades de las ARN polimerasas II y III. Los resultados observados para la ARN polimerasa II presentan valores de enriquecimiento semejantes a los análisis realizados para los promotores de *refSeq* (Pruitt, Tatusova et al. 2007), ya que esta polimerasa se encarga de transcribir todos los mensajeros que se traducirán a proteínas (genes *refSeq* utilizados en la Figura 6.7). Los sitios de unión para las subunidades de la polimerasa III presentan un empobrecimiento sorprendente para las *MIIs*, no solapando ninguno de los *clusters* analizados con esta clase de *CGIs*. La ARN polimerasa III

transcribe ARNs no codificantes implicados en procesos fundamentales para la célula, como la síntesis de proteínas, el procesado del ARN, la propia transcripción o la regulación de la cromatina (Canella, Praz et al. 2010). Además, la desregulación de la transcripción de estos ARNs no codificantes se encuentra íntimamente relacionada con diversas enfermedades como el cáncer (White 2008). Por lo tanto, debido a su importante función en prácticamente todos los tipos celulares, no es de extrañar la ausencia de *MIs* y el empobrecimiento de *DMIs-M* en este tipo de promotores.

Potenciadores

Para observar la relación de las *CGIs* con los potenciadores, se han analizado las entradas de la base de datos *VISTA* (Visel, Minovitsky et al. 2007), que son secuencias potenciadoras determinadas experimentalmente *in vivo*. Sin embargo, esta base de datos sólo contiene 1,065 elementos con actividad potenciadora demostrada, por lo que para algunas clases de *CGIs* los enriquecimientos obtenidos no presentan sustento estadístico. Así que para verificar los resultados obtenidos con el conjunto de *VISTA*, se ha seleccionado un conjunto de potenciadores potencialmente activos en función de las modificaciones de histonas (Zentner, Tesar et al. 2011) y de los sitios de unión a factores de transcripción (*TFs*) que se unen a regiones potenciadoras (véase apartado 6.1.5). En ambos conjuntos, las *DMIs-U* aparecen enriquecidas en estos elementos reguladores y se confirma el empobrecimiento encontrado para las *MIs* y *DMIs-M*.

Aisladores

Como ya se comentó en el apartado 6.1.5, dada la actividad diferencial de CTCF en función de su localización, se ha comparado la presencia de CGIs asociadas a los *clusters* de CTCF solapantes con exones (regulación del *splicing* alternativo) con aquellos que se sitúan fuera de las regiones génicas (posible unión a secuencias aisladoras). Mientras que los índices de enriquecimiento fuera de las regiones génicas se comportan de manera semejante al resto de *clusters* de TFBSs, los sitios de unión a CTCFs solapantes con exones presentan un comportamiento diferente. Los *clusters* solapantes con exones presentan un enriquecimiento en DMIs-M y MIs mayor que para el resto de elementos analizados (Figura 6.8), en consonancia con lo observado para el enriquecimiento de exones en general (Figura 6.7) y con los niveles de metilación esperados en el cuerpo génico.

6.3.3 Conservación y variaciones de secuencia

Por último, se ha analizado la conservación evolutiva de las *CGI*s observando su enriquecimiento en dos conjuntos de elementos conservados. En este análisis, también se han incluido otros elementos asociados con la evolución y la conservación de secuencia, como son: regiones que presentan un sesgo mutacional hacia GC y variaciones de secuencia, tanto poblacionales como asociadas a enfermedades.

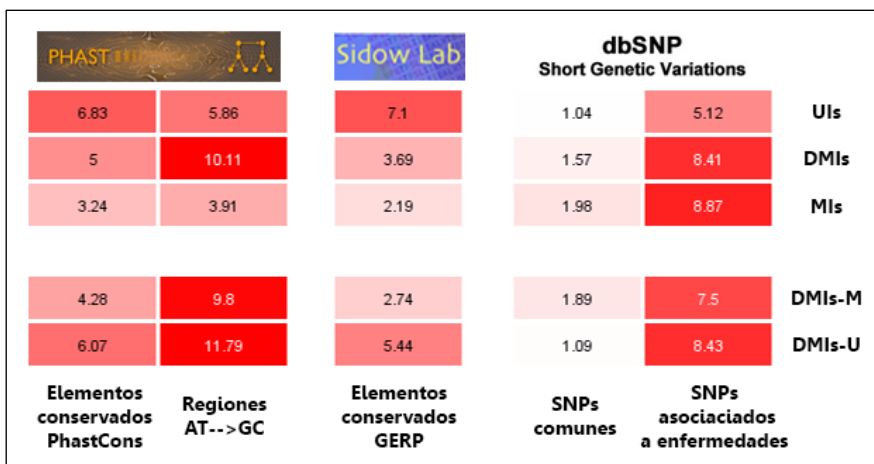


Figura 6.9. Enriquecimiento de *CGI*s en elementos conservados y variaciones de un solo nucleótido. En la tabla se muestran los índices de enriquecimiento de las clases de *CGI*s (filas) para diferentes regiones de interés evolutivo y variaciones de un solo nucleótido (columnas). De izquierda a derecha se muestran elementos conservados *phastCons* (Siepel, Bejerano et al. 2005), regiones sometidas a un sesgo mutacional hacia GC (Duret and Galtier 2009, Capra, Hubisz et al. 2013), elementos conservados *GERP* (Davydov, Goode et al. 2010), polimorfismos de un solo nucleótido (*SNPs*) y variaciones de un solo nucleótido asociadas a enfermedades (Sherry, Ward et al. 2001). La escala de color oscila entre 0 y 10: el color verde (0 a <1) simboliza empobrecimiento de la clase de *CGI* en el elemento estudiado y el color rojo (>1 a 10) enriquecimiento. Los colores próximos al blanco (1) serían casos de distribución aleatoria con respecto al elemento estudiado. Los valores de enriquecimiento/empobrecimiento con valores-*p*

superiores a 0.01 se marcan con un asterisco. Para una descripción de los distintos elementos ver apartado 6.1.5.

Elementos conservados

Se han utilizado dos conjuntos de regiones evolutivamente conservadas: *phastCons* (Siepel, Bejerano et al. 2005) y *GERP* (Davydov, Goode et al. 2010). Ambos conjuntos presentan resultados semejantes (Figura 6.9). Aunque todas las clases de *CGIs* presentan enriquecimientos considerables con los elementos conservados, se observa una tendencia a un mayor enriquecimiento con estos elementos cuanto menor sea el porcentaje de tejidos metilados. Estos resultados refuerzan la hipótesis propuesta al comentar los datos composicionales. Es decir, a pesar de la existencia de otras fuerzas evolutivas que intervienen en la conservación de estas regiones (Cohen, Kenigsberg et al. 2011), la ausencia del sesgo mutacional causado por la metilación parece ser la fuerza dominante (véase apartado 6.3.1).

Regiones con sesgo mutacional hacia GC

Las regiones en el genoma humano con sesgo mutacional hacia GC utilizadas se tomaron de las tablas de la UCSC (<http://genome.ucsc.edu>), generadas con el programa *phastBias* incluido en el paquete *PHAST* (Hubisz, Pollard et al. 2011). Todas las clases de *CGIs* se encuentran enriquecidas en estas regiones (Figura 6.9). Este sesgo mutacional podría estar provocado por un mecanismo de conversión génica, como se ha descrito para *MIs* (Cohen, Kenigsberg et al. 2011). Sin embargo, en el caso de las *DMI*s su enriquecimiento es excepcionalmente alto comparado con el de las *MIs*, que presentan un enriquecimiento

semejante en exones (Figura 6.7). Este sesgo hacia GC también se observa indirectamente en las diferencias composicionales encontradas entre *MIs* y *DMIs* (Tabla 6.4). Recientemente se ha propuesto un mecanismo epigenético capaz de mantener bajos niveles de metilación en la línea germinal en las regiones diferencialmente metiladas, de tal forma que las mutaciones espontáneas causadas por la metilación no se hereden de generación en generación (Cohen, Kenigsberg et al. 2011), y por lo tanto mantengan elevadas densidades de CpGs. A falta de incluir en el estudio datos de metilación de la línea germinal femenina, se ha observado que menos del 50% de las *DMIs* identificadas presentan estados de metilación U o T en el esperma. Estos resultados parecen indicar que además del mecanismo propuesto por Cohen y colaboradores, podría existir algún otro mecanismo que indirectamente mantenga en equilibrio la composición de CpGs en estas regiones mediante un sesgo mutacional hacia GC.

Variaciones de un solo nucleótido

Las variantes poblacionales suelen encontrarse en regiones poco conservadas del genoma (Castle 2011). Además, las tasas más altas de *SNPs* se han encontrado asociadas a regiones hipermetiladas (Qu, Hashimoto et al. 2012), lo que concuerda con los enriquecimientos observados en elementos conservados, ya que las *MIs* y *UIs* son las clases de *CGIs* con menor y mayor conservación respectivamente (aunque ambas presentan un elevado enriquecimiento en estos elementos conservados). Para comprobar esta asociación, se ha estudiado el enriquecimiento de *SNPs* comunes (frecuencia del alelo menor $\geq 1\%$ en la población, *dbSNP Common*) presentes en la base de

datos *dbSNP* (Sherry, Ward et al. 2001). Además, también se ha estudiado el enriquecimiento de otro subconjunto de variaciones de *dbSNP* (*dbSNP Flagged*) asociadas a *loci* incluidos en *LSDB* (Horaitis, Talbot et al. 2007) o en *OMIM* (Hamosh, Scott et al. 2005). Como era de esperar, se observan niveles de enriquecimiento marginales para los *SNPs* comunes en todas las clases de *CGIs*, mientras que el enriquecimiento de variaciones con asociación clínica es elevado en todas ellas. La reducida asociación de *SNPs* comunes en las *CGIs*, junto con el elevado enriquecimiento de variaciones asociadas a enfermedades, revela una gran importancia de estas regiones en el correcto funcionamiento celular.

6.4 FUNCIONES REGULADAS POR LAS *CGIs*

La metilación diferencial en los genes suele encontrarse asociada a las regiones promotoras y a los exones (apartado 6.3.2). En el caso de los exones, ya se ha comentado que las *MIs* podrían actuar como estabilizadoras de la transcripción, mientras que las *DMIs* podrían estar regulando procesos de *splicing* alternativo. La función de las *CGIs* en los promotores está bien caracterizada, permitiendo o inhibiendo la unión del complejo regulador de la ARN polimerasa (Bell, Pai et al. 2011). Aunque la función principal de la metilación se centra en la regulación de las interacciones de diferentes moléculas con el ADN, estas interacciones pueden regular diferentes funciones o procesos según los genes o *TFBSs* a los que se asocien. Aprovechando que la presencia de *CGIs* en las regiones promotoras se encuentra directamente relacionada con la

regulación de la transcripción, se han estudiado las funciones y procesos regulados por las diferentes clases de *CGIs* a partir del enriquecimiento en términos *GO* (*Gene Ontology* (Ashburner, Ball et al. 2000)) de los promotores con los que solapan. Los enriquecimientos en términos *GO* se han calculado comparando genes que presentan una sola clase de *CGI* frente al resto de los genes con *CGIs* en su región promotora. Estos cálculos se han realizado con la herramienta *GOrilla* (Eden, Navon et al. 2009).

Islas CpG constitutivamente no metiladas (Uls)

Las *Uls* localizadas en las regiones promotoras siempre se han asociado con genes domésticos, necesarios y ampliamente expresados en todos los tipos celulares. Los genes asociados con esta clase de *CGIs* son ricos en términos *GO*, encontrándose relacionados con 143 procesos biológicos y 12 funciones moleculares con un valor-*p* por debajo de $1e-5$. Los términos *GO* enriquecidos para las *Uls* hacen referencia a funciones indispensables en cualquier tipo celular: procesos metabólicos y catabólicos relacionados con el ARN o las proteínas, control del ciclo celular, transporte de moléculas, proteínas relacionadas con el *splicing* alternativo, etc... (Tabla suplementaria 6.1 y Tabla suplementaria 6.2).

Estos resultados concuerdan con la descripción clásica de *CGIs* no metiladas, generalmente asociadas a los promotores de genes domésticos.

Islas CpG constitutivamente metiladas (MIs)

En el caso de los genes asociados con *MIs*, para valores-*p* menores de $1e-5$ sólo encontramos 1 proceso biológico enriquecido (Tabla suplementaria 6.3) relacionado con la actividad de receptores olfativos. Si se relaja la significación estadística a $1e-3$, aparecen nuevos términos muy heterogéneos (cohesión de cromátidas hermanas, biosíntesis de cuerpos cetónicos o coagulación sanguínea), que en ningún caso parecen términos asociados a los tipos celulares o estadios celulares analizados en este estudio (al igual que la actividad de receptores olfativos).

Por lo tanto, las *MIs* parecen encontrarse asociadas a genes no necesarios en los tipos celulares estudiados. Es de subrayar, el hecho de que estas *CGIs* identificadas como *MIs* probablemente se clasificarían dentro de la clase *DMIs-M* si se incluyeran los tipos celulares oportunos.

Islas CpG principalmente metiladas (DMIs-M)

Al igual que las *MIs* (que como se ha comentado, probablemente acabarían incluyéndose en esta clase si se introdujeran nuevos tipos celulares), las *DMIs-M* presentan pocos términos *GO* asociados. En este caso, la totalidad de los procesos biológicos enriquecidos se encuentran asociados a las funciones básicas de los espermatozoides: reproducción, meiosis y genes implicados en el procesado de pi-ARNs (Tabla suplementaria 6.4). En relación con esto último, se ha observado que la metilación aberrante en los promotores de los genes que se asocian con el procesado de pi-ARNs provoca infertilidad en ratones (Heyn, Ferreira et al. 2012). Los espermatozoides presentan la mayor fracción de *CGIs*

diferencialmente metiladas con respecto al resto de tejidos analizados (Figura 6.4), además de ser el tipo celular responsable del 26% de las *DMIs-M* identificadas (Figura 6.10). Estas diferencias con respecto al resto de tejidos analizados explicarían que los términos enriquecidos para esta clase de *DMIs* hagan referencia a funciones específicas de los espermatozoides.

