UNIVERSIDAD
DE GRANADA

TESIS DOCTORAL

Programa de Doctorado en Biología
Fundamental y de Sistemas

Metilación diferencial en el genoma humano y su asociación con la transcripción

Ricardo Lebrón Aguilar

El doctorando, **Ricardo Lebrón Aguilar**, y los directores de la Tesis, **José L. Oliver Jiménez** y **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, 25 de marzo de 2019

Directores de la Tesis

Fdo.: José L. Oliver Jiménez

OLIVER JIMENEZ JOSE LUTGARDO - 74588605L

Digitally signed by OLIVER JIMENEZ JOSE LUTGARDO - 74588605L
Date: 2019.03.25 10:40:31 +01'00'

Fdo.: Michael Hackenberg

HACKENBERG MICHAEL - X4798161Q

Firmado digitalmente por HACKENBERG MICHAEL - X4798161Q
Fecha: 2019.03.25 17:31:40 +01'00'

Doctorando

Fdo.: Ricardo Lebrón Aguilar

LEBRON AGUILAR RICARDO - 79221253F

Firmado digitalmente por LEBRON AGUILAR RICARDO - 79221253F
Fecha: 2019.03.25 08:31:53 +01'00'

"The important thing in science

is not so much to obtain new facts

as to discover new ways of thinking about them"

Sir William Lawrence Bragg

Agradecimientos

Hay tantas personas a las que me gustaría agradecer tantas cosas, que tendría que escribir un libro solo para eso. Así que me gustaría comenzar estos agradecimientos con una disculpa a aquellas personas a las que no incluya aquí o a las que incluya, pero mi gratitud no haga justicia a sus actos. Estoy seguro de que muchas personas que han pasado por mi vida han dejado huella en mí y merecerían ser mencionadas.

En primer lugar, me gustaría agradecer a mis Directores de Tesis Doctoral, José L. Oliver y Michael Hackenberg, por todo el esfuerzo y paciencia que me han dedicado, así como por la confianza que siempre han depositado en mí. Gracias también por ser siempre tan familiares y cercanos conmigo. También me gustaría extender mi gratitud a mis compañeros de grupos, Cristina, Ernesto, José Mª, Stavros, Nicolas y David. En estos años de convivencia hemos compartido muchos buenos momentos dignos de recordar. Quiero agradecer especialmente a Ernesto su apoyo en estos últimos meses que tan duros me han resultado. También quiero agradecer a las personas que ya no están en el grupo, pero con las que compartí parte de esta experiencia que es nacer y crecer como investigador. Gracias Fran, gracias Guillermo, gracias Antonio. A Guillermo quiero agradecerle que siempre haya sido como un hermano mayor y como un mentor para mí.

Toda mi gratitud al Departamento de Genética de la Universidad de Granada. Han sido siempre muy acogedores y familiares conmigo. Son personas estupendas con las que merece la pena cruzarse de vez en cuando y compartir desde tu curiosidad científica hasta tus inquietudes personales.

Una mención muy distinguida merece Miguel Burgos. Por más que me esforzara nunca podría recompensar todo lo que ha hecho por mí. Gracias por apoyarme siempre y acogerme como a uno más en tu familia. Eres mucho más que un tutor o un compañero para mí. Gracias también Rosie Burgos, por su cálida amistad, su apoyo y por echarme una mano con las partes en inglés de esta memoria. Y cómo no, quiero agradecer a Paco Barrionuevo su cariño y preocupación hacia mí. Sé que le he dado muchos quebraderos de cabeza, pero él siempre me ha tendido la mano cuando más lo necesitaba.

Gracias a mi familia, en especial a mis padres y a mi hermano. Siempre me han apoyado, a pesar de que mis inquietudes y aspiraciones siempre se hayan interpuesto. Lamento no haberos dedicado suficiente tiempo, pero prometo que eso va a cambiar.

A mis amigos, en especial a Carmen y a Fran, les quiero agradecer que hayan compartido conmigo este largo camino. Estos 12 últimos años han cambiado drásticamente lo que suponía que iba a ser mi vida y sé que sin vosotros nunca hubiese alcanzado mis objetivos. Si hoy he llegado hasta aquí es, sin duda, gracias a vosotros.

Me he reservado lo mejor para el final. La persona a la que más quiero es también la persona que más me ha apoyado siempre. Mi pareja siempre ha estado ahí, en los buenos y en los malos momentos. Sé que no he podido prestarle la atención que se merece en estos últimos años y también sé que en más de una ocasión sus ilusiones se han hecho pedazos por estar demasiado ocupado. Hemos pasado momentos muy duros, pero sé que nos esperan otros muy dulces. Te quiero.

Abstract

A human being is composed of more than 400 cell types, which differ in the specific set of genes they transcribe, despite having the same genomic sequence. The differences between cell types lie in the specific epigenetic information accompanying the genome and in the transcription factors present in the cell.

In adult human cells, cytosine methylation occurs primarily at CpG sites and is probably the most important epigenetic mark, as it contributes to transcription regulation while remaining stable throughout the cell lineage, and changing during cell fate establishment. According to the traditional paradigm, methylation in the promoter is associated with the repression of transcription, although there are cases in which it is associated with the activation of transcription or in which transcription is independent of methylation. On the other hand, the effect of methylation on transcription regulation is not limited to promoters, but also to other regions such as enhancers and the gene body are also involved.

For more than a decade, it has been possible to detect the level of methylation of each cytosine in the genome, thanks to the emergence of a mass sequencing technique known as Whole-Genome Bisulfite Sequencing (WGBS). However, there are many sources of error that affect the reliability of the results obtained, causing erroneous detections in the methylation level of some cytosines and even loss of information in certain regions. As a response to these problems, many researchers choose to average the methylation levels of CpG sites within the regions of interest, assuming that errors will compensate for each other and therefore sacrificing the high resolution this technique offers.

Nevertheless, the average methylation of a region is not always relevant and can even lead to erroneous conclusions. It has recently been described that only 16.6% of CpG sites on promoters have an effect upon transcription when their methylation changes. This evidences the need to develop methods that allow a more reliable detection of the methylation levels of each cytosine.

The first objective of this Doctoral Thesis was to design and implement a protocol for obtaining methylation maps, from WGBS reads, in an attempt to solve all known problems: i) eliminating low quality positions or those that have been entered during library preparation, as well as duplicate reads, ii) correcting problems arising from the alignment of reads, iii) discarding positions and reads affected by bias in methylation and iv) distinguishing between C/T substitutions and non-methylated cytosines.

During the development of this protocol, a type of bias caused by the use of new genomic assembly models was discovered. The last two versions of the human genome assembly include alternative haplotypes, which attempt to collect structural and sequence variations from different human populations or ethnicities, in order to prevent reads from these haplotypes from misaligning in other regions of the genome. However, it has not been evaluated whether this inclusion might cause any problems.

In this Doctoral Thesis, it is described for the first time that the use of the new assembly models causes the loss of reads from polymorphic loci as a consequence of an increase in the percentage of reads with ambiguous alignment. To recover these reads and assign them to the consensus assembly, a two-stage alignment strategy was designed: i) all reads face full assembly and, ii) those whose alignment has been proved ambiguous during the first stage are confronted with a version of the assembly without alternative haplotypes. Finally, the unique-alignment reads from both alignments are brought together and will be used in later stages of the protocol.

Once the protocol was mature enough, it was decided to implement it as an open-source program, which received the name of MethFlow. The workflow of this program starts from WGBS reads in FASTQ format and ends with the obtaining of methylation maps after going through several stages which deal with biases and contaminations using third-party programs combined with our own code. The most important stages are those in the two-stage alignment, in which Bismark is used following the strategy described above, and the detection of methylation levels from corrected alignments by using MethylExtract because it is capable of distinguishing C/T substitutions of non-methylated cytosines.

One of the major problems that the scientific community faces today is the lack of reproducibility of results. To ensure this reproducibility, the MethFlow architecture was designed based on: i) containers generated from a configuration file, which indicates the version of each program, its installation process and configuration, and ii) a sophisticated framework for complex pipelines, providing comprehensive control and a thorough record of the executed processes. Finally, MethFlow was provided with a modular structure, so that later modules could be added to perform related tasks, such as analyzing changes in methylation or its association with transcription.

Once a suitable tool was available to study the methylation levels of individual cytosines, it was hypothesized that, depending on the genomic context in which it occurs and the type of transcription factors involved, methylation may contribute to the positive or negative regulation of transcription or have no effect.

To prove this hypothesis it was necessary to: i) obtain a collection of human methylation maps that would collect as many cell types and individuals as possible, ii) characterize the differences in methylation due to cell type and individual and, iii) study the association with transcription of methylation changes in individual CpG sites and their possible impact on regulatory elements of transcription.

The Roadmap Epigenomics, ENCODE and Enhancing GTEx projects have public sets of WGBS reads for a wide range of human samples. Using MethFlow, methylation maps for 86 human samples from 52 cell types of 29 individuals were obtained. From 51 of the 86 samples, transcription profiles were also obtained through ENCODE DATA. These methylation maps and transcription profiles were fundamental in characterizing methylation changes in the human genome and their association with transcription.

Each cell type has a characteristic methylation pattern, partly inherited from the stem cell that precedes it in its lineage and partly modified during the cell differentiation process. Similarly, the same cell type may have certain differences in methylation between individuals due to genetic and environmental factors. Both types of variability in methylation can be expected to have different biological implications.

To study the variability of methylation, samples were compared in pairs and then those changes in methylations that were characteristic of the cell type or the individual were chosen. A method for detecting Differentially Methylated CpGs (DMCs) based on the Fisher's Exact Test was developed and incorporated into MethFlow as a module. Two types of DMCs were then defined: i) intra-individual DMCs, whose methylation varies between different cell types of the same individual, and ii) inter-individual DMCs, whose methylation varies between

individuals for a given cell type. Once as many sets of DMCs as pairs could be formed following these two definitions, strict sets of intra-individual DMCs and inter-individual DMCs were defined: i) for each sample, those DMCs common to all their peer comparisons (intra-individual or inter-individual, as appropriate) were selected and ii) all the selected DMCs were brought together in a single set.

It was then necessary to design a method to study enrichment in DMCs of a given set of genomic elements. Since the distribution of CpG sites in the genome is not random, enrichment was defined as the ratio between the percentage of CpG sites that are DMCs within the set of genomic elements and the percentage of CpG sites that are DMCs outside the genome.

After applying these methods and definitions to previously obtained methylation maps, it was found that 3,303,077 (12.19%) and 329,974 (1.22%) of the CpG sites of the human genome are, respectively, intra-individual DMC and inter-individual DMC. The main genomic elements related to the regulation of transcription (promoters, enhancers and transcription factors binding sites) do not show remarkable differences in intra-individual DMCs and inter-individual DMCs. However, open chromatin regions were found to be enriched in intra-individual DMCs, but impoverished in inter-individual DMCs. DMCs are under-represented in promoters, while they are over-represented in enhancers, suggesting that most methylation changes (both between cell types and between individuals) occur in enhancers. Transcription factors binding sites are also enriched in DMCs, regardless of the type of transcription factor involved. On the other hand, the proportion of DMCs decreases as the distance to the nearest transcription start site decreases, and it increases as the distance to the nearest transcription end site decreases.

As it was previously mentioned, only 16.6% of CpG sites on promoters have an effect on transcription when their methylation changes. Recently, so-called "CpG traffic lights" (CpG-TLs) have been described, which are individual CpG sites whose level of methylation is associated with the transcription rate of a nearby gene. These biological markers are well-suited to test the hypothesis which

suggests that the sign of the association between methylation and transcription depends on the genomic context in which methylation occurs and the type of transcription factors involved.

Other authors had detected CpG-TLs in the human genome, using the Spearman's correlation coefficient and selecting only those results with negative association. However, this test is sensitive to outliers. In order to reduce this problem and increase the reliability of the results, in this Doctoral Thesis a method to detect CpG-TLs was developed using a combination of the Spearman's correlation coefficient and the Kruskal-Wallis test. Two classes of CpG-TLs were also distinguished: i) reds, when the association is negative, and ii) greens, when the association is positive. This method is available as a MethFlow module.

After applying these methods and definitions to previously obtained methylation maps and transcription profiles, it was found that the number of green CpG-TLs is almost twice the number of red CpG-TLs: 126,959 (0.49%) and 66,746 (0.26%), respectively, on the CpG sites of the human genome. Red and green CpG-TLs are both over-represented in promoters and enhancers. This suggests that both have mechanisms to activate or repress transcription via methylation, probably due to different combinations of transcription factors binding sites. In sites recognized by transcription factors with greater affinity for non-methylated sites, both red and green CpG-TLs are over-represented. On the contrary, in sites recognized by transcription factors with greater affinity for methylated sites, red CpG-TLs are under-represented while green CpG-TLs are over-represented. This second type of transcription factors are fundamental in mammalian development and some are even able to recruit enzymes that remodel methylation. In terms of their distribution around genes, the proportion of green CpG-TLs decreases as the distance to the transcription starting site is reduced, while the proportion of red CpG-TLs increases.

The NGSmethDB methylation database contains an extensive collection of methylation maps for different species, cell types and individuals. In order to optimize the storage and consultation of the large volume of data produced throughout this Doctoral Thesis, including methylation, DMCs and CpG-TLs maps, it was decided to completely redesign this database. To accelerate comparisons between samples, it was decided to migrate the data to the MongoDB database system and store them in a hierarchical structure of JSON documents (a standard format that allows exchanging tagged and hierarchical data between different programming languages), where: i) each assembly has its own database, ii) each chromosome has its own collection of JSON documents, iii) each CpG site has its own JSON document and, iv) each sub-document contains a type of biological information (methylation, differential methylation or association with transcription). In the case of methylation maps, each sub-document is divided into three levels: i) the individual, ii) the sample and, iii) the type of data. Several ways of access, comparison and visualization of the data contained in the NGSmethDB were implemented, among which the following stand out: i) its programmatic access through the HTTPS protocol through a RESTful API server and ii) its connectivity to UCSC Genome Browser through Track Hubs.

In this Doctoral Thesis the reliability in detecting the methylation levels of individual cytosines from WGBS reads has been significantly improved, taking into account all the sources of error known today. This has allowed to test the hypothesis which argues that the sign of the association between methylation and transcription depends on the genomic context in which methylation occurs and the type of transcription factors that are involved. In the light of the results obtained, it has not been possible to refute this hypothesis. An unexpected finding has been that the positive association between methylation and transcription appears to be more frequent than it had been previously described, becoming even more frequent than cases with negative association. In relation to this, in transcription factors binding sites with greater affinity for methylated sites, CpG-TLs green are over-represented, but CpG-TLs red are under-represented. These positive associations may be due to a hitherto unknown transcription regulation

mechanism, but there are also likely to be cases where hydroxymethylation is positively associated with transcription, as the WGBS technique is unable to discriminate between methylation and hydroxymethylation. In further studies, OxBS-seq or TAB-seq techniques should be used in order to clarify the true nature of green CpG-TLs.

Keywords:

methylation, differential methylation, transcription, transcription factors, CpG traffic-lights

Resumen

Un ser humano se compone de más de 400 tipos celulares, los cuales difieren en el conjunto específico de genes que transcriben, pese a tener la misma secuencia genómica. Las diferencias entre tipos celulares radican en la información epigenética específica que acompaña al genoma y en los factores de transcripción presentes en la célula.

En células humanas adultas, la metilación de la citosina ocurre fundamentalmente en sitios CpG y es probablemente la marca epigenética más importante, ya que contribuye a la regulación de la transcripción, se mantiene estable a lo largo del linaje celular y se modifica durante el establecimiento del destino celular. Según el paradigma tradicional, la metilación en el promotor está asociada con la represión de la transcripción, si bien existen casos en los que se asocia con la activación de la transcripción o en los que la transcripción es independiente de la metilación. Por otra parte, el efecto de la metilación sobre la regulación de la transcripción no se

limita a los promotores, sino que otras regiones como los potenciadores y el cuerpo génico también están implicadas.

Desde hace más de una década, es posible detectar el nivel de metilación de cada citosina del genoma, gracias a la aparición de una técnica de secuenciación masiva conocida como Whole-Genome Bisulfite Sequencing (WGBS). Sin embargo, existen fuentes de error que afectan a la fiabilidad de los resultados obtenidos, provocando detecciones erróneas en el nivel de metilación de algunas citosinas e incluso pérdidas de información en ciertas regiones. Debido a estos problemas, muchos investigadores optan por promediar los niveles de metilación de los sitios CpG dentro de las regiones de interés, suponiendo que los errores se compensarán entre sí y sacrificando la alta resolución que ofrece esta técnica.

Sin embargo, la metilación promedio de una región no siempre es relevante e incluso puede llevar a conclusiones erróneas. Se ha descrito recientemente que solo el 16,6% de los sitios CpG en promotores ejercen un efecto sobre la transcripción cuando cambia su metilación. Esto pone de manifiesto la necesidad de desarrollar métodos que permitan una detección lo más fiable posible de los niveles de metilación de cada citosina.

Se estableció como primer objetivo de esta Tesis Doctoral diseñar e implementar un protocolo de obtención de mapas de metilación, a partir de lecturas de WGBS, para intentar resolver todos los problemas conocidos: i) eliminando posiciones con baja calidad o que se han introducido durante la preparación de la biblioteca, así como lecturas duplicadas, ii) corrigiendo problemas derivados del alineamiento de las lecturas, iii) descartando posiciones y lecturas afectadas por sesgos en la metilación y iv) distinguiendo entre sustituciones C/T y citosinas no metiladas.

Durante el desarrollo de este protocolo, se descubrió un tipo de sesgo ocasionado por el uso de nuevos modelos de ensamblado genómico. Las dos últimas versiones del ensamblado del genoma humano incluyen haplotipos alternativos, que tratan de recoger las variaciones estructurales y de secuencia de distintas poblaciones o etnias humanas, para evitar que las lecturas procedentes de estos haplotipos

alineen incorrectamente en otras regiones del genoma. Sin embargo, no se había evaluado si esta inclusión podría acarrear algún problema.

En esta Tesis Doctoral, se describe por primera vez que el uso de los nuevos modelos de ensamblado provoca la pérdida de lecturas procedentes de loci polimórficos, como consecuencia de un incremento en el porcentaje de lecturas con alineamiento ambiguo. Para recuperar estas lecturas y asignarlas al ensamblado consenso, se diseñó una estrategia de alineamiento en dos etapas: i) todas las lecturas se enfrentan al ensamblado completo y ii) aquellas cuyo alineamiento ha resultado ambiguo durante la primera etapa se enfrentan a una versión del ensamblado sin haplotipos alternativos. Finalmente, se reúnen las lecturas con alineamiento único procedentes de ambos alineamientos, las cuales se utilizarán en posteriores etapas del protocolo.

Una vez el protocolo estuvo maduro, se decidió implementarlo como un programa de código abierto, que recibió el nombre de MethFlow. El flujo de trabajo de este programa parte de lecturas de WGBS en formato FASTQ y finaliza con la obtención de mapas de metilación, atravesando por diversas etapas de tratamiento de sesgos y contaminaciones en las que se utilizan programas de terceros combinados con código propio. Las etapas más importantes son el alineamiento en dos etapas, en la que se utiliza Bismark siguiendo la estrategia antes descrita, y la detección de los niveles de metilación a partir de los alineamientos corregidos, en la que se utiliza MethylExtract por ser capaz de distinguir sustituciones C/T de citosinas no-metiladas.

Uno de los mayores problemas a los que se enfrenta hoy en día la comunidad científica es la falta de reproducibilidad de los resultados. Para garantizar esta reproducibilidad, la arquitectura de MethFlow se diseñó con base en: i) contenedores generados a partir de un fichero de configuración, en el que se indica la versión de cada programa, su proceso de instalación y de configuración, y ii) un sofisticado framework para pipelines complejas, que ofrece un control y un

registro exhaustivo de los procesos ejecutados. Por último, se dotó a MethFlow de una estructura modular, de manera que más tarde se pudieran añadir módulos que desempeñen tareas relacionadas, como analizar cambios en la metilación o su asociación con la transcripción.

Una vez se dispuso de una herramienta adecuada para estudiar los niveles de metilación de citosinas individuales, se planteó la hipótesis de que, dependiendo del contexto genómico en que se produzca y del tipo de factores de transcripción que intervengan, la metilación puede contribuir a la regulación positiva o negativa de la transcripción o no tener efecto. Para poner a prueba esta hipótesis fue necesario: i) obtener una colección de mapas de metilación humanos que recogiese el mayor número posible de tipos celulares y de individuos, ii) caracterizar las diferencias de metilación debidas al tipo celular y al individuo y iii) estudiar la asociación con la transcripción de los cambios de metilación en sitios CpG individuales y su posible impacto sobre elementos reguladores de la transcripción.

Los proyectos Roadmap Epigenomics, ENCODE y Enhancing GTEx disponen de conjuntos públicos de lecturas de WGBS para un amplio abanico de muestras humanas. Utilizando MethFlow, se obtuvieron los mapas de metilación para 86 muestras humanas, procedentes de 52 tipos celulares de 29 individuos. De 51 de las 86 muestras se obtuvieron también los perfiles de transcripción a través de ENCODE DATA. Estos mapas de metilación y perfiles de transcripción resultaron fundamentales para caracterizar los cambios de metilación en el genoma humano y su asociación con la transcripción.

Cada tipo celular posee un patrón de metilación característico, en parte heredado de la célula madre que le precede en su linaje y en parte modificado durante el proceso de diferenciación celular. De igual manera, un mismo tipo celular puede presentar ciertas diferencias de metilación entre individuos debido a factores genéticos y ambientales. Cabe esperar que ambos tipos de variabilidad en la metilación tengan distintas implicaciones biológicas.

Para estudiar la variabilidad de la metilación, se decidió seguir una estrategia de comparación de las muestras por pares y posteriormente seleccionar aquellos cambios de metilación que fuesen característicos del tipo celular o del individuo. Se desarrolló un método de detección de CpGs diferencialmente metilados (DMCs) basado en el test exacto de Fisher y se incorporó a MethFlow como módulo. A continuación, se definieron dos tipos de DMCs: i) DMCs intra-individuales, cuya metilación varía entre diferentes tipos celulares de un mismo individuo, y ii) DMCs inter-individuales, cuya metilación varía entre individuos para un tipo celular dado. Una vez obtenidos tantos conjuntos de DMCs como parejas fue posible formar siguiendo estas dos definiciones, se definieron sendos conjuntos estrictos de DMCs intra-individuales y DMCs inter-individuales: i) para cada muestra, se seleccionaron aquellos DMCs comunes a todas sus comparaciones por pares (intra-individuales o inter-individuales, según proceda) y ii) se reunieron en un único conjunto todos los DMCs seleccionados.

A continuación, fue necesario diseñar un método para estudiar la riqueza en DMCs de un conjunto de elementos genómicos dado. Puesto que la distribución de los sitios CpG en el genoma no es aleatoria, se definió la riqueza como el cociente entre el porcentaje de sitios CpG que son DMCs dentro del conjunto de elementos genómicos y el porcentaje de sitios CpG que son DMCs fuera del mismo.

Tras aplicar estos métodos y definiciones a los mapas de metilación previamente obtenidos, se encontró que 3.303.077 (12,19%) y 329.974 (1,22%) de los sitios CpG del genoma humano son, respectivamente, DMC intra-individuales y DMC inter-individuales. Los principales elementos genómicos relacionados con la regulación de la transcripción (promotores, potenciadores y sitios de unión a factores de transcripción) no presentan diferencias destacables en DMCs intra-individuales y DMCs inter-individuales. Sin embargo, se encontró que las regiones de cromatina abierta están enriquecidas en DMCs intra-individuales, pero empobrecidas en

DMCs inter-individuales. Los promotores son pobres en DMCs, mientras que los potenciadores son ricos, lo cual sugiere que la mayoría de cambios de metilación (tanto entre tipos celulares como entre individuos) ocurren en potenciadores. Los sitios de unión a factores de transcripción también son ricos en DMCs, independientemente del tipo de factor de transcripción del que se trate. Por otra parte, la proporción de DMCs disminuye a medida que decrece la distancia al sitio de inicio de la transcripción más próximo y aumenta a medida que decrece la distancia al sitio de fin de la transcripción más próximo.

Como ya se ha mencionado, solo el 16,6% de los sitios CpG en promotores ejercen un efecto sobre la transcripción cuando cambia su metilación. Recientemente, se han descrito los llamados "semáforos CpG" (CpG-TLs), los cuales son sitios CpG individuales cuyo nivel de metilación está asociado con la tasa de transcripción de un gen cercano. Estos marcadores biológicos son muy adecuados para poner a prueba la hipótesis de que el signo de la asociación entre la metilación y la transcripción depende del contexto genómico en que se produce la metilación y del tipo de factores de transcripción implicados.

Otros autores habían detectado CpG-TLs en el genoma humano, utilizando el coeficiente de correlación de Spearman y seleccionando solo aquellos resultados con asociación negativa. Sin embargo, este test es sensible a los valores atípicos. Para reducir este problema y aumentar la fiabilidad de los resultados, en esta Tesis Doctoral se desarrolló un método de detección de CpG-TLs utilizando una combinación del coeficiente de correlación de Spearman y el test de Kruskal-Wallis. También se distinguieron dos clases de CpG-TLs: i) rojos, cuando la asociación es negativa, y ii) verdes, cuando la asociación es positiva. Este método está disponible como módulo de MethFlow.

Tras aplicar estos métodos y definiciones a los mapas de metilación y perfiles de transcripción previamente obtenidos, se encontró que la cantidad de CpG-TLs verdes es casi el doble que la de los CpG-TLs rojos: 126.959 (0,49%) y 66.746 (0,26%), respectivamente, de los sitios CpG del genoma humano. Los promotores y potenciadores son ricos en CpG-TLs, tanto rojos como verdes. Esto sugiere que ambos disponen de mecanismos para activar o reprimir la transcripción vía metilación, probablemente debido a diferentes combinaciones de sitios de unión a factores de transcripción. Mientras que los sitios de unión a factores de transcripción con mayor afinidad por sitios no metilados son ricos en CpG-TLs rojos y verdes, los sitios de unión a factores de transcripción con mayor afinidad por sitios metilados son pobres en CpG-TLs rojos y ricos en CpG-TLs verdes. Este segundo tipo de factores de transcripción son fundamentales en el desarrollo y algunos son capaces de reclutar enzimas que remodelan la metilación. En cuanto a su distribución en torno a los genes, la proporción de CpG-TLs verdes disminuye a medida que decrece la distancia al sitio de inicio de la transcripción, mientras que la proporción de CpG-TLs rojos aumenta.

La base de datos dedicada a la metilación NGSmethDB contiene una amplia colección de mapas de metilación para diferentes especies, tipos celulares e individuos. Con el fin de optimizar el almacenamiento y consulta del gran volumen de datos producidos a lo largo de esta Tesis Doctoral, entre los que se incluyen mapas de metilación, de DMCs y de CpG-TLs, se decidió rediseñar por completo esta base de datos. Para agilizar las comparaciones entre muestras, se optó por migrar los datos al sistema de bases de datos MongoDB y almacenarlos en una estructura jerárquica de documentos JSON (un formato estándar que permite intercambiar datos etiquetados y jerarquizados entre distintos lenguajes de programación), donde: i) cada ensamblado posee su propia base de datos, ii) cada cromosoma posee su propia colección de documentos JSON, iii) cada sitio CpG posee su propio documento JSON y iv) cada subdocumento contiene un tipo de información biológica (metilación, metilación diferencial o asociación con la

transcripción). En el caso de los mapas de metilación, cada subdocumento se divide en tres niveles: i) el individuo, ii) la muestra y iii) el tipo de dato. Se implementaron varias vías de acceso, comparación y visualización de los datos contenidos en la NGSmethDB, entre las que destacan: i) su acceso programático mediante el protocolo HTTPS a través de un servidor RESTful API y ii) su conectividad con UCSC Genome Browser a través de Track Hubs.

En esta Tesis Doctoral se ha mejorado notablemente la fiabilidad en la detección de los niveles de metilación de las citosinas individuales a partir de lecturas de WGBS, tomando en cuenta todas fuentes de error conocidas en la actualidad. Esto ha permitido poner a prueba la hipótesis de que el signo de la asociación entre la metilación y la transcripción depende del contexto genómico en que se produce la metilación y del tipo de factores de transcripción implicados. A la vista de los resultados obtenidos, no ha sido posible refutar esta hipótesis. Un hallazgo inesperado ha sido que la asociación positiva entre la metilación y la transcripción parece ser más frecuente de lo que previamente se había descrito, llegando incluso a ser más frecuente que los casos con asociación negativa. En relación a esto, los sitios de unión a factores de transcripción con mayor afinidad por sitios metilados son ricos en CpG-TLs verdes pero pobres en CpG-TLs rojos. Estas asociaciones positivas podrían deberse a un mecanismo de regulación de la transcripción hasta ahora desconocido, pero también es probable que en realidad se trate de casos en los que la hidroximetilación se asocia positivamente con la transcripción, ya que la técnica WGBS es incapaz de discriminar entre metilación e hidroximetilación. En futuros estudios, se deberían utilizar las técnicas OxBS-seq o TAB-seq para tratar de esclarecer la verdadera naturaleza de los CpG-TLs verdes.

Palabras clave:

metilación, metilación diferencial, transcripción, factores de transcripción, semáforos CpG

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Se estima que un ser humano adulto está compuesto por $3 \cdot 10^{13}$ células (Sender, Fuchs, & Milo, 2016), clasificadas en los cientos de tipos celulares que componen sus tejidos, órganos y sistemas (Vickaryous & Hall, 2006). Todas estas células tienen su origen en el cigoto y comparten una misma secuencia genómica, pero cada **tipo celular** se caracteriza por expresar un conjunto específico de genes, que está relacionado con otras propiedades de la célula, como su forma, su localización, su función y su relación con otras células (Schumacher et al., 2017).

Este perfil de expresión presenta cierta plasticidad, conocida como **estado celular**, que permite a la célula autorregularse y reaccionar ante su entorno (Clevers et al., 2017; Schumacher et al., 2017). Ejemplos de estado celular son el estado en fase G1 del ciclo celular o el estado de privación de nutrientes.

Durante milenios, el desarrollo embrionario ha fascinado e intrigado a generaciones de filósofos y científicos. Aristóteles acuñó el término *epigénesis* para describir la elaboración de la diversidad de partes que componen a un ser humano a partir de un huevo indiferenciado, en oposición a la teoría de la preformación (Maienschein, 2012). En la década de 1940, Conrad Waddington utilizó este término para describir su idea de un *paisaje epigenético*, una metáfora abstracta que trata de ilustrar cómo los genes influencian la decisión del destino celular (Stern, 2000). En la década de 1970, Robin Holliday ahondó en la idea del paisaje epigenético, enfatizando en cómo las células mantienen su tipo celular tras la mitosis (Holliday & Pugh, 1975). Actualmente, el término *epigenética* se utiliza para referirse a aquella información originada por un evento o una señal transitorios y que perdura a modo de **memoria celular**, distinguiéndola de los procesos celulares que ocasionan los estados transitorios (Ptashne, 2007).

Hoy se sabe que las diferencias entre tipos celulares radican en su memoria celular, establecida por los factores de transcripción pioneros durante el desarrollo (Iwafuchi-Doi & Zaret, 2014, 2016; Zaret & Mango, 2016) y constituida por marcas epigenéticas como la **metilación del ADN** y las **modificaciones de histonas** (Atlasi & Stunnenberg, 2017; D'Urso & Brickner, 2014). Todas estas marcas tienen en común que intervienen en la regulación de la transcripción, se mantienen estables a lo largo de generaciones celulares y son modificadas por señales del desarrollo o señales ambientales.

Esta Tesis Doctoral se ha centrado en el estudio de la metilación del ADN, probablemente la marca epigenética más importante. Se ha contribuido notablemente a la mejora en la detección de los niveles de metilación de citosinas individuales a partir de lecturas de WGBS, lo que ha permitido estudiar cambios de metilación en los sitios CpG del genoma humano y su asociación con la transcripción. La hipótesis de partida fue que, dependiendo del contexto genómico en que se produzca y del tipo de factor de transcripción que intervenga, la metilación puede contribuir a la regulación positiva o negativa de la transcripción o no tener efecto.

En los siguientes apartados de esta introducción, se describe la metilación del ADN en humanos, su asociación con la transcripción, los métodos más utilizados para la detección de la metilación y las fuentes de error que afectan a la detección de la metilación a partir de lecturas de WGBS.

1.1 La metilación del ADN en humanos

La metilación del ADN en el carbono 5 de la citosina produce 5-metilcitosina, una de las marcas epigenéticas más estudiadas. La primera evidencia de metilación en el ADN se remonta a 1898, cuando Ruppel aisló de la bacteria *Tubercle bacillus* un ácido nucleico inusual, ya que contenía un nucleótido metilado que no pudo identificar. En 1925, Johnson y Coghill hidrolizaron el ADN de *Tubercle bacillus* con ácido sulfúrico y detectaron una pequeña cantidad de 5-metilcitosina (Johnson & Coghill, 1925). Este descubrimiento fue criticado por basarse solo en propiedades ópticas de la molécula. En 1948, Hotchkiss aisló 5-metilcitosina a partir de ADN de timo de vaca utilizando una técnica de cromatografía en papel (Hotchkiss, 1948). Wyatt confirmó este descubrimiento en varios animales y en una planta, en 1950, y además comprobó que cada tejido tenía una cantidad de 5-metilcitosina característica (Wyatt, 1950, 1951).

La primera vez que se planteó que la metilación del ADN podría actuar como una marca epigenética fue en 1975, cuando Riggs, Holliday y Pugh propusieron que la metilación era la responsable de la inactivación del cromosoma X en las hembras de mamíferos (Holliday & Pugh, 1975; Riggs, 1975). Propusieron un modelo de propagación semiconservativa del patrón de metilación, basado en dos aspectos clave:

- Los sitios en los que el ADN se metila deben ser palindrómicos. El palíndromo en cuestión debía ser el dinucleótido CpG, ya que Doskocil y Sorm habían descubierto, en 1962, que la principal diana de la metilación en mamíferos son los sitios CpG (Doskočil & Šorm, 1962).
- Las enzimas responsables de metilar ADN no-metilado y ADN hemi-metilado (metilado en una sola hebra) deben ser diferentes.

Postularon que el primer evento de metilación debía ser mucho más difícil que el segundo. Sin embargo, una vez modificada la primera hebra, la hebra complementaria sería rápidamente modificada en el mismo sitio palindrómico. De esta manera, una marca de metilación presente en una hebra parental será copiada en la hebra hija durante la replicación del ADN, resultando en la transmisión exitosa del estado de metilación a la siguiente generación de células. Poco después, Bird y Southern confirmaron el modelo de la propagación semiconservativa de la metilación utilizando enzimas de restricción sensibles a la metilación para detectar el estado de metilación del ADN (Adrian P. Bird, 1978; Adrian P. Bird & Southern, 1978). Años después, Bestor e Ingram descubrieron y purificaron la enzima necesaria para la propagación semiconservativa de la metilación (Bestor & Ingram, 1983).

En 1981, los experimentos de Compere y Palmiter confirmaron que la metilación está implicada en la regulación de la expresión génica y en la diferenciación celular (Compere & Palmiter, 1981). Por su parte, Cattanach y Kirk descubrieron que la metilación también es responsable de la impronta genómica (Cattanach & Kirk, 1985), heredada de los parentales y mantenida en la línea somática a lo largo de la vida del individuo, y tiempo después se confirmó que la metilación de los genes improntados controla su expresión (Bell & Felsenfeld, 2000; Hark et al., 2000; Kanduri et al., 2000).