Esta clase de *DMIs*, al igual que se ha observado para *DMRs* principalmente metiladas (Ziller, Gu et al. 2013), parece implicada en la regulación de funciones tejido-específicas. Es decir, en la mayoría de tejidos las regiones promotoras de estos genes tejido-específicos se encontrarían metiladas, desmetilándose de manera específica en aquellos tejidos donde sean necesarios los productos resultantes de su transcripción.

Islas CpG principalmente no metiladas (DMIs-U)

Las *CGIs* principalmente no metiladas presentan multitud de términos enriquecidos: 71 procesos biológicos y 15 funciones moleculares (Tabla suplementaria 6.5 y Tabla suplementaria 6.6). En general, la mayoría de los términos encontrados son factores de transcripción o proteínas implicadas en la señalización celular, que regulan procesos asociados con la diferenciación celular o el desarrollo del organismo. La fracción de tejidos con diferencias de metilación significativas dentro de esta clase de *DMIs* (Figura 6.10) muestra que un gran número de las *DMIs-U* se deben a diferencias entre tejidos diferenciados y no diferenciados (*h1* y *wa09*), ya que el porcentaje de *DMIs-U* debidas a estos últimos es del 25%.

Observando estos resultados, podemos concluir que un gran número de las *DMIs-U* parecen implicadas en procesos de desarrollo y diferenciación. De esta manera, los genes implicados en el desarrollo o diferenciación de ciertos tejidos presentarán niveles bajos de metilación en sus promotores de manera específica. Los términos *GO* enriquecidos son coherentes con un mayor porcentaje de *DMIs-U* causado por una hipermetilación de los promotores de dichos genes en células madre, ya que estas no se encuentran diferenciadas y por lo tanto no presentarán transcripción activa de genes implicados en la diferenciación o el desarrollo.

| | | | | | | | | | | | |
|-------|------|----------|----------|-------|-------|-------|-------|-------|-------|-------|---------------|
| SPERM | HMEC | P.CORTEX | WA09 | FIBRO | BCELL | H1 | CD133 | IMR90 | HSPC | PBMC | |
| 0.26 | 0.10 | 0.088 | 0.078 | 0.076 | 0.062 | 0.062 | 0.061 | 0.059 | 0.059 | 0.027 | DMIs-M |
| 0.14 | 0.13 | 0.11 | 0.099 | 0.087 | 0.074 | 0.073 | 0.066 | 0.061 | 0.056 | 0.043 | DMIs-U |
| H1 | HMEC | WA09 | P.CORTEX | FIBRO | IMR90 | BCELL | CD133 | HSPC | SPERM | PBMC | |

Figura 6.10. Fracción de pares de tejidos diferencialmente metilados para cada tejido y clase de *DMI*. Para cada una de las clases mayoritarias de *DMIs* (*DMIs-M*, en azul y *DMIs-U*, en rojo) se muestra la fracción de pares de tejidos identificados como diferencialmente metilados en el que se encuentra cada tipo celular.

6.5 DISCUSIÓN

En este último capítulo de la tesis se ha presentado un estudio de metilación diferencial centrado en regiones con una alta densidad de CpGs (*CGIs*).

En primer lugar, se han comparado diferentes métodos estadísticos, tratando de encontrar un método que permita determinar diferencias de metilación entre regiones previamente definidas. El método estadístico desarrollado, junto con el algoritmo *CpGcluster* mejorado, ha permitido identificar y caracterizar diferentes clases de *CGIs* a partir de su metilación en 11 tipos celulares sanos, entre las que destacan las *DMIs*.

A diferencia de las regiones diferencialmente metiladas (*DMRs*) identificadas en otros estudios (Ziller, Gu et al. 2013), las *DMIs* presentan una alta densidad de CpGs, que es una característica esencial de las regiones determinantes de la metilación (*MDRs*) (Lienert, Wirbelauer et al. 2011). Además, los resultados funcionales obtenidos justifican el uso de algoritmos (como *CpGcluster*) que preseleccionen regiones con elevadas densidades de CpGs antes de afrontar estudios de metilación diferencial. Las diferencias más importantes encontradas entre *DMIs* y *DMRs* se resumen a continuación:

- (1) Las *DMIs* presentan un elevado enriquecimiento tanto en promotores (con un solapamiento del 15%) como en exones (20% de solapamiento), mientras que el porcentaje de solapamiento de las *DMRs* con estos elementos no supera en ninguno de estos casos el 5%.

- (2) Las *DMRs* presentan un solapamiento muy elevado con regiones intrónicas (aproximadamente un 40%); sin embargo las *DMIs* se encuentran empobrecidas en estos elementos.
- (3) Otra diferencia importante entre las *DMRs* y las *DMIs* se centra en el tipo de *DMIs* asociadas a los *TFBSs*. Mientras que el 60% de los *TFBSs* de *ENCODE* se asocian a *DMRs* principalmente metiladas, las *DMIs-M* sólo cubren un 1% de los *clusters* analizados en este estudio. Además, sólo el 30% de las *DMIs-M*, en comparación con el 93% de las *DMIs-U*, solapan con algún *TFBSs*.

También se ha observado cierta relación entre el número de tejidos no metilados que presentan las *CGIs* y su enriquecimiento en elementos conservados. Esta relación podría deberse a la simple probabilidad de encontrarse no metiladas en la línea germinal o a una mayor conservación de la función que desempeña cada clase de *CGI*. Además, el enriquecimiento marginal en *SNPs*, junto con el elevado enriquecimiento en variaciones asociadas a enfermedades encontrado en todas las clases de *CGIs*, sugiere que las *CGIs* predichas por *CpGcluster* podrían utilizarse como epimarcadores de algunas patologías.

Por último, los estudios funcionales mediante términos *GO*, han puesto de manifiesto, por un lado, la asociación de las *DMIs-M* (como posiblemente también las *MIs*) con funciones tejido-específicas (tal y como ya se había observado para las *DMRs* principalmente metiladas), y, por otro, han desvelado la implicación de las *DMIs-U* en funciones relacionadas con la diferenciación celular y el desarrollo del organismo.

6.6 MATERIAL SUPLEMENTARIO

| <i>UIs</i> PROCESOS BIOLÓGICOS | | | | |
|-----------------------------------|--|----------------|----------------|----------|
| Término GO | Descripción | Valor-p | Valor-q | E |
| GO:0000956 | nuclear-transcribed mRNA catabolic process | 3.20E-09 | 5.78E-07 | 1.13 |
| GO:0006413 | translational initiation | 1.91E-06 | 1.87E-04 | 1.13 |
| GO:0000184 | nuclear-transcribed mRNA catabolic process, nonsense-mediated decay | 2.15E-06 | 2.05E-04 | 1.13 |
| GO:0006401 | RNA catabolic process | 2.20E-09 | 4.31E-07 | 1.12 |
| GO:0006402 | mRNA catabolic process | 2.22E-08 | 3.31E-06 | 1.12 |
| GO:0050658 | RNA transport | 4.25E-06 | 3.80E-04 | 1.12 |
| GO:0050657 | nucleic acid transport | 4.25E-06 | 3.77E-04 | 1.12 |
| GO:0051236 | establishment of RNA localization | 4.25E-06 | 3.74E-04 | 1.12 |
| GO:0007050 | cell cycle arrest | 6.67E-06 | 5.41E-04 | 1.12 |
| GO:0000278 | mitotic cell cycle | 1.18E-13 | 4.53E-11 | 1.11 |
| GO:0008380 | RNA splicing | 1.14E-09 | 2.41E-07 | 1.11 |
| GO:0006412 | translation | 2.96E-09 | 5.51E-07 | 1.11 |
| GO:0000375 | RNA splicing, via transesterification reactions | 2.19E-06 | 2.08E-04 | 1.11 |
| GO:0000377 | RNA splicing, via transesterification reactions with bulged adenosine as nucleophile | 4.52E-06 | 3.95E-04 | 1.11 |
| GO:0000398 | mRNA splicing, via spliceosome | 4.52E-06 | 3.92E-04 | 1.11 |
| GO:0010467 | gene expression | 1.88E-19 | 2.33E-16 | 1.1 |
| GO:0016071 | mRNA metabolic process | 1.02E-16 | 7.12E-14 | 1.1 |
| GO:0044265 | cellular macromolecule catabolic process | 2.61E-15 | 1.54E-12 | 1.1 |
| GO:0006396 | RNA processing | 5.89E-15 | 3.14E-12 | 1.1 |
| GO:0007049 | cell cycle | 1.64E-14 | 7.95E-12 | 1.1 |
| GO:0016482 | cytoplasmic transport | 1.31E-12 | 4.18E-10 | 1.1 |
| GO:0006397 | mRNA processing | 3.05E-11 | 7.75E-09 | 1.1 |
| GO:0006511 | ubiquitin-dependent protein catabolic process | 3.40E-08 | 4.87E-06 | 1.1 |
| GO:0045786 | negative regulation of cell cycle | 2.56E-07 | 2.95E-05 | 1.1 |
| GO:0044770 | cell cycle phase transition | 5.89E-07 | 6.34E-05 | 1.1 |
| GO:0044772 | mitotic cell cycle phase transition | 7.11E-07 | 7.50E-05 | 1.1 |
| GO:0051169 | nuclear transport | 3.81E-06 | 3.43E-04 | 1.1 |

ISLAS CpG DIFERENCIALMENTE METILADAS

| | | | | |
|------------|--|----------|----------|------|
| GO:0006913 | nucleocytoplasmic transport | 5.67E-06 | 4.80E-04 | 1.1 |
| GO:0034470 | ncRNA processing | 5.90E-06 | 4.92E-04 | 1.1 |
| GO:0010498 | proteasomal protein catabolic process | 8.66E-06 | 6.81E-04 | 1.1 |
| GO:0046907 | intracellular transport | 4.09E-18 | 3.52E-15 | 1.09 |
| GO:1902582 | single-organism intracellular transport | 1.59E-15 | 9.86E-13 | 1.09 |
| GO:0009057 | macromolecule catabolic process | 3.37E-13 | 1.18E-10 | 1.09 |
| GO:0034655 | nucleobase-containing compound catabolic process | 2.63E-09 | 4.98E-07 | 1.09 |
| GO:0051603 | proteolysis involved in cellular protein catabolic process | 3.03E-08 | 4.39E-06 | 1.09 |
| GO:0043632 | modification-dependent macromolecule catabolic process | 6.37E-08 | 8.69E-06 | 1.09 |
| GO:0019941 | modification-dependent protein catabolic process | 9.23E-08 | 1.15E-05 | 1.09 |
| GO:0006605 | protein targeting | 4.95E-06 | 4.22E-04 | 1.09 |
| GO:0009141 | nucleoside triphosphate metabolic process | 8.56E-06 | 6.79E-04 | 1.09 |
| GO:0006886 | intracellular protein transport | 7.56E-10 | 1.66E-07 | 1.08 |
| GO:0044764 | multi-organism cellular process | 1.67E-08 | 2.60E-06 | 1.08 |
| GO:0016032 | viral process | 1.95E-08 | 2.99E-06 | 1.08 |
| GO:0044403 | symbiosis, encompassing mutualism through parasitism | 1.95E-08 | 2.95E-06 | 1.08 |
| GO:0044270 | cellular nitrogen compound catabolic process | 8.26E-08 | 1.07E-05 | 1.08 |
| GO:0046700 | heterocycle catabolic process | 8.26E-08 | 1.06E-05 | 1.08 |
| GO:0007169 | transmembrane receptor protein tyrosine kinase signaling pathway | 2.99E-07 | 3.41E-05 | 1.08 |
| GO:0006281 | DNA repair | 5.67E-07 | 6.15E-05 | 1.08 |
| GO:0016568 | chromatin modification | 3.07E-06 | 2.82E-04 | 1.08 |
| GO:0007346 | regulation of mitotic cell cycle | 8.08E-06 | 6.45E-04 | 1.08 |
| GO:0044267 | cellular protein metabolic process | 2.83E-24 | 4.52E-21 | 1.07 |
| GO:0045184 | establishment of protein localization | 1.56E-12 | 4.83E-10 | 1.07 |
| GO:0015031 | protein transport | 1.91E-12 | 5.77E-10 | 1.07 |
| GO:0051726 | regulation of cell cycle | 4.54E-09 | 7.80E-07 | 1.07 |
| GO:0044419 | interspecies interaction between organisms | 5.91E-08 | 8.15E-06 | 1.07 |
| GO:0033043 | regulation of organelle organization | 9.00E-08 | 1.14E-05 | 1.07 |
| GO:0070647 | protein modification by small protein conjugation or removal | 9.00E-08 | 1.13E-05 | 1.07 |
| GO:1901361 | organic cyclic compound catabolic process | 1.22E-07 | 1.46E-05 | 1.07 |
| GO:0019439 | aromatic compound catabolic process | 1.78E-07 | 2.12E-05 | 1.07 |
| GO:0030030 | cell projection organization | 1.44E-06 | 1.47E-04 | 1.07 |

ISLAS CpG DIFERENCIALMENTE METILADAS

| | | | | |
|------------|---|----------|----------|------|
| GO:0032269 | negative regulation of cellular protein metabolic process | 7.40E-06 | 5.95E-04 | 1.07 |
| GO:0032446 | protein modification by small protein conjugation | 9.02E-06 | 7.05E-04 | 1.07 |
| GO:0006464 | cellular protein modification process | 7.17E-14 | 3.20E-11 | 1.06 |
| GO:0036211 | protein modification process | 7.17E-14 | 3.08E-11 | 1.06 |
| GO:0051649 | establishment of localization in cell | 2.43E-12 | 7.15E-10 | 1.06 |
| GO:0044248 | cellular catabolic process | 2.12E-11 | 5.50E-09 | 1.06 |
| GO:0006793 | phosphorus metabolic process | 1.47E-10 | 3.43E-08 | 1.06 |
| GO:0006796 | phosphate-containing compound metabolic process | 2.44E-10 | 5.47E-08 | 1.06 |
| GO:1901575 | organic substance catabolic process | 1.44E-09 | 2.93E-07 | 1.06 |
| GO:0033554 | cellular response to stress | 4.21E-09 | 7.36E-07 | 1.06 |
| GO:0019637 | organophosphate metabolic process | 1.70E-06 | 1.69E-04 | 1.06 |
| GO:0031401 | positive regulation of protein modification process | 2.41E-06 | 2.26E-04 | 1.06 |
| GO:0032270 | positive regulation of cellular protein metabolic process | 2.91E-06 | 2.69E-04 | 1.06 |
| GO:1901135 | carbohydrate derivative metabolic process | 3.64E-06 | 3.31E-04 | 1.06 |
| GO:0006974 | cellular response to DNA damage stimulus | 5.89E-06 | 4.95E-04 | 1.06 |
| GO:0044260 | cellular macromolecule metabolic process | 9.11E-35 | 1.02E-30 | 1.05 |
| GO:0006139 | nucleobase-containing compound metabolic process | 7.79E-20 | 1.09E-16 | 1.05 |
| GO:0090304 | nucleic acid metabolic process | 1.30E-17 | 1.04E-14 | 1.05 |
| GO:0016070 | RNA metabolic process | 4.71E-15 | 2.63E-12 | 1.05 |
| GO:0043412 | macromolecule modification | 1.50E-11 | 4.10E-09 | 1.05 |
| GO:0006996 | organelle organization | 9.09E-11 | 2.21E-08 | 1.05 |
| GO:0010604 | positive regulation of macromolecule metabolic process | 1.74E-10 | 3.96E-08 | 1.05 |
| GO:0009056 | catabolic process | 7.91E-10 | 1.70E-07 | 1.05 |
| GO:0071702 | organic substance transport | 5.15E-08 | 7.20E-06 | 1.05 |
| GO:0031324 | negative regulation of cellular metabolic process | 9.63E-08 | 1.18E-05 | 1.05 |
| GO:0032268 | regulation of cellular protein metabolic process | 1.87E-07 | 2.20E-05 | 1.05 |
| GO:0051246 | regulation of protein metabolic process | 1.90E-07 | 2.22E-05 | 1.05 |
| GO:0010605 | negative regulation of macromolecule metabolic process | 4.68E-07 | 5.18E-05 | 1.05 |
| GO:1902589 | single-organism organelle organization | 1.43E-06 | 1.48E-04 | 1.05 |
| GO:0051247 | positive regulation of protein metabolic process | 1.59E-06 | 1.59E-04 | 1.05 |
| GO:0031399 | regulation of protein modification process | 2.10E-06 | 2.05E-04 | 1.05 |

ISLAS CpG DIFERENCIALMENTE METILADAS

| | | | | |
|------------|---|----------|----------|------|
| GO:0044237 | cellular metabolic process | 6.80E-34 | 3.80E-30 | 1.04 |
| GO:0043170 | macromolecule metabolic process | 4.02E-26 | 1.12E-22 | 1.04 |
| GO:0006807 | nitrogen compound metabolic process | 4.35E-19 | 4.87E-16 | 1.04 |
| GO:0034641 | cellular nitrogen compound metabolic process | 1.66E-18 | 1.69E-15 | 1.04 |
| GO:0046483 | heterocycle metabolic process | 2.90E-18 | 2.70E-15 | 1.04 |
| GO:0006725 | cellular aromatic compound metabolic process | 7.71E-17 | 5.75E-14 | 1.04 |
| GO:1901360 | organic cyclic compound metabolic process | 1.52E-16 | 9.99E-14 | 1.04 |
| GO:0071840 | cellular component organization or biogenesis | 7.43E-15 | 3.77E-12 | 1.04 |
| GO:0016043 | cellular component organization | 3.74E-14 | 1.74E-11 | 1.04 |
| GO:1901576 | organic substance biosynthetic process | 9.73E-14 | 3.88E-11 | 1.04 |
| GO:0044249 | cellular biosynthetic process | 1.20E-13 | 4.47E-11 | 1.04 |
| GO:0009058 | biosynthetic process | 1.23E-13 | 4.42E-11 | 1.04 |
| GO:0019538 | protein metabolic process | 5.42E-13 | 1.84E-10 | 1.04 |
| GO:0009059 | macromolecule biosynthetic process | 7.36E-13 | 2.42E-10 | 1.04 |
| GO:0034645 | cellular macromolecule biosynthetic process | 8.57E-12 | 2.39E-09 | 1.04 |
| GO:0044271 | cellular nitrogen compound biosynthetic process | 1.30E-09 | 2.68E-07 | 1.04 |
| GO:0018130 | heterocycle biosynthetic process | 1.74E-09 | 3.48E-07 | 1.04 |
| GO:0009893 | positive regulation of metabolic process | 2.23E-09 | 4.30E-07 | 1.04 |
| GO:0031325 | positive regulation of cellular metabolic process | 3.00E-09 | 5.50E-07 | 1.04 |
| GO:0034654 | nucleobase-containing compound biosynthetic process | 4.06E-09 | 7.20E-07 | 1.04 |
| GO:1901362 | organic cyclic compound biosynthetic process | 8.35E-09 | 1.39E-06 | 1.04 |
| GO:0019438 | aromatic compound biosynthetic process | 2.24E-08 | 3.29E-06 | 1.04 |
| GO:0032774 | RNA biosynthetic process | 3.99E-08 | 5.65E-06 | 1.04 |
| GO:0009892 | negative regulation of metabolic process | 1.51E-06 | 1.52E-04 | 1.04 |
| GO:0051128 | regulation of cellular component organization | 2.84E-06 | 2.65E-04 | 1.04 |
| GO:0010557 | positive regulation of macromolecule biosynthetic process | 6.63E-06 | 5.41E-04 | 1.04 |
| GO:0008152 | metabolic process | 2.97E-27 | 1.10E-23 | 1.03 |
| GO:0044238 | primary metabolic process | 4.88E-26 | 1.09E-22 | 1.03 |
| GO:0071704 | organic substance metabolic process | 1.19E-25 | 2.21E-22 | 1.03 |
| GO:0060255 | regulation of macromolecule metabolic process | 4.82E-12 | 1.38E-09 | 1.03 |
| GO:0031323 | regulation of cellular metabolic | 1.92E-11 | 5.10E-09 | 1.03 |

| | | | | |
|------------|--|----------|----------|------|
| | process | | | |
| GO:0080090 | regulation of primary metabolic process | 3.46E-11 | 8.59E-09 | 1.03 |
| GO:0019222 | regulation of metabolic process | 1.26E-10 | 3.00E-08 | 1.03 |
| GO:2000112 | regulation of cellular macromolecule biosynthetic process | 4.86E-09 | 8.22E-07 | 1.03 |
| GO:0010556 | regulation of macromolecule biosynthetic process | 9.51E-09 | 1.56E-06 | 1.03 |
| GO:0019219 | regulation of nucleobase-containing compound metabolic process | 1.37E-08 | 2.22E-06 | 1.03 |
| GO:0051171 | regulation of nitrogen compound metabolic process | 1.59E-08 | 2.53E-06 | 1.03 |
| GO:0044710 | single-organism metabolic process | 1.64E-08 | 2.58E-06 | 1.03 |
| GO:0051252 | regulation of RNA metabolic process | 6.75E-08 | 9.09E-06 | 1.03 |
| GO:0031326 | regulation of cellular biosynthetic process | 6.78E-08 | 9.02E-06 | 1.03 |
| GO:0010468 | regulation of gene expression | 7.11E-08 | 9.35E-06 | 1.03 |
| GO:0009889 | regulation of biosynthetic process | 9.80E-08 | 1.19E-05 | 1.03 |
| GO:2001141 | regulation of RNA biosynthetic process | 4.27E-07 | 4.77E-05 | 1.03 |
| GO:0006355 | regulation of transcription, DNA-templated | 5.01E-07 | 5.49E-05 | 1.03 |
| GO:0051234 | establishment of localization | 6.57E-07 | 6.99E-05 | 1.03 |
| GO:0048523 | negative regulation of cellular process | 1.31E-06 | 1.37E-04 | 1.03 |
| GO:0006810 | transport | 1.49E-06 | 1.51E-04 | 1.03 |
| GO:0006351 | transcription, DNA-templated | 2.13E-06 | 2.05E-04 | 1.03 |
| GO:0048519 | negative regulation of biological process | 4.82E-06 | 4.14E-04 | 1.03 |
| GO:0048522 | positive regulation of cellular process | 6.16E-06 | 5.06E-04 | 1.03 |
| GO:0050794 | regulation of cellular process | 3.04E-07 | 3.43E-05 | 1.02 |
| GO:0009987 | cellular process | 8.28E-14 | 3.43E-11 | 1.01 |
| GO:0008150 | biological_process | 5.90E-06 | 4.89E-04 | 1.01 |

Tabla suplementaria 6.1. Procesos biológicos enriquecidos en los genes cuyos promotores (R13) presentan asociadas UIs. En la tabla se muestran los procesos biológicos con *valores-p* inferiores a $1e-5$. La tabla se encuentra ordenada de mayor a menor enriquecimiento (E). Además, se incluyen los *valores-q* corregidos a partir de los *valores-p* para análisis múltiples según el método de Benjamini y Hochberg (Benjamini and Hochberg 1995). Los enriquecimientos se han obtenido a partir del programa *GOrilla* (Eden, Navon et al. 2009).ç

UIs
FUNCIONES MOLECULARES

| Término GO | Descripción | Valor-p | Valor-q | E |
|-------------------|------------------------------------|----------------|----------------|----------|
| GO:0003735 | structural constituent of ribosome | 9.18E-06 | 2.76E-03 | 1.11 |
| GO:0003723 | RNA binding | 4.86E-11 | 4.39E-08 | 1.07 |
| GO:0019899 | enzyme binding | 1.13E-08 | 5.84E-06 | 1.06 |
| GO:0005515 | protein binding | 7.05E-30 | 2.55E-26 | 1.04 |
| GO:1901265 | nucleoside phosphate binding | 5.65E-07 | 2.27E-04 | 1.04 |
| GO:0000166 | nucleotide binding | 5.90E-07 | 2.13E-04 | 1.04 |
| GO:0016740 | transferase activity | 1.96E-06 | 6.42E-04 | 1.04 |
| GO:1901363 | heterocyclic compound binding | 1.20E-11 | 1.44E-08 | 1.03 |
| GO:0097159 | organic cyclic compound binding | 1.79E-10 | 1.29E-07 | 1.03 |
| GO:0003824 | catalytic activity | 5.57E-10 | 3.35E-07 | 1.03 |
| GO:0003676 | nucleic acid binding | 7.05E-08 | 3.18E-05 | 1.03 |
| GO:0005488 | binding | 1.36E-15 | 2.46E-12 | 1.02 |

Tabla suplementaria 6.2. Funciones moleculares enriquecidas en los genes cuyos promotores (R13) presentan asociadas *UIs* (Véase la leyenda de la Tabla suplementaria 6.1 para los detalles).

MIIs
FUNCIONES MOLECULARES

| Término GO | Descripción | Valor-p | Valor-q | E |
|-------------------|-----------------------------|----------------|----------------|----------|
| GO:0004984 | olfactory receptor activity | 3.65E-11 | 1.32E-07 | 7.71 |

Tabla suplementaria 6.3. Funciones moleculares enriquecidas en los genes cuyos promotores (R13) presentan asociadas *MIIs* (Véase la leyenda de la Tabla suplementaria 6.1 para los detalles).

DMIs-M
PROCESOS BIOLÓGICOS

| Términos GO | Descripción | Valor-p | Valor-q | E |
|--------------------|---|----------------|----------------|----------|
| GO:0034587 | piRNA metabolic process | 1.31E-06 | 7.31E-03 | 9.73 |
| GO:0007126 | meiosis | 1.78E-07 | 1.99E-03 | 4.11 |
| GO:0048610 | cellular process involved in reproduction | 2.50E-06 | 9.32E-03 | 2.05 |

Tabla suplementaria 6.4. Procesos biológicos enriquecidos en los genes cuyos promotores (R13) presentan asociadas *DMIs-M* (Véase la leyenda de la Tabla suplementaria 6.1 para los detalles).