Todos estos descubrimientos sentaron las bases para el estudio de las implicaciones de la metilación en la diferenciación y el mantenimiento de la identidad celular, constituyendo un mecanismo estable para la herencia de los patrones de expresión génica característicos del tipo celular a lo largo de sucesivas

mitosis (Cheedipudi, Genolet, & Dobreva, 2014; Henikoff & Greally, 2016; M. Kim & Costello, 2017; H. Luo et al., 2017; Shipony et al., 2014).

1.1.1 Mantenimiento, establecimiento y reversión

El mantenimiento del patrón de metilación de la célula madre en sus células hijas es posible gracias a la acción de la ADN metiltransferasa DNMT1 (Bestor & Ingram, 1983), que metila eficazmente los sitios CpG que están hemi-metilados (es decir, metilados en la hebra original pero no en la nueva hebra) y deja intactos los sitios CpG no-metilados (Hermann, Goyal, & Jeltsch, 2004). Este mecanismo de mantenimiento no actúa en todo momento, sino que es regulado por varios factores y solamente actúa durante la fase S, sincronizándose con la replicación del ADN (Probst, Dunleavy, & Almouzni, 2009; Smith & Meissner, 2013). Los principales reguladores de la DNMT1 son UHRF1 (Arita, Ariyoshi, Tochio, Nakamura, & Shirakawa, 2008; Bostick et al., 2007), que se une a sitios CpG hemi-metilados y recluta a la DNMT1, y la modificación de histonas H3K9me, que se une a UHRF1 y regula la estabilidad de la DNMT1 durante la fase S (Rothbart et al., 2012).

Aunque los patrones de metilación de las células se pueden propagar de forma estable, también sufren cambios durante el desarrollo y a lo largo de la vida del individuo. Durante la formación de los gametos y en las etapas tempranas del desarrollo, se eliminan casi por completo los patrones de metilación de las células (excepto la impronta genómica) y se establece un nuevo patrón de metilación que se irá modificando a medida que se produzca el desarrollo de cada uno de los linajes celulares (Dean, 2014; Smith & Meissner, 2013).

Tan importante es la metilación para el desarrollo embrionario, que las mutaciones que provocan la pérdida de función en alguna de las enzimas responsables del mantenimiento y establecimiento de la metilación son letales (E. Li, Bestor, & Jaenisch, 1992; Okano, Bell, Haber, & Li, 1999). Las enzimas responsables de

establecer la metilación *de novo* son las ADN metiltransferasas DNMT3A y DNMT3B, que forman un complejo con su homólogo sin actividad enzimática DNMT3L (Jia, Jurkowska, Zhang, Jeltsch, & Cheng, 2007; Okano et al., 1999). A pesar de que DNMT3L carece de dominio catalítico, juega un papel fundamental en el establecimiento de metilación, ya que regula el reclutamiento de DNMT3A y DNMT3B en las regiones que se van a metilar (Ooi et al., 2007). En roedores, se ha descrito DNMT3C, una ADN metiltransferasa que metila *de novo* los promotores de elementos retrotransponibles jóvenes en la línea germinal masculina y es necesaria para la fertilidad en ratón (Barau et al., 2016). Hasta la fecha no se ha descrito la existencia de DNMT3C en humanos.

En cuanto a la reversión de la metilación, si el mecanismo de mantenimiento de la metilación no actúa en una región durante sucesivas mitosis, dicha región acaba perdiendo su metilación. Sin embargo, existen también mecanismos enzimáticos de eliminación de la metilación. Las enzimas TET1, TET2 y TET3 pueden oxidar a la 5-metilcitosina hasta tres veces consecutivas, produciendo 5-hidroximetilcitosina, 5-formilcitosina y 5-carboxilcitosina (Y. F. He et al., 2011; Ito et al., 2010, 2011; Tahiliani et al., 2009). Las formas oxidadas de la 5-metilcitosina carecen de mecanismos de mantenimiento y se pierden en sucesivas mitosis, aunque la 5-formilcitosina y la 5-carboxilcitosina también pueden ser eliminadas mediante el mecanismos de escisión de bases por TDG y posterior reparación del ADN (Cortellino et al., 2011; Y. F. He et al., 2011; Maiti & Drohat, 2011).

1.1.2 Localización y funciones

En mamíferos, la principal diana de la metilación son los sitios CpG. Entre el 60% y el 80% (dependiendo del tipo celular) de los 29 millones de sitios CpG del genoma humano están metilados (Lister et al., 2009). Sin embargo, algunos tipos celulares y tejidos presentan también metilación en otros contextos de secuencia (CHG y CHH, donde H es A, T o C), como le ocurre a las células madre embrionarias, a los oocitos y al cerebro (Laurent et al., 2010; Lister et al., 2013, 2009; Tomizawa et al., 2011; Xie et al., 2012; Ziller et al., 2013). En esta Tesis Doctoral, el

estudio de los cambios de metilación en el genoma humano y su asociación con la transcripción se ha limitado al contexto CpG, al ser el único comparable en un amplio abanico de muestras humanas.

En los genomas de mamífero, los dinucleótidos CpG están infrarrepresentados. Sin embargo, existen regiones densas en sitios CpG, conocidas como islas CpG, que solapan con las regiones promotoras de más del 70% de los genes (Deaton & Bird, 2011; Illingworth et al., 2010). A pesar de su alta densidad en sitios CpG, la mayoría de islas CpG están no-metiladas y su metilación suele estar asociada con la regulación negativa de la iniciación de la transcripción. El que la mayoría de estudios sobre la metilación se hayan centrado en las islas CpG en promotores, ha provocado que la metilación se asocie casi exclusivamente con la inhibición de la transcripción.

Sin embargo, hoy se sabe que la función de la metilación depende del contexto genómico en que se produzca (Peter A. Jones, 2012). Por ejemplo, existen también islas CpG en las regiones 3' de algunos genes, cuya metilación no está asociada con la represión de la transcripción (Larsen, Solheim, & Prydz, 1993) e incluso en algunos casos puede estar asociada con la activación de la transcripción (D.-H. Yu et al., 2013). En la sección 1.2, se describe en más profundidad la asociación entre la metilación y la transcripción, dependiendo del contexto genómico en que ocurra la metilación.

En los cuerpos génicos el nivel de metilación medio es más alto que en el resto del genoma (entre el 80% y el 90%), aunque son pobres en sitios CpG (Lister et al., 2009). El nivel de metilación de los exones es más alto que el de los intrones, produciéndose cambios abruptos en las fronteras exón-intrón. Se ha descrito que la metilación interviene en la regulación del *splicing* a través del reclutamiento en exones de la proteína MeCP2 (Laurent et al., 2010; Maunakea, Chepelev, Cui, & Zhao, 2013).

La metilación también juega un papel importante en la estabilidad cromosómica y genómica. La metilación de los elementos repetidos de los centrómeros contribuye a la correcta segregación de los cromosomas (Moarefi & Chédin, 2011). Por otra parte, la metilación de los elementos transponibles provoca su silenciamiento (A. De Mendoza et al., 2018; Yoder, Walsh, & Bestor, 1997). En genomas de mamífero, estos elementos transponibles representan más del 35% del genoma y en ellos se concreta la mayor parte de la metilación.

1.2 Asociación entre metilación y transcripción

Según el paradigma tradicional, cuando se metila una isla CpG ubicada en el promotor de un gen activo, el gen se vuelve inactivo (véase Figura 1) (Deaton & Bird, 2011; Illingworth et al., 2010). De igual manera, la metilación de los elementos transponibles provoca su silenciamiento (A. De Mendoza et al., 2018; Yoder et al., 1997). Dado que la mayor parte de los sitios CpG en los genomas de mamífero están metilados, se ha propuesto que la metilación podría contribuir a reducir el ruido transcripcional (A. P. Bird, 1993).

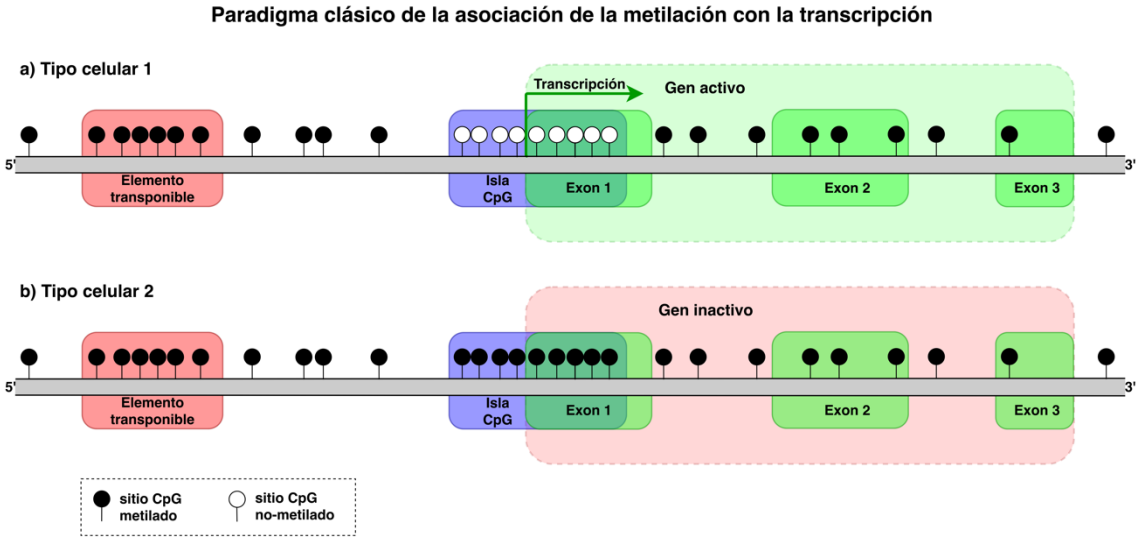

Figura 1. Paradigma clásico de la asociación de la metilación con la transcripción

La mayoría de sitios CpG están metilado, exceptuando las islas CpG que están ubicadas en el promotor de un gen activo (a). Si la isla CpG se metilada, el gen se vuelve inactivo (b).

Sin embargo, la metilación de islas CpG en las regiones 3' de algunos genes, no está asociada con la represión de la transcripción (Larsen et al., 1993) e incluso puede en algunos casos se asocia con la activación de la transcripción (D.-H. Yu et al., 2013). También se ha demostrado que la transcripción de algunos genes es independiente del estado de metilación de la isla CpG en su promotor (A P Bird, 1984). Por otra parte, los promotores con baja densidad en sitios CpG frecuentemente están metilados, sin que se vea afectada su actividad transcripcional (Eckhardt et al., 2006; Weber et al., 2007).

Por otra parte, se ha comprobado que el efecto inhibitorio de la metilación en el promotor de algunos genes se puede ver contrarrestado por la influencia de un potenciador (Boyes & Bird, 1992). Paradójicamente, los potenciadores presentan niveles intermedios de metilación (Sharifi-Zarchi et al., 2017) y necesitan estar sometidos a un ciclo constante de metilación e hidroximetilación para permanecer activos (Rinaldi et al., 2016).

En la década de 1980, se describió una asociación positiva entre la metilación del cuerpo génico y la actividad transcripcional del gen (Wolf, Jollyt, Lunnen, Friedmannt, & Migeon, 1984). En la actualidad, esta asociación ha sido ampliamente confirmada (Aran, Toperoff, Rosenberg, & Hellman, 2011; Hellman & Chess, 2007), dando lugar a la conocida como *paradoja de la metilación* (P. A. Jones, 1999). Según esta paradoja, la metilación inhibe la etapa de iniciación de la transcripción, pero activa la etapa de elongación.

Sin embargo, la asociación positiva entre metilación y transcripción no se limita al cuerpo génico. En el promotor de algunos genes específicos de tejido, está presente la secuencia CRE (TGA**CG**TCA). Para que estos genes se transcriban, es necesario que se metile el sitio CpG de la secuencia CRE y que se una el factor de transcripción C/EBPα (Rishi et al., 2010). Otro ejemplo es el gen FoxA2, un

regulador principal del desarrollo del endodermo y de las células beta pancreáticas. El promotor de este gen presenta una isla CpG, que está metilada en los tejidos en los que se expresa FoxA2 y no-metilada en los tejidos en los que no se expresa (Bahar Halpern, Vana, & Walker, 2014).

Si bien existen aún muchas incógnitas sobre los mecanismos que relacionan la metilación con la transcripción, cada vez está más claro que la relación entre ambas es compleja y dependiente de factores de transcripción concretos que interactúan con sitios CpG individuales concretos.

1.2.1 Interacciones entre los factores de transcripción y la metilación

Los factores de transcripción pueden regular positiva o negativamente, recibiendo el nombre de activadores y represores, respectivamente (Lambert et al., 2018). En el paradigma clásico de la metilación, se asume que la metilación bloquea la unión de los factores de transcripción. Suponiendo un promotor que tenga un único sitio de unión para un factor de transcripción concreto (véase Figura 2):

- Si el factor de transcripción es activador y el sitio de unión está no-metilado, el activador se unirá y se iniciará la transcripción (figura 2a). En cambio, si el sitio de unión está metilado, no se unirá y el gen permanecerá inactivo.
- Si el factor de transcripción es inhibidor y el sitio de unión está no-metilado, el represor se unirá y bloqueará la transcripción (figura 2c). En cambio, si el sitio de unión está metilado, no se unirá y el gen permanecerá activo.

Sin embargo, recientemente este marco teórico ha aumentado considerablemente. Se ha descrito que un tercio de los factores de transcripción humanos, se unen de manera específica a secuencias metiladas (Yin et al., 2017). A estos factores de transcripción se les llama *Methyl-Plus*, mientras que a los factores de transcripción que se unen de manera específica a secuencias no-metiladas se les llama *Methyl-Minus*. Muchos de los factores de transcripción Methyl-Plus están entre los reguladores del desarrollo más importantes (Bürglin, 2011).

Asumiendo de nuevo el mismo modelo simplificado de promotor, la Figura 2 muestra la relación entre estos tipos de factores de transcripción y la metilación:

- Si el factor de transcripción es activador y Methyl-Minus, la metilación del sitio de unión provoca la inactivación de la transcripción (figura 2a).
- Si el factor de transcripción es activador y Methyl-Plus, la metilación del sitio de unión provoca la activación de la transcripción (figura 2b).
- Si el factor de transcripción es represor y Methyl-Minus, la metilación del sitio de unión provoca la activación de la transcripción (figura 2c).
- Si el factor de transcripción es represor y Methyl-Plus, la metilación del sitio de unión provoca la inactivación de la transcripción (figura 2d).

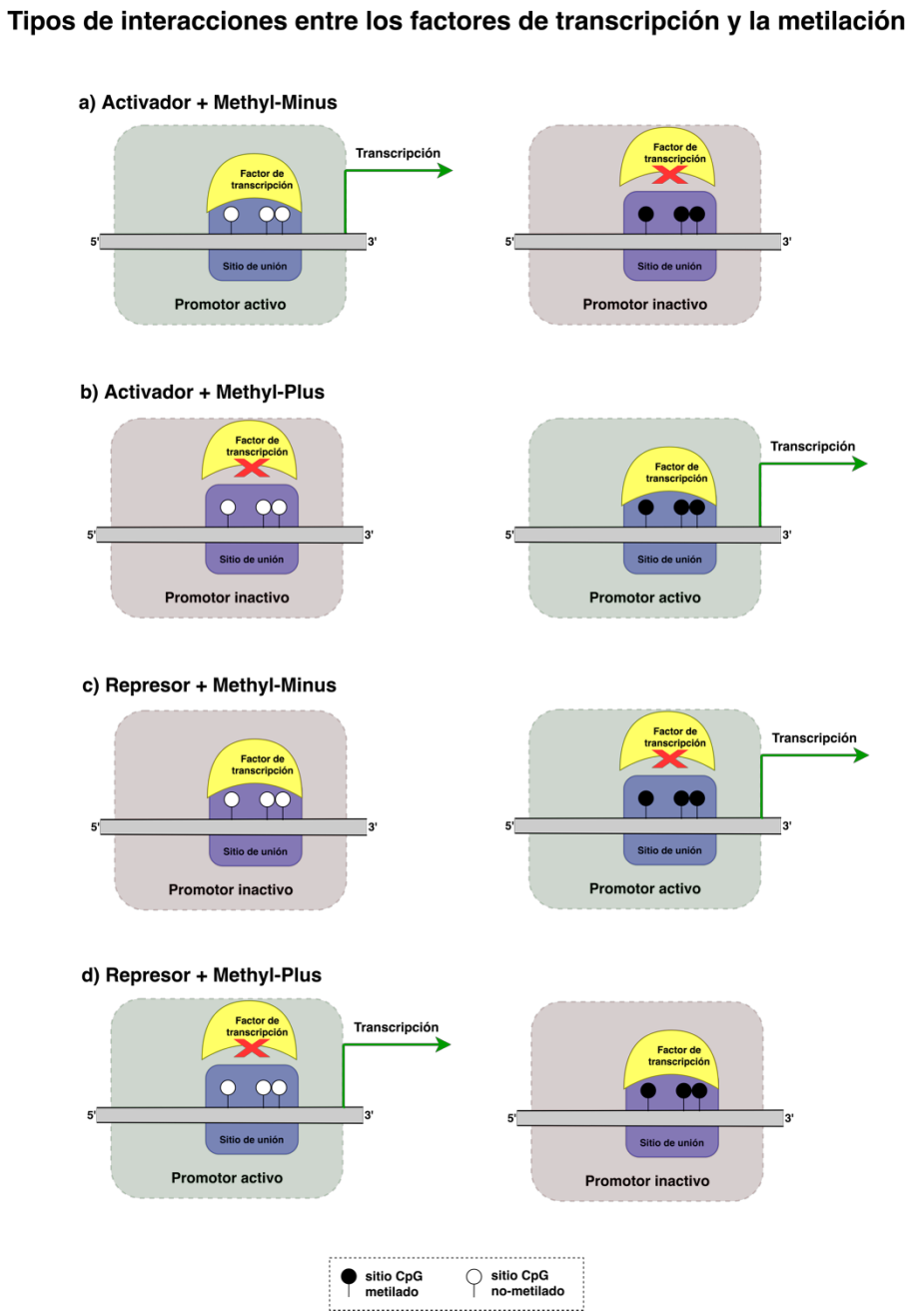

Figura 2. Tipos de interacciones entre los factores de transcripción y la metilación

Dependiendo de si el factor de transcripción es activador o represor y de si se une preferentemente a sitios no-metilados (Methyl-Minus) o a sitios metilados (Methyl-Plus) el efecto de la metilación sobre la transcripción será diferente.

Algunos factores de transcripción, conocidos como pioneros, son capaces de modificar el estado de metilación en torno a su sitio de unión, gracias al reclutamiento de modificadores epigenéticos (Iwafuchi-Doi, 2019; Iwafuchi-Doi & Zaret, 2014, 2016; Zaret & Mango, 2016). Por ejemplo, se ha demostrado que los

factores pioneros FOXA y Pax7 provocan la desmetilación los promotores de algunos genes durante el desarrollo (Donaghey et al., 2018; Mayran et al., 2018; Zhang et al., 2016).

1.2.2 Semáforos CpG

En los estudios tradicionales de asociación entre la metilación y la transcripción, se ha utilizado la metilación promedio de todos los sitios CpG que componen el promotor. Sin embargo, no todos los sitios CpG en promotores tienen el mismo nivel de metilación e incluso se ha demostrado que un único sitio CpG diferencialmente metilado es suficiente para provocar un efecto en la tasa de transcripción del gen ESR1 (Fürst, Kliem, Meyer, & Ulbrich, 2012).

Se ha descrito recientemente que solo el 16,6% de los sitios CpG en promotores ejercen un efecto sobre la transcripción cuando cambia su nivel de metilación (Harbers et al., 2014). Son los llamados *semáforos CpG*, sitios CpG cuyo nivel de metilación correlaciona negativamente con la tasa de transcripción de un gen cercano (Harbers et al., 2014; Lioznova et al., 2019). En esta Tesis Doctoral, se ha extendido este concepto también a sitios CpG cuya correlación es positiva. La Figura 3 representa esquemáticamente la definición de semáforo CpG utilizada a lo largo de esta Tesis Doctoral, así como su tipo en función del signo de la correlación (rojo si es negativa y verde si es positiva).

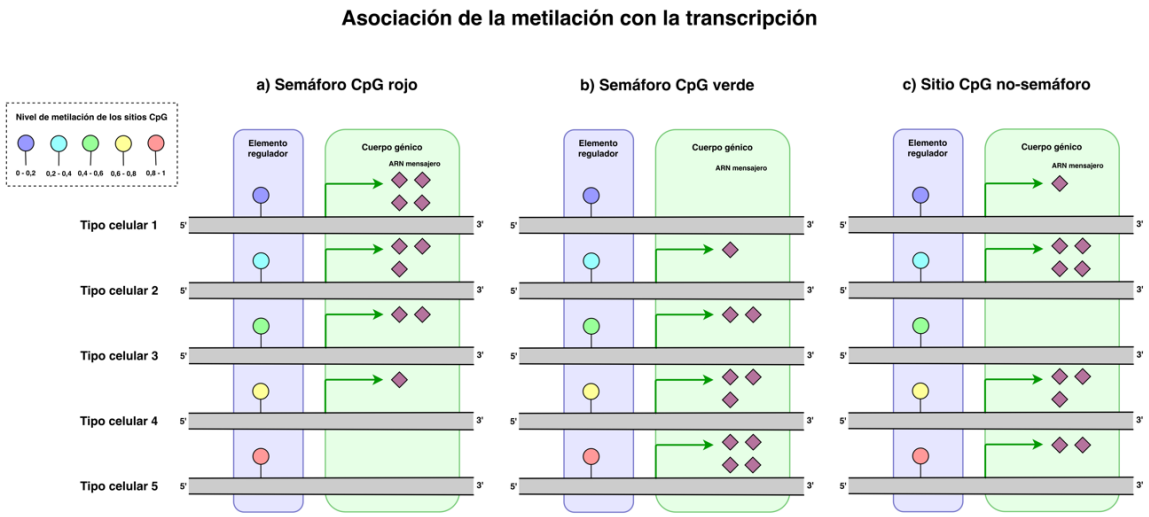

Figura 3. Asociación de la metilación con la transcripción

Cuando la metilación de un semáforo CpG rojo aumenta, disminuye la tasa de transcripción del gen (a). En cambio, cuando la metilación de un semáforo CpG verde aumenta, aumenta la tasa de transcripción del gen (b). La figura c representa el caso en que no existe asociación entre el nivel de metilación del sitio CpG y la transcripción del gen.

1.3 Métodos de detección de la metilación

Los primeros estudios de la metilación del ADN se centraron en cuantificar la cantidad total de 5-metilcitosina del genoma y en determinar el estado de metilación de genes de interés (Lisanti et al., 2013). Actualmente, los métodos de secuenciación masiva permite obtener mapas genómicos de metilación a la resolución de citosinas individuales (Laird, 2010). En las siguientes secciones, se describen los principales métodos de detección de la metilación, basado en la digestión enzimática, el enriquecimiento por afinidad y la conversión con bisulfito.

En esta Tesis Doctoral, se han utilizado exclusivamente datos procedentes del método de secuenciación WGBS (Urich, Nery, Lister, Schmitz, & Ecker, 2015), debido a su capacidad para detectar la metilación de la práctica totalidad de las citosinas del genoma humano.

1.3.1 Basados en la digestión enzimática

Los métodos basados en el uso de enzimas de restricción aprovechan las propiedades de los isoesquizómeros. Se dice que dos enzimas de restricción son isoesquizómeros cuando tienen la misma secuencia de reconocimiento y el mismo punto de corte, pero presentan diferente susceptibilidad al estado de metilación del ADN. Las enzimas de restricción sensibles a la metilación (MREs), como BstUI, HpaII, NotI y SmaI solo cortan secuencias no-metiladas (Yong, Hsu, & Chen, 2016).

La técnica de secuenciación MRE-seq consiste en la digestión del ADN genómico con una MRE, la selección por tamaño de los fragmentos resultantes y posterior secuenciación con técnicas de secuenciación masiva (D. Li, Zhang, Xing, & Wang, 2015; Maunakea et al., 2010). Aunque esta técnica permite estimar niveles de metilación en las secuencias de reconocimiento, estas representan una fracción reducida del genoma.

Relacionado con el anterior, existe un método conocido como *Comprehensive High-throughput Array for Relative Methylation* (CHARM). Se basa en utilizar McrBc, una enzima de restricción que digiere ADN metilado (Sutherland, Coe, & Raleigh, 1992), seleccionar los fragmentos generados por tamaño e hibridarlos en un array (Irizarry et al., 2008).

1.3.2 Basados en el enriquecimiento por afinidad

Los métodos basados en enriquecimiento por afinidad utilizan proteínas con dominios de unión a sitios CpG metilados (MBD) o anticuerpos específicos frente a la 5-metilcitosina para enriquecer la biblioteca de ADN en regiones metiladas.

Se han desarrollado métodos basado en el uso de MBD, que acoplan la obtención de fragmentos de ADN ricos en regiones metiladas con técnicas de secuenciación

masiva (como MBDCap-seq) o con hibridación en array (como MBD-chip) (Brinkman et al., 2010).

Por otra parte, se ha desarrollado métodos basado en la precipitación inmunológica de fragmentos de ADN utilizando anticuerpos frente a la 5-metilcitosina (*Methylated DNA ImmunoPrecipitation* o MeDIP). Los fragmentos precipitados se puede analizar mediante ténicas de secuenciación masiva (como MeDIP-seq) o por hibridación en array (como MeDIP-chip) (Weng, Huang, & Yan, 2009).

Estas técnicas permiten medir enriquecimientos en regiones metiladas a lo largo del genoma, pero no pueden medir niveles de metilación. La resolución de MeDIP-seq es de 100-300 pb y no puede discriminar entre contextos de metilación, pero es económica y requiere una cantidad de ADN mucho menor que otras muestras (1 ng, en lugar de los 5 μg que necesita la técnica WGBS), lo que la convierte en una candidata interesante para estudiar tipos celulares raros o microdisecciones de tejidos (Clark et al., 2012; Taiwo et al., 2012; Zhao, Whyte, Hopkins, Kirk, & Prather, 2014). Otro problema de estas técnicas es que son sensibles a la densidad en sitios CpG y las variaciones en el número de copias (Bock et al., 2010; Down et al., 2008).

1.3.3 Basados en la conversión con bisulfito

El tratamiento de ADN genómico con bisulfito sódico desamina las citosinas no-metiladas, convirtiéndolas en uracilos (que serán reemplazados por timinas durante la amplificación por PCR), mientras que las citosinas metiladas permanecen inalteradas (Frommer et al., 1992). Los métodos basados en la conversión con bisulfito proporcionan resolución de citosinas individuales y en principio de utilizaron para investigar la metilación de secuencias de interés, acoplando la conversión con bisulfito con la técnica de secuenciación de Sanger (Stocks et al., 1989). Actualmente, estos métodos no se limitan a estudiar *loci* específicos, sino que permite estudiar la metilación en genoma completo.

La técnica *Whole-Genome Bisulfite Sequencing* (WGBS) teóricamente es capaz de detectar el estado de metilación de todas las citosinas del genoma (Cokus et al., 2008; Lister et al., 2008; Urich et al., 2015). En esta técnica, el ADN genómico se purifica, se fragmenta utilizando técnicas de sonicación, se reparan los extremos con bases no-metiladas y se añaden a ambos extremos de los fragmentos adaptadores metilados. Finalmente, los fragmentos se seleccionan por tamaño, se tratan con bisulfito sódico, se amplifican por PCR y se secuencian mediante técnicas de secuenciación masiva. La principal ventaja de WGBS es su capacidad para detectar la metilación de cualquier citosina, independientemente del contexto de secuencia o de la región genómica a la que pertenezca. Esto incluye regiones con baja densidad en sitios CpG que no se pueden estudiar con otras técnicas, como las regiones intergénicas, los dominios parcialmente metilados o los elementos reguladores distales. Sin embargo, tiene el inconveniente de que necesita gran cantidad de ADN genómico (al menos 5 μg). WGBS se ha convertido en la técnica estándar de varios proyectos internacionales, como Roadmap Epigenomics (Kundaje et al., 2015; Leung et al., 2015), ENCODE (Bernstein et al., 2012), Enhancing GTEx (Van Wittenberghe et al., 2017) y BLUEPRINT (Adams et al., 2012).

Para reducir los costes de la investigación de la metilación en genomas de mamíferos, Meissner *et al.* desarrollaron la técnica conocida como *Reduced-Representation Bisulfite Sequencing* (RRBS). Esta técnica integra la digestión mediante la enzima de restricción MspI, la conversión con bisulfito y la secuenciación masiva de fragmentos dentro de un rango de tamaños (normalmente, entre 40 y 220 pbs) (Meissner et al., 2008). Permite cubrir la mayor parte de las islas CpG, que representan el 1-3% del genoma. Esto abarata significativamente los costes con respecto a WGBS, pero se pierde la información de regiones de interés, como los reguladores distales, y además sigue necesitando una cantidad relativamente alta de ADN genómico (0,01 – 0,3 μg) (Bock et al., 2011; Smith, Gu, Bock, Gnirke, & Meissner, 2009).

Por último, existen también métodos que combinan la conversión con bisulfito y la hibradación en array. El protocolo de *Illumina HumanMethylation450 BeadChip* (HM450K) se basa en la conversión con bisulfito del ADN genómico, su amplificación por PCR y posterior hibridación en arrays que contienen sondas prediseñadas, que permiten distinguir si los sitios CpG están metilados o no-metilados. El array HM450K es un método económico y permite detectar los niveles de metilación de 450.000 sitios CpGs del genoma humano, incluyendo islas CpG, promotores, regiones 5'-UTRs, primeros exones, cuerpos génicos y regiones 3'-UTRs (Dedeurwaerder et al., 2011; Sandoval et al., 2011). El consorcio The Cancer Genome Atlas ha utilizado MH450K para caracterizar la metilación de más de 7.500 muestras procedentes de 200 tipos diferentes de cáncer (Creighton et al., 2013; Hammerman et al., 2012; Pidsley et al., 2016; Stirzaker, Taberlay, Statham, & Clark, 2014). Recientemente, se ha desarrollado una versión mejorada de HM450K, conocida como *Infinium MethylationEPIC Bead-Chip* (EPIC). Este nuevo array cubre más del 90% de los sitios CpG representados en HM450K y otros 350.000 sitios CpG en regiones identificadas como potenciadores por los proyectos FANTOM5 (Lizio et al., 2015) y ENCODE (Siggens & Ekwall, 2014).

1.4 Fuentes de error en WGBS

Existen multitud de fuentes de error que afectan a la fiabilidad de los resultados obtenidos del análisis de lecturas de WGBS, provocando detecciones erróneas en el nivel de metilación de algunas citosinas e incluso pérdidas de información en ciertas regiones (Barturen, Oliver, & Hackenberg, 2017; Olova et al., 2018). El origen de cada una de estas fuentes de error varía, pudiendo originarse en la preparación de la biblioteca, en el proceso de secuenciación o en el análisis bioinformático de las lecturas.

Estos errores pueden dificultar poner a prueba hipótesis sobre la implicación de la metilación en procesos biológicos o en enfermedad, ya que estos errores se extienden a los análisis aguas abajo. Por ejemplo, se podría deducir erróneamente

que ciertas regiones polimórficas no sufren tantos cambios de metilación en una determinada patología, cuando lo que realmente está ocurriendo es que hay un sesgo de cobertura en estas regiones como consecuencia de una fuente de error.

En esta Tesis Doctoral, se ha mejorado notablemente la fiabilidad en la detección de los niveles de metilación de las citosinas individuales a partir de lecturas de WGBS, tomando en cuenta todas las fuentes de error que se describen a continuación.

1.4.1 Posiciones con baja calidad

Una menor calidad de secuenciación (*PHRED score*) se traduce en una mayor probabilidad de que la base secuenciada sea errónea (Cock et al., 2010). A medida que avanza la síntesis de ADN en el proceso de secuenciación, la calidad de las bases secuenciadas decae, lo que provoca que a menudo se tengan que procesar las lecturas para eliminar los extremos 3' con baja calidad (Fuller et al., 2009; Taub, Corrada Bravo, & Irizarry, 2010). No eliminar estas posiciones puede ocasionar varios perjuicios (Krueger, Kreck, Franke, & Andrews, 2012):

- Provocan un incremento del porcentaje de lecturas no-alineadas y de los alineamientos erróneos, como consecuencia de los erróneos de secuenciación.
- Dificultan la detección de las secuencias contaminantes (como los adaptadores). Se deben eliminar los extremos 3' con baja calidad antes de intentar detectar y eliminar las secuencias contaminantes.
- Debido a los anteriores, la detección del nivel de metilación de algunas citosinas podría verse afectada.

El extremo 5' de algunas lecturas puede presentar menor calidad que su región central, pero es necesario conservar este extremo intacto para más tarde poder detectar y corregir el sesgo de metilación (Lin et al., 2013). Por último, algunas lecturas pueden presentar posiciones con baja calidad en su región central,

debiendo ser excluidas durante el proceso de detección de los niveles de metilación (Barturen, Rueda, Oliver, & Hackenberg, 2014).

1.4.2 Secuencias contaminantes

Las secuencias contaminantes son todas aquellas secuencias artificiales que se introducen en los extremos de los fragmentos de ADN durante la preparación de la biblioteca y que terminan por incluirse en las lecturas tras la secuenciación. Existen varios tipos de secuencias contaminantes que afectan a las lecturas de WGBS:

- **Adaptadores.** Son oligonucleótidos compuestos por una secuencia de unión a la celda de flujo (secuencia P5) seguida de un cebador en el extremo 5' y otra secuencia de unión a la celda de flujo en el extremos 3' (secuencia P7) (Martin, 2011). La secuencia P5 y el cebador nunca se incluyen en las lecturas, pero parte o la totalidad de la secuencia P7 puede aparecer en el extremo 3' de la lectura, si el fragmento de ADN es más corto que el número de ciclos de secuenciación.
- **BS-seq tags.** Son secuencias añadidas a continuación del cebador y que permiten distinguir si la lectura procede de la hebra de ADN convertida por bisulfito o de su complementaria inversa (Chen, Cokus, & Pellegrini, 2010), producida durante la amplificación por PCR (Gibbs, 1990).
- **Bases introducidas en los extremos.** El protocolo de WGBS de Illumina incluye una etapa de sonicación del ADN, que provoca su ruptura en fragmentos con extremos protuberantes. Para reparar estos extremos, se añaden nucleótidos no-metilados. Esto provoca que la metilación detectada en los extremos de las lecturas sea menor a la detectada en la región central, contribuyendo al llamado sesgos de metilación o M-bias (Hansen, Langmead, & Irizarry, 2012).

1.4.3 Pérdida de lecturas en *loci* polimórficos

Las dos últimas versiones del ensamblado del genoma humano incluyen haplotipos alternativos (Church et al., 2011), que tratan de recoger las variaciones estructurales y de secuencia de distintas poblaciones o etnias humanas, para evitar

que las lecturas procedentes de estos haplotipos alineen incorrectamente en otras regiones del genoma (M. L. Mendoza et al., 2015).

Durante el desarrollo de esta Tesis Doctoral, se descubrió que el uso de los nuevos modelos de ensamblado durante el alineamiento de las lecturas provoca la pérdida de lecturas procedentes de *loci* polimórficos, como consecuencia de un incremento en el porcentaje de lecturas con alineamiento ambiguo (véase sección 1.4.3).