| <i>DMIs-U</i> PROCESOS BIOLÓGICOS | | | | |
|--------------------------------------|--|----------------|----------------|----------|
| Término GO | Descripción | Valor-p | Valor-q | E |
| GO:0021515 | cell differentiation in spinal cord | 1.64E-06 | 3.52E-04 | 3.57 |
| GO:0007631 | feeding behavior | 5.77E-07 | 1.57E-04 | 2.89 |
| GO:0030326 | embryonic limb morphogenesis | 7.08E-07 | 1.88E-04 | 2.62 |
| GO:0035113 | embryonic appendage morphogenesis | 7.08E-07 | 1.84E-04 | 2.62 |
| GO:0048732 | gland development | 2.34E-06 | 4.94E-04 | 2.44 |
| GO:0048562 | embryonic organ morphogenesis | 2.02E-07 | 6.86E-05 | 2.42 |
| GO:0045165 | cell fate commitment | 1.70E-07 | 5.94E-05 | 2.4 |
| GO:0035107 | appendage morphogenesis | 6.20E-06 | 1.05E-03 | 2.34 |
| GO:0035108 | limb morphogenesis | 6.20E-06 | 1.03E-03 | 2.34 |
| GO:0030182 | neuron differentiation | 2.19E-10 | 1.44E-07 | 2.33 |
| GO:0009952 | anterior/posterior pattern specification | 8.64E-08 | 3.11E-05 | 2.31 |
| GO:2000027 | regulation of organ morphogenesis | 8.34E-07 | 2.12E-04 | 2.29 |
| GO:0001501 | skeletal system development | 4.52E-06 | 8.15E-04 | 2.15 |
| GO:0007389 | pattern specification process | 5.32E-14 | 1.19E-10 | 2.13 |
| GO:0003002 | regionalization | 2.90E-09 | 1.47E-06 | 2.13 |
| GO:0048598 | embryonic morphogenesis | 2.98E-12 | 3.70E-09 | 2.04 |
| GO:0009887 | organ morphogenesis | 1.31E-11 | 1.47E-08 | 2.04 |
| GO:0009888 | tissue development | 7.97E-13 | 1.27E-09 | 1.97 |
| GO:0048729 | tissue morphogenesis | 1.61E-08 | 6.43E-06 | 1.91 |
| GO:0002009 | morphogenesis of an epithelium | 9.14E-07 | 2.27E-04 | 1.85 |
| GO:0045596 | negative regulation of cell differentiation | 3.30E-09 | 1.54E-06 | 1.84 |
| GO:0007186 | G-protein coupled receptor signaling pathway | 9.51E-09 | 3.94E-06 | 1.81 |
| GO:0044700 | single organism signaling | 8.62E-11 | 6.88E-08 | 1.77 |
| GO:0023052 | signaling | 8.62E-11 | 6.42E-08 | 1.77 |
| GO:0007267 | cell-cell signaling | 1.60E-10 | 1.12E-07 | 1.77 |
| GO:0048468 | cell development | 3.46E-08 | 1.33E-05 | 1.77 |
| GO:0051093 | negative regulation of developmental process | 3.42E-09 | 1.53E-06 | 1.75 |
| GO:0009653 | anatomical structure morphogenesis | 1.58E-16 | 4.41E-13 | 1.69 |

| | | | | |
|------------|---|----------|----------|------|
| GO:0048731 | system development | 2.95E-09 | 1.43E-06 | 1.68 |
| GO:0007268 | synaptic transmission | 5.30E-06 | 9.41E-04 | 1.67 |
| GO:0007610 | behavior | 2.38E-06 | 4.93E-04 | 1.66 |
| GO:0060284 | regulation of cell development | 3.06E-07 | 1.01E-04 | 1.63 |
| GO:0007154 | cell communication | 2.49E-09 | 1.39E-06 | 1.62 |
| GO:0000122 | negative regulation of transcription from RNA polymerase II promoter | 5.74E-07 | 1.60E-04 | 1.61 |
| GO:0051960 | regulation of nervous system development | 3.20E-06 | 6.07E-04 | 1.6 |
| GO:0048513 | organ development | 2.61E-10 | 1.62E-07 | 1.57 |
| GO:2000026 | regulation of multicellular organismal development | 7.95E-11 | 6.84E-08 | 1.55 |
| GO:0045595 | regulation of cell differentiation | 1.49E-09 | 8.75E-07 | 1.52 |
| GO:0051094 | positive regulation of developmental process | 2.86E-06 | 5.61E-04 | 1.5 |
| GO:0048856 | anatomical structure development | 2.07E-18 | 1.16E-14 | 1.49 |
| GO:0045892 | negative regulation of transcription, DNA-templated | 3.16E-07 | 1.01E-04 | 1.49 |
| GO:0045944 | positive regulation of transcription from RNA polymerase II promoter | 3.47E-06 | 6.35E-04 | 1.48 |
| GO:0030154 | cell differentiation | 8.07E-09 | 3.47E-06 | 1.45 |
| GO:0051253 | negative regulation of RNA metabolic process | 1.62E-06 | 3.54E-04 | 1.44 |
| GO:0051239 | regulation of multicellular organismal process | 4.70E-11 | 4.38E-08 | 1.43 |
| GO:0050793 | regulation of developmental process | 2.65E-09 | 1.41E-06 | 1.43 |
| GO:0003008 | system process | 7.71E-06 | 1.25E-03 | 1.43 |
| GO:0042127 | regulation of cell proliferation | 4.49E-07 | 1.40E-04 | 1.42 |
| GO:0009890 | negative regulation of biosynthetic process | 9.18E-07 | 2.23E-04 | 1.42 |
| GO:0031327 | negative regulation of cellular biosynthetic process | 1.05E-06 | 2.50E-04 | 1.42 |
| GO:0010629 | negative regulation of gene expression | 3.27E-06 | 6.09E-04 | 1.42 |
| GO:0048869 | cellular developmental process | 3.25E-11 | 3.30E-08 | 1.41 |
| GO:0051172 | negative regulation of nitrogen compound metabolic process | 2.48E-06 | 5.05E-04 | 1.41 |
| GO:0032501 | multicellular organismal process | 2.96E-13 | 5.51E-10 | 1.39 |
| GO:0045934 | negative regulation of nucleobase-containing compound metabolic process | 6.95E-06 | 1.14E-03 | 1.39 |
| GO:0044707 | single-multicellular organism process | 1.39E-12 | 1.94E-09 | 1.38 |
| GO:0010628 | positive regulation of gene expression | 2.63E-06 | 5.25E-04 | 1.38 |
| GO:0006357 | regulation of transcription from RNA polymerase II promoter | 1.09E-06 | 2.53E-04 | 1.37 |
| GO:0032502 | developmental process | 1.91E-18 | 2.14E-14 | 1.36 |
| GO:0044767 | single-organism developmental | 1.06E-16 | 3.96E-13 | 1.36 |

| | | process | | |
|------------|---|----------|----------|------|
| GO:0048522 | positive regulation of cellular process | 5.05E-07 | 1.53E-04 | 1.22 |
| GO:0048523 | negative regulation of cellular process | 3.19E-06 | 6.15E-04 | 1.22 |
| GO:0048518 | positive regulation of biological process | 5.36E-07 | 1.58E-04 | 1.21 |
| GO:2001141 | regulation of RNA biosynthetic process | 5.70E-06 | 9.79E-04 | 1.21 |
| GO:0048519 | negative regulation of biological process | 5.39E-06 | 9.41E-04 | 1.2 |
| GO:0051252 | regulation of RNA metabolic process | 8.37E-06 | 1.34E-03 | 1.2 |
| GO:0050896 | response to stimulus | 9.34E-06 | 1.47E-03 | 1.14 |
| GO:0050789 | regulation of biological process | 5.55E-07 | 1.59E-04 | 1.11 |
| GO:0050794 | regulation of cellular process | 1.27E-06 | 2.89E-04 | 1.11 |
| GO:0065007 | biological regulation | 1.30E-06 | 2.92E-04 | 1.1 |
| GO:0044699 | single-organism process | 4.89E-08 | 1.82E-05 | 1.09 |

Tabla suplementaria 6.5. Procesos biológicos enriquecidos en los genes cuyos promotores (R13) presentan asociadas DMIs-U (Véase la leyenda de la **Tabla suplementaria 6.1** para los detalles).

| <i>DMIs-U</i> | | | | |
|-----------------------|---|----------------|----------------|----------|
| FUNCIONES MOLECULARES | | | | |
| Término GO | Descripción | Valor-p | Valor-q | E |
| GO:0000976 | transcription regulatory region sequence-specific DNA binding | 2.61E-07 | 6.74E-05 | 2.2 |
| GO:0043565 | sequence-specific DNA binding | 1.08E-21 | 3.89E-18 | 2.1 |
| GO:0000981 | sequence-specific DNA binding RNA polymerase II transcription factor activity | 1.83E-09 | 7.34E-07 | 2.09 |
| GO:0004930 | G-protein coupled receptor activity | 1.94E-09 | 7.01E-07 | 2.04 |
| GO:0044212 | transcription regulatory region DNA binding | 1.07E-07 | 3.53E-05 | 1.81 |
| GO:0000975 | regulatory region DNA binding | 2.41E-07 | 7.27E-05 | 1.78 |
| GO:0001067 | regulatory region nucleic acid binding | 2.41E-07 | 6.71E-05 | 1.78 |
| GO:0038023 | signaling receptor activity | 1.34E-11 | 1.21E-08 | 1.77 |
| GO:0004888 | transmembrane signaling receptor activity | 4.39E-10 | 3.17E-07 | 1.76 |
| GO:0003700 | sequence-specific DNA binding transcription factor activity | 4.47E-16 | 8.08E-13 | 1.72 |
| GO:0001071 | nucleic acid binding transcription factor activity | 5.07E-16 | 6.10E-13 | 1.72 |
| GO:0004872 | receptor activity | 6.45E-10 | 3.88E-07 | 1.62 |
| GO:0004871 | signal transducer activity | 7.58E-10 | 3.91E-07 | 1.55 |

| | | | | |
|------------|-------------------------------|----------|----------|------|
| GO:0060089 | molecular transducer activity | 7.58E-10 | 3.42E-07 | 1.55 |
| GO:0003677 | DNA binding | 1.38E-06 | 3.33E-04 | 1.25 |

Tabla suplementaria 6.6. Funciones moleculares enriquecidas en los genes cuyos promotores (R13) presentan asociadas *DMIs-U* (Véase la leyenda de la Tabla suplementaria 6.1 para los detalles).

CONCLUSIONES

1. En esta Tesis Doctoral se ha desarrollado un procedimiento para identificar regiones genómicas diferencialmente metiladas, que puedan servir como marcadores epigenéticos. Esto ha supuesto el desarrollo y la implementación de: 1) dos herramientas bioinformáticas para la obtención de metilomas de alta calidad en genoma completo; 2) una base de datos para el almacenamiento, visualización y gestión de los metilomas generados; 3) un algoritmo capaz de identificar islas CpG estadísticamente significativas; y 4) un método estadístico fiable para la identificación de islas diferencialmente metiladas, a partir del análisis comparado de múltiples metilomas.
2. *NGSmethPipe*, el primer programa desarrollado en esta Tesis, permite pre-procesar y alinear las lecturas cortas procedentes de protocolos de secuenciación masiva de ADN tratado con bisulfito. Los controles de calidad implementados consiguen una alta proporción de lecturas correctamente alineadas, sin aumentar en exceso los alineamientos erróneos. Esto, unido a la

paralelización de todo el proceso, convierten a *NGSmethPipe* en una herramienta eficaz para alinear este tipo de librerías.

3. Por su parte, *MethylExtract* permite inferir simultáneamente los niveles de metilación y las variantes de un solo nucleótido a partir de una misma librería. Asimismo, este programa implementa estrictos controles para minimizar los errores de secuenciación y los fallos del tratamiento con bisulfito. Esto lo convierte en una herramienta fiable y versátil para la obtención de metilomas de alta resolución en genoma completo.
4. *NGSmethDB* es una base de datos relacional de metilomas completos, todos ellos procesados mediante el mismo protocolo, lo que permite comparar metilomas procedentes de diferentes estudios. Las herramientas de minería de datos implementadas y el navegador genómico de última generación que incorpora, han sido diseñados para realizar consultas complejas sobre múltiples tejidos en cualquier región del genoma. Actualmente, la base de datos contiene información para 6 especies y 114 tejidos y/o condiciones diferentes, lo que ha supuesto procesar aproximadamente 40 terabytes de librerías de secuenciación masiva. Estas características convierten a *NGSmethDB* en una herramienta muy potente para el análisis de metilomas de alta resolución.
5. Sobre la base de *CpGcluster*, se ha desarrollado *WordCluster*, un nuevo algoritmo capaz de identificar islas CpG estadísticamente significativas y con una alta densidad de CpGs. *WordCluster* presenta importantes ventajas con respecto a los métodos convencionales, destacando entre ellas la delimitación de

dominios de metilación más cortos pero también más homogéneos, así como una asociación más específica con elementos reguladores y regiones evolutivamente conservadas del genoma.

6. El análisis de la metilación en las islas predichas por *WordCluster* ha permitido establecer que la combinación de la prueba exacta de Fisher y la binomial negativa es el método estadístico más apropiado para identificar islas diferencialmente metiladas.
7. Atendiendo al estado de metilación de los tejidos analizados, el 66% de las islas diferencialmente metiladas pueden ser de dos tipos: las que están generalmente metiladas pero aparecen no-metiladas en algún tejido (islas *DMI-M*), y aquellas que permanecen generalmente no-metiladas pero pueden metilarse en algún tejido (islas *DMI-U*).
8. El análisis funcional mediante el enriquecimiento en términos de ontologías ha permitido establecer que las islas de tipo *DMI-M* se asocian con funciones tejido-específicas, mientras que las islas de tipo *DMI-U* están implicadas en procesos de desarrollo y diferenciación.
9. Los estudios de enriquecimiento de las islas diferencialmente metiladas en distintos elementos genómicos han revelado importantes diferencias con respecto a las regiones diferencialmente metiladas, identificadas recientemente por otros autores. Entre estas diferencias destacan el elevado enriquecimiento de las islas diferencialmente metiladas en promotores y exones y su empobrecimiento en intrones, así como la significativa menor proporción de las islas de tipo *DMI-M*

asociadas con sitios de unión a factores de transcripción. A su vez, es de destacar que las islas de tipo *DMI-U* solapan en mucho mayor grado con los sitios de unión a factores de transcripción que las islas de tipo *DMI-M*.

10. Las importantes características biológicas encontradas en las islas diferencialmente metiladas sugieren que *WordCluster* puede ser el algoritmo adecuado para preseleccionar las regiones a analizar en futuros estudios de metilación diferencial.

CUESTIONES ABIERTAS

Entre las cuestiones que esta Tesis deja abiertas para el futuro cabe destacar las siguientes:

1. Las islas diferencialmente metiladas identificadas en esta Tesis son, por definición, densas en CpGs. Sin embargo, se sabe que los *TFs* pueden interactuar con regiones con baja densidad de CpGs (Stadler, Murr et al. 2011). Sería por tanto interesante desarrollar algoritmos que permitiesen localizar regiones diferencialmente metiladas pero con baja densidad de CpGs, y repetir con ellas los estudios de enriquecimiento y asociación llevados a cabo en las *DMIs*.
2. Además de los niveles de metilación, *NGSmethPipe* y *MethylExtract* son capaces de detectar simultáneamente las *SNVs*, a partir de la misma librería de secuenciación masiva. Sin embargo, sería interesante también poder detectar las *indels*.
3. La lectura del genoma de referencia mediante ficheros multifasta, el control del sesgo de hebra durante la detección de *SNVs*, una opción

para reducir la sobreestimación que suponen los fragmentos solapantes en lecturas *pair-end*, o automatizar la detección y descarte de los sesgos en composición nucleotídica (Schwartz, Oren et al. 2011) y/o de conversión del bisulfito (Hansen, Langmead et al. 2012, Lin, Sun et al. 2013), son todas ellas mejoras potenciales.