1.4.4 Lecturas duplicadas

Algunos errores eventuales durante el proceso de preparación de la biblioteca o durante la secuenciación pueden provocar que un mismo fragmento de ADN se secuencia dos o más veces. Es lo que se conoce como lecturas duplicadas y pueden ser de dos tipos:

- **Duplicados de PCR.** Aparecen durante la etapa de amplificación de los fragmentos de ADN mediante PCR (Gibbs, 1990) y son el tipo duplicados más habitual (Kozarewa et al., 2009). Algunos protocolos de WGBS carecen de etapa de amplificación, como PBAT (Miura, Enomoto, Dairiki, & Ito, 2012), por lo que no están afectados por este tipo de duplicados.
- **Duplicados ópticos.** Se producen durante el proceso de secuenciación, cuando el secuenciador interpreta un clúster de lecturas como dos o más clústeres. Este tipo de duplicados afecta a todos los protocolos de WGBS (Kozarewa et al., 2009).

1.4.5 Fallo en la detección de indels

Durante el alineamiento de las lecturas, la introducción de una inserción o deleción (indel) cerca del extremo de la lectura está más penalizada que la introducción de una serie de desemparejamientos consecutivos. En torno a las regiones genómicas que presentan indels, es habitual encontrar una mezcla de lecturas correcta e incorrectamente alineadas. Es necesario recurrir al realineamiento local de estas

regiones mediante el método de la máxima parsimonia para corregir este problema (Liu, Siegmund, Laird, & Berman, 2012).

1.4.6 Sesgo de metilación

El sesgo de metilación o *M-bias* ocasiona que los extremos de las lecturas presenten niveles de metilación anómalos en comparación con la región central (Hansen et al., 2012). La consecuencia más común del *M-bias* es la sobreestimación de los niveles de metilación de las citosinas que están en el extremo 5' de las lecturas, debido a que la secuenciada adaptadora P5 provoca efectos estéricos locales que dificultan la conversión con bisulfito (Liu et al., 2012).

Otros factores que contribuyen al *M-bias* son la introducción de bases no-metiladas durante la reparación de extremos cohesivos, la presencia de adaptadores y el decaimiento de la calidad de secuenciación en el extremo 3' (Lin et al., 2013). La introducción de bases no-metiladas provoca la infraestimación de los niveles de metilación en ambos extremos de las lecturas. En cambio, la presencia de adaptadores y de posiciones con baja calidad pueden provocar tanto infra- como sobreestimación de los niveles de metilación en el extremo 3'.

1.4.7 Fallo en la conversión con bisulfito

Algunos fragmentos de ADN pueden adoptar espontáneamente estructuras secundarias que dificulten el acceso del bisulfito a las bases. Esto provoca que parte o la totalidad de las citosinas no-metiladas de estos fragmentos queden sin convertir en timinas, lo que se conoce como fallo en la conversión con bisulfito (W. Guo et al., 2013).

Las lecturas afectadas por el fallo en la conversión se pueden detectar midiendo el porcentaje de citosinas metiladas en los contextos no-CpG (Barturen et al., 2014). Este método se basa en que las células diferenciadas de mamíferos tienen bajos niveles de metilación en los contextos no-CpG, pero no se debe aplicar en células que presenten niveles apreciables de metilación en estos contextos, como las

células madre embrionarias de mamífero o las células de plantas (X.-J. He, Chen, & Zhu, 2011).

1.4.8 Errores debidos a sustituciones

La existencia de variaciones de secuencia es una importante fuente de errores, ya que dos tercios de los polimorfismos de un solo nucleótido (SNPs) ocurren en contexto CpG y presentan los alelos C/T o G/A (Tomso & Bell, 2003). Una sustitución C/T se puede interpretar como una citosina no-metilada, si no se comprueba la base que ocupa esta posición en la hebra complementaria: adenina en el caso de las sustituciones C/T y guanina en el caso de las citosinas no-metiladas (Barturen et al., 2014).

Capítulo 2

Objectives

1. **To design and implement a protocol for obtaining methylation maps from WGBS reads which minimizes the impact of contaminations and biases that affect the detection of methylation levels.**

2. **To obtain methylation maps from public sets of WGBS reads coming from human samples of different cell types and different individuals.**

3. **To detect and quantify the methylation changes that occur in the CpG sites of the human genome under different conditions, such as cell type or individual.**

4. **To detect and quantify the CpG sites of the human genome whose level of methylation is associated with the transcription rate of a nearby gene (CpG "traffic lights").**

5. **To study the possible impact of differentially methylated CpGs and CpG traffic lights on genomic elements related to transcription regulation, such as promoters, transcription factors binding sites.**

6. **To compile a database that allows to access, compare and visualize the methylation maps obtained, as well as the sets of differentially methylated CpGs and CpG traffic light derived from them.**

Capítulo 3

Objetivos

1. **Diseñar e implementar un protocolo de obtención de mapas de metilación a partir de lecturas de WGBS que minimice el impacto de las contaminaciones y los sesgos que afectan a la detección de los niveles de metilación.**
2. **Obtener mapas de metilación a partir de conjuntos públicos de lecturas de WGBS procedentes de muestras humanas de distintos tipos celulares y de distintos individuos.**
3. **Detectar y cuantificar los cambios de metilación que se producen en los sitios CpG del genoma humano bajo distintas condiciones, como el tipo celular o el individuo.**
4. **Detectar y cuantificar los sitios CpG del genoma humano cuyo nivel de metilación se asocia con la tasa de transcripción de algún gen cercano (semáforos CpG).**
5. **Estudiar el posible impacto de los CpGs diferencialmente metilados y de los semáforos CpG sobre elementos genómicos relacionados con la regulación de la transcripción, como promotores, potenciadores y sitios de unión a factores de transcripción.**
6. **Compilar una base de datos que permita acceder, comparar y visualizar los mapas de metilación obtenidos, así como los conjuntos de CpGs diferencialmente metilados y de semáforos CpG derivados de los mismos.**

Capítulo 4

Material y Métodos

4.1 Conjuntos de lecturas de WGBS

Como ya se ha mencionado en la introducción, la técnica WGBS (Cokus et al., 2008; Lister et al., 2008) permite detectar el nivel de metilación de cada citosina del genoma (véase sección 1.3.3).

Los repositorios públicos **Sequence Read Archive (SRA)** (Karsch-Mizrachi, Takagi, & Cochrane, 2017) y **ENCODE DATA** (Sloan et al., 2016) disponen de conjuntos de lecturas de WGBS para un amplio abanico de muestras. Se seleccionaron y descargaron los conjuntos de lecturas de WGBS de 86 muestras humanas, procedentes de los proyectos Roadmap Epigenomics (Kundaje et al., 2015; Leung et al., 2015), ENCODE (Bernstein et al., 2012) y Enhancing GTEx (Van Wittenberghe et al., 2017). La Tabla 1 resume la información más relevante sobre estas muestras.

Tabla 1. Resumen de las muestras de las que proceden las lecturas de WGBS

Las muestras clasificadas como "órganos y tejidos" son mezclas de diferentes tipos celulares, mientras que las muestras clasificadas como "células diferenciadas" y "células madre" provienen de un único tipo celular. Los tumores no provienen directamente de biopsias, sino de cultivos in vitro estables. En el recuento de muestras

únicas se han contabilizado tantas muestras como tipos ontológicos hubiese, independientemente del individuo.

Tipo de muestra	Muestras únicas	Muestras totales	Individuos	Referencias
Órganos y Tejidos adultos	23	53	7	(Kundaje et al., 2015; Leung et al., 2015; Van Wittenberghe et al., 2017)
Órganos y Tejidos fetales	10	10	7	(Kundaje et al., 2015; Leung et al., 2015)
Células diferenciadas Adultas	7	7	4	(Bernstein et al., 2012; Kundaje et al., 2015; Leung et al., 2015)
Células diferenciadas Fetales	3	3	2	(Bernstein et al., 2012; Kundaje et al., 2015; Leung et al., 2015)
Células madre adultas	1	1	1	(Kundaje et al., 2015; Leung et al., 2015)
Células madre embrionarias	5	6	2	(Kundaje et al., 2015; Leung et al., 2015)
Tumores	6	6	6	(Bernstein et al., 2012)
Total	52	86	29	

La lista completa de muestras está disponible en la **Tabla 18** del anexo, acompañada de la **Tabla 19**, que contiene más información sobre los individuos, y la **Tabla 20**, que contiene más información sobre los órganos, tejidos, tipos celulares o tumores de los que proceden las muestras.

Estos conjuntos de lecturas fueron procesados siguiendo el protocolo descrito en la sección 4.4 e implementado en **MethFlow** (véase sección 5.1).

4.2 Ensamblados genómicos

Las dos últimas versiones del ensamblado del genoma humano incluyen haplotipos alternativos, que tratan de recoger las variaciones estructurales y de secuencia de distintas poblaciones o etnias humanas, para evitar que las lecturas procedentes de estos haplotipos alineen incorrectamente en otras regiones del genoma (Church et al., 2011; M. L. Mendoza et al., 2015).

En esta Tesis Doctoral se ha utilizado el ensamblado GRCh38/hg38 del genoma humano como referencia de todas las anotaciones utilizadas y para todos los

análisis realizados. Para el alineamiento de las lecturas de WGBS, se ha seguido una estrategia de alineamiento en dos etapas (véase sección 4.4.2) en la que se han utilizado las siguientes versiones del ensamblado GRCh38/hg38:

- **Ensamblado primario.** Se trata de un modelo de ensamblado clásico, sin haplotipos alternativos. Se utilizó la versión *analysis set* de **UCSC Genome Browser** (http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/analysisSet), que contiene:
  - Los cromosomas nucleares, scaffolds no localizados (*unlocalized*) y scaffolds no colocados (*unplaced*). Las dos principales regiones PAR del cromosoma Y y los duplicados centroméricos de los cromosomas 5, 14, 19, 21 y 22 han sido sustituidos por bloques de enes ("N") en esta versión del ensamblado.
  - El cromosoma mitocondrial de consenso y la secuencia del virus Epstein-Barr, un contaminante muy frecuente en líneas celulares inmortalizadas.
- **Ensamblado completo.** Se trata del nuevo modelo de ensamblado, con haplotipos alternativos. Se utilizó la versión *full analysis set* de **UCSC Genome Browser** (http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/analysisSet), que contiene el ensamblado primario antes descrito y los haplotipos alternativos de los cromosomas nucleares.

4.3 Anotaciones genómicas

A lo largo de esta Tesis Doctoral se han utilizado numerosos tipos de anotaciones genómicas para el ensamblado GRCh38/hg38, con la finalidad de caracterizar el impacto de los cambios en la metilación sobre distintos tipos de elementos genómicos y la posible implicación de estos cambios en la regulación de la transcripción. La Tabla 2 resume la información más relevante acerca de las anotaciones genómicas utilizadas. En la sección 10.1.1 del anexo se explica en detalle el proceso de obtención, filtrado y clasificación de estas anotaciones.

Tabla 2. Anotaciones genómicas

Cada tipo de anotación corresponde a un conjunto de elementos genómicos que comparten, a grandes rasgos, características funcionales, evolutivas o composicionales. Dentro de cada tipo se pueden distinguir anotaciones con definiciones más concretas o que difieren entre sí en su método de detección. Por ejemplo, un tipo de anotación son los cuerpos génicos, dentro de los que se pueden definir conjuntos de elementos genómicos más concretos, como los exones y los intrones. En la columna de cantidad se especifica el número de elementos genómicos que componen cada anotación. Todas estas anotaciones están en formato *BED* (https://genome.ucsc.edu/FAQ/FAQformat.html#format1).

Tipo	Nombre	Descripción	Cantidad	Referencias
Promotores	relaxed promoters	Conjunto laxo de promotores	22,848	(Fishilevich et al., 2017)
	strict promoters	Conjunto estricto de promotores	13,075	
Potenciadores	relaxed enhancers	Conjunto laxo de potenciadores	222,348	
	strict enhancers	Conjunto estricto de potenciadores	40,155	
Otras regiones reguladoras	CTCFBSs	Sitios de unión de CTCF	113,242	(Zerbino, Wilder, Johnson, Juettemann, & Flicek, 2015)
	DHSs	Sitios hipersensibles a la DNAsa I	68,902	
Sitios de unión a factores de transcripción	activators TFBSs	Sitios de unión a TFs activadores	81,559,122	(Han et al., 2018; Khan et al., 2018)
	repressors TFBSs	Sitios de unión a TFs represores	99,855,452	
	methyl-minus TFBSs	Sitios de unión a TFs con afinidad por sitios no-metilados	91,073,977	(Khan et al., 2018; Yin et al., 2017)
	methyl-plus TFBSs	Sitios de unión a TFs con afinidad por sitios metilados	173,660,354	
Cuerpos génicos	protein gene bodies	Cuerpos génicos de genes codificantes	19,427	(Frankish et al., 2019)
	RNA gene bodies	Cuerpos génicos de genes no-codificantes	36,299	
	protein exons	Exones de genes codificantes	459,706	
	RNA exons	Exones de genes no-codificantes	111,161	
	protein introns	Intrones de genes codificantes	288,952	
	RNA introns	Intrones de genes no-codificantes	60,141	
	protein TSSs	TSSs de genes codificantes	19,427	
	RNA TSSs	TSSs de genes no-codificantes	36,299	
	protein TESs	TESs de genes codificantes	19,427	
	RNA TESs	TESs de genes no-codificantes	36,299	
Islas CpG	relaxed CGIs	Conjunto laxo de islas CpGs	196,277	(Gómez-Martín, Lebrón, Oliver, & Hackenberg,
	strict CGIs	Conjunto estricto de islas CpGs	24,991	

				2018; Hackenberg et al., 2006)
	UCSC CGIs	Conjunto de islas CpGs de UCSC sin enmascarar	49,015	(Casper et al., 2018; Gardiner-Garden & Frommer, 1987)
	masked UCSC CGIs	Conjunto de islas CpGs de UCSC con repetido enmascarado	26,866	
Elementos repetidos	LTRs	Repeticiones terminales largas	665,771	(Smit, 2018)
	LINEs	Elementos nucleares largos intercalados	1,412,488	
	SINEs	Elementos nucleares cortos intercalados	1,687,605	
Elementos conservados	100 spp CEs	CEs en 100 especies de cordados	9,415,524	(Siepel et al., 2005)
	30 spp CEs	CEs en 27 primates y 3 mamíferos no primates	2,684,478	
	20 spp CEs	CEs en 20 mamíferos	2,058,460	
	7 spp CEs	CEs en 7 mamíferos	1,134,081	
Polimorfismos	common SNPs	SNPs comunes	13,520,988	(Sherry, 2001)
	flagged SNPs	SNPs clínicamente asociados	195,961	
	common indels	Indels comunes	1,649,365	
	flagged indels	Indels clínicamente asociados	30,858	

4.4 Obtención de mapas de metilación

Si bien la técnica WGBS permite detectar el nivel de metilación de cada citosina del genoma (Cokus et al., 2008; Lister et al., 2008), existen multitud de fuentes de error que afectan a la fiabilidad de los resultados obtenidos (Barturen et al., 2017; Olova et al., 2018), provocando detecciones erróneas en el nivel de metilación de algunas citosinas e incluso pérdidas de información en ciertas regiones (véase sección 1.4).

En esta Tesis Doctoral, se ha diseñado e implementado un protocolo de obtención de mapas de metilación, a partir de lecturas de WGBS, que toma en cuenta todos los sesgos y fuentes de contaminación conocidos en la actualidad (véase Figura 4). En las siguientes secciones, se describen las etapas de las que se compone el protocolo, destacando el propósito de cada una de ellas e indicando, cuando corresponda, y los programas de terceros utilizados. Una descripción detallada de las opciones de los programas de terceros utilizadas en este protocolo está disponible en la sección 10.1.2 del anexo.

La implementación de este protocolo ha dado lugar al módulo principal de MethFlow, un programa de código abierto para el análisis integral de datos de metilación (véase sección 5.1). Utilizando MethFlow, se obtuvieron los mapas de metilación de las 86 muestras humanas indicadas en la sección 4.1. Estos mapas de metilación están disponibles en la base de datos dedicada a la metilación NGSmethDB (véase sección 5.4).

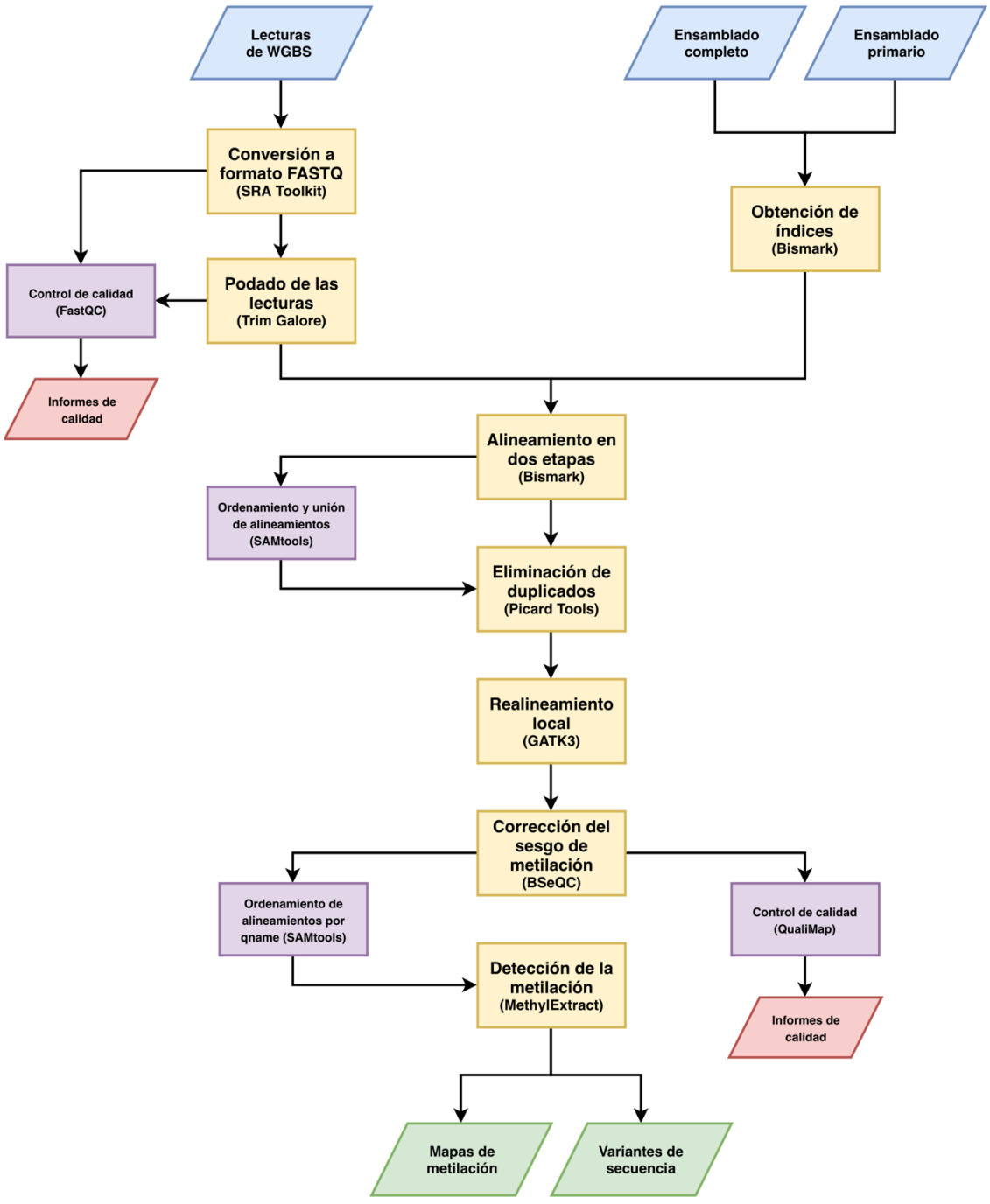

Figura 4. Protocolo de obtención de mapas de metilación

Los romboides azules representan las entradas del protocolo (lecturas de WGBS y ensamblados), los verdes las salidas principales (mapas de metilación y variantes de secuencia) y los rojos los informes de calidad. Los rectángulos amarillos representan las etapas principales del protocolo y los violetas las etapas auxiliares.

4.4.1 Tratamiento previo al alineamiento

Para detectar los niveles de metilación de las citosinas es necesario alinear las lecturas frente al ensamblado de referencia, ya que se desconoce la procedencia genómica de cada uno de los fragmentos de ADN secuenciados. Sin embargo, antes de alinear las lecturas es necesario preparar las lecturas y los ensamblados, tanto para que estén en el formato requerido por los distintos programas como para eliminar sesgos y contaminaciones.

En primer lugar, es necesario comprobar que los conjuntos de lecturas están en formato *FASTQ* (Cock et al., 2010). Los conjuntos de lecturas procedentes de **ENCODE PORTAL** (Sloan et al., 2016) ya están en formato *FASTQ*, mientras que los que proceden de **SRA** (Karsch-Mizrachi et al., 2017) están en un formato binario, llamado también *SRA*. Para convertir los conjuntos de lecturas de formato *SRA* a *FASTQ* se utiliza la herramienta *fastq-dump* de **SRA Toolkit** (Karsch-Mizrachi et al., 2017).

Una vez en formato *FASTQ*, se deben podar las lecturas para corregir los siguientes problemas:

- Posiciones con baja calidad en el extremo 3′ de las lecturas, las cuales incrementan el porcentaje de lecturas no alineadas e incorrectamente alineadas y dificultan la detección de secuencias contaminantes (véase sección 1.4.1).
- Secuencias contaminantes (como los adaptadores y las bases introducidas durante la reparación de extremos), que además de incrementar el porcentaje de lecturas no alineadas e incorrectamente alineadas, pueden provocar sesgos de metilación (véase sección 1.4.2).

Para podar las lecturas, en este protocolo se utiliza **Trim Galore** (Krueger, 2018), el cual detecta automáticamente las secuencias de los adaptadores y los elimina utilizando **cutadapt** (Martin, 2011). Adicionalmente, se comprueba la calidad de secuenciación del conjunto de lecturas antes y después del podado, utilizando **FastQC** (Andrews, 2018).

Dependiendo del programa de alineamiento que se utilice, puede ser necesario procesar el ensamblado antes del alineamiento. En este protocolo, se utiliza **Bismark** (Krueger & Andrews, 2011) con **Bowtie2** (Langmead & Salzberg, 2012) como programa de alineamiento. **Bismark** requiere la obtención de índices del ensamblado utilizando su herramienta *bismark_genome_preparation*. En primer lugar, esta herramienta convierte la secuencia del ensamblado a los dos alfabetos de tres letras: i) sustituyendo C por T y ii) sustituyendo G por A. Seguidamente, obtiene los índices de **Bismark** para ambas secuencias en alfabeto de tres letras, en el formato adecuado para **Bowtie2**.

Para poder seguir la estrategia de alineamiento en dos etapas que se describe en la siguiente sección, es necesario obtener los índices de **Bismark** para las versiones completa y primaria del ensamblado GRCh38/hg18 (véase sección 4.2 para más información sobre las versiones del ensamblado).

En la sección 10.1.2.1 del anexo, se detallan las opciones utilizadas de estos programas.

4.4.2 Alineamiento en dos etapas

Durante el desarrollo de este protocolo, se descubrió un tipo de sesgo ocasionado por el uso de nuevos modelos de ensamblado genómico (véase sección 4.2 para más información sobre el ensamblado), el cual provoca la pérdida de lecturas procedentes de *loci* polimórficos, como consecuencia de un incremento en el porcentaje de lecturas con alineamiento ambiguo. En la sección 5.1.3, se muestra una comparación de los resultados obtenidos cuando se alinean las lecturas frente a dos versiones de GRCh38/hg38: con y sin haplotipos alternativos.

Para recuperar las lecturas que se pierden en los *loci* polimórficos y asignarlas al ensamblado primario, se diseñó una estrategia de alineamiento en dos etapas (véase Figura 5):

- En la primera etapa, la totalidad de las lecturas se alinean frente al ensamblado completo (con haplotipos alternativos). Cada lectura podrá alinear en una región (lecturas con alineamiento único), en varias regiones (lecturas con alineamiento ambiguo) o no alinear en ninguna región del ensamblado (lecturas no alineadas). Las lecturas cuyo alineamiento haya resultado ambiguo durante este primer alineamiento, se reutilizan en la siguiente etapa. Por su parte, las lecturas cuyo alineamiento haya resultado único, pasan a los resultados finales.
- En la segunda etapa, las lecturas procedentes de la etapa anterior se alinean frente al ensamblado primario (sin haplotipos alternativos). Las lecturas cuyo alineamiento haya resultado único, pasan a los resultados finales junto con las de la etapa anterior.

En ambas etapas de alineamiento, se utiliza **Bismark** (Krueger & Andrews, 2011) con **Bowtie2** (Langmead & Salzberg, 2012) como programa de alineamiento. **Bismark** se encarga de convertir las lecturas a los dos alfabetos de tres letras: i) sustituyendo C por T y ii) sustituyendo G por A. Seguidamente, **Bismark** utiliza **Bowtie2** para tratar de alinear ambas versiones de las lecturas frente a una versión del ensamblado en el alfabeto de tres letras que corresponda. Se ha decidido utilizar **Bowtie2** en lugar de **Bowtie** (Langmead, Trapnell, Pop, & Salzberg, 2009) porque puede introducir gaps durante el alineamiento.

Para unir los alineamientos únicos procedentes de ambas etapas, se utiliza la herramienta *merge* de **SAMtools** (H. Li et al., 2009).

En la sección 10.1.2.2 del anexo, se detallan las opciones utilizadas de estos programas.

En la sección 5.1.3, se muestra una comparación de los resultados obtenidos mediante esta estrategia de alineamiento en dos etapas frente a la estrategia convencional de alineamiento en una etapa.

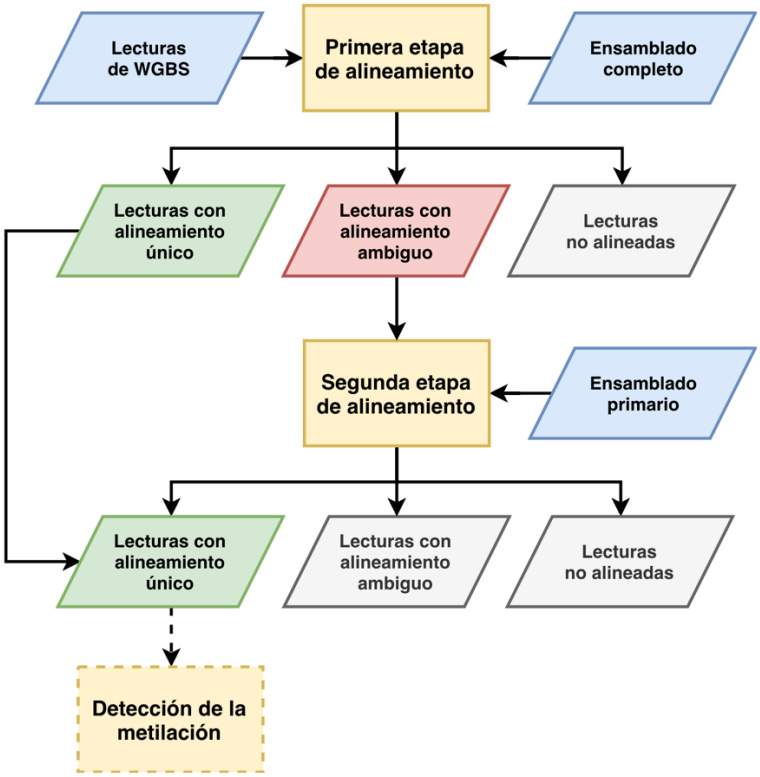

Figura 5. Estrategia de alineamiento en dos etapas

El primer alineamiento toma dos entradas: i) la totalidad de las lecturas de la muestra y ii) el ensamblado completo. Aquellas lecturas cuyo alineamiento haya resultado ambiguo durante el primer alineamiento (en rojo), se reutilizan en la siguiente etapa de alineamiento, mientras que las lecturas con alineamiento único (en verde) pasan a los resultados finales. El segundo alineamiento toma dos entradas: i) las lecturas procedentes de la etapa anterior y ii) el ensamblado primario. Las lecturas con alineamiento único pasan a los resultados finales, junto con las de la etapa anterior. Estos alineamientos serán procesados en posteriores etapas hasta la obtención del mapa de metilación. Las otras salidas producidas (en gris) se descartan.

4.4.3 Tratamiento posterior al alineamiento

Una vez se dispone de las lecturas alineadas, es necesario corregir los siguientes problemas antes de proceder a detectar los niveles de metilación de cada citosina:

- Lecturas duplicadas durante la amplificación de los fragmentos de ADN o durante su secuenciación, las cuales incrementan desproporcionadamente la cobertura de secuenciación de las regiones genómicas de las que proceden (véase sección 1.4.4). En este protocolo, se utiliza la herramienta *MarkDuplicates* de **Picard Tools** (Broad Institute, 2019) para eliminar las lecturas duplicadas.

- Indels en los extremos de las lecturas alineadas, incorrectamente interpretados como una sucesión de sustituciones (véase sección 1.4.5). En este protocolo, se utiliza una versión modificada de las herramientas *RealignerTargetCreator* y *IndelRealigner* de **GATK3** (McKenna et al., 2010), procedente de **Bis-SNP** (Liu et al., 2012), para corregir estos errores mediante un realineamiento local por el método de máxima parsimonia.
- Sesgos en el nivel de metilación de los extremos de las lecturas en comparación con la región central (véase sección 1.4.6). En este protocolo, se utiliza la herramienta *mbias* de **BSeQC** (Lin et al., 2013) para detectar y corregir el sesgo de metilación.

En la sección 10.1.2.3 del anexo, se detallan las opciones utilizadas de estos programas.

4.4.4 Detección de la metilación

Durante la detección de los niveles de metilación de cada citosina a partir de los alineamientos corregidos, es fundamental discriminar entre sustituciones C/T y citosinas no-metiladas convertidas por el bisulfito (véase sección 1.4.8). Sin embargo, la mayoría de programas para la obtención de mapas de metilación ignoran este grave problema.

En este protocolo, se utiliza **MethylExtract** (Barturen et al., 2014), que es capaz de detectar variantes de secuencia siguiendo una estrategia similar a la de **varScan** (Koboldt et al., 2009), mientras detecta el nivel de metilación de cada citosina. Además, en este protocolo se utilizan algunos de los filtros de **MethylExtract** para descartar posiciones y lecturas afectadas por los siguientes sesgos:

- Posiciones con baja calidad en el extremo 5′ o en la región central de las lecturas. Este filtro se debe aplicar *a posteriori* de la corrección de los alineamientos, ya que la región central no se puede podar y también porque el extremo 5′ es necesario para la corrección del sesgo de metilación (véase sección 1.4.1). **MethylExtract** ignora todas aquellas posiciones cuyo valor de calidad (*PHRED score*) es menor de cierto umbral.

- Lecturas afectadas por fallo en la conversión por bisulfito (véase sección 1.4.7). **MethylExtract** ignora todas aquellas lecturas cuyo porcentaje de citosinas metiladas en contextos no-CpG sea mayor de cierto umbral. Este filtro no se debe aplicar a muestras de células madre pluripotentes, ya que presentan niveles apreciables de metilación en contextos no-CpG (Ramsahoye et al., 2000; Ziller et al., 2011).

En la sección 10.1.2.4 del anexo, se detallan las opciones de **MethylExtract** utilizadas.

Los mapas de metilación obtenidos con **MethylExtract** están divididos en tres ficheros, uno para cada contexto de secuencia de las citosinas:

- **Mapa de metilación para los sitios CpG.** En células humanas diferenciadas, la metilación se localiza principalmente en este contexto (Lister et al., 2009). En esta Tesis Doctoral, se ha limitado el estudio de los cambios de metilación a los sitios CpG.
- **Mapas de metilación para los contextos CHG y CHH** (donde H es A, T o C). Si bien estos contextos parecen menos relevantes en células humanas diferenciadas, se ha detectado una cantidad significativa de citosinas metiladas en estos contextos en células madre pluripotentes (Ramsahoye et al., 2000; Ziller et al., 2011).

MethylExtract anota también las variantes de secuencia que ha encontrado en un fichero en formato *VCF* (Banks et al., 2011).

4.5 Detección de CpGs diferencialmente metilados

En esta Tesis Doctoral, se decidió limitar el estudio de los cambios de metilación en el genoma humano al contexto CpG, al tratarse del contexto más importante en células diferenciadas (Lister et al., 2009). Para estudiar la variabilidad de la

metilación, se decidió seguir una estrategia de comparación de las muestras por pares y posteriormente seleccionar aquellos cambios de metilación que fuesen característicos del tipo celular o del individuo.

En la Figura 6 se muestra el protocolo utilizado para la detección de CpGs diferencialmente metilados (DMCs), basado en el test exacto de Fisher (Fisher, 2006):

- El protocolo toma como entrada los mapas de metilación, en contexto CpG, de dos muestras distintas (muestras 1 y 2). Estas muestras podrían ser, por ejemplo, de dos tejidos distintos del mismo individuo o muestras de dos individuos para un mismo tejido. Estos mapas de metilación están en el formato de salida de **MethylExtract** (Barturen et al., 2014) y se convierten al formato de entrada de **methylKit** (S. Li et al., 2012).
- A continuación, se importan ambos mapas de metilación en **R** (https://cran.r-project.org) utilizando la función *methRead* de **methylKit**.
- Se excluyen aquellos sitios CpG que tengan una cobertura de secuenciación menor de 10 en uno o ambos mapas de metilación. De igual manera, se excluyen aquellos sitios CpG para los que no se dispone de datos en uno de los mapas de metilación. Para aplicar estos dos filtros, se utilizan las funciones *filterByCoverage* y *unite* de **methylKit**, respectivamente.
- Para cada sitio CpG, se obtiene una tabla de contingencia de 2x2 y se le aplica el test exacto de Fisher. En esta tabla de contingencia, las columnas son el número de lecturas que evidencian metilación y no-metilación y las filas son los valores para dichas columnas en las muestras 1 y 2. Una vez aplicado este test a todos los sitios CpG, se aplica una corrección de los errores de tipo I para ensayos múltiples a los valores-P (Benjamini & Hochberg, 1995), la cual disminuye la proporción de falsos positivos. Para realizar estos cálculos, se utiliza la función *calculateDiffMeth* de **methylKit**, con las opciones *adjust="BH"* y *slim=FALSE*.
- A continuación, se seleccionan como DMCs aquellos sitios CpG que presentan un valor-P corregido menor o igual a 0,05 y una diferencia en el nivel de metilación (beta) mayor o igual al 25%. Para ello, se utiliza la función *getMethylDiff* de **methylKit**, que además es capaz de separar DMCs hipometilados e hipermetilados en la muestra 2 con respecto a la muestra 1.

- Por último, se exportan a ficheros tabulares los objetos de **R** que contienen los DMCs y se convierten a formato *BED*.

El test exacto de Fisher examina si la asociación entre dos tipos de clasificación (lecturas que evidencian metilación *vs*. no-metilación en las muestras 1 y 2) es estadísticamente significativa. En este caso, se traduce en estudiar si el nivel de metilación del sitio CpG depende de la muestra.

Este protocolo se incorporó a **MethFlow** como un módulo para el análisis de la metilación diferencial (véase sección 5.1.2).

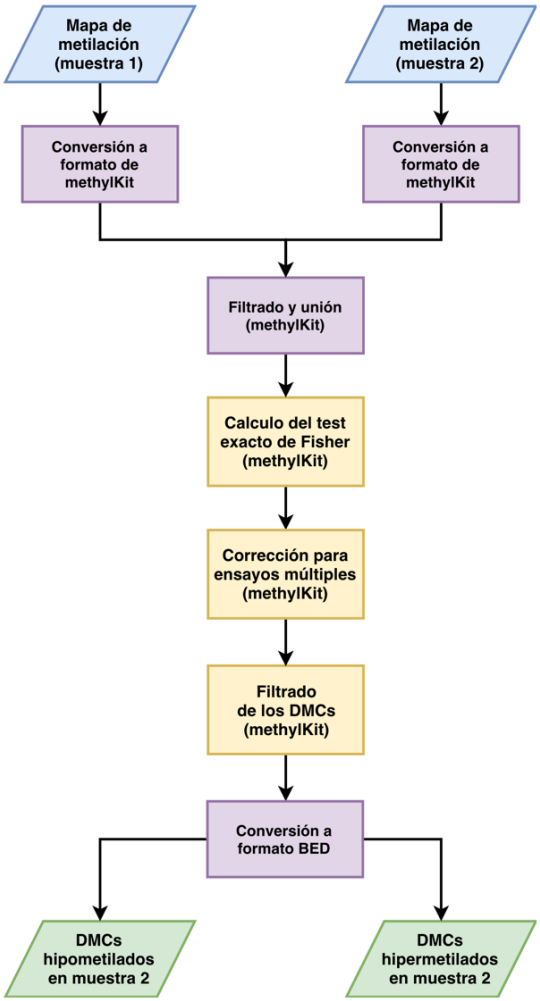

Figura 6. Protocolo de detección de CpGs diferencialmente metilados

Los romboides azules representan las entradas del protocolo (mapas de metilación de las muestras 1 y 2) y los verdes las salidas principales (conjuntos de DMCs hipometilados e hipermetilados en la muestra 2). Los rectángulos amarillos representan las etapas principales del protocolo y los violetas las etapas auxiliares.

4.5.1 Metilación diferencial intra- e inter-individual

Cada tipo celular posee un patrón de metilación característico, en parte heredado de la célula madre que le precede en su linaje (Halley-Stott & Gurdon, 2013; Jaenisch & Bird, 2003) y en parte modificado durante el proceso de diferenciación celular (J. U. Guo, Su, Zhong, Ming, & Song, 2011; M. Kim et al., 2014). De igual manera, un mismo tipo celular puede presentar ciertas diferencias de metilación entre individuos debido a factores genéticos y ambientales (Garg, Joshi, Watson, & Sharp, 2018; Hannon et al., 2018). Cabe esperar que ambos tipos de variabilidad en la metilación tengan distintas implicaciones biológicas.

Para estudiar los cambios de metilación en el genoma humano, se utilizaron algunos de los mapas de metilación de las muestras indicadas en la sección 4.1, obtenidos mediante **MethFlow** (véase sección 5.1), y se definieron dos tipos de DMCs:

- **DMCs intra-individuales (intra-DMCs)**, cuya metilación varía entre distintos órganos, tejidos y tipos celulares de un mismo individuo. En la columna *intra-DMCs* de la **Tabla 18**, se indican las muestras cuyos mapas de metilación se han utilizado para detectar este tipo de DMCs. Se han obtenido tantos conjuntos de intra-DMCs como pares posibles de muestras del mismo individuo.
- **DMCs inter-individuales (inter-DMCs)**, cuya metilación varía entre individuos para un órgano, tejido o tipo celular dado. En la columna *inter-DMCs* de la **Tabla 18**, se indican las muestras cuyos mapas de metilación se han utilizado para detectar este tipo de DMCs. Se han obtenido tantos conjuntos de inter-DMCs como pares posibles de muestras del mismo órgano, tejido o tipo celular.

Para obtener los conjuntos de DMCs de los pares de muestras indicados, se utilizó el módulo de metilación diferencial de **MethFlow** (véase sección 5.1.2), el cual sigue el protocolo descrito en la sección 4.5. Se limitó el estudio a los sitios CpG que forman parte de los autosomas. Todos estos conjuntos de DMCs están disponibles en la base de datos dedicada a la metilación **NGSmethDB** (véase sección 5.4).

A partir de los conjuntos de DMCs por pares de muestras, se obtuvieron dos conjuntos estrictos para análisis posteriores:

- **Conjunto estricto de intra-DMCs.** En primer lugar, se seleccionaron los conjuntos de intra-DMCs para aquellos individuos de los que se dispone de 5 o más muestras. Muestra a muestra, se seleccionaron aquellos sitios CpG que son intra-DMCs para esta muestra frente a todas las demás. Por último, se reunieron en un único conjunto todos los sitios CpG seleccionados. Estos son los intra-DMCs cuya metilación es específica del órgano, tejido o tipo celular.
- **Conjunto estricto de inter-DMCs.** En primer lugar, se seleccionaron los conjuntos de inter-DMCs para aquellos órganos, tejidos o tipos celulares de los que se dispone de 5 o más muestras. Muestra a muestra, se seleccionaron aquellos sitios CpG que son inter-DMCs para esta muestra frente a todas las demás. Por último, se reunieron en un único conjunto todos los sitios CpG seleccionados. Estos son los inter-DMCs cuya metilación es específica del individuo.

4.6 Detección de semáforos CpG

Si bien los estudios tradicionales de asociación con la transcripción se basan en el nivel de metilación promedio de los promotores, hoy se sabe que solo el 16,6% de los sitios CpG en promotores ejercen un efecto sobre la transcripción cuando cambia su metilación (Harbers et al., 2014) (véase sección 1.2).

Recientemente, se han descrito los llamados "semáforos CpG" (CpG-TLs), los cuales son sitios CpG individuales cuyo nivel de metilación está asociado con la tasa de transcripción de un gen cercano (Harbers et al., 2014; Lioznova et al., 2019). Estos marcadores se detectan a partir de vectores de metilación-transcripción, donde cada vector corresponde a un par CpG-gen y cada valor procede de una muestra distinta. Para detectar si existe una asociación significativa entre la metilación del CpG y la transcripción del gen en cada uno de estos pares, estos

autores aplican el método no-paramétrico de correlación de Spearman (Spearman, 1987).

Sin embargo, el coeficiente de correlación de Spearman es sensible a los valores atípicos (Vanhove, 2016), por lo que no es suficiente para obtener resultados fiables. Los coeficientes de correlación son populares entre los investigadores porque permiten resumir la relación entre dos variables. Sin embargo, un determinado coeficiente de correlación puede representar infinitos patrones de dispersión de las dos variables, por lo que es necesario representar gráficamente ambas variables para saber de que patrón se trata. La Figura 7 y la Figura 8 muestran dos ejemplos en los que dos variables (por ejemplo, metilación y transcripción) representadas por 50 puntos (muestras) presentan un coeficiente de correlación de Spearman de 0,7 y -0,7, respectivamente. Como se puede observar, existen una infinidad de situaciones posibles que pueden llevar a que dos variables presenten un determinado coeficiente de correlación. Cada una de estas situaciones puede tener una explicación biológica diferente y algunas de ellas se tratarán de simples artefactos técnicos. Dado que es impracticable y poco objetivo analizar *de visu* la dispersión de la metilación y la transcripción en 518.293.855 pares CpG-gen, es necesario aplicar un segundo test estadístico que permita filtrar los resultados obtenidos mediante el test de correlación de Spearman.

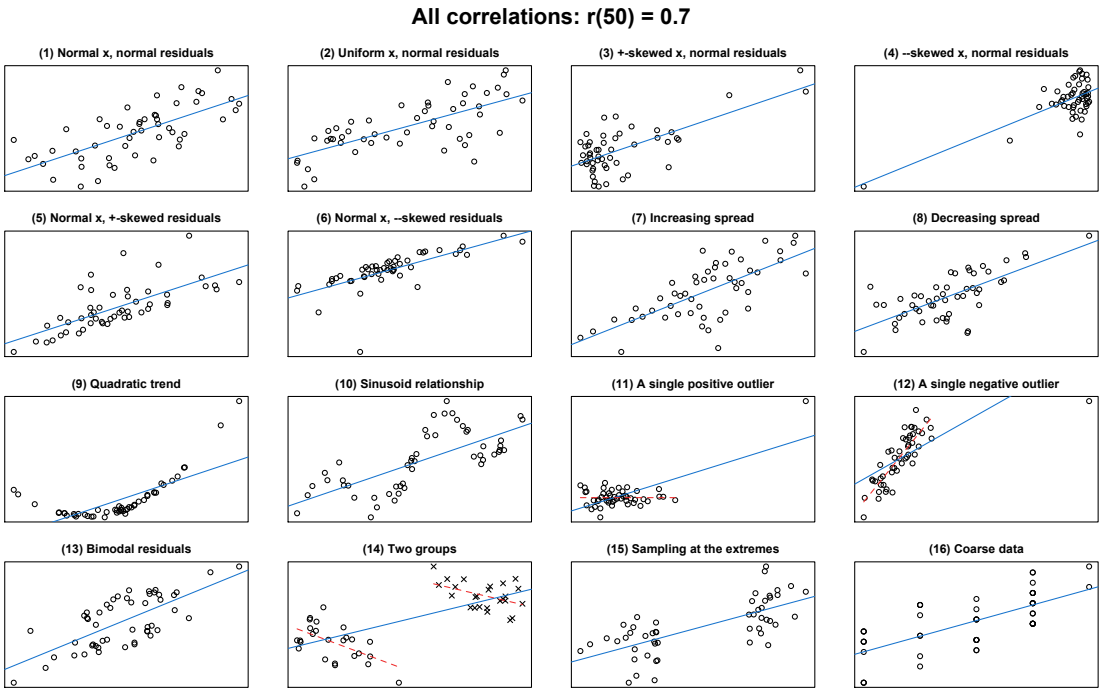

Figura 7. Ejemplos de patrones de dispersión con la misma correlación positiva

Todas estas distribuciones están formadas por 50 valores y su coeficiente de correlación de Spearman es 0,7. Mientras que en las distribuciones (1) y (2) las variables guardan una relación lineal siguiendo una distribución normal o uniforme, respectivamente, el resto de distribuciones muestran relaciones no-lineales o relaciones lineales con ciertos sesgos. Estas distribuciones se han obtenido con una función de R desarrollada por Vanhove (Vanhove, 2016).

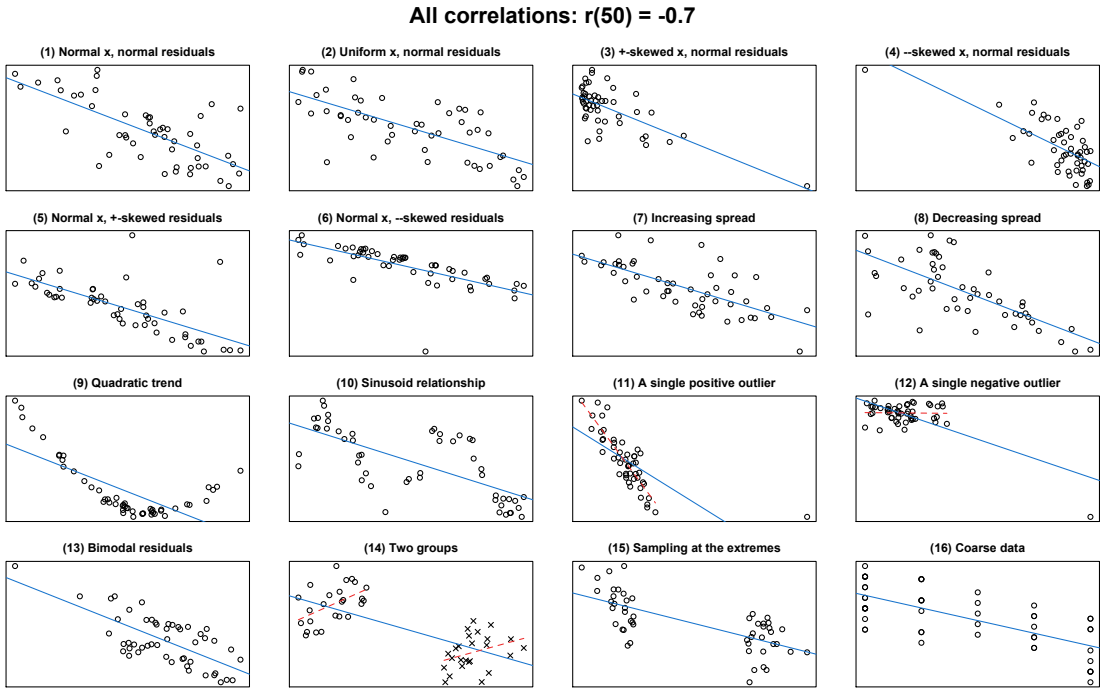

Figura 8. Ejemplos de patrones de dispersión con la misma correlación negativa

Todas estas distribuciones están formadas por 50 valores y su coeficiente de correlación de Spearman es -0,7. Mientras que en las distribuciones (1) y (2) las variables guardan una relación lineal siguiendo una distribución normal o uniforme, respectivamente, el resto de distribuciones muestran relaciones no-lineales o relaciones lineales con ciertos sesgos. Estas distribuciones se han obtenido con una función de R desarrollada por Vanhove (Vanhove, 2016).

En esta Tesis Doctoral, se ha desarrollado un método de detección de CpG-TLs que reduce este problema y aumenta la fiabilidad de los resultados, utilizando una combinación del coeficiente de correlación de Spearman y el test no-paramétrico de Kruskal-Wallis (Kruskal & Wallis, 1952). De esta manera, se puede evaluar la asociación de la metilación con la transcripción desde dos perspectivas complementarias:

- Por una parte, el coeficiente de correlación de Spearman ordena las parejas de valores metilación-transcripción de cada muestra y determina si al ordenar de forma ascendente la metilación, la transcripción asciende (asociación positiva) o desciende (asociación negativa) de forma monótona.
- Por otra parte, los valores de transcripción se pueden dividir en dos grupos: i) valores de transcripción de las muestras en las que el sitio CpG está metilado y ii) valores de transcripción de las muestras en las que el sitio CpG está no-metilado. Utilizando el test de Kruskal-Wallis es posible comprobar si las muestras en las

que el sitio CpG está metilado presentan una distribución de valores de transcripción diferente de las muestras en las que el sitio CpG está no-metilado.

El protocolo de detección de CpG-TLs se incorporó a **MethFlow** como un módulo para el análisis de la asociación con la transcripción (véase seción 5.1.2). En la Figura 9 se muestran las etapas de este protocolo, descritas a continuación.

En primer lugar, es necesario disponer de los mapas de metilación y de los perfiles de transcripción para una colección amplia de diferentes tipos celulares e individuos:

- En esta Tesis Doctoral, se han utilizado mapas de metilación del contexto CpG obtenidos mediante **MethFlow** (véase sección 5.1). Estos mapas están en el formato de salida de **MethylExtract** (Barturen et al., 2014). Para la detección de los CpG-TLs caracterizados en esta Tesis Doctoral, se utilizaron los mapas de metilación de las muestras indicadas en la columna *CpG-TLs* de la **Tabla 18**. Se limitó el estudio a los sitios CpG que forman parte de los autosomas.
- Los perfiles de expresión deben pertenecer a las mismas muestras biológicas de las que se han obtenido los mapas de metilación y estar en el formato descrito en la sección 10.1.3 del anexo. Para la detección de los CpG-TLs caracterizados en esta Tesis Doctoral, se utilizaron los perfiles de expresión de las muestras indicadas en la columna *CpG-TLs* de la **Tabla 18**, obtenidos a través de **ENCODE PORTAL** (Sloan et al., 2016) (véase sección 4.1). Se limitó el estudio a los genes que forman parte de los autosomas.

Durante la primera etapa del protocolo, estos mapas de metilación y perfiles de transcripción se filtran y transforman en matrices:

- Se filtra cada mapa de metilación para eliminar aquellos sitios CpG cuya cobertura de secuenciación es menor de 10. Tras esto, se excluyen del análisis los sitios CpG para los que se dispone de datos en menos de la mitad de las muestras. Por último, se construye una matriz de metilación, donde cada fila es un sitio CpG y cada columna una muestra. Esta matriz se comprime con **bgzip** y se indiza con **tabix** (H. Li, 2011).

- Los perfiles de transcripción se filtran siguiendo los criterios de GTEx para **eQTLs** (Aguet et al., 2017), es decir, que se excluyen todos los genes que tienen un nivel de transcripción menor de 0,1 TPM (transcriptos por millón) (Wagner, Kin, & Lynch, 2012) o menor de 6 lecturas en menos del 20% de las muestras. Por último, se construye una matriz de transcripción, donde cada fila es un gen y cada columna una muestra.

A continuación, se recorre la matriz de transcripción gen a gen:

- Para cada gen, se obtienen todos los sitios CpG que estén ubicados desde 1 Mpb aguas arriba del sitio de inicio de la transcripción (TSS) hasta 1 Mpb aguas abajo del sitio de fin de la transcripción (TES). Con cada uno de ellos se forma un par CpG-gen.
- Para cada par CpG-gen, se obtiene un vector de metilación-transcripción con los valores de cada una de las muestras. Cada par CpG-gen se evalúa en las siguientes etapas para determinar si el sitio CpG es un CpG-TL para ese gen.
- También se divide la distribución de los valores de transcripción en tres, en función del estado de metilación del sitio CpG:
    - **Estado metilado (M)**, si el nivel de metilación es mayor o igual a 0,8.
    - **Estado intermedio (I)**, si el nivel de metilación es menor que 0,8 y mayor que 0,2.
    - **Estado no-metilado (U)**, si el nivel de metilación es menor que 0,2.

Para llevar a cabo estas operaciones, es necesario consultar la matriz de metilación utilizando el paquete **pytabix** (https://github.com/slowkow/pytabix) de **Python** (https://www.python.org).

Seguidamente, a cada uno de los pares CpG-gen se le aplican dos tests estadísticos:

- **El coeficiente de correlación de Spearman (rho).** Para ello, se utiliza la función *stats.mstats.spearmanr* del paquete **SciPy** (https://www.scipy.org) de **Python**.
- **El test de Kruskal-Wallis (H).** Se aplica este test a las distribuciones de transcripción de las dos clases de metilación más extremas (M vs. U). Si no hay ninguna muestra en una de estas clases, se tomará la clase intermedia (M vs. I o I vs. U). Por último, si todas las muestras están agrupadas dentro de una misma clase (M, I o U), se descarta el par CpG-gen. Para calcular el test H, se utiliza la función *stats.kruskal* del paquete **SciPy** de **Python**.

Una vez que se han procesado todos los posibles pares CpG-gen, se aplica la corrección de los errores de tipo I para ensayos múltiples a los valores-P de ambos tests. Para ello, se utiliza la función *sandbox.stats.multicomp.multipletests* del paquete **StatsModels** (https://www.statsmodels.org) de **Python**.

Por último, se seleccionan como CpG-TLs aquellos pares CpG-gen cuyos valores-P corregidos para ambos tests sean menor o igual a 0,05. Dependiendo del signo de rho, se distinguen dos tipos de CpG-TLs:

- **CpG-TLs rojos**, cuando rho es negativo y, por tanto, la asociación de la metilación con la transcripción es también negativa.
- **CpG-TLs verdes**, cuando rho es positivo y, por tanto, la asociación de la metilación con la transcripción es también positiva.

Los conjuntos de CpG-TLs rojos y verdes obtenidos a partir de las muestras indicadas en la columna *CpG-TLs* de la **Tabla 18**, están disponibles en la base de datos dedicada a la metilación **NGSmethDB** (véase sección 5.4).

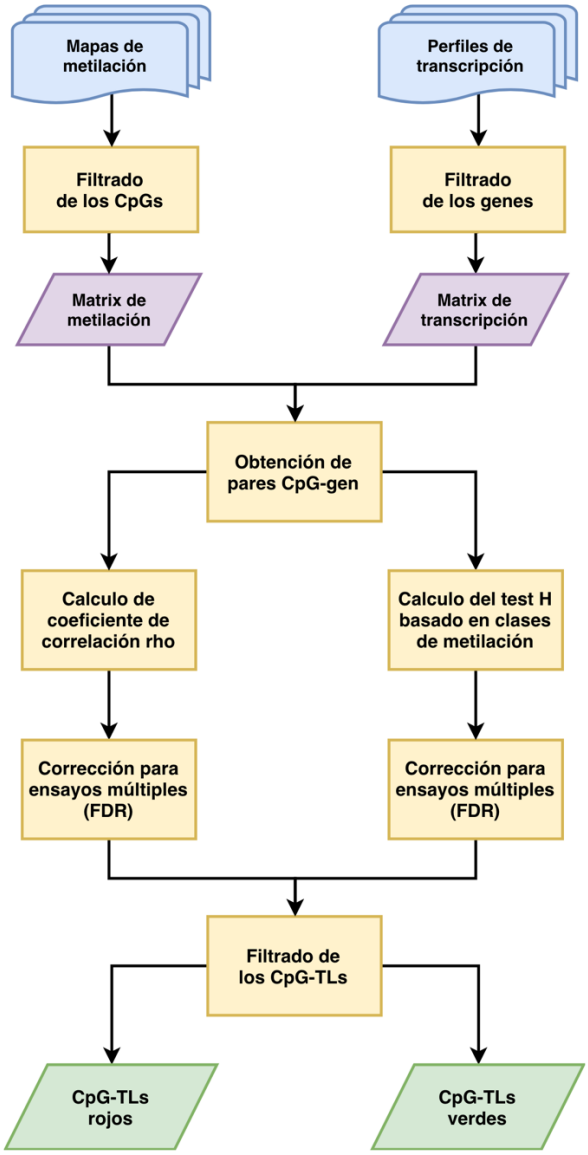

Figura 9. Protocolo de detección de semáforos CpG

En azul se representan las entradas del protocolo (mapas de metilación y perfiles de transcripción), en verde las salidas principales (conjuntos de CpG-TLs rojos y verdes). Los rectángulos amarillos representan las etapas principales del protocolo y los violetas las salidas intermedias.

4.7 Análisis de subconjuntos de sitios CpG

Los sitios CpG no se distribuyen aleatoriamente a lo largo del genoma, sino que existen elementos genómicos que tienen una densidad especialmente elevada en sitios CpGs en comparación con el resto del genoma, como los retrotransposones Alu (Y. Luo, Lu, & Xie, 2014), las islas CpG (Deaton & Bird, 2011) y muchos

promotores (Saxonov, Berg, & Brutlag, 2006). Por tanto, la densidad en CpG-TLs que tiene un elemento puede estar condicionada por su densidad en CpG. Sin embargo, que un elemento genómico tenga una alta densidad en CpG-TLs no quiere decir que la proporción de sus sitios CpG que son CpG-TLs sea también alta. Esto es aplicable a cualquier otro subconjunto de sitios CpG, como los DMCs.

Una hipótesis plausible es que la proporción de CpG-TLs sea biológicamente más relevante que la densidad. Por ejemplo, un elemento genómico con baja densidad en sitios CpG puede ser también poco denso en CpG-TLs y, sin embargo, que la mayoría o la totalidad de sus sitios CpG sean CpG-TLs implicados en la unión a factores de transcripción.

En esta Tesis Doctoral, se utilizaron medidas basadas en la proporción de DMCs o CpG-TLs con diferentes propósitos:

- Analizar la riqueza en DMCs o en CpG-TLs de diferentes elementos genómicos.
- Analizar la distribución de los DMCs y los CpG-TLs con respecto al TSS y al TES.

4.7.1 Análisis de riqueza

En esta Tesis Doctoral, se desarrolló un método propio para estudiar la riqueza en DMCs o en CpG-TLs de diferentes elementos genómicos, cuya medida de la riqueza se definió así:

$$riqueza = \frac{Y_{in}/(Y_{in} + N_{in})}{Y_{out}/(Y_{out} + N_{out})}$$

Donde Y_{in} es el número de CpGs etiquetados dentro del elemento, N_{in} el número de CpGs no-etiquetados dentro del elemento, Y_{out} el número de CpGs etiquetados fuera del elemento y N_{out} el número de CpGs no-etiquetados fuera del elemento. Se entiende por "CpGs etiquetados" aquellos sitios CpG que formen parte del subconjunto que se esté estudiando (por ejemplo, CpG-TLs) y por "CpGs no-etiquetados" al resto de sitios CpG.

Esta medida de riqueza es fácil de interpretar y de comparar:

- Los valores de riqueza menores que 1 indican que el elemento genómico está empobrecido en CpGs etiquetados. Por ejemplo, un valor de 0,5 indica que el elemento genómico tiene una proporción de CpGs etiquetados dos veces menor que el resto del genoma.
- Los valores de riqueza mayores que 1 indican que el elemento genómico está enriquecido en CpGs etiquetados. Por ejemplo, un valor de 2 indica que el elemento genómico tiene una proporción de CpGs etiquetados dos veces mayor que el resto del genoma.

La significación estadística del valor de riqueza se calcula mediante el uso de muestreos aleatorios con reemplazamiento o bootstrapping (Efron, 1979):

- Se generan 10^3 conjuntos de sitios CpG etiquetados al azar. Dado que se trata de muestreos aleatorios con reemplazamiento, la probabilidad de ser etiquetado de cada sitio CpG es independiente a la del resto.
- Se calcula el valor de riqueza para cada uno de estos conjuntos aleatorios. Estos valores de riqueza constituyen la distribución esperada de valores de riqueza.
- Por último, se calcula el Z-score y el valor-P de la riqueza observada frente a la distribución de riquezas esperadas.

La Figura 10 describe el funcionamiento de este método de enriquecimiento. Este método se incorporó al módulo de herramientas de **MethFlow** (véase sección 5.1.2).

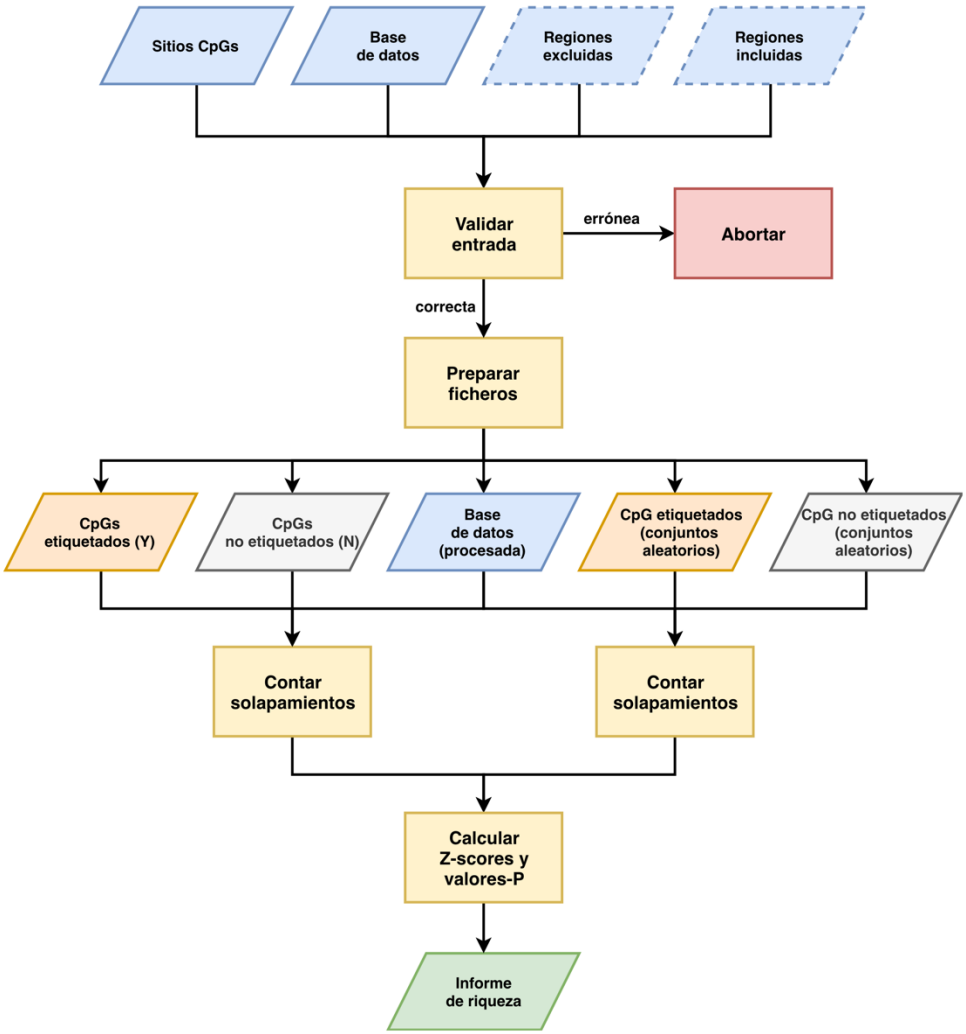

Figura 10. Análisis de riqueza en subconjuntos de sitios CpG

Este método, toma dos entradas obligatorias: i) las coordenadas de cada sitio CpG en formato *BED*, incluyendo una columna adicional que indica si el sitio CpG está etiquetado o no; ii) una base de datos de anotaciones genómicas, compuesta por un directorio con un fichero *BED* para cata tipo de elemento genómico que se desee estudiar. Además, puede tomar otras dos entradas opcionales: i) regiones que obligatoriamente deben excluirse del análisis; ii) regiones a las que obligatoriamente debe circunscribirse el análisis. Primero, se validan los ficheros de entrada y, si todo está correcto, se preparan los ficheros para las siguientes etapas. Durante el proceso, se eliminan las regiones redundantes de la anotación y se generan conjuntos aleatorios de sitios CpGs etiquetados y no etiquetados (por defecto, 1000 conjuntos). Estos conjuntos aleatorios se utilizarán para calcular la significación estadística (Z-score y valor-P) de la riqueza, para cada uno de los elementos genómicos. Por último, se genera una tabla de enriquecimientos con la riqueza calculada para cada anotación, su Z-score, su valor-P, si es significativo o no, si hay enriquecimiento (riqueza > 1) o empobrecimiento (riqueza < 1) cuando la riqueza es significativa y los valores utilizados por los cálculos.

4.7.2 Análisis de distancia

En torno al TSS, la densidad de los sitios CpG aumenta como consecuencia de que muchos promotores contienen una isla CpG (Deaton & Bird, 2011; Saxonov et al., 2006). También algunos TES solapan con islas CpG en genes implicados en el desarrollo (D.-H. Yu et al., 2013). Por ello, no sería extraño que hubiese una mayor densidad de DMCs o de CpG-TLs cerca del TSS o del TES, ya que son regiones ricas en sitios CpG. Para evitar este problema, se midió la proporción de los DMCs o CpG-TLs a una determinada distancia del TSS o TES con la siguiente fórmula:

$$R_d = \frac{S_d}{T_d} \times 100$$

Donde R_d es la proporción de DMCs o CpG-TLs a la distancia *d*, S_d es el número de DMCs o CpG-TLs que están a la distancia *d* y T_d es el número total de sitios CpGs que están a la distancia *d*, expresada en pares de bases (pb).

Siguiendo este método, se analizó la distribución de distancias al TSS y al TES de los DMCs y los CpG-TLs de la siguiente manera:

- Los DMCs no están a priori asociados a ningún gen. Se podrían estudiar las distancias de cada DMC con todos los TSSs o TESs que estén en el mismo cromosoma, pero el resultado sería difícil de interpretar biológicamente. A modo de simplificación, se estudió la distancia de cada DMC a su TSS más próximo y a su TES más próximo (valor S_d), siempre que no esté a una distancia mayor a 10 kpb. Se procedió de igual manera con el total de sitios CpG (valor T_d), calculando su distancia al TSS más próximo y al TES más próximo hasta un máximo de 10 kbp.
    - Los CpG-TLs, en cambio, están asociados a un gen. Se estudió la distancia de cada CpG-TL al TSS y al TES de su gen asociado (valor S_d), descartando aquellas distancias que fueran mayores a 10 kpb. Se procedió de igual manera con todos los pares CpG-gen estudiados (valor T_d), resultaran o no en la detección de un CpG-TL.

Capítulo 5

Resultados

5.1 MethFlow: una herramienta para el análisis de la metilación

A lo largo de este Tesis Doctoral, se desarrolló una herramienta para el análisis de datos de metilación, conocida como **MethFlow**, que integra:

- El protocolo de obtención de mapas de metilación a partir de lecturas de WGBS, descrito en la sección 4.4. Este protocolo incorpora una serie de mejoras que reducen los errores en la detección de la metilación de citosinas individuales. Destaca su estrategia de alineamiento en dos etapas.
- El método de detección de DMCs entre pares de muestras, descrito en la sección 4.5. Este método se ha utilizado para la detección de DMCs intra-individuales e inter-individuales.
- El método de detección de CpG-TLs, descrito en la sección 4.6. Se trata de sitios CpG individuales, cuya metilación se asocia con la transcripción de uno o más genes.

A continuación, se describe la implementación de MethFlow y los módulos de los que está compuesto. También se muestra la comparación entre el alineamiento

convencional y el alineamiento en dos etapas, que justificó su incorporación al protocolo de obtención de mapas de metilación.

5.1.1 Implementación

MethFlow está formado por *pipelines*, organizadas en módulos. Para implementarla, se ha utilizado **Nextflow** (Di Tommaso et al., 2017), un sofisticado *framework* de código abierto para el desarrollo de *pipelines* complejas, entendiéndose como tales aquellas en las que:

- La ejecución de algunos procesos está condicionada por las opciones escogidas por el usuario o por el tipo de entrada.
- El flujo de datos entre procesos está ramificado, creando un árbol de procesos en forma de grafo dirigido acíclico.
- Se toman decisiones automáticas sobre qué procesos se pueden ejecutar en paralelo y cuáles deben esperar hasta que su entrada esté lista o hasta que los recursos computacionales que necesita estén disponibles.

Nextflow simplifica considerablemente el desarrollo de *pipelines* con estas necesidades y además las dota de características que garantizan la reproducibilidad de los resultados y un control exhaustivo de los procesos:

- El uso de ficheros de configuración fáciles de leer y de modificar, que contienen las opciones de los trabajos que se van a ejecutar.
- Cada proceso se ejecuta dentro de un entorno aislado, evitando que interfiera con otros procesos.
- Una vez finalizado el proceso, el entorno se destruye, pero deja traza de toda su actividad: entrada, salida, errores, comandos ejecutados, variables de entorno, etc.
- Permite limitar el uso de memoria y del procesador, así como el tiempo máximo de computo, tanto para el trabajo en su conjunto como para procesos concretos.
- La *pipeline* se puede interrumpir cuando sea necesario y reanudarla más tarde, continuando el trabajo por el proceso en el que se interrumpió.
- Si un proceso falla por superar el tiempo máximo de computo, el programa lo reintenta un determinado número de veces (ajustable por el usuario) antes de cancelarlo.

- Si un proceso falla por algún problema en los datos de entrada o en las opciones, se pueden realizar los cambios oportunos y el programa detectará automáticamente que procesos puede tomar de la caché y cuales debe repetir. Esto se puede utilizar también para ejecutar nuevos trabajos similares a trabajos previos, aprovechando las etapas ya completadas cuyas entradas y opciones no se hayan visto modificadas.