4. Igualmente interesante sería la automatización del proceso de incorporación de nuevos metilomas a *NGSmethDB*, lo que facilitaría mantener en continuo crecimiento el número de metilomas en la base de datos.
5. Las diferencias de asociación encontradas entre las *DMIs* identificadas en esta Tesis y las *DMRs* definidas por otros autores (Ziller, Gu et al. 2013), junto con la conocida capacidad de los *TFs* para interaccionar con regiones hipermetiladas de baja densidad de CpGs (Stadler, Murr et al. 2011), permiten plantearse la pregunta de si la metilación presenta una función activa solamente en aquellas regiones con una densidad de CpGs suficiente como para suponer un impedimento estérico en las interacciones de los *TFs* con el ADN. Para responder a esta cuestión, serían necesarios análisis más específicos que permitiesen constatar las diferencias existentes entre las regiones diferencialmente metiladas de alta y baja densidad de CpGs.
6. Las diferencias encontradas en la función y en la asociación diferencial de cada tipo de *DMIs* (*DMIs-U* y *DMIs-M*) sugieren, por último, que la metilación podría estar regulando de manera tejido-específica la unión de los *TFs* a regiones críticas para la célula. Una posibilidad es que la metilación pudiera inhibir la diferenciación de células madre bloqueando la unión de los *TFs* implicados en la

diferenciación celular (*DMIs-U* en *TFBSs*), o bien evitando la interferencia con la transcripción de genes indispensables (*DMIs-M* y *MIs* en exones). Estas cuestiones podrían abordarse mediante el diseño experimental adecuado, con el fin de comprobar, por un lado, si la metilación en regiones densas de CpGs es capaz de evitar la interacción de los *TFs* con el ADN, y por otro si las regiones densas en CpGs hipermetilados son capaces por sí mismas de inhibir la diferenciación celular.

7. Quizás el reto más importante para el futuro será aplicar todos estos desarrollos a los epigenomas que se empiezan a obtener ahora a partir de células únicas, ya que la heterogeneidad entre las células que componen un mismo tejido (Jaffe and Irizarry 2014) puede ser un factor importante a tener en cuenta.

LISTADO DE FIGURAS

Figura 1.1. Número de publicaciones sobre epigenética entre los años 1994 y 2013. La gráfica representa el número de publicaciones que contienen los términos “*Epigenetic*” o “*Epigenomic*” incluidas en la base de datos *PubMed* (<http://www.ncbi.nlm.nih.gov/pubmed>) durante los últimos 20 años (1994-2013). Además, se incluyen algunos de los hitos, tanto técnicos como científicos, más importantes en la investigación epigenética: la aparición del primer método de secuenciación masiva (“*454 sequencing*” en el año 2000), la fundación del consorcio *ENCODE* (Consortium 2004), la publicación de los resultados del proyecto piloto de *ENCODE* (Birney, Stamatoyannopoulos et al. 2007), la incorporación de “*Illumina*” al mercado de la secuenciación masiva (2007), la fundación del consorcio “*ROADMAP Epigenomics*” (Bernstein, Stamatoyannopoulos et al. 2010) y la publicación en septiembre de 2012 de 29 artículos (<http://www.nature.com/encode/>) con los resultados obtenidos a partir del proyecto *ENCODE*.

Figura 1.2. CGIs predichas por *CpGcluster* en el gen *PIK3R2*. En la figura se muestra el gen *PIK3R2* (anotación de RefSeq (Pruitt, Tatusova et al. 2007)), las *CGIs* predichas por *CpGcluster* (Hackenberg, Previti et al. 2006, Hackenberg, Carpena et al. 2011) y los niveles de metilación para los CpGs de 4 tejidos diferentes (Hackenberg, Barturen et al. 2011, Geisen, Barturen et al. 2014). En verde se representan los CpGs no metilados y en rojo los metilados. Las islas predichas en el gen *PIK3R2* ilustran diferentes tipos que pueden clasificarse en función de su estado de metilación y/o de su localización: (a) y (c) son islas no metiladas en todos los tejidos, normalmente asociadas a las regiones promotoras de genes domésticos (a), pero también pueden encontrarse asociadas a otros elementos genómicos como exones (c); sin embargo, (b) y (d) son islas diferencialmente metiladas, reguladoras potenciales de la interacción de otras moléculas con el ADN.

Figura 1.3. Representación esquemática de una región diferencialmente metilada. Los círculos representan citosinas en dinucleótidos CpG para dos muestras diferentes (Tejidos A y B); los círculos rojos simbolizan citosinas metiladas y los verdes no metiladas.

Figura 1.4. Representación esquemática de la interacción de factores de transcripción con el ADN. En la figura se muestran dos regiones hipermetiladas, una con baja densidad de CpGs (izquierda) y otra con alta densidad (derecha). Los círculos representan citosinas en dinucleótidos CpG (los rojos simbolizan citosinas metiladas y los verdes no metiladas). Por otro lado, el factor de transcripción (*TF*) con un sitio de unión a estas regiones se representa como una elipse verde.

Figura 3.1. Preprocesado de las lecturas. La figura muestra los pasos necesarios para preparar las lecturas antes de ser alineadas. En primer lugar se elimina la región de menor calidad de las lecturas (A), recortando antes del primer nucleótido con una calidad de secuenciación menor de 2 (#) y posteriormente se busca de manera iterativa la secuencia del adaptador (B).

Figura 3.2. Tipos de secuenciación masiva con tratamiento de bisulfito: *MethylC-Seq* y *BS-Seq*. Tras la desnaturalización y el tratamiento con bisulfito del ADN se pierde la complementariedad de hebra, ya que las citosinas no metiladas se convierten en uracilos (coloreadas en verde). Durante la amplificación del ADN, los uracilos serán sustituidos por timinas. En la figura se muestran las lecturas resultantes de ambos protocolos: (A) El protocolo *MethylC-Seq* genera la librería de lecturas de manera direccional, donde encontraremos las secuencias tratadas con bisulfito para la cadena Watson (*BSW*) y para la cadena Crick (*BSC*); (B) En el protocolo *BS-Seq* se realizan dos *PCRs* consecutivas, que resultan en la presencia tanto de las lecturas provenientes de *BSW* y *BSC* como de sus reversas complementarias (*BSWRC* y *BSCRC*).

Figura 3.3. Alineamiento de lecturas tratadas con bisulfito basado en el modelo de referencia con tres letras. Para calcular los perfiles de metilación, en primer lugar las lecturas deben ser alineadas frente a un genoma de referencia. Una de las técnicas más utilizadas para lidiar con la reducción de complejidad causada por el bisulfito, es convertir tanto el genoma de referencia como las lecturas a un alfabeto de tres letras. En el protocolo *MethylC-Seq* (A), las lecturas vienen de las cadenas Watson (*BSW*) y Crick (*BSC*) convertidas por el bisulfito. Por lo tanto, todas las citosinas sin convertir en las lecturas se cambian por timinas y se las trata de alinear con la hebra directa de la referencia C>T y con la complementaria inversa de la referencia G>A (flechas azules). Mientras que en el protocolo *BS-Seq* (B), las lecturas pueden alinearse con las dos hebras de ambas referencias (flechas azules y amarillas): referencia C>T (lecturas *BSW* y *BSWRC* en la **Figura 3.2**) y referencia G>A (lecturas *BSC* y *BSCRC* en la **Figura 3.2**). En caso de encontrar el alineamiento correcto, se revierten los cambios tanto en la referencia como en la lectura y se infieren los valores de metilación de manera directa: un desparejamiento C/T indica una citosina no metilada (coloreadas en verde) y una citosina en ambas secuencias significa que la citosina se encuentra metilada (coloreadas en rojo). En el caso de las lecturas provenientes de

las secuencias complementarias a las tratadas con bisulfito (sólo en el protocolo *BS-Seq*), el desemparejamiento G/A significaría la existencia de una citosina no metilada en la hebra complementaria.

Figura 3.4. Desambiguación mediante extensión de la semilla. El alineamiento se realiza mediante *Bowtie*, usando una semilla por defecto de 26 pb y con 1 desemparejamiento como máximo en la región semilla. En primer lugar, se preseleccionan los dos mejores alineamientos para cada secuencia de referencia (en este caso *BSW* y *BSC*), tomándose los alineamientos con menor número de desemparejamientos dentro de la semilla y con menores valores de calidad en los desemparejamientos a lo largo de la lectura. Entonces, se mide la longitud desde el final de la semilla hasta el próximo desemparejamiento, seleccionando como mejor alineamiento aquel que presente la distancia más larga (el (b) en el ejemplo de la figura). Las citosinas convertidas se representan en azul, mientras que los desemparejamientos se encuentran coloreados en rojo, así como las distancias desde el final de la semilla al siguiente desemparejamiento.

Figura 3.5. Eficiencia en el recorte de adaptadores. Se muestra el porcentaje de lecturas recortadas (conjunto de datos descritos en el apartado 3.2.1.1, donde todas las lecturas poseen el adaptador en 3') por calidad (azul), por calidad y adaptador (verde) o sólo adaptador (rojo). Se muestran dos metodologías diferentes, (A) *NGSmethPipe* y (B) búsqueda iterativa del adaptador sin previo recorte por calidad.

Figura 3.6. Comparación entre los alineamientos de lecturas únicas y la desambiguación de multilecturas mediante la extensión de la semilla. Se han simulado y alineado 2,000,000 de lecturas, tal y como se describe en el apartado 3.2.1.1, para el *contig GL000022.1* del cromosoma 2 de hg19. En azul se representan el número de alineamientos correctos (eje izquierdo de ordenadas) y en rojo el porcentaje de lecturas erróneamente alineadas (eje derecho de ordenadas). Estos valores se representan para diferentes longitudes de la semilla y dos metodologías diferentes: un método basado en la extensión de la semilla del alineamiento (cuadrados) y otro método que sólo selecciona lecturas con alineamientos únicos (círculos), ambos implementados por *NGSmethPipe*.

Figura 3.7. Comparación entre *NGSmethPipe* y *Bismark*. Se han simulado y alineado 2,000,000 de lecturas, tal y como se describe en el apartado 3.2.1.1, para el *contig GL000022.1* del cromosoma 2 de hg19. En azul se representan el número de alineamientos correctos (eje izquierdo de ordenadas) y en rojo el porcentaje de lecturas erróneamente alineadas (eje derecho de ordenadas). Estos valores se representan para diferentes longitudes de la semilla y dos

programas diferentes *NGSmethPipe* (cuadrados) y *Bismark* (círculos). En ambos programas se han utilizado sus respectivos parámetros por defecto.

Figura 3.8. Representación de las proporciones de CpGs metilados y no metilados para cada muestra. Se representan las células madre (cuadrados rojos) y los tejidos frescos (círculos verdes) formando un subconjunto rodeado por un círculo rojo, a su vez las líneas celulares diferenciadas (triángulos azules) se encuentran rodeados por un círculo azul. También se incluyen en la figura una línea de células madre en la que se ha inducido la diferenciación (*wa09fibro*, rombo marrón) y una línea tumoral (*hcc1954*, estrella amarilla).

Figura 3.9. Distribuciones de los niveles de metilación en CpGs para diferentes tejidos. Se representan las proporciones de CpGs para 10 grupos de niveles de metilación en 6 tejidos, 2 para cada clase de las previamente descritas: tejidos frescos (*cd133hsc* y *hspc*) en verde, líneas celulares de fibroblastos (*fibro* e *imr90*) en azul y líneas celulares de células madre (*h1* y *wa09*) en rojo.

Figura 6.1. Diferencias absolutas de metilación para cada CGI identificada como diferencialmente metilada por cada método estadístico. Las diferencias de metilación se calcularon entre las líneas celulares *h1* e *imr90* (Lister, Pelizzola et al. 2009), incluyéndose en el análisis sólo aquellos contextos CpG con una profundidad mayor a 5. Los resultados mostrados en (F) y (G), se obtuvieron aplicando dos test estadísticos conjuntamente (el análisis basado en la binomial negativa junto con cada una de las variantes de la prueba exacta de Fisher). En estos análisis conjuntos, ambos valores- p deben mantenerse por debajo del umbral seleccionado (0.05). La línea horizontal es el límite superior de las diferencias que pueden caer dentro de una misma clase, ya que 0.2 es la amplitud de las clases descritas en Material y métodos (apartado 6.1.1).

Figura 6.2. Curva ROC de los diferentes análisis estadísticos. Se representa la sensibilidad (S_n) frente a $1 -$ especificidad ($1 - S_p$) de la detección de diferencias de metilación para todos los pares de tejidos incluidos en el estudio (Apartado 6.1). Los valores- p utilizados en los análisis estadísticos varían entre $1e-10$ y 1. En la parte inferior derecha de la imagen se muestra una ampliación de la zona resaltada para los análisis que muestran mejores resultados (Binomial negativa, Fisher y Binomial negativa + Fisher). La línea negra discontinua muestra una predicción aleatoria.

Figura 6.3. CGIs con metilación significativamente diferente en función del análisis estadístico utilizado. Las imágenes A y B muestran CGIs identificadas como diferencialmente metiladas por la binomial negativa y no por la prueba exacta de Fisher. Las imágenes C y D, muestran ejemplos del caso contrario. Las franjas azules representan las CGIs, las 2 primeras

pistas muestran valores de metilación para todos los CpGs presentes en las islas, mientras que las 2 pistas inferiores representan CpGs con una profundidad mínima de 5 (profundidad mínima utilizada en los análisis estadísticos). Los tejidos incluidos en todos los casos son *h1* e *imr90* (ver apartado 6.1), donde *h1* es la primera pista representada e *imr90* la segunda en los pares descritos previamente. En las tablas bajo las imágenes se incluye el identificador de la isla (Isla CpG) y los parámetros utilizados en cada uno de los análisis estadísticos, así como el valor-*p* resultante en cada caso.