El entorno aislado en que se ejecutan los procesos es una imagen de contenedor **Docker** (https://www.docker.com), que está disponible en **Docker Hub** (https://hub.docker.com) y que MethFlow descargará automáticamente en su primera ejecución. En esta imagen de contenedor se han incluido todos los *scripts* propios y *software* de terceros que necesitan los protocolos de MethFlow, descritos en las secciones 4.4, 4.5 y 4.6. De cada uno de los programas y paquetes incluidos en la imagen del contenedor se conoce la versión, el proceso de instalación y configuración. Además, MethFlow contiene un fichero *Dockerfile* a partir del cual se puede volver a generar una imagen de contenedor idéntica siempre que sea necesario. De esta manera, se garantiza que el *software* utilizado sea siempre el mismo y que los resultados obtenidos sean reproducibles en otros ordenadores. Gracias a esta imagen de contenedor, las dos únicas dependencias de MethFlow son **Java 8 o superior** (https://www.java.com) y **Docker 18 o superior** (https://docs.docker.com/install).

MethFlow es un programa de código abierto y está disponible en el repositorio *rlebron-bioinfo/methflow* de **GitHub** (https://github.com/rlebron-bioinfo/methflow). Este repositorio contiene toda la documentación necesaria para la instalación y la utilización de los módulos de MethFlow, así como tutoriales con ejemplos de uso rápido.

5.1.2 Módulos

MethFlow está compuesto por cinco módulos especializado en distintas tareas e interconectados entre sí (véase Figura 11):

- **Módulo principal.** Fue el primer módulo que se desarrolló y se utiliza para la obtención de mapas de metilación a partir de lecturas de WGBS. Está contenido dentro del directorio *meth* en el repositorio de MethFlow y se dispone de dos *pipelines*, que son versiones alternativas del protocolo descrito en la sección 4.4:
    - **Versión para ensamblados clásicos.** Esta versión es la que se debe utilizar cuando el ensamblado que se esté utilizando carezca de haplotipos alternativos. En lugar de utilizar la estrategia de alineamiento en dos etapas, utiliza el alineamiento convencional en una etapa. Está contenida dentro del directorio *meth/one-stage* en el repositorio de MethFlow.
    - **Versión para nuevos ensamblados.** Esta versión es la que se debe utilizar cuando el ensamblado que se esté utilizando disponga de haplotipos alternativos. Utiliza la estrategia de alineamiento en dos etapas. Está contenida dentro del directorio *meth/two-stage* en el repositorio de MethFlow. Esta versión es la que se ha utilizado para la obtención de mapas de metilación en esta Tesis Doctoral.
- **Módulo de metilación diferencial.** Se utiliza para la detección de DMCs en pares de muestras, siguiendo el protocolo descrito en la sección 4.5. Toma como entrada un directorio con mapas de metilación obtenidos mediante el módulo principal y un fichero tabular con dos columnas (una para cada muestra del par) en las que se indican los pares de muestras que se quieren comparar y ejecuta en paralelo tantas comparaciones como permitan los recursos máximos especificados. Está contenido dentro del directorio *dmcs* del repositorio de MethFlow.
- **Módulo de asociación con la transcripción.** Se utiliza para la detección de CpG-TLs a partir de mapas de metilación obtenidos mediante el módulo principal y datos externos de transcripción en el formato especificado en la sección 10.1.3 del anexo. Este módulo sigue el protocolo descrito en la sección 4.6. Está contenido dentro del directorio *tls* del repositorio de MethFlow.
- **Módulo de descargas.** Se utiliza para descargar mapas de metilación de la base de datos dedicada a la metilación NGSmethDB (véase sección 5.4), para más tarde utilizarlos en la detección de DMCs. Para descargar los datos, se conecta con el servidor RESTful API de NGSmethDB mediante HTTPS, por lo que requiere una

conexión a internet (véase sección 5.4.2). Está contenido dentro del directorio *dump* del repositorio de MethFlow.

- **Módulo de herramientas.** Incluye varias herramientas que permiten, entre otras cosas: i) calcular la riqueza en DMCs o CpG-TLs de distintos tipos de anotaciones génicas, ii) calcular la proporción de DMCs o CpG-TLs en torno al TSS o al TES y iii) convertir los mapas de metilación a formatos utilizados por otros programas, como el formato de **methylKit** (S. Li et al., 2012) o el formato *bedMethyl* de ENCODE (Sloan et al., 2016). Está contenido dentro del directorio *tools* del repositorio de MethFlow.

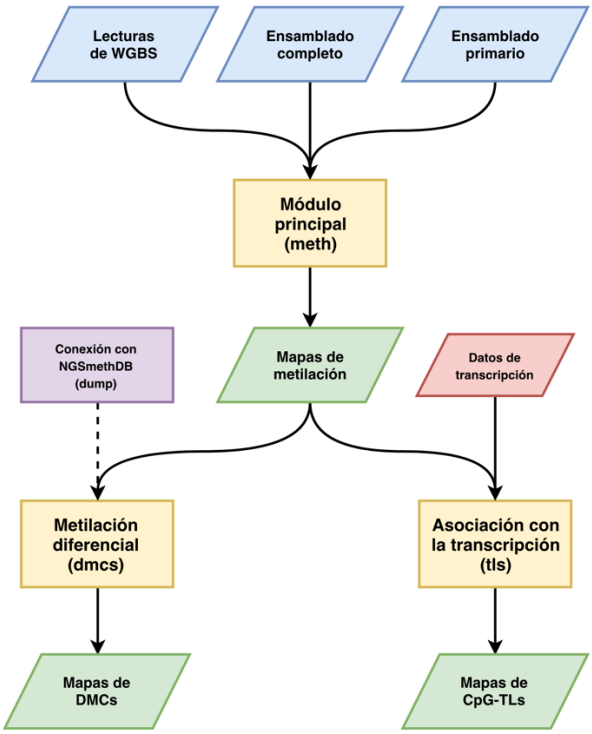

Figura 11. Estructura y flujo de datos de MethFlow

Se indican en azul las entradas al módulo principal y en rojo las entradas adicionales que necesitan otros módulos. En violeta se indica la descarga de mapas de metilación de NGSmethDB, que se pueden utilizar como entrada del módulo de metilación diferencial. En verde se indican las salidas de MethFlow: i) mapas de metilación, ii) mapas de DMCs y iii) mapas de CpG-TLs. Se indican en amarillo los módulos más importantes y en violeta los módulos secundarios. El módulo de herramientas no se ha representado en la figura.

5.1.3 Recuperación de lecturas en *loci* polimórficos

Durante el desarrollo del protocolo de obtención de mapas de metilación (véase sección 4.4), se descubrió un tipo de sesgo ocasionado por el uso de nuevos modelos de ensamblado genómico, que contienen secuencias adicionales que representan haplotipos alternativos de distintas poblaciones y etnias humanas (Church et al., 2011). Las secuencias de estos haplotipos alternativos reducen el porcentaje de alineamientos incorrectos, ya que en ausencia de estas secuencias las lecturas procedentes de haplotipos alternativos alinearían en otras regiones del genoma (M. L. Mendoza et al., 2015).

Sin embargo, una comparación con 64 muestras de 33 tipos celulares y 10 individuos (véase **Tabla 21**) realizada en esta Tesis Doctoral, ha demostrado que el uso de ensamblados con haplotipos alternativos reduce el porcentaje de lecturas con alineamiento único (véase Figura 12), a la par que aumenta el porcentaje de lecturas con alineamiento ambiguo (véase Figura 13).

Tras aplicar la estrategia de alineamiento en dos etapas (véase sección 4.4.2), los porcentajes de lecturas con alineamiento único y de lecturas con alineamiento ambiguo vuelven a ser muy similares a los que se obtienen cuando se prescinde de los haplotipos alternativos. En la **Tabla 21**, se indican los porcentajes de lecturas con alineamiento único y con alineamiento ambiguo para cada muestra frente al ensamblado primario (sin haplotipos alternativos), frente al ensamblado completo (con haplotipos alternativos) y utilizando la estrategia de alineamiento en dos etapas.

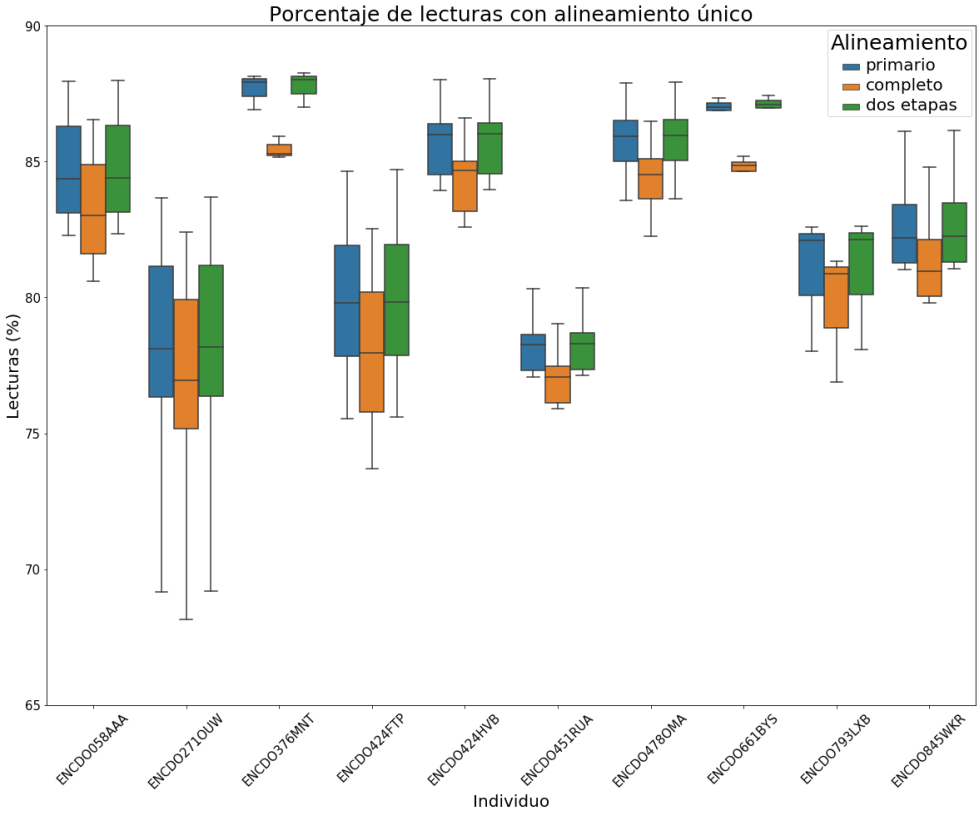

Figura 12. Porcentaje de lecturas con alineamiento único

Cada gráfico de caja representa la distribución del porcentaje de lecturas con alineamiento único para un individuo. Se puede observar que el efecto de los haplotipos alternativos sobre el alineamiento sigue la misma tendencia para todos los individuos y que, de igual manera, en todos los casos el alineamiento en dos etapas devuelve el porcentaje de lecturas con alineamiento único a valores muy similares a lo que se obtendrían prescindiendo de los haplotipos alternativos.

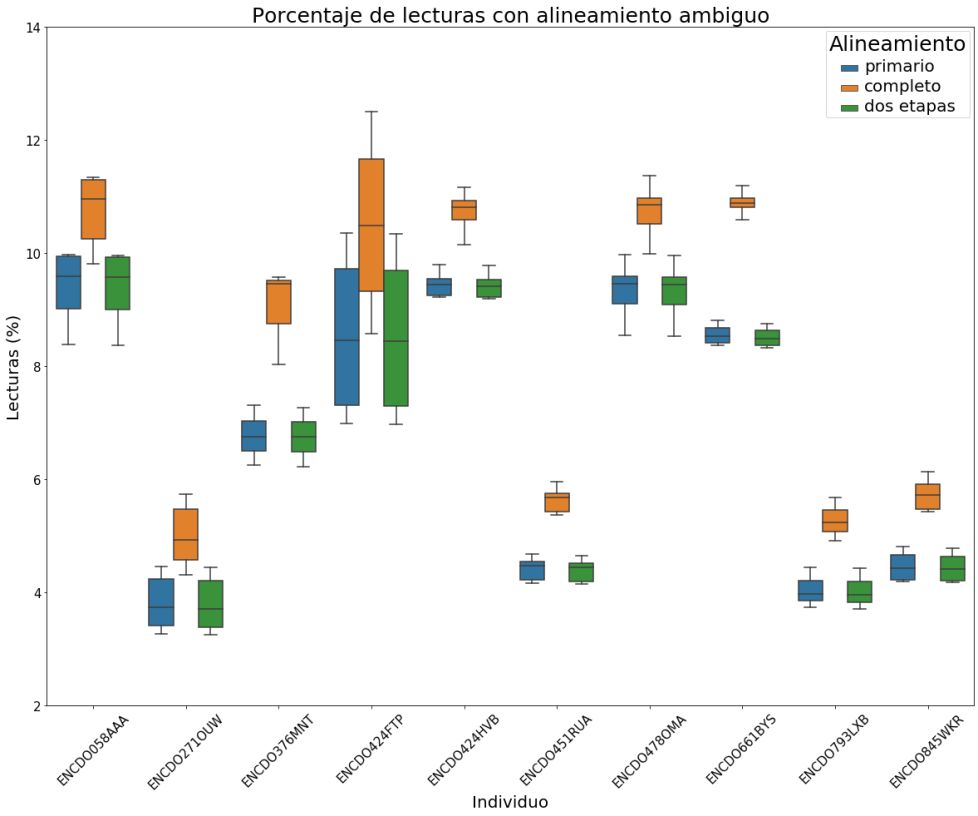

Figura 13. Porcentaje de lecturas con alineamiento ambiguo

Cada gráfico de caja representa la distribución del porcentaje de lecturas con alineamiento ambiguo para un individuo. Se puede observar que el efecto de los haplotipos alternativos sobre el alineamiento sigue la misma tendencia para todos los individuos y que, de igual manera, en todos los casos el alineamiento en dos etapas devuelve el porcentaje de lecturas con alineamiento ambiguo a valores muy similares a lo que se obtendrían prescindiendo de los haplotipos alternativos.

En el alineamiento en dos etapas, el porcentaje de lecturas ambiguas recuperadas en la segunda etapa de alineamiento osciló entre el 12% y el 29%, dependiendo de la muestra (véase Figura 14). De las lecturas recuperadas, entre el 97% y el 99% (dependiendo de la muestra) se asignaron correctamente en las regiones de los *loci* alternativos en el ensamblado primario (véase Figura 15). En la **Tabla 21**, se indican los porcentajes de lecturas ambiguas recuperadas y de lecturas correctamente asignadas tras el alineamiento en dos etapas para cada muestra.

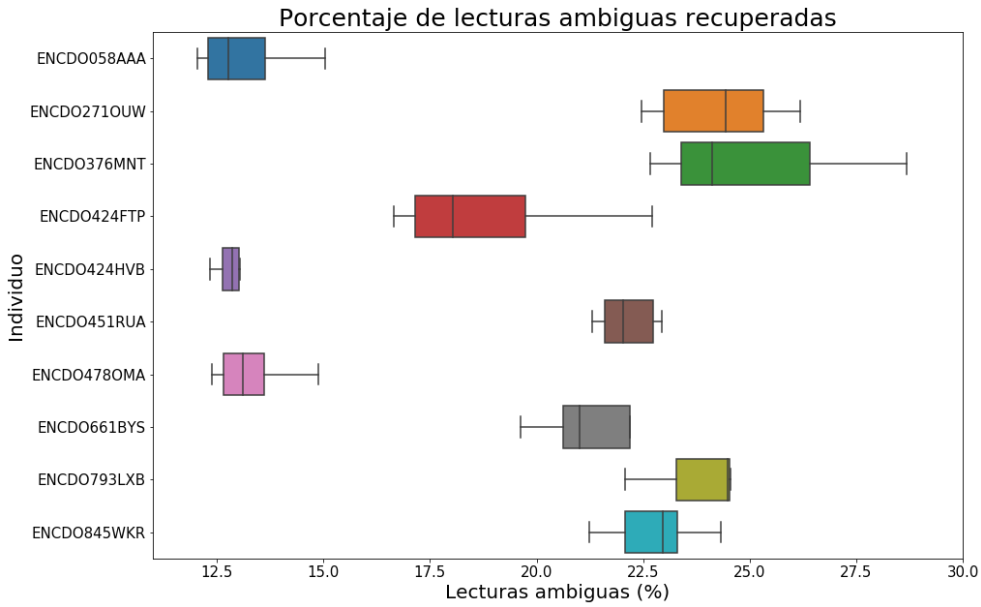

Figura 14. Porcentaje de lecturas ambiguas recuperadas

Cada gráfico de caja representa la distribución del porcentaje de lecturas ambiguas recuperadas tras el alineamiento en dos etapas para un individuo.

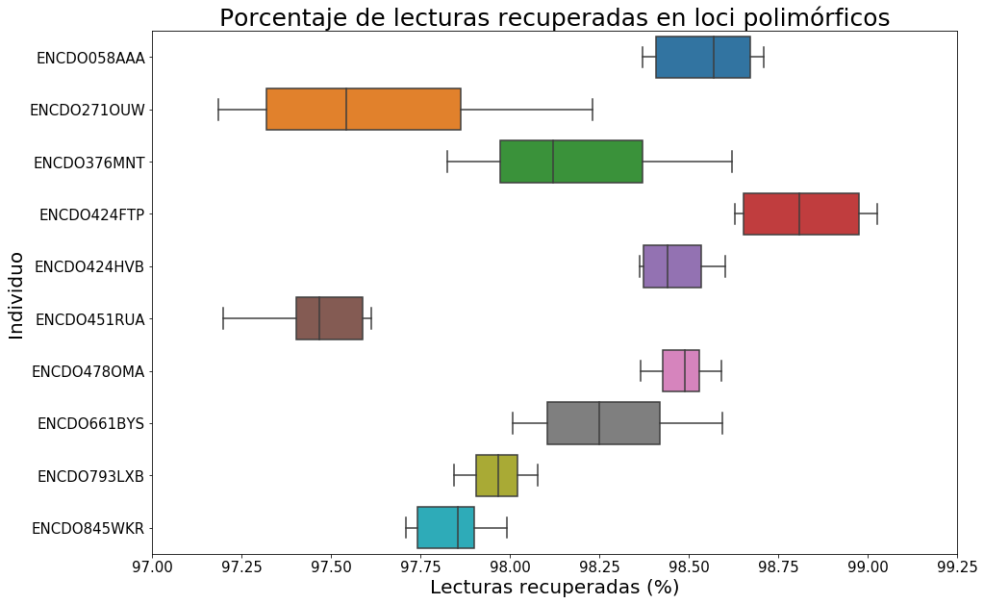

Figura 15. Porcentaje de lecturas recuperadas en loci polimórficos

Cada gráfico de caja representa la distribución del porcentaje de lecturas recuperadas correctamente asignadas en las regiones de los *loci* alternativos en el ensamblado primario.

5.2 DMCs intra- e inter-individuales

En esta sección, se muestran y describen los resultados obtenidos tras la detección y caracterización de los DMCs intra-individuales e inter-individuales. En la sección 6.2, se discuten en profundidad las posibles implicaciones funcionales de estos marcadores.

Utilizando el módulo de metilación diferencial de MethFlow (véase sección 5.1.2), se analizaron:

- **221 pares de muestras para la detección de DMCs intra-individuales.** En la **Tabla 22** del anexo, se indican los pares de muestras comparadas, así como el número y porcentaje de DMCs intra-individuales detectados en cada par. A partir de estas DMCs intra-individuales, se obtuvo un conjunto de 3.303.077 DMCs intra-individuales estrictas, que representan el 12,19% de los sitios CpG del ensamblado GRCh38/hg38 del genoma humano.
- **61 pares de muestras para la detección de DMCs inter-individuales.** En la **Tabla 23** del anexo, se indican los pares de muestras comparadas, así como el número y porcentaje de DMCs inter-individuales detectados en cada par. A partir de estas DMCs inter-individuales, se obtuvo un conjunto de 329.974 DMCs inter-individuales estrictas, que representan el 1,22% de los sitios CpG del ensamblado GRCh38/hg38 del genoma humano.

Los resultados mostrados en las siguientes secciones corresponden a sendos conjuntos estrictos de DMCs intra-individuales e inter-individuales.

5.2.1 Riqueza en elementos genómicos

Siguiendo el método descrito en la sección 4.7.1, se midió la riqueza en DMCs intra-individuales e inter-individuales de distintos tipos de elementos genómicos, descritos en detalle en la sección 4.3.

La **Tabla 3** y la **Tabla 4** muestran la riqueza en DMCs intra-individuales e inter-individuales, respectivamente, de los distintos tipos de elementos genómicos. Estos elementos genómicos presentan alguno de estos tres estados de riqueza:

- **Enriquecido (R).** El elemento genómico tiene una proporción de DMCs mayor de lo esperado por azar.
- **Aleatorio (A).** El elemento genómico tiene una proporción de DMCs indistinguible de la que presenta el resto del genoma.
- **Empobrecido (P).** El elemento genómico tiene una proporción de DMCs menor de lo esperado por azar.

Tabla 3. Riqueza en DMCs intra-individuales de diferentes elementos genómicos

R.O.: riqueza observada; R.E.: riqueza esperada; R: enriquecido; P: empobrecido; A: aleatorio

Tipo	Nombre	R.O.	R.E. (media)	R. E. (SD)	Z-score	Valor-P	Estado
Promotores	relaxed promoters	0,424	0,999	0,002	-323,435	0,001	P
	strict promoters	0,442	0,999	0,002	-312,038	0,001	P
Potenciadores	relaxed enhancers	1,515	1,000	0,001	376,926	0,001	R
	strict enhancers	1,449	1,000	0,003	178,140	0,001	R
Otras regiones reguladoras	CTCFBSs	0,856	0,999	0,002	-63,010	0,001	P
	DHSs	1,159	1,000	0,004	41,010	0,001	R
Cuerpos génicos	protein gene bodies	0,898	1,000	0,002	-67,390	0,001	P
	RNA gene bodies	1,068	1,000	0,001	46,942	0,001	R
	protein exons	0,511	1,000	0,002	-291,142	0,001	P
	RNA exons	0,715	1,001	0,003	-91,492	0,001	P
	protein introns	0,990	1,000	0,002	-6,561	0,001	P
	RNA introns	1,124	1,000	0,002	75,846	0,001	R
	protein TSSs	0,070	0,999	0,046	-20,181	0,001	P
	RNA TSSs	0,561	1,022	0,059	-7,779	0,001	P
	protein TESs	1,107	1,057	0,164	0,307	0,272	A
	RNA TESs	1,088	1,025	0,081	0,776	0,248	A
Islas CpG	relaxed CGIs	0,177	1,000	0,002	-490,302	0,001	P
	strict CGIs	0,028	1,000	0,002	-428,412	0,001	P
	UCSC CGIs	0,103	0,999	0,002	-468,197	0,001	P
	masked UCSC CGIs	0,060	0,999	0,002	-472,455	0,001	P
Elementos repetidos	LTRs	0,978	1,001	0,002	-13,353	0,001	P

	LINEs	1,350	1,000	0,002	196,834	0,001	R
	SINEs	0,525	1,000	0,001	-360,806	0,001	P
Elementos conservados	100 spp CEs	0,994	1,000	0,002	-3,486	0,001	P
	30 spp CEs	1,017	1,000	0,002	9,590	0,001	R
	20 spp CEs	1,042	1,000	0,002	18,973	0,001	R
	7 spp CEs	1,041	1,000	0,002	24,187	0,001	R
Polimorfismos	common SNPs	2,195	1,001	0,002	505,372	0,001	R
	flagged SNPs	0,714	1,006	0,016	-17,915	0,001	P
	common indels	1,159	1,001	0,016	10,085	0,001	R
	flagged indels	0,434	1,017	0,051	-11,496	0,001	P

Tabla 4. Riqueza en DMCs inter-individuales de diferentes elementos genómicos

R.O.: riqueza observada; R.E.: riqueza esperada; R: enriquecido; P: empobrecido; A: aleatorio

Tipo	Nombre	R.O.	R.E. (media)	R. E. (SD)	Z-score	Valor-P	Estado
Promotores	relaxed promoters	0,328	0,998	0,007	-95,758	0,001	P
	strict promoters	0,349	0,998	0,007	-99,331	0,001	P
Potenciadores	relaxed enhancers	1,126	1,000	0,005	23,531	0,001	R
	strict enhancers	1,137	0,999	0,006	22,013	0,001	R
Otras regiones reguladoras	CTCFBSs	0,747	1,001	0,008	-31,833	0,001	P
	DHSs	0,901	1,000	0,011	-8,704	0,001	P
Cuerpos génicos	protein gene bodies	0,879	0,998	0,005	-22,576	0,001	P
	RNA gene bodies	1,055	1,001	0,006	8,756	0,001	R
	protein exons	0,457	0,996	0,006	-93,470	0,001	P
	RNA exons	0,794	1,002	0,010	-19,926	0,001	P
	protein introns	0,984	0,999	0,005	-3,023	0,001	P
	RNA introns	1,098	1,001	0,007	14,581	0,001	R
	protein TSSs	0,081	0,983	0,155	-5,831	0,001	P
	RNA TSSs	0,619	1,017	0,197	-2,025	0,001	A
	protein TESs	1,039	0,922	0,468	0,250	0,430	A
	RNA TESs	0,856	1,073	0,358	-0,605	0,246	A
Islas CpG	relaxed CGIs	0,218	1,000	0,004	-186,507	0,001	P
	strict CGIs	0,041	1,000	0,007	-135,368	0,001	P
	UCSC CGIs	0,114	1,001	0,005	-162,354	0,001	P
	masked UCSC CGIs	0,082	1,000	0,006	-152,361	0,001	P
Elementos repetidos	LTRs	1,402	1,001	0,006	69,370	0,001	R
	LINEs	1,322	1,002	0,005	58,490	0,001	R
	SINEs	0,932	1,000	0,004	-15,959	0,001	P
Elementos conservados	100 spp CEs	0,620	0,996	0,008	-47,730	0,001	P

	30 spp CEs	0,580	0,997	0,008	-52,305	0,001	P
	20 spp CEs	0,588	0,996	0,008	-53,406	0,001	P
	7 spp CEs	0,614	0,997	0,007	-52,333	0,001	P
Polimorfismos	common SNPs	11,950	1,000	0,009	1196,864	0,001	R
	flagged SNPs	0,477	1,010	0,080	-6,680	0,001	P
	common indels	1,642	0,989	0,047	13,897	0,001	R
	flagged indels	0,293	1,057	0,167	-4,560	0,001	P

5.2.1.1 Sitios de unión a factores de transcripción

También se estudió la riqueza en DMCs para los sitios de unión de un total de 108 factores de transcripción, que se han clasificado en activadores o represores y Methyl-Minus o Methyl-Plus. En la sección 1.2.1, se explican las diferencias entre tipos de factores de transcripción y su efecto sobre la transcripción en función del estado de metilación de su sitio de unión.

Las siguientes tablas resumen el número de factores de transcripción activadores (Tabla 5), represores (Tabla 6), Methyl-Minus (Tabla 7) y Methyl-Plus (Tabla 8) que presentan sitios de unión enriquecidos, empobrecidos o con una distribución aleatoria de DMCs.

Para ningún tipo de factor de transcripción se hallaron diferencias significativas en la riqueza de DMCs intra-individuales e inter-individuales. El valor-P calculado mediante el test exacto de Fisher (Fisher, 2006) para las dos clases extremas (enriquecidos y empobrecidos) se indica en la leyenda de cada tabla.

Para consultar cada factor de transcripción, véanse la **Tabla 24** y la **Tabla 25**, que muestran respectivamente la riqueza en DMCs intra-individuales e inter-individuales de los sitios de unión de cada factor de transcripción.

Tabla 5. Riqueza en DMCs de los sitios de unión a factores de transcripción activadores

Aplicando el test exacto de Fisher (Fisher, 2006) a las dos clases extremas (enriquecidos y empobrecidos), no se encontraron diferencias significativas entre DMCs intra-individuales e inter-individuales (valor-P = 0,714).

Estado de riqueza	DMCs intra-individuales	DMCs inter-individuales
Enriquecidos	15 (78,95%)	13 (68,42%)
Aleatorios	0 (0,00%)	1 (5,26%)
Empobrecidos	4 (21,05%)	5 (26,32%)

Tabla 6. Riqueza en DMCs de los sitios de unión a factores de transcripción represores

Aplicando el test exacto de Fisher (Fisher, 2006) a las dos clases extremas (enriquecidos y empobrecidos), no se encontraron diferencias significativas entre DMCs intra-individuales e inter-individuales (valor-P = 1).

Estado de riqueza	DMCs intra-individuales	DMCs inter-individuales
Enriquecidos	17 (65,38%)	17 (65,38%)
Aleatorios	1 (3,85%)	2 (7,69%)
Empobrecidos	8 (30,77%)	7 (26,92%)

Tabla 7. Riqueza en DMCs de los sitios de unión a factores de transcripción Methyl-Minus

Aplicando el test exacto de Fisher (Fisher, 2006) a las dos clases extremas (enriquecidos y empobrecidos), no se encontraron diferencias significativas entre DMCs intra-individuales e inter-individuales (valor-P = 0,7885).

Estado de riqueza	DMCs intra-individuales	DMCs inter-individuales
Enriquecidos	23 (67,65%)	20 (58,82%)
Aleatorios	0 (0,00%)	6 (17,65%)
Empobrecidos	11 (32,35%)	8 (23,53%)

Tabla 8. Riqueza en DMCs de los sitios de unión a factores de transcripción Methyl-Plus

Aplicando el test exacto de Fisher (Fisher, 2006) a las dos clases extremas (enriquecidos y empobrecidos), no se encontraron diferencias significativas entre DMCs intra-individuales e inter-individuales (valor-P = 0,4938).

Estado de riqueza	DMCs intra-individuales	DMCs inter-individuales
Enriquecidos	33 (84,62%)	30 (76,92%)
Aleatorios	0 (0,00%)	6 (15,38%)
Empobrecidos	6 (15,38%)	3 (7,69%)

5.2.2 Distribución en torno a los genes

Siguiendo el método descrito en la sección 4.7.2, se midió la proporción de DMCs intra-individuales e inter-individuales en torno al TSS y al TES del gen más

próximo. Se estudiaron por separado los genes codificantes (Figura 16) y los genes no-codificantes (Figura 17).

La proporción de DMCs intra-individuales e inter-individuales disminuye a medida que decrece la distancia al TSS más próximo, tanto en genes codificantes como no-codificantes. En cambio, la proporción de DMCs intra-individuales al principio disminuye conforme se aproxima el TES y aumenta súbitamente cuando el TES está muy próximo, tanto en genes codificantes como no-codificantes. En cambio, la proporción de DMCs inter-individuales tan solo decae lentamente conforme se aproxima al TES, tanto en genes codificantes como no-codificantes.

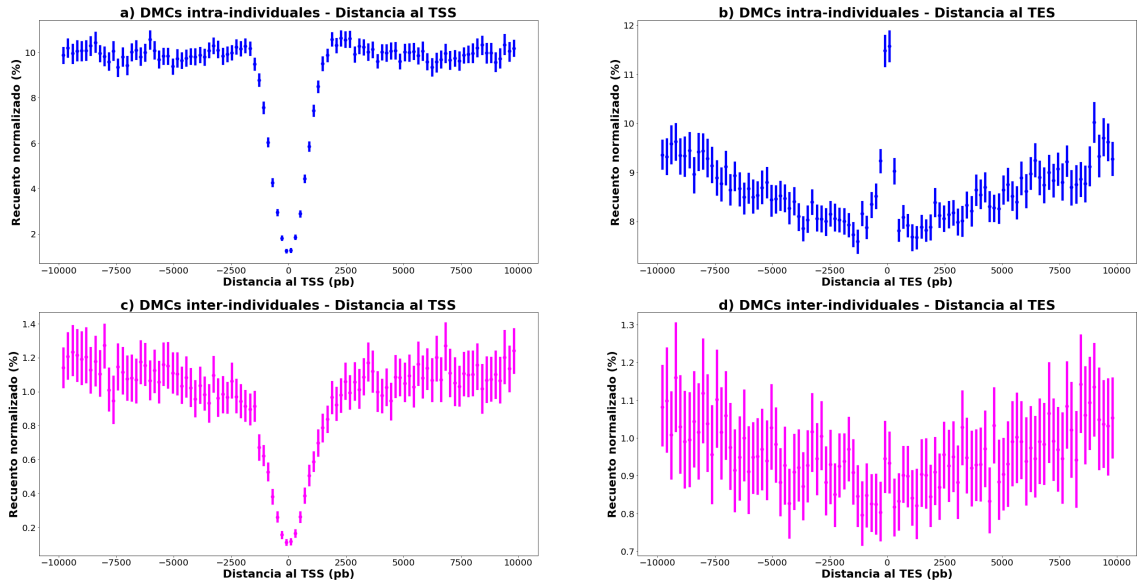

Figura 16. Proporción de DMCs en torno al TSS y al TES del gen codificante más próximo

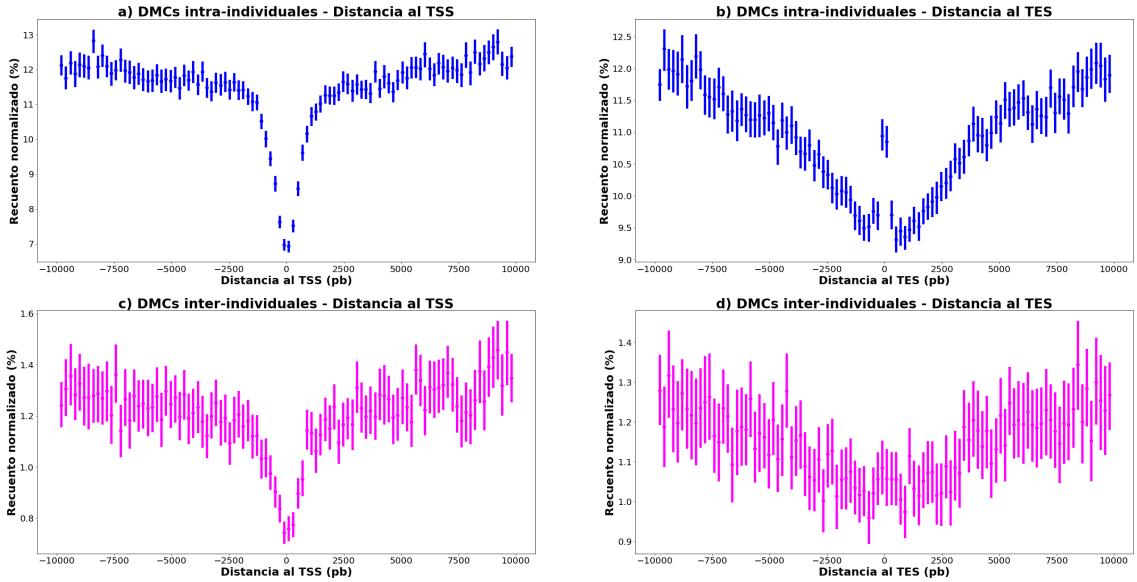

Figura 17. Proporción de DMCs en torno al TSS y al TES del gen no-codificante más próximo

5.3 Semáforos CpG

En esta sección, se muestran y describen los resultados obtenidos tras la detección y caracterización de los CpG-TLs rojos y verdes. En la sección 6.3, se discuten en profundidad las posibles implicaciones funcionales de estos marcadores.

Utilizando el módulo de asociación con la transcripción de MethFlow (véase sección 5.1.2), se analizó la asociación metilación-transcripción de 518.293.855 pares CpG-gen, correspondientes a 25.770.588 sitios CpG y 20.038 genes. En 280.424 pares CpG-gen se encontró asociación significativa entre el nivel de metilación del sitio CpG y la tasa de transcripción del gen, suponiendo un total de 193.705 CpG-TLs (0,75% de los sitios CpG analizados).

Como ya se ha descrito en la sección 1.2.2, los CpG-TLs se dividen en dos tipos, dependiendo del signo de su correlación metilación-transcripción:

- **Rojos**, cuando un incremento de la metilación correlaciona con un decremento de la transcripción. Presentan una correlación metilación-transcripción negativa.
- **Verdes**, cuando un incremento de la metilación correlaciona con un incremento de la transcripción. Presentan una correlación metilación-transcripción positiva.

En total, se detectaron:

- 93.346 pares CpG-gen (0,02% de los pares CpG-gen analizados) con asociación negativa, suponiendo un total de 66.746 CpG-TLs rojos (0,26% de los sitios CpG analizados).
- 187.078 pares CpG-gen (0,04% de los pares CpG-gen analizados) con asociación positiva, suponiendo un total de 126.959 CpG-TLs verdes (0,49% de los sitios CpG analizados).

5.3.1 Ejemplos de semáforos CpG

A continuación, se describen un ejemplo de CpG-TL rojo (Figura 18) y un ejemplo de CpG-TL verde (Figura 19).

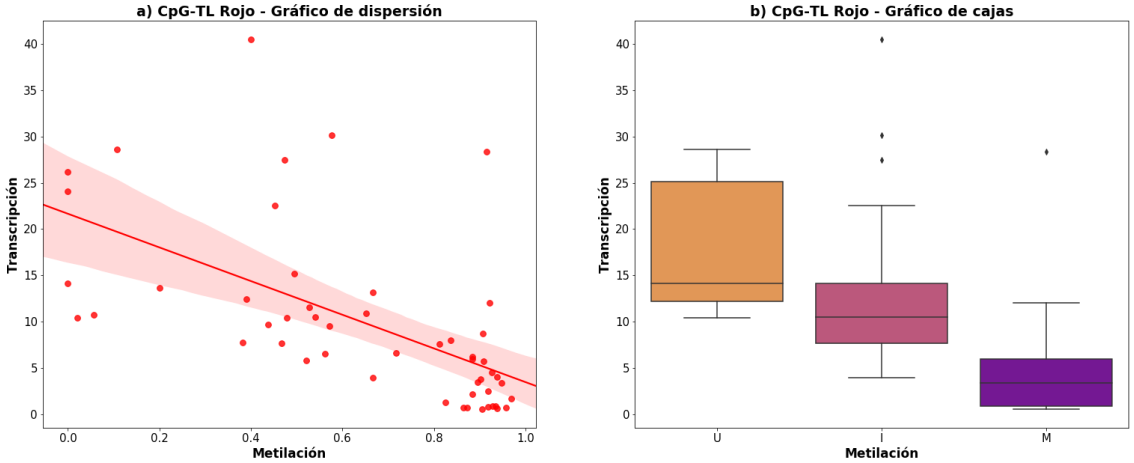

Figura 18. Ejemplo de CpG-TL rojo

La metilación del sitio CpG localizado en las coordenadas chr1:38666789-38666790 del ensamblado GRCh38/hg38 se asocia negativamente con la transcripción del gen ENSG00000196449.3 de GENCODE 29 (Frankish et al., 2019). El test rho devolvió un valor de -0,7096, con un valor-P de $5{,}5405 \times 10^{-9}$ y un valor-P ajustado mediante FDR de $2{,}5938 \times 10^{-6}$. El test h devolvió un valor de 26,1861, con un valor-P de $2{,}0595 \times 10^{-6}$ y un valor-P ajustado mediante FDR de 0,0238. **Gráfico de dispersión (a).** Cada uno de los puntos representa el nivel de metilación y el nivel de transcripción para una muestra, siendo en total 51 muestras. La línea roja indica el ajuste lineal de la transcripción frente a la metilación y las bandas rojas claras la zona con un 95% de confianza. **Gráfico de cajas (b).** Las categorías U, I y M comprenden a las muestras en las que este CpG está no-metilado (de 0 a 0,2), con metilación intermedia (de 0,2 a 0,8) o metilado (de 0,8 a 1), respectivamente. Se aprecia como la categoría U presenta mayores niveles de transcripción que I y este a su vez mayores niveles de transcripción que M.

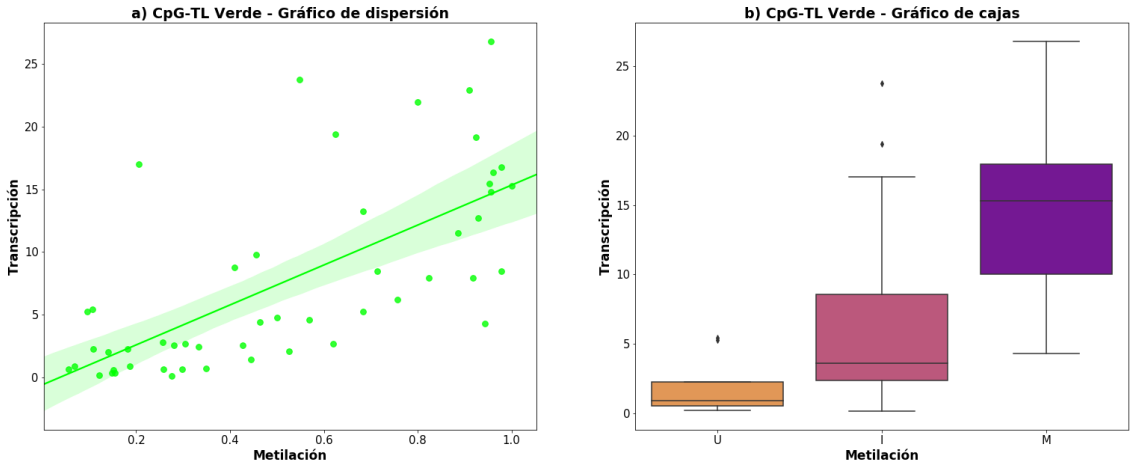

Figura 19. Ejemplo de CpG-TL verde

La metilación del sitio CpG localizado en las coordenadas chr2:20335270-20335271 del ensamblado GRCh38/hg38 se asocia positivamente con la transcripción del gen ENSG00000118961.14 de GENCODE 29 (Frankish et al., 2018). El test rho devolvió un valor de 0,7209, con un valor-P de 2,4354x10-9 y un valor-P ajustado mediante FDR de 1,1479x10-6. El test h devolvió un valor de 24,2229, con un valor-P de 5,4963x10-6 y un valor-P ajustado mediante FDR de 0,0289. **Gráfico de dispersión (a).** Cada uno de los puntos representa el nivel de metilación y el nivel de transcripción para una muestra, siendo en total 51 muestras. La línea verde indica el ajuste lineal de la transcripción frente a la metilación y las bandas verdes claras la zona con un 95% de confianza. **Gráfico de cajas (b).** Las categorías U, I y M comprenden a las muestras en las que este CpG está no-metilado (de 0 a 0,2), con metilación intermedia (de 0,2 a 0,8) o metilado (de 0,8 a 1), respectivamente. Se aprecia como la categoría U presenta menores niveles de transcripción que I y este a su vez menores niveles de transcripción que M.

5.3.2 Riqueza en elementos genómicos

Siguiendo el método descrito en la sección 4.7.1, se midió la riqueza en CpG-TLs rojos y verdes de distintos tipos de elementos genómicos, descritos en detalle en la sección 4.3.

La Tabla 9 y la Tabla 10 muestran la riqueza en CpG-TLs rojos y verdes, respectivamente, de los distintos tipos de elementos genómicos. Estos elementos genómicos presentan alguno de estos tres estados de riqueza:

- **Enriquecido (R).** El elemento genómico tiene una proporción de CpG-TLs mayor de lo esperado por azar.
- **Aleatorio (A).** El elemento genómico tiene una proporción de CpG-TLs indistinguible de la que presenta el resto del genoma.
- **Empobrecido (P).** El elemento genómico tiene una proporción de CpG-TLs menor de lo esperado por azar.

Tabla 9. Riqueza en CpG-TLs rojos de diferentes elementos genómicos

R.O.: riqueza observada; R.E.: riqueza esperada; R: enriquecido; P: empobrecido; A: aleatorio

Tipo	Nombre	R.O.	R.E. (media)	R. E. (SD)	Z-score	Valor-P	Estado
Promotores	relaxed promoters	1,958	0,999	0,014	68,436	0,001	R
	strict promoters	1,956	1,002	0,015	61,933	0,001	R
Potenciadores	relaxed enhancers	2,085	1,002	0,009	115,014	0,001	R
	strict enhancers	2,709	0,996	0,014	120,813	0,001	R
Otras regiones reguladoras	CTCFBSs	1,319	1,005	0,021	15,186	0,001	R
	DHSs	1,691	1,008	0,026	26,405	0,001	R
Cuerpos génicos	protein gene bodies	1,255	1,004	0,011	23,528	0,001	R
	RNA gene bodies	1,099	1,005	0,013	7,047	0,001	R
	protein exons	1,789	1,001	0,014	58,017	0,001	R
	RNA exons	1,757	1,006	0,025	29,979	0,001	R
	protein introns	1,152	1,004	0,009	16,109	0,001	R
	RNA introns	1,010	1,004	0,013	0,438	0,350	A
	protein TSSs	1,158	1,061	0,361	0,267	0,350	A
	RNA TSSs	0,918	1,103	0,484	-0,382	0,448	A
	protein TESs	1,656	1,252	1,467	0,275	0,353	A
	RNA TESs	1,191	1,042	0,881	0,169	0,325	A
Islas CpG	relaxed CGIs	1,055	1,005	0,013	3,789	0,001	R
	strict CGIs	0,862	1,006	0,016	-9,079	0,001	P
	UCSC CGIs	1,020	1,007	0,013	1,016	0,100	A
	masked UCSC CGIs	1,117	1,007	0,014	7,639	0,001	R
Elementos repetidos	LTRs	0,949	1,005	0,019	-2,931	0,001	P
	LINEs	0,890	1,008	0,012	-9,797	0,001	P
	SINEs	0,599	1,000	0,010	-42,216	0,001	P
Elementos conservados	100 spp CEs	1,288	1,001	0,016	17,441	0,001	R
	30 spp CEs	1,377	0,997	0,018	20,704	0,001	R
	20 spp CEs	1,417	0,999	0,017	24,788	0,001	R
	7 spp CEs	1,441	1,000	0,016	27,106	0,001	R
Polimorfismos	common SNPs	4,702	0,996	0,019	191,204	0,001	R
	flagged SNPs	0,943	0,980	0,175	-0,212	0,494	A
	common indels	3,826	0,997	0,076	37,032	0,001	R
	flagged indels	1,401	1,055	0,410	0,844	0,275	A

Tabla 10. Riqueza en CpG-TLs verdes de diferentes elementos genómicos

R.O.: riqueza observada; R.E.: riqueza esperada; R: enriquecido; P: empobrecido; A: aleatorio

Tipo	Nombre	R.O.	R.E. (media)	R. E. (SD)	Z-score	Valor-P	Estado
Promotores	relaxed promoters	1,620	1,001	0,008	76,189	0,001	R
	strict promoters	1,674	0,999	0,011	60,693	0,001	R
Potenciadores	relaxed enhancers	1,901	1,001	0,008	106,169	0,001	R
	strict enhancers	2,161	0,999	0,015	76,622	0,001	R
Otras regiones reguladoras	CTCFBSs	1,620	1,001	0,017	37,271	0,001	R
	DHSs	1,541	1,001	0,018	30,645	0,001	R
Cuerpos génicos	protein gene bodies	1,525	1,003	0,008	61,960	0,001	R
	RNA gene bodies	0,927	0,995	0,010	-7,080	0,001	P
	protein exons	1,870	1,001	0,009	98,622	0,001	R
	RNA exons	1,561	1,003	0,010	55,039	0,001	R
	protein introns	1,330	1,002	0,008	38,900	0,001	R
	RNA introns	0,842	0,994	0,011	-13,924	0,001	P
	protein TSSs	0,406	1,014	0,158	-3,841	0,001	P
	RNA TSSs	0,724	0,921	0,275	-0,715	0,226	A
	protein TESs	4,352	1,262	0,801	3,856	0,001	R
	RNA TESs	0,626	1,083	0,605	-0,755	0,246	A
Islas CpG	relaxed CGIs	1,089	1,001	0,006	14,649	0,001	R
	strict CGIs	1,019	1,000	0,012	1,473	0,050	A
	UCSC CGIs	1,162	1,000	0,009	18,043	0,001	R
	masked UCSC CGIs	1,289	1,002	0,010	28,168	0,001	R
Elementos repetidos	LTRs	0,786	0,999	0,008	-26,743	0,001	P
	LINEs	0,845	1,001	0,009	-16,587	0,001	P
	SINEs	0,584	1,001	0,008	-55,480	0,001	P
Elementos conservados	100 spp CEs	1,398	1,003	0,009	45,609	0,001	R
	30 spp CEs	1,513	1,000	0,008	65,541	0,001	R
	20 spp CEs	1,512	1,000	0,010	53,278	0,001	R
	7 spp CEs	1,517	0,999	0,009	56,723	0,001	R
Polimorfismos	common SNPs	2,054	1,005	0,015	69,770	0,001	R
	flagged SNPs	1,382	1,005	0,092	4,112	0,001	R
	common indels	0,893	0,993	0,058	-1,723	0,026	A
	flagged indels	2,284	0,969	0,202	6,517	0,001	R

5.3.2.1 Sitios de unión a factores de transcripción

También se estudió la riqueza en CpG-TLs para los sitios de unión de un total de 108 factores de transcripción, que se han clasificado en activadores o represores y Methyl-Minus o Methyl-Plus. En la sección 1.2.1, se explica las diferencias entre tipos de factores de transcripción y su efecto sobre la transcripción en función del estado de metilación de su sitio de unión.

Las siguientes tablas resumen el número de factores de transcripción activadores (Tabla 11), represores (Tabla 12), Methyl-Minus (Tabla 13) y Methyl-Plus (Tabla 14) que presentan sitios de unión enriquecidos, empobrecidos o con una distribución aleatoria de CpG-TLs.

Se encontró que los factores de transcripción Methyl-Plus (es decir, con afinidad por sitios metilados), presentan diferencias significativas en su riqueza en CpG-TLs rojos y verdes (valor-P = 0,0002). Concretamente, la mayoría de ellos presentan sitios de unión pobres en CpG-TLs rojos y ricos en CpG-TLs verdes (Tabla 14). Este valor-P se calculó aplicando el test exacto de Fisher (Fisher, 2006) para las dos clases extremas (enriquecidos y empobrecidos). El valor-P para los otros tipos de factores de transcripción se indica en la leyenda de las tablas antes citadas.

Para consultar cada factor de transcripción, véanse la **Tabla 26** y la **Tabla 27**, que muestran respectivamente la riqueza en CpG-TLs rojos y verdes de los sitios de unión de cada factor de transcripción.

Tabla 11. Riqueza en CpG-TLs de los sitios de unión a factores de transcripción activadores

Aplicando el test exacto de Fisher (Fisher, 2006) a las dos clases extremas (enriquecidos y empobrecidos), no se encontraron diferencias significativas entre CpG-TLs rojos y verdes (valor-P = 1).

Estado de riqueza	CpG-TLs rojos	CpG-TLs verdes
Enriquecidos	4 (21,05%)	5 (26,32%)
Aleatorios	9 (47,37%)	8 (42,11%)
Empobrecidos	6 (31,58%)	6 (31,58%)

Tabla 12. Riqueza en CpG-TLs de los sitios de unión a factores de transcripción represores

Aplicando el test exacto de Fisher (Fisher, 2006) a las dos clases extremas (enriquecidos y empobrecidos), no se encontraron diferencias significativas entre CpG-TLs rojos y verdes (valor-P = 1).

Estado de riqueza	CpG-TLs rojos	CpG-TLs verdes
Enriquecidos	7 (26,92%)	8 (30,77%)
Aleatorios	12 (46,15%)	9 (34,62%)
Empobrecidos	7 (26,92%)	9 (34,62%)

Tabla 13. Riqueza en CpG-TLs de los sitios de unión a factores de transcripción Methyl-Minus

Aplicando el test exacto de Fisher (Fisher, 2006) a las dos clases extremas (enriquecidos y empobrecidos), no se encontraron diferencias significativas entre CpG-TLs rojos y verdes (valor-P = 0.7338).

Estado de riqueza	CpG-TLs rojos	CpG-TLs verdes
Enriquecidos	21 (61,76%)	19 (55,88%)
Aleatorios	6 (17,65%)	11 (32,35%)
Empobrecidos	7 (20,59%)	4 (11,76%)

Tabla 14. Riqueza en CpG-TLs de los sitios de unión a factores de transcripción Methyl-Plus

Aplicando el test exacto de Fisher (Fisher, 2006) a las dos clases extremas (enriquecidos y empobrecidos), se encontraron diferencias significativas entre CpG-TLs rojos y verdes (valor-P = 0.0002).

Estado de riqueza	CpG-TLs rojos	CpG-TLs verdes
Enriquecidos	5 (12,82%)	21 (53,85%)
Aleatorios	15 (38,46%)	11 (28,20%)
Empobrecidos	19 (48,72%)	7 (17,95%)

5.3.3 Distribución en torno a los genes

Siguiendo el método descrito en la sección 4.7.2, se midió la proporción de CpG-TLs rojos y verdes en torno al TSS y al TES de su gen asociado. Se estudiaron por separado los genes codificantes (Figura 20) y los genes no-codificantes (Figura 21).

La proporción de CpG-TLs rojos aumenta a medida que decrece la distancia al TSS de su gen asociado, tanto en genes codificantes como no-codificantes. En cambio, la proporción de CpG-TLs verdes disminuye a medida que decrece la distancia al TES de su gen asociado, tanto en genes codificantes como no-codificantes, aunque la bajada es mucho más pronunciada en genes codificantes. 5 kpb aguas arriba del TSS hay una subida drástica de CpG-TLs verdes en genes no-codificantes. En el

caso de los genes codificantes, la proporción de CpG-TLs rojos y verdes aumenta a medida que se aproxima al TES y disminuye súbitamente cuando el TES está muy próximo. En el caso de los genes no-codificantes, no parece que haya una tendencia bien definida en torno al TES.

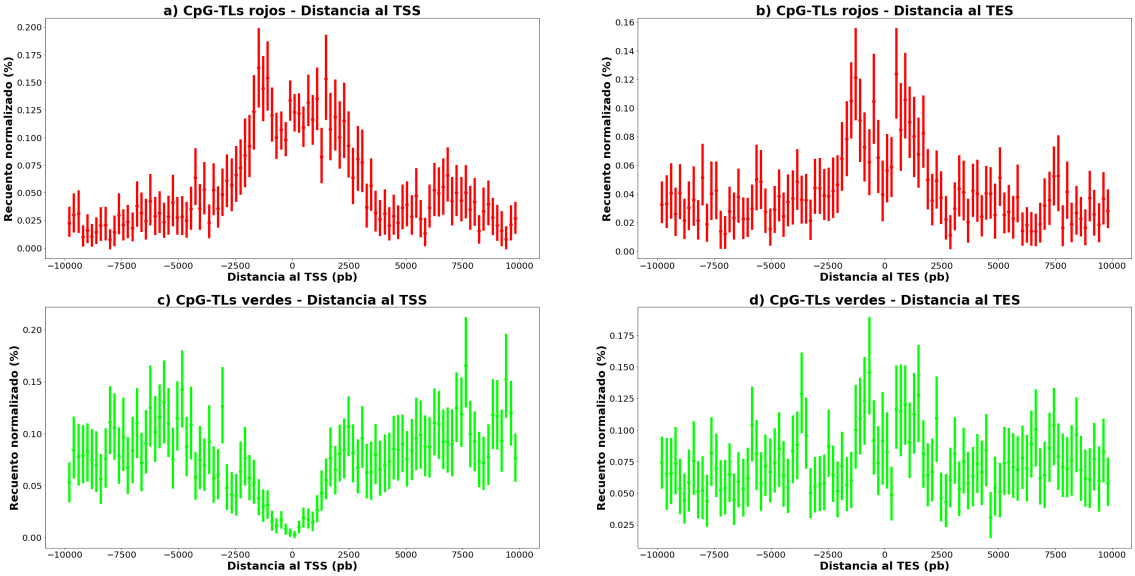

Figura 20. Proporción de CpG-TLs en torno al TSS y al TES de su gen asociado (solo codificantes)

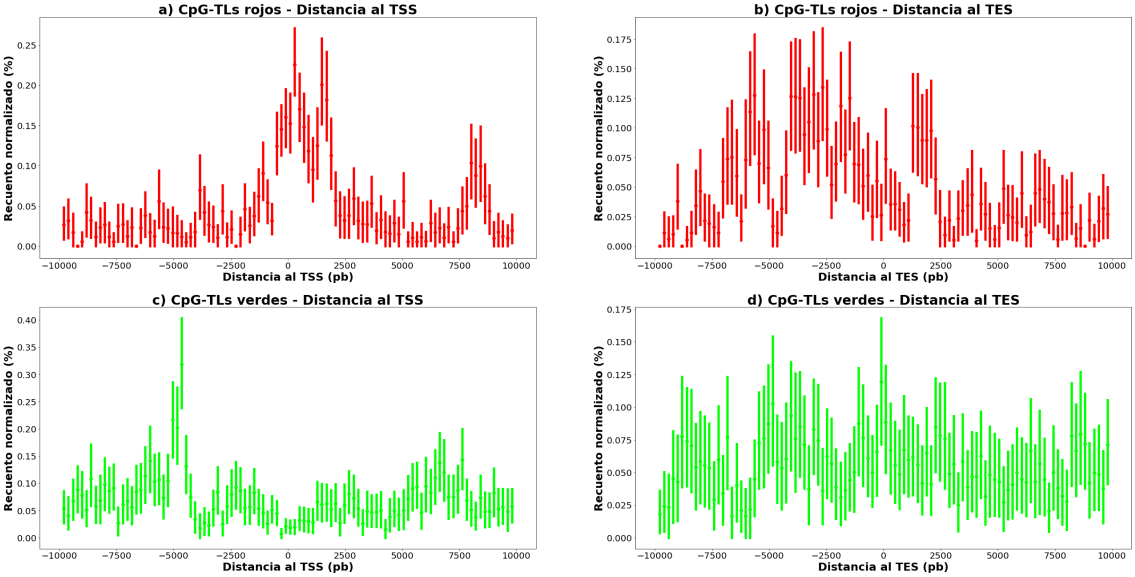

Figura 21. Proporción de CpG-TLs en torno al TSS y al TES de su gen asociado (solo no-codificantes)

5.4 NGSmethDB: una base de datos dedicada a la metilación

NGSmethDB (https://bioinfo2.ugr.es/NGSmethDB) es una base de datos especializada en mapas de metilación en genoma completo y marcadores biológicos relacionados, que se ha ido ampliando y mejorando a lo largo de los años (Geisen, Barturen, Alganza, Hackenberg, & Oliver, 2014; Hackenberg, Barturen, & Oliver, 2011; Ricardo Lebrón et al., 2017).

A lo largo de esta Tesis Doctoral, se rediseñó por completo esta base de datos (Ricardo Lebrón et al., 2017), con el objetivo de optimizar el almacenamiento y la consulta del gran volumen de datos producidos mediante MethFlow (véase sección 5.1). En las siguientes secciones, se describe el contenido de la base de datos, la nueva estructura del back-end, las vías de acceso a los datos y se muestran algunos ejemplos de visualización en UCSC Genome Browser (Casper et al., 2018).

5.4.1 Contenido de la base de datos

Todos los mapas de metilación, de DMCs y de CpG-TLs obtenidos mediante MethFlow a lo largo de esta Tesis Doctoral, están actualmente disponibles en NGSmethDB:

- Mapas de metilación en genoma completo para 86 muestras humanas, procedentes de 52 tipos celulares de 29 individuos (https://bioinfo2.ugr.es/NGSmethDB/single-cytosine-methylation). En la sección 4.4, se describe el protocolo de obtención de estos mapas de metilación.
- 221 mapas de DMCs intra-individuales y 61 mapas de DMCs inter-individuales, obtenidos por comparación de pares de muestras, así como sendos mapas de los conjuntos estrictos de DMCs intra-individuales y DMCs inter-individuales (https://bioinfo2.ugr.es/NGSmethDB/differential-methylation). En la sección 4.5, se describe el protocolo de detección de DMCs.

- Un mapa de CpG-TLs que muestra la localización de cada CpG-TL rojo y cada CpG-TL verde, así como un arco que los conecta con su gen asociado (https://bioinfo2.ugr.es/NGSmethDB/cpg-traffic-lights). Este mapa está en el formato *interact* (BED5+13) (https://genome.ucsc.edu/goldenpath/help/interact.html) de UCSC Genome Browser (Casper et al., 2018). Los CpG-TLs se representan como segmentos y arcos del color correspondiente (rojo o verde), donde cada CpG-TL tiene dos segmentos: el sitio CpG y el gen asociado. En la sección 4.6, se describe el protocolo de detección de CpG-TLs.

Todos estos mapas corresponden al ensamblado GRCh38/hg38 del genoma humano. En la base de datos NGSmethDB también están disponibles mapas de segmentos homogéneos de metilación (Ricardo Lebrón et al., 2017), mapas de metilación para el ensamblado GRCh37/hg19 del genoma humano y mapas de metilación para otras 7 especies, aunque no son objeto de estudio en esta Tesis Doctoral.

5.4.2 Estructura del *back-end*

Con el fin de optimizar el almacenamiento y consulta del gran volumen de datos producidos a lo largo de esta Tesis Doctoral, se decidió rediseñar por completo el back-end de esta base de datos (Ricardo Lebrón et al., 2017). Este *back-end* está formado por dos componentes:

- Una instancia del sistema de bases de datos **MongoDB** (https://www.mongodb.com), un sistema NoSQL que, en lugar de utilizar el esquema tradicional basado en tablas relacionales, utiliza documentos en formato *JSON* (ECMA International, 2013). Se trata de un formato estándar que permite un esquema dinámico de datos etiquetados y jerarquizados, intercambiable entre distintos lenguajes de programación.
- Un **servidor RESTful API** (Fielding, 2000) que permite consultar programáticamente a través del protocolo HTTPS los datos contenidos en

MongoDB y descargarlos en formato *JSON*, *CSV* o *TSV*. Este servidor RESTful API se implementó en JavaScript, utilizando el framework Node.js (https://nodejs.org), el cual permite ejecutar código JavaScript en el servidor y aporta la mayoría de funcionalidades necesarias para desarrollar un back-end.

Para agilizar las comparaciones entre muestras, se optó por almacenar los datos siguiendo esta estructura jerárquica:

i. Cada ensamblado se almacenó en una base de datos.
ii. Cada cromosoma, scaffold o contig se almacenó en una colección de documentos JSON. Estas colecciones se agrupan dentro de bases de datos y cada una de ellas consiste en una lista de documentos JSON independientes.
iii. Cada sitio CpG se almacenó en un documento JSON, dentro de la colección del cromosoma, scaffold o contig correspondiente. Cada uno de estos documentos contiene las coordenadas del sitio CpG y una serie de subdocumentos.
iv. Cada subdocumento contiene un tipo diferente de información biológica: metilación, metilación diferencial o asociación con la transcripción.

En el caso de los mapas de metilación, cada subdocumento se divide en tres niveles jerárquicos:

i. El individuo del que procede la muestra.
ii. El órgano, tejido, tipo celular o tumor del que procede la muestra.
iii. El tipo de dato: número de lecturas que evidencian metilación, cobertura total, nivel de metilación (beta) y datos por separado de cada hebra.

En lo concerniente a la metilación diferencial, cada subdocumento contiene la siguiente información:

- Si es o no una DMC intra-individual. En caso afirmativo, se anotan los pares de muestras para los que es un DMC intra-individual, su diferencia de metilación, su valor-P y su valor-P corregido en cada par de muestras y si forma o no parte del conjunto estricto.
- Si es o no una DMC inter-individual. En caso afirmativo, se anotan los pares de muestras para los que es un DMC inter-individual, su diferencia de metilación, su valor-P y su valor-P corregido en cada par de muestras y si forma o no parte del conjunto estricto.

En lo concerniente a la asociación con la transcripción, cada subdocumento contiene la siguiente información:

- Si es o no un CpG-TL rojo. En caso afirmativo, se anotan los genes para los que es un CpG-TL rojo, su valor de rho y de H, así como sus respectivos valores-P y valores-P corregidos.
- Si es o no un CpG-TL verde. En caso afirmativo, se anotan los genes para los que es un CpG-TL verde, su valor de rho y de H, así como sus respectivos valores-P y valores-P corregidos.

5.4.3 Vías de acceso a los datos

NGSmethDB dispone de múltiples vías de acceso a los datos, representadas en la Figura 22:

- **Acceso HTTPS.** Mediante el protocolo HTTPS se pueden enviar consultas al servidor RESTful API, desde un navegador o programáticamente, y descargar los datos recibidos en formato JSON, CSV o TSV. Por esta vía se puede acceder a todos los datos de metilación, DMCs y CpG-TLs correspondientes a la región genómica especificada en la consulta.
- **Cliente RESTful API.** Se trata de un programa desarrollado en **Python** (https://www.python.org) que accede programáticamente al servidor RESTful API y se descarga los datos correspondientes a las regiones especificadas por el usuario en el fichero BED que toma como entrada. Este programa es de código abierto y está disponible en el repositorio *rlebron-bioinfo/ngsmethdb* (https://github.com/rlebron-bioinfo/ngsmethdb) de **GitHub**. También está disponible en la página web de NGSmethDB como standalone o como máquina virtual para **VirtualBox** (https://www.virtualbox.org).
- **Formulario web.** Este formulario realiza consultas a través del servidor RESTful API y muestras los datos recibidos en forma de tabla en la página web de NGSmethDB. Solo permite acceder a los datos de metilación.

- **Ficheros comprimidos (dumps).** Son ficheros tabulares comprimidos con **bzip2** (http://www.bzip.org) y que contienen una copia completa de los mapas de metilación, de DMCs y de CpG-TLs en formato *BED* (https://genome.ucsc.edu/FAQ/FAQformat.html#format1).
- **Acceso a través de UCSC.** Los mapas de metilación, de DMCs y de CpG-TLs están disponibles como **UCSC Track Hubs** (Raney et al., 2014), lo que permite que sean accesibles a través de las herramientas **Genome Browser** (James Kent et al., 2002), **Table Browser** (Karolchik, 2003) y **Data Integrator** (Hinrichs et al., 2016). A través de UCSC es posible, además, enviar los datos de NGSmethDB que se estén consultando a plataformas bioinformáticas como **Galaxy** (Afgan et al., 2018), **GenomeSpace** (Qu et al., 2016) y **GREAT** (McLean et al., 2010). Los UCSC Track Hubs están disponibles a través de los siguientes enlaces:
    - Mapas de metilación: https://bioinfo2.ugr.es/NGSmethDB_hub/hg38/hub.txt
    - Mapas de DMCs: https://bioinfo2.ugr.es/DMCdb_hub/hg38/hub.txt
    - Mapas de CpG-TLs: https://bioinfo2.ugr.es/TLdb_hub/hg38/hub.txt

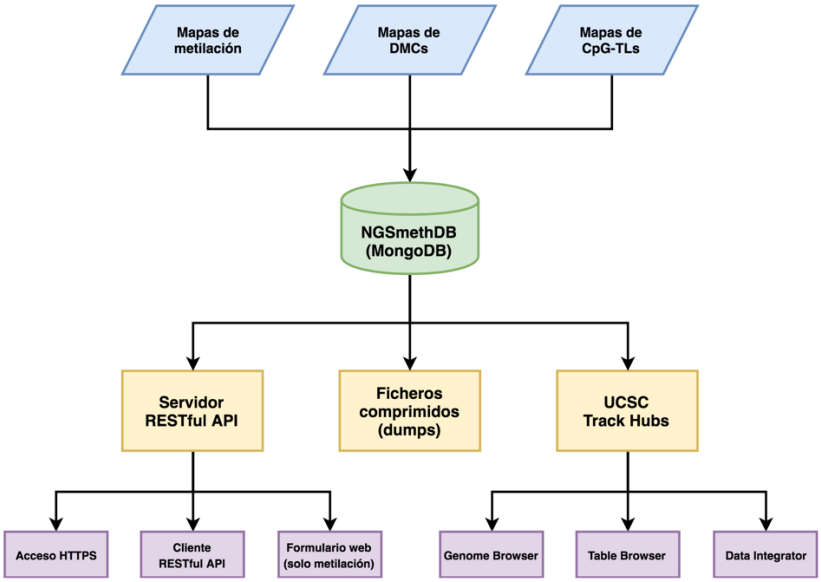

Figura 22. Estructura y flujo de datos de NGSmethDB

Los datos de los mapas de metilación, de DMCs, de CpG-TLs obtenidos mediante MethFlow (en azul) se incorporan a la instancia del sistema de bases de datos MongoDB de NGSmethDB (en verde). Estos datos se pueden consultar a través de tres vías principales (en amarillo): i) un servidor RESTful API, ii) una colección de ficheros comprimidos que contiene una copia completa de los mapas y iii) UCSC Track Hubs. Al servidor RESTful API y a los UCSC Track Hubs se puede acceder por más de una vía (en violeta).

Como ya se ha descrito en la sección 5.1.2, los mapas de metilación de NGSmethDB también se pueden obtener a través del módulo de descarga de MethFlow.

5.4.4 Visualización en UCSC Genome Browser

La mejor forma de visualizar los mapas de metilación, DMCs y CpG-TLs es a través de UCSC Genome Browser (Casper et al., 2018; James Kent et al., 2002), ya que permite la comparación *de visu* de estos mapas frente a miles de anotaciones genómicas de terceros. También es posible realizar comparaciones cuantitativas gracias a las herramientas Table Browser (Karolchik, 2003) y Data Integrator (Hinrichs et al., 2016) de UCSC.

A continuación, se muestran y describen tres secciones de UCSC Genome Browser a modo de ejemplo:

- En la Figura 23, se muestran los niveles de metilación de varios tejidos en torno al gen ATG4C.
- En la Figura 24, se muestran DMCs intra-individuales e inter-individuales en testículo.
- En la Figura 25, se muestran CpG-TLs en promotores y potenciadores.