Figura 6.4. Árbol de distancias entre tipos celulares basado en la fracción de CGIs diferencialmente metiladas entre pares de tejidos. Se ha utilizado el programa *MEGA6* (Tamura, Stecher et al. 2013) y el método *UPGMA* (Nei and Kumar 2000).

Figura 6.5. Clases de CGIs y porcentajes de solapamiento con los genes y con las regiones intergénicas. En la gráfica se muestra el número y porcentaje de islas CpG clasificadas (Apartado 6.1.4) como no metiladas (*UIs*), metiladas (*MI*s), diferencialmente metiladas (*DMI*s) y no asignadas a ninguna de las clases (*NA*). En el interior de cada porción, otro gráfico circular muestra el solapamiento de cada clase con regiones génicas de *refSeq* (Pruitt, Tatusova et al. 2007), donde la región génica se ha definido como el cuerpo génico \pm 500 pb de entorno génico (Apartado 6.1.5).

Figura 6.6. Número de tejidos metilados y no metilados en las DMI. En el eje de abscisas se representan el número de tejidos no metilados y en el de ordenadas los metilados. Los colores simbolizan el número de *DMI*s, tal y como se muestra en la leyenda situada a la derecha de la figura.

Figura 6.7. Enriquecimiento de CGIs en regiones génicas. En la tabla se muestran los índices de enriquecimiento de las clases de *CGIs* (filas) para diferentes regiones génicas (columnas). De izquierda a derecha: región génica, región promotora R13, exones, intrones y una región en torno al final de la transcripción etiquetada como R8 (estas regiones se definen en el apartado 6.1.5). La escala de color oscila entre 0 y 10, el color verde (0 a <1) simboliza empobrecimiento en el elemento estudiado y el color rojo (>1 a 10) enriquecimiento. Los colores próximos al blanco (1) serían casos de distribución aleatoria con respecto al elemento estudiado. Todos los valores de enriquecimientos/empobrecimiento presentan valores-*p* por debajo de 0.01.

Figura 6.8. Enriquecimiento de CGIs en distintos elementos reguladores. En la tabla se muestran los índices de enriquecimiento de las clases de *CGIs* (filas) para diversos elementos reguladores (columnas). De izquierda a derecha: *clusters* de sitios de unión a factores de transcripción, *clusters* de sitios de hipersensibilidad a la DNasa, dos conjuntos de *clusters* de

subunidades de la ARN polimerasa II y III, potenciadores de la base de datos *VISTA* (Visel, Minovitsky et al. 2007), potenciadores potencialmente activos (Zentner, Tesar et al. 2011), sitios de unión del factor de transcripción CTCF fuera de regiones génicas y sitios de unión del factor de transcripción CTCF solapantes con regiones exónicas. Los datos, salvo los potenciadores *VISTA*, se han extraído del proyecto *ENCODE* (Consortium, Bernstein et al. 2012). La escala de color oscila entre 0 y 10: el color verde (0 a <1) simboliza empobrecimiento de la clase de *CGI* en el elemento estudiado y el color rojo (>1 a 10) enriquecimiento. Los colores próximos al blanco (1) serían casos de distribución aleatoria con respecto al elemento estudiado. Los valores de enriquecimiento/empobrecimiento con valores-*p* superiores a 0.01 se marcan con un asterisco. La descripción de los conjuntos de datos utilizados puede encontrarse en el apartado 6.1.5.

Figura 6.9. Enriquecimiento de *CGIs* en elementos conservados y variaciones de un solo nucleótido. En la tabla se muestran los índices de enriquecimiento de las clases de *CGIs* (filas) para diferentes regiones de interés evolutivo y variaciones de un solo nucleótido (columnas). De izquierda a derecha se muestran elementos conservados *phastCons* (Siepel, Bejerano et al. 2005), regiones sometidas a un sesgo mutacional hacia GC (Duret and Galtier 2009, Capra, Hubisz et al. 2013), elementos conservados *GERP* (Davydov, Goode et al. 2010), polimorfismos de un solo nucleótido (*SNPs*) y variaciones de un solo nucleótido asociadas a enfermedades (Sherry, Ward et al. 2001). La escala de color oscila entre 0 y 10: el color verde (0 a <1) simboliza empobrecimiento de la clase de *CGI* en el elemento estudiado y el color rojo (>1 a 10) enriquecimiento. Los colores próximos al blanco (1) serían casos de distribución aleatoria con respecto al elemento estudiado. Los valores de enriquecimiento/empobrecimiento con valores-*p* superiores a 0.01 se marcan con un asterisco. Para una descripción de los distintos elementos ver apartado 6.1.5.

Figura 6.10. Fracción de pares de tejidos diferencialmente metilados para cada tejido y clase de *DMI*. Para cada una de las clases mayoritarias de *DMIs* (*DMIs-M*, en azul y *DMIs-U*, en rojo) se muestra la fracción de pares de tejidos identificados como diferencialmente metilados en el que se encuentra cada tipo celular.

LISTADO DE TABLAS

Tabla 3.1. Características de los alineadores para lecturas tratadas con bisulfito. La fila “Lecturas de entrada” indica el tipo de lecturas que admite cada método: BS (basadas en secuencias) y BC (basadas en color). El asterisco (*) en la fila de multiprocesadores, significa que estos programas realizan el procesamiento múltiple sólo durante el alineamiento con *Bowtie*. La fila del “alineamiento basado en semilla”, especifica los métodos que realizan su alineamiento mediante el método de semilla. La fila “Q” indica qué métodos utilizan la calidad de las secuencias durante el alineamiento. La fila “Simples (no direccional)” especifica los métodos que alinean lecturas simples para el protocolo *BS-Seq* (no direccional), ya que todos alinean lecturas singulares para el protocolo *MethylC-Seq* (direccional). Por último, las filas “Emparejadas” indican los métodos que alinean lecturas emparejadas provenientes de los protocolos *MethylC-Seq* (direccional) y *BS-Seq* (no direccional).

Tabla 3.2. Comparación entre *NGSmethPipe* y *Bismark*, para la misma longitud de semilla (l=26) y el mismo número de alineamientos correctos (aprox. 1.65e6).

Tabla 3.3. Resultados obtenidos con *NGSmethPipe* y *MethylExtract* para una serie de datos analizados en esta tesis. Los conjuntos de datos que se muestran provienen de diversas publicaciones: *bcell*, *cd133hsc* y *hspc* (Hodges, Molaro et al. 2011); *fibro*, *wa09fibro* y *wa09* (Laurent, Wong et al. 2010); *h1* e *imr90* (Lister, Pelizzola et al. 2009); *hmec* y *hcc1954* (Hon, Hawkins et al. 2012); *prefrontalcortex_hs1570* (Zeng, Konopka et al. 2012). Todos ellos presentan valores de metilación para al menos el 75% de sus CpGs, con una profundidad mínima de 5 lecturas. Para cada conjunto de datos se muestra la media y desviación estándar (SD) para la cobertura de secuenciación, tanto de CpGs como de cualquier posición del genoma, y para los niveles de metilación en CpGs. También se muestra el porcentaje de CpGs metilados (niveles de metilación ≥ 0.8) y no metilados (niveles de metilación ≤ 0.2), así como el número de variaciones detectadas frente a la referencia, con su porcentaje sobre el total de posiciones en el genoma entre paréntesis. Para evitar confusiones, se han mantenido los nombres originales asignados por los respectivos autores.

Tabla 6.1. Clasificación discreta de los niveles de metilación.

Tabla 6.2. Tipos celulares y TFBSs utilizados para definir el conjunto de potenciadores potencialmente activos.

Tabla 6.3. Diferencias de metilación por pares de tejidos. En la tabla se muestran: el número de comparaciones con datos suficientes para los 66 pares de tejidos analizados en las CGIs, el número de comparaciones con un valor-p significativo para ambos análisis estadísticos y las comparaciones con diferencias significativas para uno de los análisis, pero no significativas para el otro. Además, para cada uno de estos casos se muestran las medias pesadas (X) y desviaciones estándar (SD) de las diferencias absolutas de metilación.

Tabla 6.4. Estadística básica de la composición de las clases de CGIs. En la tabla se muestran la media y la desviación estándar de las características composicionales de las clases de CGIs descritas (Apartado 6.1.4). Los valores que se muestran son: el número de CGIs de cada clase (#), su longitud, la proporción de CpGs observados/esperados (*Ratio [O/E]*), la fracción de G+C (GC) y la densidad de CpGs (medida como el número de CpGs dividido por la longitud de la isla).

Tabla suplementaria 6.1. Procesos biológicos enriquecidos en los genes cuyos promotores (R13) presentan asociadas UIs. En la tabla se muestran los procesos biológicos con *valores-p* inferiores a $1e-5$. La tabla se encuentra ordenada de mayor a menor enriquecimiento (E). Además, se incluyen los *valores-q* corregidos a partir de los *valores-p* para análisis múltiples según el método de Benjamini y Hochberg (Benjamini and Hochberg 1995). Los enriquecimientos se han obtenido a partir del programa *GORilla* (Eden, Navon et al. 2009).

Tabla suplementaria 6.2. Funciones moleculares enriquecidas en los genes cuyos promotores (R13) presentan asociadas UIs

Tabla suplementaria 6.3. Funciones moleculares enriquecidas en los genes cuyos promotores (R13) presentan asociadas MIs.

Tabla suplementaria 6.4. Procesos biológicos enriquecidos en los genes cuyos promotores (R13) presentan asociadas DMIs-M.

Tabla suplementaria 6.5. Procesos biológicos enriquecidos en los genes cuyos promotores (R13) presentan asociadas DMIs-U.

Tabla suplementaria 6.6. Funciones moleculares enriquecidas en los genes cuyos promotores (R13) presentan asociadas DMIs-U.

REFERENCIAS

Abecasis, G. R., D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles and G. A. McVean (2010). "A map of human genome variation from population-scale sequencing." Nature **467**(7319): 1061-1073.

Abecasis, G. R., A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth and G. A. McVean (2012). "An integrated map of genetic variation from 1,092 human genomes." Nature **491**(7422): 56-65.

Amoreira, C., W. Hindermann and C. Grunau (2003). "An improved version of the DNA Methylation database (MethDB)." Nucleic Acids Res **31**(1): 75-77.

Antequera, F. and A. Bird (1993). "Number of CpG islands and genes in human and mouse." Proc Natl Acad Sci U S A **90**(24): 11995-11999.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Barrett, T., D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muetter and R. Edgar (2009). "NCBI GEO: archive for high-throughput functional genomic data." Nucleic Acids Res **37**(Database issue): D885-890.

Barrett, T., S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis and A. Soboleva (2013). "NCBI GEO: archive for functional genomics data sets--update." Nucleic Acids Res **41**(Database issue): D991-995.

Barturen, G., A. Rueda, J. L. Oliver and M. Hackenberg (2013). "MethylExtract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data." F1000Res **2**(217): 217.

Bell, A. C., A. G. West and G. Felsenfeld (1999). "The protein CTCF is required for the enhancer blocking activity of vertebrate insulators." Cell **98**(3): 387-396.

Bell, J. T., A. A. Pai, J. K. Pickrell, D. J. Gaffney, R. Pique-Regi, J. F. Degner, Y. Gilad and J. K. Pritchard (2011). "DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines." Genome Biol **12**(1): R10.

Bell, J. T. and T. D. Spector (2012). "DNA methylation studies using twins: what are they telling us?" Genome Biol **13**(10): 172.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." J. Roy. Statist. Soc. Ser. B **57**: 289–300.

Bernstein, B. E., J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen and J. A. Thomson (2010). "The NIH Roadmap Epigenomics Mapping Consortium." Nat Biotechnol **28**(10): 1045-1048.

Bird, A., M. Taggart, M. Frommer, O. J. Miller and D. Macleod (1985). "A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA." Cell **40**(1): 91-99.

Bird, A. P. (1980). "DNA methylation and the frequency of CpG in animal DNA." Nucleic Acids Res **8**(7): 1499-1504.

Birney, E., J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo,

A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korb, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W. K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C. N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C. L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S. E. Antonarakis, Y. Fu, E. D. Green, U. Karaoz, A. Siepel, J. Taylor, L. A. Liefer, K. A. Wetterstrand, P. J. Good, E. A. Feingold, M. S. Guyer, G. M. Cooper, G. Asimenos, C. N. Dewey, M. Hou, S. Nikolaev, J. I. Montoya-Burgos, A. Loytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N. R. Zhang, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. A. Stone, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, A. Sidow, N. D. Trinklein, Z. D. Zhang, L. Barrera, R. Stuart, D. C. King, A. Ameur, S. Enroth, M. C. Bieda, J. Kim, A. A. Bhinge, N. Jiang, J. Liu, F. Yao, V. B. Vega, C. W. Lee, P. Ng, A. Shahab, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. A. Singer, T. A. Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J. C. Fowler, P. Couttet, A. W. Bruce, O. M. Dovey, P. D. Ellis, C. F. Langford, D. A. Nix, G. Euskirchen, S. Hartman, A. E. Urban, P. Kraus, S. Van Calcar, N. Heintzman, T. H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C. K. Glass, M. G. Rosenfeld, S. F. Aldred, S. J. Cooper, A. Halees, J. M. Lin, H. P. Shulha, X. Zhang, M. Xu, J. N. Haidar, Y. Yu, Y. Ruan, V. R. Iyer, R. D. Green, C. Wadelius, P. J. Farnham, B. Ren, R. A. Harte, A. S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A. S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R. M. Kuhn, D. Karolchik, L. Armengol, C. P. Bird, P. I. de Bakker, A. D. Kern, N. Lopez-Bigas, J. D. Martin, B. E. Stranger, A. Woodroffe, E. Davydov, A. Dimas, E. Eyra, I. B. Hallgrimsdottir, J. Huppert, M. C. Zody, G. R. Abecasis, X. Estivill, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. A. Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. A. Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson,

M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu and P. J. de Jong (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Bock, C., E. M. Tomazou, A. B. Brinkman, F. Muller, F. Simmer, H. Gu, N. Jager, A. Gnirke, H. G. Stunnenberg and A. Meissner (2010). "Quantitative comparison of genome-wide DNA methylation mapping technologies." Nat Biotechnol **28**(10): 1106-1114.