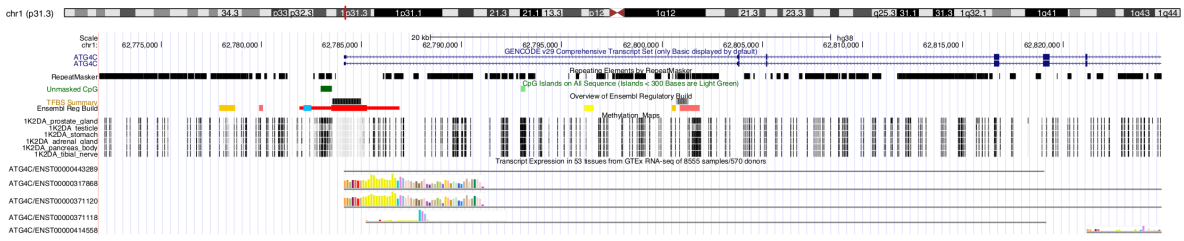

Figura 23. Sección del UCSC Track Hub de metilación de NGSmethDB

Se muestran los mapas de metilación de la glándula prostática, el testículo, el estómago, la glándula suprarrenal, el páncreas y el nervio tibial del individuo 1K2DA en una región en torno al promotor del gen ATG4C, una cisteína proteasa que interviene en la autofagia. En la región marcada como promotor (en rojo) en la pista Ensembl Regulatory Build (Zerbino et al., 2015), se aprecia que los niveles de metilación de todas las muestras del individuo 1K2DA son muy bajos, mientras que en las regiones colindantes vuelven a ser altos. Aguas arriba del promotor hay elementos repetidos, según la pista de RepeatMasker (Smit, 2018), y esa

región está mayoritariamente metilada en todas las muestras. Es interesante resaltar que según la pista de GTEx (Aguet et al., 2017), el gen ATG4C se expresa en la mayoría de tejidos. URL:

http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&hubUrl=http://bioinfo2.ugr.es/NGSmethDB_hub/hg38/hub.txt&hgS_loadUrlName=http://bioinfo2.ugr.es/NGSmethDB_hub/hg38/ATG4C_1K2DA_meth&hgS_doLoadUrl=submit

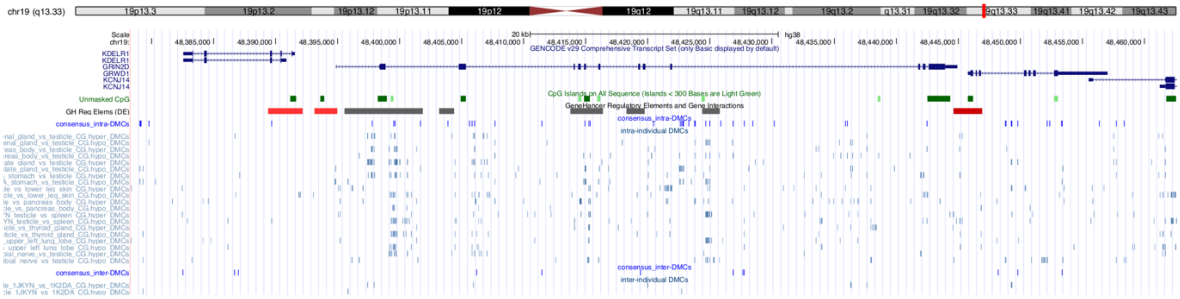

Figura 24. Sección del UCSC Track Hub de DMCs de NGSmethDB

Se muestran los mapas de intra-DMCs e inter-DMCs en las que está implicado el testículo. Las intra-DMCs y las inter-DMCs tienden localizarse en los potenciadores, marcados en gris en la pista de GeneHancer (Fishilevich et al., 2017), aunque el número de inter-DMCs mostrado es mucho menor. También es interesante resaltar que tienden a evitar los promotores, marcados en rojo en la pista de GeneHancer (Fishilevich et al., 2017), y también las islas CpG de UCSC (Casper et al., 2018), marcadas en verde. URL:

http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&hubUrl=http://bioinfo2.ugr.es/DMCdb_hub/hg38/hub.txt&hgS_loadUrlName=http://bioinfo2.ugr.es/DMCdb_hub/hg38/GRIN2D_testicle_dmcs&hgS_doLoadUrl=submit

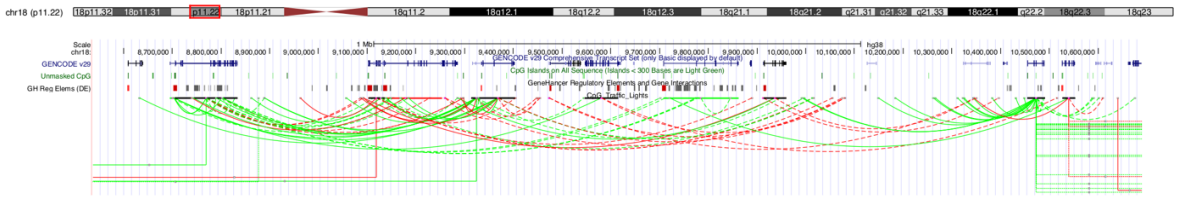

Figura 25. Sección del UCSC Track Hub de CpG-TLs de NGSmethDB

Se muestra el mapa de interacciones de los CpG-TLs rojos y los CpG-TLs verdes con sus genes asociados. La mayoría de CpG-TLs (sean verdes o rojos) se localizan en los promotores (en rojo) y en los potenciadores (en gris) de la pista de GeneHancer (Fishilevich et al., 2017). URL: http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&hubUrl=http://bioinfo2.ugr.es/TLdb_hub/hg38/hub.txt&hgS_loadUrlName=http://bioinfo2.ugr.es/TLdb_hub/hg38/chr18_tls&hgS_doLoadUrl=submit

Capítulo 6

Discusión

Gracias a las notables mejoras en la detección de los niveles de metilación de las citosinas individuales puestas a punto en esta Tesis Doctoral, ha sido posible desarrollar marcadores biológicos para investigar la metilación diferencial en el genoma humano y su asociación con la transcripción. Esto ha permitido cuestionar la hipótesis de que el signo de la asociación entre la metilación y la transcripción depende del contexto genómico en que se produce la metilación y del tipo de factores de transcripción implicados, la cual no ha podido rechazarse a la vista de los resultados obtenidos.

En las siguientes secciones, se describen las principales características e innovaciones del software desarrollado a lo largo de esta Tesis Doctoral (MethFlow y NGSmethDB), así como las posibles implicaciones funcionales de los marcadores biológicos investigados (DMCs y CpG-TLs).

6.1 Características e innovaciones de MethFlow

La técnica de secuenciación masiva conocida como WGBS (Cokus et al., 2008; Lister et al., 2008; Urich et al., 2015) permite desde hace más de una década detectar el nivel de metilación de cada citosina del genoma. Sin embargo, los resultados obtenidos mediante esta técnica están afectados por diversas fuentes de error (véase sección 1.4), llevando a muchos investigadores a tener que promediar los niveles de metilación en las regiones de interés para mitigar estos errores, aunque implique sacrificar la alta resolución que ofrece esta técnica. Sin embargo, la metilación promedio de una región no siempre es relevante e incluso puede llevar a conclusiones erróneas.

Por ello, se estableció como primer objetivo de esta Tesis Doctoral diseñar e implementar un protocolo de obtención de mapas de metilación, a partir de lecturas de WGBS, para intentar resolver todos los problemas conocidos. Este protocolo se basó parcialmente en las soluciones propuestas por Barturen *et al.* para muchas de estas fuentes de error (Barturen et al., 2017, 2014; Hackenberg, Barturen, & L., 2012).

Una vez el protocolo estuvo maduro, se decidió implementarlo como un programa de código abierto, que recibió el nombre de MethFlow (véase sección 5.1). El flujo de trabajo de este programa parte de lecturas de WGBS en formato FASTQ y finaliza con la obtención de mapas de metilación, atravesando por diversas etapas de tratamiento de errores en las que se utilizan programas de terceros combinados con código propio (véase sección 5.1.1).

Si bien existen otros programas y protocolos para la detección de los niveles de metilación de sitios CpG individuales, no toman en cuenta algunas fuentes de error. En la Tabla 15 se muestra una comparación de MethFlow frente algunos programas o combinaciones de programas frecuentemente utilizados para la obtención de mapas de metilación.

Tabla 15. Comparación de programas y protocolos de obtención de mapas de metilación

Tratamientos / Análisis	MethFlow	NGSmethPipe y MethylExtract	TrimGalore, Bismark y methylKit	TrimGalore, Bismark y RnBeads	ENCODE	nf-core / methylSeq

Posiciones con baja calidad	✓	✓	✓	✓	✓	✓
Secuencias contaminantes	✓	✓	✓	✓	✓	✓
Pérdida de lecturas en *loci* polimórficos	✓	✗	✗	✗	✗	✗
Lecturas duplicadas	✓	✓	✓	✓	✓	✓
Fallo en la detección de indels	✓	✗	✗	✗	✗	✗
Sesgo de metilación	✓	✗	✗	✗	✗	✓
Fallo en la conversión con bisulfito	✓	✓	✗	✗	✗	✗
Errores debidos a sustituciones	✓	✓	✗	✗	✗	✓
Metilación diferencial	✓	✗	✓	✓	✗	✗
Asociación con la transcripción	✓	✗	✗	✗	✗	✗
Referencias	(R. Lebrón, Barturen, Gómez-Martín, Oliver, & Hackenberg, 2016; Ricardo Lebrón et al., 2017)	(Barturen et al., 2014; Hackenberg et al., 2012)	(Krueger, 2018; Krueger & Andrews, 2011; S. Li et al., 2012)	(Krueger, 2018; Krueger & Andrews, 2011; Müller et al., 2019)	(Bernstein et al., 2012; Sloan et al., 2016)	(Ewels, 2019)

Algunas características exclusivas de MethFlow lo diferencian del resto de programas y protocolos de obtención de mapas de metilación:

- Es el único protocolo que toma en cuenta la pérdida de lecturas en *loci* polimórficos, descubierta durante el desarrollo de esta Tesis Doctoral. En la sección 6.1.1, se discute la estrategia utilizada para solucionar este problema.

- Es el único protocolo que toma en cuenta los fallos en la detección de indels, gracias al uso de una versión modificada de GATK3 (Broad Institute, 2017; Liu et al., 2012).
- Si bien nf-core/methylSeq (Ewels, 2019) corrige parcialmente el sesgo de metilación durante el podado de las lecturas, MethFlow lo corrige por completo, gracias al uso de BSeQC (Lin et al., 2013).
- MethFlow dispone de módulos para la detección de DMCs y CpG-TLs, siendo el primer programa público para la detección de CpG-TLs.
- Permite descargar directamente mapas de metilación disponibles en la base de datos dedicada a la metilación NGSmethDB (Ricardo Lebrón et al., 2017) (véase sección 5.4) y utilizarlos para detectar DMCs.

Sin embargo, MethFlow carece de algunas características interesantes presentes en RnBeads 2.0 (Müller et al., 2019):

- Admite como entrada mapas de metilación procedentes de varios protocolos de secuenciación masiva (entre ellos, WGBS) y arrays, aunque no es capaz de obtener estos mapas de metilación a partir de datos crudos.
- Permite inferir el tipo celular, así como la edad y el sexo del individuo a partir de los niveles de metilación de las citosinas.

Uno de los mayores problemas a los que se enfrenta hoy en día la comunidad científica es la falta de reproducibilidad de los resultados (Baker, 2016). Para garantizar esta reproducibilidad, el diseño de la arquitectura de MethFlow se basó en los siguientes componentes:

- Contenedores generados a partir de un fichero de configuración, en el que se indica la versión de cada programa, su proceso de instalación y de configuración.
- Un sofisticado *framework* para *pipelines* complejas, que ofrece un control y un registro exhaustivo de los procesos ejecutados e incluso permite reanudarlos si se detiene la ejecución voluntariamente o debido a algún error.

Véase la sección 5.1.1 para más información sobre la arquitectura de MethFlow y sus características.

Una extensa colección de mapas de metilación, DMCs y CpG-TLs obtenidos mediante MethFlow están públicamente disponibles en la base de datos dedicada a la metilación NGSmethDB (Ricardo Lebrón et al., 2017) (véase sección 5.4).

6.1.1 Estrategia de alineamiento en dos etapas

El Human Genome Project (Collins, Lander, Rogers, & Waterson, 2004) produjo un ensamblado de referencia del genoma humano de alta calidad, que ha permitido conocer mejor la variabilidad genética de las poblaciones y etnias humanas (Campbell et al., 2015; Sudmant et al., 2015), estudiar el efecto de las variaciones de secuencia sobre la transcripción (Aguet et al., 2017; Van Wittenberghe et al., 2017) y la implicación de las marcas epigenéticas en la memoria celular (Bernstein et al., 2012; Kundaje et al., 2015; Leung et al., 2015). Sin embargo, este ensamblado de referencia se construyó colapsando las secuencias de miles de individuos en una única secuencia consenso. Esto provoca algunos inconvenientes en el análisis de lecturas cortas (M. L. Mendoza et al., 2015).

Con el objetivo de superar las limitaciones del modelo clásico de ensamblado, el Genome Reference Consortium inició la inclusión de parches de haplotipos alternativos en la versión del ensamblado GRCh37/hg19 (Church et al., 2011), los cuales tratan de recoger las variaciones estructurales y de secuencia de distintas poblaciones y etnias humanas, evitando así que las lecturas procedentes de estos haplotipos alineen incorrectamente en otras regiones del genoma. Sin embargo, no se había evaluado si esta inclusión podría acarrear algún problema.

En esta Tesis Doctoral, se describe por primera vez que el uso de los nuevos modelos de ensamblados provoca la pérdida de lecturas procedentes de *loci* polimórficos, como consecuencia de un incremento en el porcentaje de lecturas con alineamiento ambiguo (véase sección 5.1.3).

Para recuperar estas lecturas y asignarlas al ensamblado consenso, se diseñó una estrategia de alineamiento en dos etapas:

i. Todas las lecturas se enfrentan al ensamblado completo.
ii. Aquellas cuyo alineamiento ha resultado ambiguo durante la primera etapa, se enfrentan a una versión del ensamblado sin haplotipos alternativos (ensamblado primario).

Finalmente, se reúnen las lecturas con alineamiento único procedentes de ambos alineamientos, las cuales se utilizarán en posteriores etapas del protocolo. Véase la sección 4.4.2 para más información sobre la estrategia de alineamiento en dos etapas.

Tras probar con los conjuntos de lecturas de varias muestras la estrategia de alineamiento en dos etapas frente a la estrategia convencional de alineamiento con el ensamblado completo y con el ensamblado primario, se comprobó que:

- La estrategia de alineamiento en dos etapas reduce prácticamente por completo la pérdida de lecturas en *loci* polimórficos.
- Las lecturas recuperadas se asignan correctamente a los *loci* correspondiente en el ensamblado consenso.
- El porcentaje de lecturas con alineamiento ambiguo se reduce hasta prácticamente el mismo porcentaje que se obtiene cuando se prescinde de los haplotipos alternativos y en paralelo aumenta el porcentaje de lecturas con alineamiento único.

Por tanto, la estrategia de alineamiento en dos etapas permite sacar partido a las ventajas del uso de haplotipos alternativos, sin el inconveniente de la pérdida de lecturas en *loci* polimórficos.

6.2 Posibles implicaciones funcionales de los DMCs

La metilación de los sitios CpG probablemente sea el componente más importante de la memoria celular, ya que a diferencia del resto de marcas epigenéticas se mantiene estable en todas las fases del ciclo celular (H. Luo et al., 2017). Cada tipo

celular posee un patrón de metilación característico, en parte heredado de la célula madre que le precede en su linaje (Halley-Stott & Gurdon, 2013; Jaenisch & Bird, 2003) y en parte modificado durante el proceso de diferenciación celular (J. U. Guo et al., 2011; M. Kim et al., 2014).

Por otra parte, un mismo tipo celular puede presentar ciertas diferencias de metilación entre individuos debido a factores genéticos y ambientales (Garg et al., 2018; Hannon et al., 2018). Incluso los gemelos recién nacidos presentan diferencias en los patrones de metilación en varios tejidos debido a pequeñas diferencias en el ambiente intrauterino (Ollikainen et al., 2010).

Con el objetivo de medir y caracterizar la variabilidad de la metilación en el genoma humano, en esta Tesis Doctoral se siguió una estrategia consistente en la comparación de pares de muestras humanas y posterior selección de aquellos cambios de metilación que fuesen característicos del tipo celular o del individuo (véase sección 4.5.1). Los resultados obtenidos mediante este método se describen en la sección 5.2. Con base en estos resultados, se discuten a continuación las posibles implicaciones de los DMCs intra-individuales e inter-individuales en diferentes procesos fisiológicos y patológicos.

Se encontró que 3.303.077 (12,19%) y 329.974 (1,22%) de los sitios CpG del genoma humano son, respectivamente, DMCs intra-individuales y DMCs inter-individuales. El hecho de que se hayan detectado diez veces más DMCs intra-individuales que DMCs inter-individuales se puede deber, en parte, a que el número de pares de muestras comparadas ha sido mayor en el primer caso (221 pares de muestras en comparaciones intra-individuales y 61 pares de muestras en comparaciones inter-individuales). En su estudio clásico de metilación diferencial en el genoma humano, Ziller *et al.* encontraron que un 21,8% de los sitios CpG son DMCs (Ziller et al., 2013), casi el doble de DMCs de los detectados en esta Tesis Doctoral. Es probable que una parte de los DMCs detectados por Ziller *et al.* sean en realidad artefactos debidos a las fuentes de error de la técnica WGBS, ya que no tuvieron en cuenta la mayoría de ellas.

Los DMCs intra-individuales e inter-individuales afectan de la misma manera a los principales elementos genómicos relacionados con la regulación de la transcripción: promotores, potenciadores y sitios de unión a factores de transcripción. Sin embargo, se encontró que las regiones de cromatina abierta están enriquecidas en DMCs intra-individuales, pero empobrecidas en DMCs inter-individuales.

Los promotores son pobres en DMCs, mientras que los potenciadores son ricos, lo cual sugiere que la mayoría de cambios de metilación (tanto entre tipos celulares como entre individuos) ocurren en potenciadores. La diferencia entre ambos es considerable, ya que los promotores (del conjunto estricto) presentan un 44,2% y un 34,9%, respectivamente, de los DMCs intra-individuales e inter-individuales que cabrían esperar por azar, mientras que los potenciadores (del conjunto estricto) presentan un 144,9% y un 113,7%, respectivamente. Estos resultados son coherentes con los obtenidos por Ziller *et al.*, encontrando estos autores que la mayoría de los DMCs se ubican en elementos reguladores distales, particularmente potenciadores y sitios de unión a factores de transcripción (Ziller et al., 2013). Zhou *et al.* obtuvieron unos resultados similares, encontrando que las regiones diferencialmente metiladas específicas de tejido están fuertemente asociadas con potenciadores en rata, ratón y humano (Zhou et al., 2017).

Los sitios de unión para la mayoría de factores de transcripción también son ricos en DMCs, independientemente del tipo de factor de transcripción del que se trate: el 74,6% de los factores de transcripción tienen sitios de unión ricos en DMCs intra-individuales, mientras que el 67,8% tienen sitios de unión ricos en DMCs inter-individuales. Zhou *et al.* encontraron que algunos patrones de metilación que no están conservados entre rata, ratón y humano se pueden explicar por ganancias y pérdidas de sitios de unión a factores de transcripción específicos de tejido, sugiriendo que son fundamentales en el establecimiento del patrón de metilación específico de cada tejido (Zhou et al., 2017). Mediante análisis computacionales, Li *et al.* detectaron que los patrones de metilación en los sitios de unión a factores de transcripción son específicos del tipo celular (Xuan Lin et al., 2019), lo cual es coherente con que sean ricos en DMCs. Estos autores, además, encontraron que un

mismo factor de transcripción puede presentar diferentes motivos de unión al ADN, dependiendo del perfil de metilación.

En cuanto al cuerpo génico, cabe destacar que los sitios de inicio de la transcripción (TSSs) y los exones son pobres en DMCs, pero los sitios de fin de la transcripción (TESs) presentan una distribución aleatoria de DMCs. En los genes codificantes, los TSSs presentan un 7% y un 8,1%, respectivamente, de los DMCs intra-individuales e inter-individuales que cabrían esperar por azar, mientras que los exones presentan un 51,1% y un 45,7%, respectivamente. En el caso de los genes no-codificantes, los TSSs presentan un 56,1% de los DMCs intra-individuales que cabrían esperar por azar y una distribución aleatoria de DMCs inter-individuales, mientras que los exones presentan un 71,5% y un 79,4% de los DMCs intra-individuales e inter-individuales que cabrían esperar por azar. Los exones están habitualmente metilados y se sabe que los cambios de metilación en exones alternativos afectan al splicing, pero no los cambios en exones constitutivos (Shayevitch, Askayo, Keydar, & Ast, 2018). El hecho de que la mayoría de exones no cambien su metilación y permanezcan metilados podría explicar que sean pobres en DMCs.

La proporción de DMCs (tanto intra-individuales como inter-individuales) disminuye a medida que decrece la distancia al TSS más próximo y aumenta a medida que decrece la distancia al TES más próximo. El hecho de que la proporción de DMCs sea más baja cerca del TSS podría tener relación con que más de un 70% de los TSSs estén ubicados en islas CpG, las cuales suelen permanecer no-metiladas (Saxonov et al., 2006). Esto explicaría también que, independientemente de la definición que se utilice, las islas CpG sean pobres en DMCs: entre el 2,8% y el 17,7% de los DMCs intra-individuales esperados por azar y entre el 4,1% y el 21,8% de los DMCs inter-individuales esperados por azar. Por otra parte, existen también islas CpG en regiones intragénicas e intergénicas, pero a menudo están metiladas (Jeziorska et al., 2017). Por tanto, no cabe esperar que

este tipo de islas CpG presenten cambios de metilación entre tipos celulares o entre individuos.

En cuanto a los elementos repetidos, los LINEs son ricos en DMCs (tanto entre tipos celulares como entre individuos), mientras que los SINES son pobres. Es interesante que los LINEs sean ricos en DMCs inter-individuales, ya que se ha descrito que la metilación de LINE-1 (la familia de LINEs más importante del genoma humano) está asociada positivamente con un estilo de vida saludable y asociada negativamente con el porcentaje de grasa corporal en individuos jóvenes sanos (Marques-Rocha et al., 2016). Por otra parte, las LTRs son pobres en DMCs intra-individuales y ricas en DMCs inter-individuales.

Los elementos conservados son pobres en DMCs inter-individuales, pero no en DMCs intra-individuales. Se ha descrito que entre el 11% y el 37% de los patrones de metilación específicos de tejido están en regiones conservadas (Zhou et al., 2017). Sin embargo, los elementos ultra-conservados no-codificantes (UCNEs) muestran una variabilidad inter-individual en su metilación mayor que LINE-1, a pesar de que la similitud de secuencia de estas regiones entre humano y pollo es del 95% (Colwell et al., 2018). Es posible que la relación entre los elementos conservados y la variabilidad de su metilación dependa en gran medida de la naturaleza de estos elementos.

En cuanto a los polimorfismos, las DMCs intra-individuales e inter-individuales están sobre-representados en polimorfismos comunes (tanto SNPs como indels) e infra-representados en polimorfismos asociados a enfermedades. Destaca la elevada sobre-representación de DMCs en los SNPs: 219,5% y 1195% de los DMCs intra-individuales e inter-individuales, respectivamente, esperados por azar. Es posible que los individuos analizados sean heterocigóticos para muchos de los sitios CpG. Esto es plausible si se toma en cuenta que la mayoría de SNPs ocurren en contexto CpG (Tomso & Bell, 2003).

6.3 Posibles implicaciones funcionales de los semáforos CpG

Aunque el paradigma tradicional postula que la metilación completa de islas CpG en promotores bloquea la iniciación de la transcripción debido a la unión de MBD2 a sus sitios CpG metilados (Deaton & Bird, 2011; Illingworth et al., 2010; Magdinier & Wolffe, 2002), hoy se sabe que solo la metilación de algunos de estos sitios CpG ejerce un efecto sobre la transcripción (Harbers et al., 2014).

Recientemente, se han descrito los llamados "semáforos CpG" (CpG-TLs), los cuales son sitios CpG individuales cuyo nivel de metilación está asociado con la tasa de transcripción de un gen cercano (Harbers et al., 2014; Lioznova et al., 2019). Estos marcadores biológicos son muy adecuados para poner a prueba la hipótesis de que el signo de la asociación entre la metilación y la transcripción depende del contexto genómico en que se produce la metilación y del tipo de factores de transcripción implicados.

Otros autores han investigado la presencia de CpG-TLs en el genoma humano, aplicando el coeficiente de correlación de Spearman (Spearman, 1987) a vectores de metilación-transcripción, correspondiendo cada vector a un par CpG-gen y cada valor a una muestra distinta (Harbers et al., 2014; Lioznova et al., 2019). De esta manera, estudiaron todos los pares CpG-gen en los que el sitio CpG está ubicado entre 10 kpb aguas arriba del TSS y el TES (Lioznova et al., 2019). En el primer estudio, encontraron que los niveles de metilación del 16,6% de las citosinas y los perfiles de transcripción de genes cercanos presentaban una correlación negativa significativa (Harbers et al., 2014). En el segundo estudio, demostraron que los niveles de metilación de los CpG-TLs son mejores marcadores del nivel de transcripción del gen asociado que el promedio de metilación en promotores o en cuerpos génicos (Lioznova et al., 2019). Encontraron

33.276 CpG-TLs (0,13% de los sitios CpG del genoma humano) asociados negativamente a 7.997 genes.

En esta Tesis Doctoral, se ha desarrollado un método de detección de CpG-TLs que reduce el impacto de los valores atípicos y aumenta la fiabilidad de los resultados, utilizando una combinación del coeficiente de correlación de Spearman y el test de Kruskal-Wallis (Kruskal & Wallis, 1952) (véase sección 4.6). Se amplió la región estudiada en torno a los genes, desde 1 Mpb aguas arriba del TSS hasta 1 Mpb aguas abajo del TES, y se distinguieron dos clases de CpG-TLs: i) rojos, cuando la asociación es negativa, ii) verdes, cuando la asociación es positiva.

Tras aplicar este método mejorado, se encontró que la abundancia de CpG-TLs verdes es casi el doble que la de los CpG-TLs rojos: 126.959 (0,49%) y 66.746 (0,26%), respectivamente, de los sitios CpG del genoma humano. El número de CpG-TLs rojos es el doble del encontrado por Lioznova *et al.* (33.276 CpG-TLs), pero se debe tener en cuenta que estos autores han estudiado una región en torno a los genes más reducida (Lioznova et al., 2019). Es probable que una parte de los CpG-TLs verde se traten en realidad de casos en los que la hidroximetilación se asocia positivamente con la transcripción (Green et al., 2016), ya que la técnica WGBS es incapaz de discriminar entre metilación e hidrometilación (Huang et al., 2010).

Los promotores y potenciadores son ricos en CpG-TLs, tanto rojos (195,6% y 270,9% de lo esperado por azar, respectivamente) como verdes (167,4% y 216,1% de lo esperado por azar, respectivamente). Esto sugiere que ambos disponen de mecanismos para activar o reprimir la transcripción vía metilación, probablemente debido a diferentes combinaciones de sitios de unión a factores de transcripción. También los sitios de unión a CTCF (posibles aisladores) y las regiones de cromatina abierta son ricas en CpG-TLs, tanto rojos (131,9% y 169,1% de lo esperado por azar, respectivamente) como verdes (162% y 154,1% de lo esperado por azar, respectivamente). Estos resultados son coherentes con los obtenidos por Lioznova *et al.*, los cuales encontraron que los CpG-TLs (rojos) están enriquecidos en todos los tipos de regiones reguladoras (Lioznova et al., 2019). En contraste con el modelo tradicional, según el cual cada potenciador es específico de un tipo

celular, se ha descrito que los potenciadores pueden actuar en muchos tipos celulares diferentes y que su metilación define gradientes de expresión en los genes regulados, en lugar de respuestas de todo o nada (Aran, Sabato, & Hellman, 2013).

Sin embargo, los resultados que se han obtenido acerca de la implicación de los CpG-TLs en los sitios de unión a factores de transcripción (TFBSs) son muy diferentes a los publicados por otros autores. Harbers *et al.* observaron que los CpG-TLs estaban fuertemente excluidos de los TFBSs (Harbers et al., 2014). Esta exclusión resultó ser más fuerte en los sitios de unión a represores que en los sitios de unión a activadores o a factores de transcripción multifuncionales. Dentro del TFBS, la exclusión resultó ser más fuerte en las posiciones principales que en las posiciones periféricas. Por ello, concluyeron que la regulación de la unión de factores de transcripción mediante la metilación directa y selectiva de sus sitios de unión debe estar restringida a casos especiales y no puede considerarse como un mecanismo general de la regulación de la transcripción. Más tarde, Lioznova *et al.* describieron que la unión de los factores de transcripción NRF1, ETS, STAT y la familia de factores de transcripción IRF está regulada por la metilación de sus sitios de unión o de regiones próximas a los mismos (Lioznova et al., 2019).

En cambio, en esta Tesis Doctoral se encontró que el 23,7% de los activadores y el 28,8% de los represores presentan sitios de unión ricos en CpG-TLs, mientras que el 58,8% de los factores de transcripción con preferencia por sitios no-metilados (Methyl-Minus) y el 33,3% de los factores de transcripción con preferencia por sitios metilados (Methyl-Plus) presentan sitios de unión ricos en CpG-TLs. Estas discrepancias con los estudios anteriores se pueden deber a dos diferencias metodológicas fundamentales: i) al ampliar la región estudiada, no solo se han tenido en cuenta el promotor y el cuerpo génico, sino también regiones relativamente distales, y ii) en los estudios previos se han excluido los CpG-TLs con asociación positiva.

Se encontró que, mientras que la mayoría de factores de transcripción Methyl-Minus presentan sitios de unión ricos en CpG-TLs rojos (61,8% de los factores de transcripción) y verdes (55,9% de los factores de transcripción), la mayoría de factores de transcripción Methyl-Plus presentan sitios de unión ricos en CpG-TLs verdes (53,8% de los factores de transcripción) y solo una minoría presentan sitios de unión ricos en CpG-TLs rojos (12,8% de los factores de transcripción). Esto sugiere que muchos de los factores de transcripción Methyl-Plus podrían ser activadores. Es interesante destacar que los factores Methyl-Plus son fundamentales en el desarrollo y algunos (como FOXA y Pax7) son capaces de reclutar enzimas que revierten la metilación (Bürglin, 2011; Donaghey et al., 2018; Mayran et al., 2018; Yin et al., 2017; Zhang et al., 2016).

En cuanto a su distribución en torno a los genes, la proporción de CpG-TLs verdes disminuye a medida que decrece la distancia al TSS, mientras que la proporción de CpG-TLs rojos aumenta. Dado que el primer estudio de CpG-TLs se limitó a una región próxima al TSS, es posible que esto llevara a los autores a pensar que las asociaciones positivas son menos relevantes (al ser baja su proporción en torno al TSS) y que por ello las excluyesen del estudio (Harbers et al., 2014). Estos resultados sugieren que los CpG-TLs rojos podrían estar más relacionados con la regulación de la iniciación de la transcripción (tal como postula el paradigma tradicional de la metilación), mientras que los CpG-TLs verdes podrían estar más relacionados con la regulación de la elongación o de la terminación de la transcripción. De hecho, se conoce que la metilación del cuerpo génico está asociada positivamente con la elongación de la transcripción y que algunas islas CpG en 3’ de los genes también presentan una asociación positiva con la transcripción (Aran et al., 2011; Hellman & Chess, 2007; Wolf et al., 1984; D.-H. Yu et al., 2013). Sin embargo, los exones son ricos tanto en CpG-TLs rojos (178,9% de lo esperado por azar) como en CpG-TLs verdes (187% de lo esperado por azar), por lo que la relación entre la metilación del cuerpo génico y la transcripción podría ser compleja.

6.4 Características e innovaciones de NGSmethDB

Las técnicas de secuenciación masiva han permito que la disponibilidad de datos públicos de diferentes tipos de información biológica (genotipo, metilación, modificaciones de histonas, perfiles de expresión, estructura 3D de la cromatina, etc.) aumenten exponencialmente (Levy & Myers, 2016). Sin embargo, este mismo crecimiento ha traído consigo una enorme heterogeneidad de protocolos de secuenciación y de métodos de análisis bioinformáticos, dificultando que las comparaciones entre muestras publicadas por distintos autores.

La base de datos dedicada a la metilación NGSmethDB (https://bioinfo2.ugr.es/NGSmethDB) (Geisen et al., 2014; Hackenberg et al., 2011; Ricardo Lebrón et al., 2017) contiene una amplia colección de mapas de metilación en genoma completo para diferentes especies, tipos celulares e individuos, obtenidos mediante un mismo protocolo de análisis que reduce todas las fuentes de error conocidas que afectan a las lecturas de WGBS (véase sección 1.4).

Si bien existen otras bases de datos dedicadas a la metilación, no todas ellas utilizan un mismo protocolo bioinformático para analizar las muestras ni tratan las fuentes de error que afectan a la detección de la metilación. Por otra parte, algunas bases de datos no disponen de datos de metilación en genoma completo, sino solo en *loci* concretos, y otras solo disponen de datos de muestras patológicas, como muestras de tumores. En la Tabla 16 se muestra una comparación de la NGSmethDB frente a otras bases de datos dedicadas a la metilación.

Tabla 16. Comparación entre bases de datos de metilación

Características	NGSmethDB	MethBase	MethDB	DiseaseMeth	CMS	ENCODE PORTAL
Tratamiento de fuentes de error	✓	✓	✗	✗	✗	✓
Mismo protocolo de análisis	✓	✓	✓	✗	✓	✓
Genoma completo	✓	✓	✗	✓	✓	✓
Muestras fisiológicas	✓	✓	✓	✗	✗	✓
Muestras patológicas	✓	✓	✓	✓	✓	✓
DMCs / DMRs	✓	✓	✗	✗	✓	✗
CpG-TLs	✓	✗	✗	✗	✗	✗
Acceso programático	✓	✗	✗	✗	✗	✓
Ficheros de descarga	✓	✗	✗	✓	✗	✓
Visualización	✓	✓	✗	✓	✓	✓
Referencias	(Geisen et al., 2014; Hackenberg et al., 2011; Ricardo Lebrón et al., 2017)	(Song et al., 2013)	(Amoreira, Hindermann, & Grunau, 2003; Grunau, 2002; Negre & Grunau, 2006)	(Lv et al., 2012; Xiong et al., 2017)	(Gu et al., 2013)	(Bernstein et al., 2012; Sloan et al., 2016)

Algunas características exclusivas de la NGSmethDB la diferencian del resto de bases de datos dedicadas a la metilación:

- Si bien otras bases de datos como MethBase (Song et al., 2013) y ENCODE PORTAL (Bernstein et al., 2012; Sloan et al., 2016) toman en cuenta algunas fuentes de error, como las posiciones con baja calidad, los adaptadores y las lecturas duplicadas, la NGSmethDB es la única base de datos que toma en cuenta todas las fuentes de error conocidas actualmente, gracias al protocolo implementado en MethFlow (véase sección 6.1).
- Ninguna otra base de datos dispone en la actualidad de mapas de DMCs en genoma completo (véase sección 6.2), sino que se limitan a estudiar la metilación diferencial en ciertas regiones, como los promotores o las islas CpG.

- Ninguna otra base de datos dispone en la actualidad de mapas de CpG-TLs, un importante tipo de marcadores biológicos para estudiar la asociación de la metilación con la transcripción (véase sección 6.3).
- Si bien ENCODE PORTAL dispone también de un acceso programático, este no permite consultar datos de sitios CpG individuales, sino solamente obtener los metadatos de la muestra y el mapa de metilación completo. El acceso programático de la NGSmethDB permite un acceso detallado a cada sitio CpG y la información que se dispone del mismo (véase la sección 5.4.3).

Con el fin de optimizar el almacenamiento y la consulta del gran volumen de datos producidos a lo largo de esta Tesis Doctoral, entre los que se incluyen mapas de metilación, de DMCs y de CpG-TLs, se decidió rediseñar por completo la base de datos NGSmethDB (Ricardo Lebrón et al., 2017).

Para agilizar las comparaciones entre muestras, se optó por migrar los datos al sistema de bases de datos MongoDB y almacenarlos en una estructura jerárquica de documentos JSON (un formato estándar que permite intercambiar datos etiquetas y jerarquizados entre distintos lenguajes de programación). En la sección 5.4.2, se describe en detalla la estructura jerárquica utilizada por el back-end de la NGSmethDB.