Bonasio, R., S. Tu and D. Reinberg (2010). "Molecular signals of epigenetic states." Science **330**(6004): 612-616.

Canella, D., V. Praz, J. H. Reina, P. Cousin and N. Hernandez (2010). "Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells." Genome Res **20**(6): 710-721.

Capra, J. A., M. J. Hubisz, D. Kostka, K. S. Pollard and A. Siepel (2013). "A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes." PLoS Genet **9**(8): e1003684.

Castle, J. C. (2011). "SNPs occur in regions with less genomic sequence conservation." PLoS One **6**(6): e20660.

Cock, P. J., C. J. Fields, N. Goto, M. L. Heuer and P. M. Rice (2010). "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants." Nucleic Acids Res **38**(6): 1767-1771.

Cohen, N. M., E. Kenigsberg and A. Tanay (2011). "Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection." Cell **145**(5): 773-786.

Cokus, S. J., S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini and S. E. Jacobsen (2008). "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning." Nature **452**(7184): 215-219.

Consortium, E. P. (2004). "The ENCODE (ENCyclopedia Of DNA Elements) Project." Science **306**(5696): 636-640.

- Consortium, E. P., B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter and M. Snyder (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.
- Cooper, D. N., M. H. Taggart and A. P. Bird (1983). "Unmethylated domains in vertebrate DNA." Nucleic Acids Res **11**(3): 647-658.
- Cooper, G. M., E. A. Stone, G. Asimenos, N. C. S. Program, E. D. Green, S. Batzoglou and A. Sidow (2005). "Distribution and intensity of constraint in mammalian genomic sequence." Genome Res **15**(7): 901-913.
- Chen, P. Y., S. J. Cokus and M. Pellegrini (2010). "BS Seeker: precise mapping for bisulfite sequencing." BMC Bioinformatics **11**: 203.
- Davydov, E. V., D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow and S. Batzoglou (2010). "Identifying a high fraction of the human genome to be under selective constraint using GERP++." PLoS Comput Biol **6**(12): e1001025.
- Dindot, S. V., R. Person, M. Strivens, R. Garcia and A. L. Beaudet (2009). "Epigenetic profiling at mouse imprinted gene clusters reveals novel epigenetic and genetic features at differentially methylated regions." Genome Research **19**(8): 1374-1383.
- Doi, A., I. H. Park, B. Wen, P. Murakami, M. J. Aryee, R. Irizarry, B. Herb, C. Ladd-Acosta, J. Rho, S. Loewer, J. Miller, T. Schlaeger, G. Q. Daley and A. P. Feinberg (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**(12): 1350-1353.
- Duret, L. and N. Galtier (2009). "Biased gene conversion and the evolution of mammalian genomic landscapes." Annu Rev Genomics Hum Genet **10**: 285-311.
- Eden, E., R. Navon, I. Steinfeld, D. Lipson and Z. Yakhini (2009). "GORilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists." BMC Bioinformatics **10**: 48.
- Ewing, B. and P. Green (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Res **8**(3): 186-194.

Ewing, B., L. Hillier, M. C. Wendl and P. Green (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." Genome Res **8**(3): 175-185.

Frith, M. C., R. Mori and K. Asai (2012). "A mostly traditional approach improves alignment of bisulfite-converted DNA." Nucleic Acids Res **40**(13): e100.

Frommer, M., L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy and C. L. Paul (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**(5): 1827-1831.

Gardiner-Garden, M. and M. Frommer (1987). "CpG islands in vertebrate genomes." J Mol Biol **196**(2): 261-282.

Geisen, S., G. Barturen, A. M. Alganza, M. Hackenberg and J. L. Oliver (2014). "NGSmethDB: an updated genome resource for high quality, single-cytosine resolution methylomes." Nucleic Acids Res **42**(1): D53-59.

Glass, J. L., R. F. Thompson, B. Khulan, M. E. Figueroa, E. N. Olivier, E. J. Oakley, G. Van Zant, E. E. Bouhassira, A. Melnick, A. Golden, M. J. Fazzari and J. M. Greally (2007). "CG dinucleotide clustering is a species-specific property of the genome." Nucleic Acids Res **35**(20): 6798-6807.

Gottlieb, G. (1991). "Experiential canalization of behavioral development: Results." Developmental Psychology **27**(1): 35-39.

Grunau, C., E. Renault, A. Rosenthal and G. Roizes (2001). "MethDB--a public database for DNA methylation data." Nucleic Acids Res **29**(1): 270-274.

Gu, F., M. S. Doderer, Y. W. Huang, J. C. Roa, P. J. Goodfellow, E. L. Kizer, T. H. Huang and Y. Chen (2013). "CMS: a web-based system for visualization and analysis of genome-wide methylation data of human cancers." PLoS One **8**(4): e60980.

Gu, H., C. Bock, T. S. Mikkelsen, N. Jager, Z. D. Smith, E. Tomazou, A. Gnirke, E. S. Lander and A. Meissner (2010). "Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution." Nat Methods **7**(2): 133-136.

- Hackenberg, M., G. Barturen, P. Carpena, P. L. Luque-Escamilla, C. Previti and J. L. Oliver (2010). "Prediction of CpG-island function: CpG clustering vs. sliding-window methods." BMC Genomics **11**: 327.
- Hackenberg, M., G. Barturen and J. L. Oliver (2011). "NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data." Nucleic Acids Res **39**(Database issue): D75-79.
- Hackenberg, M., G. Barturen and J. L. Oliver (2012). DNA methylation Profiling from High-Throughput Sequencing Data. DNA Methylation, InTech - Open Access Publisher, ISBN 979-953-307-453-4. **in press**.
- Hackenberg, M., P. Carpena, P. Bernaola-Galvan, G. Barturen, A. M. Alganza and J. L. Oliver (2011). "WordCluster: detecting clusters of DNA words and genomic elements." Algorithms Mol Biol **6**: 2.
- Hackenberg, M., C. Previti, P. L. Luque-Escamilla, P. Carpena, J. Martinez-Aroza and J. L. Oliver (2006). "CpGcluster: a distance-based algorithm for CpG-island detection." BMC Bioinformatics **7**(1): 446.
- Hackenberg, M., N. Rodriguez-Ezpeleta and A. M. Aransay (2011). "miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments." Nucleic Acids Res **39**(Web Server issue): W132-138.
- Hackenberg, M., A. Rueda, P. Carpena, P. Bernaola-Galvan, G. Barturen and J. L. Oliver (2012). "Clustering of DNA words and biological function: a proof of principle." J Theor Biol **297**: 127-136.
- Hackenberg, M., M. Sturm, D. Langenberger, J. M. Falcon-Perez and A. M. Aransay (2009). "miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments." Nucleic Acids Res **37**(Web Server issue): W68-76.
- Hach, F., F. Hormozdiari, C. Alkan, I. Birol, E. E. Eichler and S. C. Sahinalp (2010). "mrsFAST: a cache-oblivious algorithm for short-read mapping." Nat Methods **7**(8): 576-577.
- Hamosh, A., A. F. Scott, J. S. Amberger, C. A. Bocchini and V. A. McKusick (2005). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders." Nucleic Acids Res **33**(Database issue): D514-517.

Hansen, K. D., B. Langmead and R. A. Irizarry (2012). "BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions." Genome Biol **13**(10): R83.

Harris, E. Y., N. Ponts, K. G. Le Roch and S. Lonardi (2012). "BRAT-BW: efficient and accurate mapping of bisulfite-treated reads." Bioinformatics **28**(13): 1795-1796.

Harris, R. A., T. Wang, C. Coarfa, R. P. Nagarajan, C. Hong, S. L. Downey, B. E. Johnson, S. D. Fouse, A. Delaney, Y. Zhao, A. Olshen, T. Ballinger, X. Zhou, K. J. Forsberg, J. Gu, L. Echipare, H. O'Geen, R. Lister, M. Pelizzola, Y. Xi, C. B. Epstein, B. E. Bernstein, R. D. Hawkins, B. Ren, W. Y. Chung, H. Gu, C. Bock, A. Gnirke, M. Q. Zhang, D. Haussler, J. R. Ecker, W. Li, P. J. Farnham, R. A. Waterland, A. Meissner, M. A. Marra, M. Hirst, A. Milosavljevic and J. F. Costello (2010). "Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications." Nat Biotechnol **28**(10): 1097-1105.

He, X., S. Chang, J. Zhang, Q. Zhao, H. Xiang, K. Kusonmano, L. Yang, Z. S. Sun, H. Yang and J. Wang (2008). "MethyCancer: the database of human DNA methylation and cancer." Nucleic Acids Res **36**(Database issue): D836-841.

Hellman, A. and A. Chess (2007). "Gene body-specific methylation on the active X chromosome." Science **315**(5815): 1141-1143.

Heyn, H., H. J. Ferreira, L. Bassas, S. Bonache, S. Sayols, J. Sandoval, M. Esteller and S. Larriba (2012). "Epigenetic disruption of the PIWI pathway in human spermatogenic disorders." PLoS One **7**(10): e47892.

Hodges, E., A. Molaro, C. O. Dos Santos, P. Thekkat, Q. Song, P. J. Uren, J. Park, J. Butler, S. Rafii, W. R. McCombie, A. D. Smith and G. J. Hannon (2011). "Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment." Mol Cell **44**(1): 17-28.

Hon, G. C., R. D. Hawkins, O. L. Caballero, C. Lo, R. Lister, M. Pelizzola, A. Valsesia, Z. Ye, S. Kuan, L. E. Edsall, A. A. Camargo, B. J. Stevenson, J. R. Ecker, V. Bafna, R. L. Strausberg, A. J. Simpson and B. Ren (2012). "Global DNA hypomethylation coupled to repressive chromatin domain

formation and gene silencing in breast cancer." Genome Research **22**(2): 246-258.

Hon, G. C., N. Rajagopal, Y. Shen, D. F. McCleary, F. Yue, M. D. Dang and B. Ren (2013). "Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues." Nat Genet **45**(10): 1198-1206.

Horaitis, O., C. C. Talbot, Jr., M. Phommavanh, K. M. Phillips and R. G. Cotton (2007). "A database of locus-specific databases." Nat Genet **39**(4): 425.

Hsieh, F., S. C. Chen and K. Pollard (2009) "A Nearly Exhaustive Search for CpG Islands on Whole Chromosomes." The International Journal of Biostatistics **5**, 1-24.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik and M. Clamp (2002). "The Ensembl genome database project." Nucleic Acids Res **30**(1): 38-41.

Hubisz, M. J., K. S. Pollard and A. Siepel (2011). "PHAST and RPHAST: phylogenetic analysis with space/time models." Brief Bioinform **12**(1): 41-51.

Irizarry, R. A., C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J. B. Potash, S. Sabunciyan and A. P. Feinberg (2009). "The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores." Nat Genet **41**(2): 178-186.

Jacquier, A. (2009). "The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs." Nat Rev Genet **10**(12): 833-844.

Jaffe, A. E. and R. A. Irizarry (2014). "Accounting for cellular heterogeneity is critical in epigenome-wide association studies." Genome Biol **15**(2).

Jones, P. A. (2012). "Functions of DNA methylation: islands, start sites, gene bodies and beyond." Nat Rev Genet **13**(7): 484-492.

Jorde, L. B. and S. P. Wooding (2004). "Genetic variation, classification and 'race'." Nat Genet **36**(11 Suppl): S28-33.

Karolchik, D., A. S. Hinrichs and W. J. Kent (2009). "The UCSC Genome Browser." Curr Protoc Bioinformatics **Chapter 1**: Unit1 4.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler and D. Haussler (2002). "The human genome browser at UCSC." Genome Research **12**(6): 996-1006.

Kim, T. H., Z. K. Abdullaev, A. D. Smith, K. A. Ching, D. I. Loukinov, R. D. Green, M. Q. Zhang, V. V. Lobanenko and B. Ren (2007). "Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome." Cell **128**(6): 1231-1245.

Klose, R. J. and A. P. Bird (2006). "Genomic DNA methylation: the mark and its mediators." Trends Biochem Sci **31**(2): 89-97.

Krueger, F. and S. R. Andrews (2011). "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications." Bioinformatics **27**(11): 1571-1572.

Krueger, F., B. Kreck, A. Franke and S. R. Andrews (2012). "DNA methylome analysis using short bisulfite sequencing data." Nat Methods **9**(2): 145-151.

Laird, P. W. (2010). "Principles and challenges of genomewide DNA methylation analysis." Nat Rev Genet **11**(3): 191-203.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczký, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S.

Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissole, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kasprzyk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi and Y. J. Chen (2001). "Initial sequencing and analysis of the human genome." *Nature* **409**(6822): 860-921.

Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." *Genome Biol* **10**(3): R25.

Laurent, L., E. Wong, G. Li, T. Huynh, A. Tsigos, C. T. Ong, H. M. Low, K. W. Kin Sung, I. Rigoutsos, J. Loring and C. L. Wei (2010). "Dynamic changes in the human methylome during differentiation." Genome Res **20**(3): 320-331.

Li, E., C. Beard and R. Jaenisch (1993). "Role for DNA methylation in genomic imprinting." Nature **366**(6453): 362-365.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, S., F. E. Garrett-Bakelman, A. Akalin, P. Zumbo, R. Levine, B. L. To, I. D. Lewis, A. L. Brown, R. J. D'Andrea, A. Melnick and C. E. Mason (2013). "An optimized algorithm for detecting and annotating regional differential methylation." BMC Bioinformatics **14 Suppl 5**: S10.

Li, Y., J. Zhu, G. Tian, N. Li, Q. Li, M. Ye, H. Zheng, J. Yu, H. Wu, J. Sun, H. Zhang, Q. Chen, R. Luo, M. Chen, Y. He, X. Jin, Q. Zhang, C. Yu, G. Zhou, Y. Huang, H. Cao, X. Zhou, S. Guo, X. Hu, X. Li, K. Kristiansen, L. Bolund, J. Xu, W. Wang, H. Yang, J. Wang, R. Li, S. Beck and X. Zhang (2010). "The DNA methylome of human peripheral blood mononuclear cells." PLoS Biol **8**(11): e1000533.

Lienert, F., C. Wirbelauer, I. Som, A. Dean, F. Mohn and D. Schubeler (2011). "Identification of genetic elements that autonomously determine DNA methylation states." Nat Genet **43**(11): 1091-1097.

Lin, X., D. Sun, B. Rodriguez, Q. Zhao, H. Sun, Y. Zhang and W. Li (2013). "BSeQC: quality control of bisulfite sequencing experiments." Bioinformatics **29**(24): 3227-3229.

Lister, R. and J. R. Ecker (2009). "Finding the fifth base: genome-wide sequencing of cytosine methylation." Genome Res **19**(6): 959-966.

Lister, R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar and J. R. Ecker (2008). "Highly integrated single-base resolution maps of the epigenome in Arabidopsis." Cell **133**(3): 523-536.

Lister, R., M. Pelizzola, R. H. Downen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren and J.

R. Ecker (2009). "Human DNA methylomes at base resolution show widespread epigenomic differences." Nature **462**(7271): 315-322.

Lister, R., M. Pelizzola, Y. S. Kida, R. D. Hawkins, J. R. Nery, G. Hon, J. Antosiewicz-Bourget, R. O'Malley, R. Castanon, S. Klugman, M. Downes, R. Yu, R. Stewart, B. Ren, J. A. Thomson, R. M. Evans and J. R. Ecker (2011). "Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells." Nature **471**(7336): 68-73.

Liu, Y., K. D. Siegmund, P. W. Laird and B. P. Berman (2012). "Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data." Genome Biol **13**(7): R61.

Lv, J., H. Liu, J. Su, X. Wu, B. Li, X. Xiao, F. Wang, Q. Wu and Y. Zhang (2012). "DiseaseMeth: a human disease methylation database." Nucleic Acids Res **40**(Database issue): D1030-1035.

Maitra, A., D. E. Arking, N. Shivapurkar, M. Ikeda, V. Stastny, K. Kassaei, G. Sui, D. J. Cutler, Y. Liu, S. N. Brimble, K. Noaksson, J. Hyllner, T. C. Schulz, X. Zeng, W. J. Freed, J. Crook, S. Abraham, A. Colman, P. Sartipy, S. Matsui, M. Carpenter, A. F. Gazdar, M. Rao and A. Chakravarti (2005). "Genomic alterations in cultured human embryonic stem cells." Nat Genet **37**(10): 1099-1103.

Maunakea, A. K., R. P. Nagarajan, M. Bilenky, T. J. Ballinger, C. D'Souza, S. D. Fouse, B. E. Johnson, C. Hong, C. Nielsen, Y. Zhao, G. Turecki, A. Delaney, R. Varhol, N. Thiessen, K. Shchors, V. M. Heine, D. H. Rowitch, X. Xing, C. Fiore, M. Schillebeeckx, S. J. Jones, D. Haussler, M. A. Marra, M. Hirst, T. Wang and J. F. Costello (2010). "Conserved role of intragenic DNA methylation in regulating alternative promoters." Nature **466**(7303): 253-257.

Meyer, L. R., A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, B. J. Raney, A. Pohl, V. S. Malladi, C. H. Li, B. T. Lee, K. Learned, V. Kirkup, F. Hsu, S. Heitner, R. A. Harte, M. Haeussler, L. Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, T. R. Dreszer, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler and W. J. Kent (2013). "The UCSC Genome Browser database: extensions and updates 2013." Nucleic Acids Res **41**(Database issue): D64-69.

Moarefi, A. H. and F. Chedin (2011). "ICF syndrome mutations cause a broad spectrum of biochemical defects in DNMT3B-mediated de novo DNA methylation." J Mol Biol **409**(5): 758-772.

Molaro, A., E. Hodges, F. Fang, Q. Song, W. R. McCombie, G. J. Hannon and A. D. Smith (2011). "Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates." Cell **146**(6): 1029-1041.

Negre, V. and C. Grunau (2006). "The MethDB DAS server: adding an epigenetic information layer to the human genome." Epigenetics **1**(2): 101-105.

Nei, M. and S. Kumar (2000). Molecular evolution and phylogenetics. New York ; Oxford, Oxford University Press.

Oberdoerffer, S. (2012). "A conserved role for intragenic DNA methylation in alternative pre-mRNA splicing." Transcription **3**(3): 106-109.

Olson, S. A. (2002). "EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite." Brief Bioinform **3**(1): 87-91.

Ondov, B. D., C. Cochran, M. Landers, G. D. Meredith, M. Dudas and N. H. Bergman (2010). "An alignment algorithm for bisulfite sequencing using the Applied Biosystems SOLiD System." Bioinformatics **26**(15): 1901-1902.

Pedersen, B., T. F. Hsieh, C. Ibarra and R. L. Fischer (2011). "MethylCoder: software pipeline for bisulfite-treated sequences." Bioinformatics **27**(17): 2435-2436.

Pollard, K. S., M. J. Hubisz, K. R. Rosenbloom and A. Siepel (2010). "Detection of nonneutral substitution rates on mammalian phylogenies." Genome Res **20**(1): 110-121.

Ponger, L. and D. Mouchiroud (2002). "CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences." Bioinformatics **18**(4): 631-633.

Pruitt, K. D., T. Tatusova and D. R. Maglott (2007). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic Acids Res **35**(Database issue): D61-65.

Qu, W., S. Hashimoto, A. Shimada, Y. Nakatani, K. Ichikawa, T. L. Saito, K. Ogoshi, K. Matsushima, Y. Suzuki, S. Sugano, H. Takeda and S. Morishita (2012). "Genome-wide genetic variations are highly correlated with proximal DNA methylation patterns." Genome Res **22**(8): 1419-1425.

Rice, P., I. Longden and A. Bleasby (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-277.

Robertson, K. D. (2005). "DNA methylation and human disease." Nat Rev Genet **6**(8): 597-610.

Robinson, M. D., A. L. Statham, T. P. Speed and S. J. Clark (2010). "Protocol matters: which methylome are you actually studying?" Epigenomics **2**(4): 587-598.

Rosenbloom, K. R., C. A. Sloan, V. S. Malladi, T. R. Dreszer, K. Learned, V. M. Kirkup, M. C. Wong, M. Maddren, R. Fang, S. G. Heitner, B. T. Lee, G. P. Barber, R. A. Harte, M. Diekhans, J. C. Long, S. P. Wilder, A. S. Zweig, D. Karolchik, R. M. Kuhn, D. Haussler and W. J. Kent (2013). "ENCODE data in the UCSC Genome Browser: year 5 update." Nucleic Acids Res **41**(Database issue): D56-63.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander and D. Altshuler (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." Nature **409**(6822): 928-933.

Salser, W. (1978). "Globin mRNA sequences: analysis of base pairing and evolutionary implications." Cold Spring Harb Symp Quant Biol **42 Pt 2**: 985-1002.

Saxonov, S., P. Berg and D. L. Brutlag (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**(5): 1412-1417.

Schultz, M. D., R. J. Schmitz and J. R. Ecker (2012). "'Leveling' the playing field for analyses of single-base resolution DNA methylomes." Trends Genet **28**(12): 583-585.

Schwartz, S., E. Meshorer and G. Ast (2009). "Chromatin organization marks exon-intron structure." Nat Struct Mol Biol **16**(9): 990-995.

Schwartz, S., R. Oren and G. Ast (2011). "Detection and removal of biases in the analysis of next-generation sequencing reads." PLoS One **6**(1): e16685.

Sharp, A. J., E. Stathaki, E. Migliavacca, M. Brahmachary, S. B. Montgomery, Y. Dupre and S. E. Antonarakis (2011). "DNA methylation profiles of human active and inactive X chromosomes." Genome Res **21**(10): 1592-1600.

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nat Biotechnol **26**(10): 1135-1145.

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski and K. Sirotkin (2001). "dbSNP: the NCBI database of genetic variation." Nucleic Acids Res **29**(1): 308-311.

Shi, J., J. Hu, Q. Zhou, Y. Du and C. Jiang (2013). "PEpiD: a prostate epigenetic database in mammals." PLoS One **8**(5): e64289.

Shukla, S., E. Kavak, M. Gregory, M. Imashimizu, B. Shutinoski, M. Kashlev, P. Oberdoerffer, R. Sandberg and S. Oberdoerffer (2011). "CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing." Nature **479**(7371): 74-79.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller and D. Haussler (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." Genome Res **15**(8): 1034-1050.

Skinner, M. E., A. V. Uzilov, L. D. Stein, C. J. Mungall and I. H. Holmes (2009). "JBrowse: a next-generation genome browser." Genome Research **19**(9): 1630-1638.

Smit, A. F. A., R. Hubley and P. Green (1996-2010) "RepeatMasker Open-3.0."

Smith, A. D., W. Y. Chung, E. Hodges, J. Kendall, G. Hannon, J. Hicks, Z. Xuan and M. Q. Zhang (2009). "Updates to the RMAP short-read mapping software." Bioinformatics **25**(21): 2841-2842.

Smith, Z. D. and A. Meissner (2013). "DNA methylation: roles in mammalian development." Nat Rev Genet **14**(3): 204-220.

Stadler, M. B., R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Scholer, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari and D. Schubeler (2011). "DNA-binding factors shape the mouse methylome at distal regulatory regions." Nature **480**(7378): 490-495.

Stein, L. D., C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J. E. Stajich, T. W. Harris, A. Arva and S. Lewis (2002). "The generic genome browser: a building block for a model organism system database." Genome Res **12**(10): 1599-1610.

Suzuki, M. M. and A. Bird (2008). "DNA methylation landscapes: provocative insights from epigenomics." Nat Rev Genet **9**(6): 465-476.

Takai, D. and P. A. Jones (2002). "Comprehensive analysis of CpG islands in human chromosomes 21 and 22." Proc Natl Acad Sci U S A **99**(6): 3740-3745.

Tamura, K., G. Stecher, D. Peterson, A. Filipski and S. Kumar (2013). "MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0." Mol Biol Evol **30**(12): 2725-2729.

Treangen, T. J. and S. L. Salzberg (2012). "Repetitive DNA and next-generation sequencing: computational challenges and solutions." Nat Rev Genet **13**(1): 36-46.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D.

Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferreira, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu (2001). "The sequence of the human genome." *Science* **291**(5507): 1304-1351.

- Visel, A., S. Minovitsky, I. Dubchak and L. A. Pennacchio (2007). "VISTA Enhancer Browser--a database of tissue-specific human enhancers." Nucleic Acids Res **35**(Database issue): D88-92.
- Waddington, C. H. (2012). "The epigenotype. 1942." Int J Epidemiol **41**(1): 10-13.
- Wang, H., M. T. Maurano, H. Qu, K. E. Varley, J. Gertz, F. Pauli, K. Lee, T. Canfield, M. Weaver, R. Sandstrom, R. E. Thurman, R. Kaul, R. M. Myers and J. A. Stamatoyannopoulos (2012). "Widespread plasticity in CTCF occupancy linked to DNA methylation." Genome Res **22**(9): 1680-1688.
- Westesson, O., M. Skinner and I. Holmes (2013). "Visualizing next-generation sequencing data with JBrowse." Brief Bioinform **14**(2): 172-177.
- White, R. J. (2008). "RNA polymerases I and III, non-coding RNAs and cancer." Trends Genet **24**(12): 622-629.
- Wiench, M., S. John, S. Baek, T. A. Johnson, M. H. Sung, T. Escobar, C. A. Simmons, K. H. Pearce, S. C. Biddie, P. J. Sabo, R. E. Thurman, J. A. Stamatoyannopoulos and G. L. Hager (2011). "DNA methylation status predicts cell type-specific enhancer activity." EMBO J **30**(15): 3028-3039.
- Wong, N. C., L. H. Wong, J. M. Quach, P. Canham, J. M. Craig, J. Z. Song, S. J. Clark and K. H. Choo (2006). "Permissive transcriptional activity at the centromere through pockets of DNA hypomethylation." PLoS Genet **2**(2): e17.
- Wu, H., B. Caffo, H. A. Jaffee, R. A. Irizarry and A. P. Feinberg (2010). "Redefining CpG islands using hidden Markov models." Biostatistics **11**(3): 499-514.
- Xi, Y. and W. Li (2009). "BSMAP: whole genome bisulfite sequence MAPping program." BMC Bioinformatics **10**: 232.
- Xin, Y., B. Chanrion, A. H. O'Donnell, M. Milekic, R. Costa, Y. Ge and F. G. Haghghi (2012). "MethylomeDB: a database of DNA methylation profiles of the brain." Nucleic Acids Res **40**(Database issue): D1245-1249.

Yoder, J. A., C. P. Walsh and T. H. Bestor (1997). "Cytosine methylation and the ecology of intragenomic parasites." Trends Genet **13**(8): 335-340.

Zeng, J., G. Konopka, B. G. Hunt, T. M. Preuss, D. Geschwind and S. V. Yi (2012). "Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution." Am J Hum Genet **91**(3): 455-465.

Zentner, G. E., P. J. Tesar and P. C. Scacheri (2011). "Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions." Genome Res **21**(8): 1273-1283.

Zhang, Y., H. Liu, J. Lv, X. Xiao, J. Zhu, X. Liu, J. Su, X. Li, Q. Wu, F. Wang and Y. Cui (2011). "QDMR: a quantitative method for identification of differentially methylated regions by entropy." Nucleic Acids Res **39**(9): e58.

Zhu, J., F. He, S. Hu and J. Yu (2008). "On the nature of human housekeeping genes." Trends Genet **24**(10): 481-484.

Ziller, M. J., H. Gu, F. Muller, J. Donaghey, L. T. Tsai, O. Kohlbacher, P. L. De Jager, E. D. Rosen, D. A. Bennett, B. E. Bernstein, A. Gnirke and A. Meissner (2013). "Charting a dynamic DNA methylation landscape of the human genome." Nature **500**(7463): 477-481.