También se implementaron varias vías de acceso, comparación y visualización de los datos contenidos en la NGSmethDB (véase sección 5.4.3), entre las que destacan su acceso programático mediante el protocolo HTTPS a través de un servidor RESTful API (Fielding, 2000) y su conectividad con UCSC Genome Browser (Casper et al., 2018) a través de Track Hubs (Raney et al., 2014).

Gracias a sus características e innovaciones, NGSmethDB es una base de datos muy adecuada para investigar nuevos marcadores biológicos y la implicación de la metilación en distintos procesos fisiológicos o patológicos.

Capítulo 7

Conclusions

1. **Thanks to the new open-source tool MethFlow, one can take advantage of the WGBS technique assets without the inconveniences caused by biases and contaminations. This allows high quality methylation maps to be obtained and related biological markers (DMCs and CpG-TLs) to be detected. Its design ensures reproducible results while being easy to use.**

2. **The use of assemblies with alternative haplotypes results in the loss of reads from polymorphic loci, due to an increase in ambiguous alignments. Thanks to MethFlow's two-stage alignment strategy, it is possible to retrieve these reads and assign them to the consensus assembly.**

3. **The combination of the Spearman correlation coefficient and the Kruskal-Wallis test improves the detection of CpG-TLs by decreasing the sensitivity to outliers and increasing the reliability of the results.**

4. **Since the distribution of CpG sites in the genome is not random, it was necessary to define the enrichment in DMCs or CpG-TLs as the quotient between the percentages of these markers inside and outside the genomic element.**

5. **The main genomic elements related to the regulation of transcription do not present remarkable differences in intra-individual DMCs and inter-individual DMCs. However, in open chromatin regions intra-individual DMCs are over-represented and inter-individual DMCs, under-represented.**

6. **In the human genome, the percentage of CpG-TLs positively associated with transcription (green) is almost twice of those negatively associated (red). It is likely that some of these CpG-TLs are due to positive associations between hydromethylation and transcription, as the WGBS technique does not make a distinction between 5-methylcytosine and 5-hydroxymethylcytosine.**

7. **In enhancers, both DMCs and CpG-TLs are over-represented, while in promoters DMCs are over-represented but CpG-TLs are under-represented. This seems to indicate that their methylation is associated with changes in transcription, but also that most changes in methylation occur in enhancers.**

8. **In most transcription factor binding sites, DMCs are over-represented. In contrast, the enrichment in CpG-TLs depends on the type of transcription factor: in transcription factors with a preference for non-methylated binding sites, red and green CpG-TLs, are over-represented, while in transcription factors with a preference for methylated binding sites, green CpG-TLs are over-represented but red CpG-TLs are under-represented.**

9. **The proportion of DMCs decreases as the distance to the transcription start site is reduced and increases with proximity to the transcription end site. However, CpG-TLs only show significant changes near the starting site of the transcription: the proportion of green CpG-TLs decreases with proximity to this site, while the proportion of red CpG-TLs increases.**

10. **An extensive collection of methylation DMCs and CpG-TLs maps obtained through MethFlow are publicly available in the database dedicated to NGSmethDB methylation. These maps can be consulted, compared and viewed through several channels, including programmatic access with a RESTful API server and connectivity with the UCSC Genome Browser via Track Hubs.**

Capítulo 8

Conclusiones

1. **Gracias a la nueva herramienta de código abierto MethFlow, se pueden aprovechar las ventajas de la técnica WGBS sin los inconvenientes ocasionados por sesgos y contaminaciones, permitiendo obtener mapas de metilación de alta calidad y detectar marcadores biológicos relacionados (DMCs y CpG-TLs). Su diseño garantiza la reproducibilidad de los resultados, a la par que mantiene la facilidad de uso.**

2. **La utilización de ensamblados con haplotipos alternativos provoca la pérdida de lecturas procedentes de loci polimórficos, debido a un incremento de los alineamientos ambiguos. Gracias a la estrategia de alineamiento en dos etapas de MethFlow, es posible recuperar estas lecturas y asignarlas al ensamblado consenso.**

3. **La combinación del coeficiente de correlación de Spearman y del test de Kruskal-Wallis mejora la detección de CpG-TLs, al disminuir la sensibilidad a los valores atípicos y aumentar la fiabilidad de los resultados.**

4. **Dado que la distribución de los sitios CpG en el genoma no es aleatoria, fue necesario definir la riqueza en DMCs o en CpG-TLs como el cociente entre los porcentajes de estos marcadores dentro y fuera del elemento genómico.**

5. **Los principales elementos genómicos relacionados con la regulación de la transcripción no presentan diferencias destacables en DMCs intra-individuales y DMCs inter-**

individuales. Sin embargo, las regiones de cromatina abierta son ricas en DMCs intra-individuales y pobres en DMCs inter-individuales.

6. **En el genoma humano, el porcentaje de CpG-TLs asociados positivamente con la transcripción (verdes) es casi el doble del que presentan los que están asociados negativamente (rojos). Es probable que algunos de estos CpG-TLs se deban a asociaciones positivas entre hidrometilación y transcripción, ya que la técnica WGBS no discrimina entre 5-metilcitosina y 5-hidroximetilcitosina.**

7. **Los potenciadores son ricos en DMCs y en CpG-TLs, mientras que los promotores son pobres en DMCs pero ricos en CpG-TLs. Esto parece indicar que la metilación de ambos está asociada con cambios en la transcripción, pero que la mayoría de cambios en la metilación ocurre en potenciadores.**

8. **La mayoría de sitios de unión a factores de transcripción son ricos en DMCs. En cambio, la riqueza en CpG-TLs depende del tipo de factor de transcripción: los sitios de unión a factores de transcripción con preferencia por sitios no-metilados son ricos en CpG-TLs rojos y verdes, mientras que los sitios de unión a factores de transcripción con preferencia por sitios metilados son pobres en CpG-TLs rojos pero ricos en CpG-TLs verdes.**

9. **La proporción de DMCs disminuye a medida que decrece la distancia al sitio de inicio de la transcripción y aumenta con la proximidad al sitio de fin de la transcripción. En cambio, los CpG-TLs solo presentan cambios significativos cerca del sitio de inicio de la transcripción: la proporción de los CpG-TLs verdes disminuye con la proximidad a este sitio, mientras que la proporción de CpG-TLs rojos aumenta.**

10. **Una extensa colección de mapas de metilación, DMCs y CpG-TLs obtenidos mediante MethFlow están públicamente disponibles en la base de datos dedicada a la metilación NGSmethDB. Se puede acceder a estos mapas, compararlos y visualizarlos a través de varias vías, entre las que destacan el acceso programático por medio de un servidor RESTful API y su conectividad con UCSC Genome Browser por medio de Track Hubs.**

Capítulo 9

Perspectivas de futuro

Entre la infinidad de posibles mejoras y continuaciones de la investigación presentada en esta Tesis Doctoral, dos de los aspectos que resultarían cruciales para profundizar en la comprensión e interpretación de los resultados son la discriminación entre la metilación e hidrometilación y el análisis de células individuales.

La 5-hidrometilcitosina es una modificación epigenética, procedente de la oxidación enzimática de la 5-metilcitosina. Normalmente, se la considera un paso previo en la reversión de la metilación, pero también existen indicios que apuntan a que podría tener implicaciones funcionales propias (Bachman et al., 2014; Wu & Zhang, 2015). Las técnicas de detección basadas en el tratamiento con bisulfito son incapaces de discriminar entre 5-metilcitosina y 5-hidroximetilcitosina, ya que ambas son resistentes a la acción del bisulfito y se secuencian como citosina (Huang et al., 2010).

Actualmente, existen varias técnicas que permiten detectar la hidroximetilación del ADN. Una de ellas es hMeDIP, una técnica de enriquecimiento por afinidad basada en MeDIP y en la que se utiliza un anticuerpo frente a la 5-hidrometilcitosina para capturar ADN enriquecido en esta modificación (Nestor & Meehan, 2014). También se han desarrollado técnicas basadas en el tratamiento con bisulfito, como OxBS-seq y TAB-seq. En la técnica OxBS-seq, primero se trata

químicamente el ADN para oxidar selectivamente la 5-hidroximetilcitosina a 5-formilcitosina y a continuación se aplica el tratamiento con bisulfito (Booth et al., 2013). De esta manera, la 5-metilcitosina se secuencia como citosina y la 5-hidroximetilcitosina como timina. Comparando el mapa de metilación así obtenido con el mapa de metilación de WGBS, se puede inferir que posiciones estaban hidroximetiladas. Sin embargo, la técnica por excelencia para la secuenciación de la hidroximetilación es TAB-seq, ya que permite detectar directamente la 5-hidroximetilcitosina con una resolución comparable a WGBS. En esta técnica, se protege químicamente a la 5-hidroximetilcitosina, se trata el ADN con TET1 y a continuación se aplica el tratamiento con bisulfito (M. Yu, Hon, Szulwach, Song, Jin, et al., 2012; M. Yu, Hon, Szulwach, Song, Zhang, et al., 2012). De esta manera, la 5-metilcitosina se oxida hasta 5-carboxilcitosina, que se convierte en timina por el tratamiento con bisulfito, mientras que la 5-hidroximetilcitosina queda inalterada y finalmente se secuencia como citosina.

Estas técnicas pueden ser fundamentales para interpretar correctamente la asociación entre la metilación y la transcripción, ya que una proporción de los CpG-TLs podrían estar evidenciando en realidad asociaciones entre la hidroximetilación y la transcripción.

La mayoría de técnicas de detección de la metilación tienen en común que necesitan poblaciones de células y que no son adecuadas para evaluar la heterogeneidad entre células (H. Guo et al., 2015; Schwartzman & Tanay, 2015). Para solucionar estos problemas, se han desarrollado técnicas que permiten detectar la metilación de células individuales. Una de estas técnicas es *Single-Cell Reduced-Representation Bisulfite Sequencing* (scRRBS) (H. Guo et al., 2015), que consiste en aplicar la técnica RRBS a un tubo con el lisado de la célula para minimizar las pérdidas de ADN. Mediante esta técnica se puede detectar la metilación de cerca de un millón de sitios CpG en células individuales. Otra técnica es scBS-seq (Smallwood et al., 2014), que está basada en la técnica PBAT y permite detectar la metilación en el 48,4% de los sitios CpG en células individuales. Por último, existe una variante de WGBS para células individuales, conocida como scWGBS (Farlik et al., 2015).

Con la mejora y aplicación de estas técnicas, debería ser más fácil profundizar en la dinámica de la metilación en respuesta a señales ambientales y del desarrollo. Incluso la propia diferenciación celular es una fuente de heterogeneidad celular, ya que durante la salida de la pluripotencia las células experimentan oscilaciones en su metilación a escala genómica (Rulands et al., 2018). Es probable que la heterogeneidad de metilación entre células proporcione algún tipo de ventaja, como ya se ha demostrado que ocurre con la heterogeneidad en el contenido en factores de transcripción (Torres-Padilla & Chambers, 2014).

Capítulo 10

Anexos

10.1 Métodos suplementarios

10.1.1 Procesado de las anotaciones genómicas

10.1.1.1 Promotores y potenciadores

Las anotaciones de promotores y potenciadores utilizadas proceden de **GeneHancer**, una base de datos de asociaciones promotor-gen y potenciador-gen que forma parte de **GeneCards** (Fishilevich et al., 2017). Sus anotaciones se basan en varias fuentes de información:

- Sitios hipersensibles a la DNAsa I y regiones ricas en la marca de histona H3K27ac, procedentes de los resultados del proyecto ENCODE (Bernstein et al., 2012).
- La colección de elementos reguladores de **Ensembl Regulatory Build** (Zerbino et al., 2015).
- El atlas de potenciadores activos del proyecto FANTOM5 (Rackham et al., 2014).
- La colección de potenciadores validados mediante ensayos en ratones transgénicos de **VISTA Enhancer Browser** (Visel, Minovitsky, Dubchak, & Pennacchio, 2007).

- La colección de super-potenciadores de **dbSUPER** (Khan & Zhang, 2016). Se trata de clústeres de potenciadores que dirigen la expresión de genes específicos del tipo celular y son cruciales para mantener la identidad celular.
- Promotores para la polimerasa II detectados experimentalmente, procedentes de la base de datos **EPDnew** (Dreos, Ambrosini, Périer, & Bucher, 2013).
- Elementos no-codificantes ultra-conservados, procedentes de la base de datos **UCNEbase** (Dimitrieva & Bucher, 2013).

Se descargaron los conjuntos laxo y estricto de las anotaciones de **GeneHancer** a través de **UCSC Genome Browser** (Casper et al., 2018) y se separaron de la siguiente manera:

- Se tomaron como promotores aquellos elementos genómicos cuyo valor en la columna *elementType* de la anotación fuera *Promoter* o *Promoter/Enhancer*. Estos últimos son promotores que actúan a su vez como potenciadores de otros genes. Según la anotación de **GeneHancer**, la mayoría de los promotores actúan también como potenciadores: 18,135 de 22,848 en el conjunto laxo y 13,038 de 13,075 en el conjunto estricto.
- Se tomaron como potenciadores aquellos elementos genómicos cuyo valor en la columna *elementType* de la anotación fuera *Enhancer*.

10.1.1.2 Otras regiones reguladoras

Aparte de los promotores y los potenciadores, se estudiaron otros dos tipos de regiones asociadas con la regulación de la transcripción:

- Sitios de unión a CTCF (CTCFBSs), un factor de transcripción que puede bloquear la interacción entre promotor y potenciador y formar bucles en la cromatina (S. Kim, Yu, & Kaang, 2015).
- Sitios hipersensibles a la DNAsa I (DHSs), en los cuales la cromatina está abierta y se asocian con regulación positiva de la transcripción (Boyle et al., 2008; Wang et al., 2012).

Se descargaron ambas anotaciones a través del **UCSC Track Hub** (Raney et al., 2014) de **Ensembl Regulatory Build** (Zerbino et al., 2015), una colección de

regiones reguladoras determinadas experimentalmente mediante técnicas de ChIP-seq para 18 tipos celulares humanos (Bernstein et al., 2012). Se obtuvieron ambas anotaciones a partir de la pista *overview/RegBuild*, separándolas en función del prefijo que presentan en su columna *name*: *ctcf_* para los CTCFBSs y *open_* para los DHSs.

10.1.1.3 Sitios de unión a factores de transcripción

Las anotaciones de sitios de unión a factores de transcripción (TFBSs) utilizadas proceden de **JASPAR 2018** (Khan et al., 2018), una base de datos de predicciones de TFBSs para multitud de factores de transcripción (TFs) en un amplio abanico de especies. Se descargaron estas anotaciones a través del **UCSC Track Hub** (Raney et al., 2014) de **JASPAR 2018** para humanos, seleccionando aquellos TFBSs cuyo valor-P fuera menor o igual a 10^{-3} (columna *score* mayor o igual a 300) y separándolos en función del TF del que se tratara (indicado en la columna *name*).

Una vez obtenidas las anotaciones, se clasificaron en función de las características biológicas de los TFs:

- Su regulación positiva (activadores) o negativa (represores) de la transcripción (Lambert et al., 2018). Se tomó la clasificación de cada TF como activador o represor de la base de datos **TRRUST v2** (Han et al., 2018), excluyendo a los TFs bifuncionales (que actúan como activadores o represores dependiendo del contexto genómico). En total se seleccionaron 19 TFs activadores y 26 TFs represores:
    - **Activadores:** ATOH1, CREB3, CTCFL, DUX4, DLX3, ESRRG, FOXL1, GBX2, GCM2, HOXD9, LHX2, LHX4, MLXIPL, NFATC3, OTX1, PITX3, POU4F2, SPIC y ZIC1.
    - **Represores:** BHLHE41, CREB3L1, DLX4, E2F7, ELK3, EN1, ERF, EVX1, FOXG1, GFI1B, HEY2, HINFP, HOXB13, HOXC10, HOXC13, HOXA11, HOXC9, KLF14, KLF16, KLF12, TBX2, TEAD4, TGIF1, TCF12, TCFL5 y ZBTB7A.

- Su preferencia por sitios no-metilados (Methyl-Minus) o metilados (Methyl-Plus). Esta clasificación se obtuvo de un estudio sistemático reciente en el que se analizó dicha preferencia para 542 TFs humanos (Yin et al., 2017). Se han excluido aquellos TFs insensibles a la metilación o cuyo motivo de unión cambia en función del estado de metilación. En total se seleccionaron 34 TFs Methyl-Minus y 39 TFs Methyl-Plus:
    - **Methyl-Minus:** ARNTL, CREB1, CREB3, E2F4, ELF1, ELF3, ELF4, ELF5, ELK1, ELK3, ERG, ETV1, ETV2, ETV3, ETV4, ETV5, FEV, FLI1, FOXP3, GABPA, GMEB1, HES5, HES7, HEY2, HES1, HES2, MLX, MYCN, ONECUT2, RUNX3, SPDEF, SPIB, ZBED1 y ZBTB7B.
    - **Methyl-Plus:** ESR1, HOXA10, HOXA13, HOXB13, HOXC10, HOXC11, HOXC12, HOXC13, HOXD11, HOXD13, HOXA11, HOXC9, HOXD9, MEIS2, MEIS3, NFATC1, NFATC3, NR2F1, NR3C2, NR2F6, PAX9, PBX1, PKNOX1, PKNOX2, POU2F2, POU3F1, POU3F2, POU3F4, POU5F1, RARA, RFX5, RORB, RXRB, RXRG, RARB, RARG, SCRT1, SCRT2 y TBX20.

10.1.1.4 Cuerpos génicos

Las coordenadas de los distintos tipos de elementos genómicos que componen el cuerpo génico se obtuvieron a partir de la anotación génica de **GENCODE 29** (Frankish et al., 2019), en formato *GFF2/GTF* (https://www.ensembl.org/info/website/upload/gff.html):

- **Cuerpos génicos completos.** Se obtuvieron las coordenadas de los cuerpos génicos en toda su longitud a partir de aquellas filas del fichero *GTF* cuyo valor de la tercera columna (*feature*) fuera *gene*. Se separaron los cuerpos génicos de genes codificantes y no-codificantes utilizando la información adicional de la columna 9 (*attribute*): los genes cuyo valor *gene_biotype* fuera *protein_coding* o terminara en *_gene* se clasificaron como genes codificantes y el resto como no-codificantes (véase https://www.gencodegenes.org/pages/biotypes.html para más información sobre los biotipos de **GENCODE**).

- **Exones.** Se obtuvieron las coordenadas de los exones a partir de aquellas filas del fichero *GTF* cuyo valor de la tercera columna (*feature*) fuera *exon*. Se separaron los

exones de genes codificantes y no-codificantes de la misma manera en la que se procedió con los cuerpos génicos.

- **Intrones.** Se obtuvieron las coordenadas de los intrones utilizando la herramienta *subtract* de **BEDTools** (Quinlan & Hall, 2010), para sustraerle a cada cuerpo génico las coordenadas de sus exones. Se separaron en intrones de genes codificantes y no-codificantes siguiendo la clasificación del cuerpo génico.
- **Sitios de inicio y de fin de la transcripción (TSSs y TESs).** Se obtuvieron sus coordenadas a partir de la primera y la última posición del cuerpo génico, tomando en cuenta la hebra en la que se localiza el gen. Se separaron en TSSs y TESs de genes codificantes y no-codificantes siguiendo la clasificación del cuerpo génico.

10.1.1.5 Islas CpG

El genoma humano contiene unos 29 millones de sitios CpG, siendo solo el 20% de lo que cabría esperar de su contenido en G+C, como consecuencia de la elevada tasa de mutación por desaminación no-enzimática de la 5-metilcitosina (Antequera, 2003; Sved & Bird, 1990). El 25% de los sitios CpG forman parte de retrotransposones Alu (Y. Luo et al., 2014) y se estima que el 2% forman parte de regiones ricas en sitios CpG, conocidas como islas CpG (Deaton & Bird, 2011). Sin embargo, este porcentaje puede variar en función de la definición de isla CpG que se utilice (Hackenberg et al., 2010), por lo que se tomaron anotaciones obtenidas por dos métodos distintos:

- Islas CpG definidas como clústeres naturales de sitios CpG, siguiendo un método de clusterización basado en la distancia entre sitios CpG consecutivos (Hackenberg et al., 2006). Se calcularon un conjunto laxo y otro estricto de islas CpG utilizando **gCluster** (Gómez-Martín et al., 2018) frente a los cromosomas nucleares del ensamblado GRCh38/hg38. Como umbral de distancia se tomó la intersección entre las distribuciones observada y esperada de distancias entre sitios CpG consecutivos en genoma completo. Los niveles de significación utilizados fueron 10^{-5} para el conjunto laxo y 10^{-20} para el conjunto estricto.

- Islas CpG de **UCSC Genome Browser** (Casper et al., 2018), basadas en un algoritmo de aglutinación de sitios CpG y en la posterior selección de aquellos segmentos que cumplen los criterios de Gardiner-Frommer (Gardiner-Garden & Frommer, 1987). Muchas de estas islas CpG solapan total o parcialmente con retrotransposones Alu, por lo que **UCSC Genome Browser** ofrece una versión en la que se han enmascarado los elementos repetidos detectados con **RepeatMasker** (Smit, 2018) (pista *cpgIslandExt*) y otra versión sin enmascarar (pista *cpgIslandExtUnmasked*). Se utilizaron ambas versiones.

10.1.1.6 Elementos repetidos

UCSC Genome Browser (Casper et al., 2018) dispone de una pista de elementos repetidos detectados con **RepeatMasker** (Smit, 2018) (pista *rmsk*), a partir de la que se obtuvieron las coordenadas genómicas de tres clases de elementos repetidos:

- **Repeticiones terminales largas (LTRs).** Son secuencias de 250-600 pb que se repiten a ambos lados de retrovirus integrados en el genoma y de algunos retrotransposones (Schon et al., 2009). Están anotados como *LTR* en la columna *repClass* de la pista *rmsk*.
- **Elementos nucleares largos intercalados (LINEs).** Son retrotransposones de 6-8 kpb, dispersos por el genoma y que tienen actividad retrotranscriptasa propia. La familia de LINEs más importante del genoma humano es LINE-1 (Beck et al., 2010). Están anotados como *LINE* en la columna *repClass* de la pista *rmsk*.
- **Elementos nucleares cortos intercalados (SINEs).** Son retrotransposones de 50-500 pb, dispersos por el genoma y que carecen de actividad retrotranscriptasa propia. La familia de SINEs más importante del genoma humano es Alu (Bennett et al., 2008). Están anotados como *SINE* en la columna *repClass* de la pista *rmsk*.

10.1.1.7 Elementos conservados

UCSC Genome Browser (Casper et al., 2018) dispone de una serie de pistas de elementos evolutivamente conservados detectados con **PhastCons** (Siepel et al., 2005). Este método está basado en el uso de un modelo oculto de Markov (Stamp, 2004), en el que la probabilidad de que cada nucleótido esté conservado depende

de los nucleótidos adyacentes. Se obtuvieron las coordenadas genómicas de cuatro conjuntos de elementos conservados:

- Elementos conservados en 100 genomas de cordados
  (pista *phastConsElements100way*).
- Elementos conservados en 27 genomas de primates y 3 genomas de mamíferos no-primates
  (pista *phastConsElements30way*).
- Elementos conservados en 20 genomas de mamíferos
  (pista *phastConsElements20way*).
- Elementos conservados en 7 genomas de mamíferos
  (pista *phastConsElements7way*).

10.1.1.8 Polimorfismos

UCSC Genome Browser (Casper et al., 2018) dispone de una serie de pistas de polimorfismos procedentes de la base de datos **dbSNP 151** (Sherry, 2001), a partir de las que se obtuvieron las coordenadas genómicas de cuatro conjuntos de polimorfismos:

- Polimorfismos de un solo nucleótido (SNPs) comunes, cuya frecuencia del alelo menos frecuente es mayor o igual al 1% en alguna de las cinco super-poblaciones humanas (africanos, americanos mezclados, asiáticos del este, asiáticos del sur y europeos) del proyecto 1000 Genomes (Campbell et al., 2015). Se tomó esta anotación de la pista *snp151Common*, seleccionando aquellos elementos cuya columna *class* tuviera el valor *single*.
- SNPs clínicamente asociados, cuya frecuencia del alelo menos frecuente se desconoce o es menor del 1% en todas las super-poblaciones humanas del proyecto 1000 Genomes y que, además, se han asociado clínicamente con alguna enfermedad o trastorno. Se tomó esta anotación de la pista *snp151Flagged*, seleccionando aquellos elementos cuya columna *class* tuviera el valor *single*.

- Inserciones y deleciones (indels) comunes, cuya frecuencia del alelo menos frecuente es mayor o igual al 1% en alguna de las cinco super-poblaciones humanas del proyecto 1000 Genomes. Se tomó esta anotación de la pista *snp151Common*, seleccionando aquellos elementos cuya columna *class* tuviera el valor *in-del*.
- Indels clínicamente asociados, cuya frecuencia del alelo menos frecuente se desconoce o es menor del 1% en todas las super-poblaciones humanas del proyecto 1000 Genomes y que, además, se han asociado clínicamente con alguna enfermedad o trastorno. Se tomó esta anotación de la pista *snp151Flagged*, seleccionando aquellos elementos cuya columna *class* tuviera el valor *in-del*.

10.1.2 Protocolo de obtención de mapas de metilación

En esta sección, se describen las opciones de los programas de terceros utilizadas en el protocolo de obtención de mapas de metilación (véase sección 4.4).

10.1.2.1 Tratamiento previo al alineamiento

El primer bloque de etapas comprende el tratamiento previo al alineamiento de las lecturas frente al ensamblado de referencia (véase sección 4.4.1). Se divide en varias etapas, recogidas en las siguientes secciones.

10.1.2.1.1 Conversión a formato FASTQ

Para convertir los conjuntos de lecturas de formato *SRA* a *FASTQ*, se utiliza la herramienta *fastq-dump* de **SRA Toolkit** (Karsch-Mizrachi et al., 2017) con la opción *--split-files* para garantizar que los fragmentos forward y reverse de los conjuntos de lecturas paired-end van a parar a ficheros *FASTQ* distintos.

10.1.2.1.2 Control de calidad

Antes del podado de las lecturas, se realiza un control para comprobar la calidad de secuenciación (*PHRED score*), el contenido en bases en función de la posición, la presencia de posiciones con secuenciación ambigua, la distribución del

contenido en G+C y la presencia de k-meros sobre-representados, entre otros. Se utiliza **FastQC** (Andrews, 2018) con las opciones por defecto.

10.1.2.1.3 Podado de las lecturas

Para el podado de las lecturas, se utiliza **Trim Galore** (Krueger, 2018), el cual detecta automáticamente las secuencias de los adaptadores y las elimina utilizando **cutadapt** (Martin, 2011). Dependiendo del tipo de lectura y del tipo de biblioteca de secuenciación, se utilizan diferentes opciones:

- Si las lecturas son paired-end, se utiliza la opción *--paired* para procesar conjuntamente los fragmentos forward y reverse. Cuando un fragmento pierde a su pareja durante el podado, este también se descarta.
- Según el tipo de biblioteca de secuenciación, puede ser necesario eliminar un número fijo de posiciones en los extremos 5′ y 3′ de las lecturas, tras el podado de las secuencias de los adaptadores. Estas posiciones contienen, habitualmente, bases introducidas en la reparación de extremos durante la preparación de la biblioteca de secuenciación y pueden introducir sesgos en la metilación (véase sección 1.4.2). Para eliminar estas posiciones, se utilizan las opciones *--clip_r1* y *--three_prime_clip_r1* para lecturas single-end y *--clip_r1*, *--clip_r2*, *--three_prime_clip_r1* y *--three_prime_clip_r2* para las lecturas paired-end. La Tabla 17 muestra el valor recomendado para cada una de estas opciones en función del tipo de biblioteca de secuenciación. Este protocolo se ha implementado como el módulo principal de MethFlow (véase sección 5.1.2), el cual solo necesita que se le indique el tipo de biblioteca de secuenciación para elegir correctamente estas opciones.
- Se utiliza además la opción *--fastqc* para que **Trim Galore** realice un control de calidad de las lecturas tras el podado, utilizando **FastQC**.

Tabla 17. Opciones de podado y alineamiento en función del tipo de biblioteca

Estas recomendaciones se han tomado del anexo IX de la guía de usuario de **Bismark** (Krueger & Andrews, 2011) y se han actualizado siguiendo las recomendaciones de la *pipeline* **nf-core/methylseq** (Ewels, 2019).

Biblioteca de secuenciación	Podado en 5' (R1)	Podado en 5' (R2)	Podado en 3' (R1)	Podado en 3' (R2)	Alineamiento
TruSeq / EpiGnome	8	8	8	8	Por defecto
Accel-NGS / Swift	10	15	10	10	Por defecto
CEGX	6	6	2	2	Por defecto
PBAT	6	9	6	9	*--pbat*
Zymo Pico-Methyl	10	15	10	10	*--non_directional*
Single-cell / scBS-seq	6	6	6	6	*--non_directional*

10.1.2.1.4 Obtención de índices

Para poder utilizar **Bismark** (Krueger & Andrews, 2011) con **Bowtie2** (Langmead & Salzberg, 2012) como programa de alineamiento, es necesario obtener los índices del ensamblado utilizando la herramienta *bismark_genome_preparation* de **Bismark** con la opción *--bowtie2*.

10.1.2.2 Alineamiento en dos etapas

En ambas etapas de alineamiento, se utiliza **Bismark** (Krueger & Andrews, 2011) con las siguientes opciones:

- Para utilizar **Bowtie2** (Langmead & Salzberg, 2012) como programa de alineamiento, se utiliza la opción *--bowtie2*.
- Para obtener los alineamientos únicos en formato *BAM* (H. Li et al., 2009), se utiliza la opción *--bam*.
- En la primera etapa de alineamiento, para obtener las lecturas cuyo alineamiento ha resultado ambiguo en formato *FASTQ* (Cock et al., 2010), se utiliza la opción *--ambiguous*.
- Se utiliza la opción *--score_min L,0-0.2* para especificar una puntuación mínima de alineamiento más exigente que el valor por defecto que utiliza **Bowtie2** (*--score_min L,0-0.6*).
- Dependiendo del tipo de biblioteca de secuenciación, se utilizan algunas opciones adicionales (véase Tabla 17).

Para cada etapa del alineamiento, se obtiene un informe del proceso utilizando la herramienta *bismark2report* de **Bismark** con la opción *--alignment_report*. También se obtiene un resumen del proceso utilizando la herramienta *bismark2summary* de **Bismark** con las opciones por defecto.

Para unir los alineamientos únicos procedentes de ambas etapas, se utiliza la herramienta *merge* de **SAMtools** (H. Li et al., 2009) con las opciones por defecto.

10.1.2.3 Tratamiento posterior al alineamiento

El tercer bloque de etapas comprende el tratamiento posterior al alineamiento de las lecturas frente al ensamblado de referencia (véase sección 4.4.3). Se divide en varias etapas, recogidas en las siguientes secciones.

10.1.2.3.1 Ordenación por coordenadas

Para poder eliminar las lecturas duplicadas, primero es necesario ordenar los alineamientos en función de sus coordenadas genómicas. Para ello, se utiliza la herramienta *sort* de **SAMtools** (H. Li et al., 2009) con las opciones por defecto. A continuación, si la muestra que se está analizando se compone de varios conjuntos de lecturas, sus alineamientos se unen en un único fichero *BAM* utilizando la herramienta *merge* de **SAMtools** con las opciones por defecto.

10.1.2.3.2 Eliminación de duplicados

Una vez ordenados los alineamientos en función de sus coordenadas genómicas, se utiliza la herramienta *MarkDuplicates* de **Picard Tools** (Broad Institute, 2019) con la opción *REMOVE_DUPLICATES=true* para eliminar las lecturas duplicadas.

El algoritmo de eliminación de duplicados de *MarkDuplicates* consiste en detectar lecturas cuyas coordenadas del extremo 5' coincidan y, entre ellas, elegir la que tenga una calidad de secuenciación media mayor. El resto de lecturas se consideran duplicados y se eliminan. Si bien el algoritmo de eliminación de duplicados de **MethylExtract** (Barturen et al., 2014) es más sofisticado, solo se ha

implementado para lecturas single-end, por lo que se ha preferido utilizar *MarkDuplicates*.

10.1.2.3.3 Realineamiento local

Para corregir los indels en los extremos de las lecturas alineadas, se lleva a cabo un realineamiento local por el método de máxima parsimonia en las regiones afectadas. En primer lugar, se utiliza una versión modificada de la herramienta *RealignerTargetCreator* de **GATK3** (McKenna et al., 2010) con las opciones por defecto. Esta herramienta detecta las regiones afectadas por este problema y almacena sus coordenadas en un fichero que será utilizado por la siguiente herramienta. A continuación, se utiliza una versión modificada de la herramienta *IndelRealigner* de **GATK3** (McKenna et al., 2010) con las opciones por defecto, para corregir el alineamiento en las regiones afectadas.

Estas versiones modificadas de las herramientas de **GATK3** proceden de **Bis-SNP** (Liu et al., 2012) y están adaptadas para trabajar en los alfabetos de tres letras que requiere el alineamiento del ADN tratado con bisulfito.

10.1.2.3.4 Corrección del sesgo de metilación

Antes de detectar el nivel de metilación de cada citosina a partir de los alineamientos, es necesario detectar y corregir el sesgo de metilación. Por ello, se utiliza la herramienta *mbias* de **BSeQC** (Lin et al., 2013) con las siguientes opciones:

- Se establece el umbral de significación estadística para la detección de posiciones afectadas por el sesgo de metilación en 0.01, utilizando la opción *--pvalue*.
- Para activar el podado automático de las posiciones en los extremos de las lecturas afectadas por el sesgo de metilación, se utiliza la opción *--auto*.

El algoritmo de detección del sesgo de metilación de *mbias* consiste en tomar la región central de las lecturas (del 30% al 70% de la longitud) y calcular una distribución normal a partir de estos valores de metilación, la cual se utiliza para detectar si las posiciones de los extremos se desvían significativamente de esta distribución.

10.1.2.3.5 Ordenación por identificador

Para poder detectar y corregir los problemas descritos en los apartados anteriores, es necesario que los alineamientos estén ordenados en función de sus coordenadas genómicas. Sin embargo, para la detección de los niveles de metilación con **MethylExtract** (Barturen et al., 2014) es necesario que los alineamientos estén ordenados por identificador (*query name* o *qname*). Para ello, se utiliza la herramienta *sort* de **SAMtools** (H. Li et al., 2009) con la opción *-n*.

10.1.2.3.6 Control de calidad

Se realiza un control de calidad de los alineamientos utilizando las herramientas *flagstat* y *stats* de **SAMtools** (H. Li et al., 2009) con las opciones por defecto. También se utiliza la herramienta *bamqc* de **QualiMap 2** (Okonechnikov, Conesa, & García-Alcalde, 2015) con la opción *--collect-overlap-pairs* para detectar posibles solapamientos entre el fragmento forward y el fragmento reverse en lecturas paired-end.

10.1.2.4 Detección de la metilación

Para detectar los niveles de metilación de cada citosina, se utiliza **MethylExtract** (Barturen et al., 2014). Este programa es capaz de detectar variantes de secuencia siguiendo una estrategia similar a la de **varScan** (Koboldt et al., 2009), mientras detecta el nivel de metilación de cada citosina. Además, dispone de una amplia gama de filtros para evitar sesgos y contaminaciones. Se utilizan las siguientes opciones de **MethylExtract**:

- *flagW=0* para lecturas single-end y *flagW=99,147* para lecturas paired-end.
- *flagC=16* para lecturas single-end y *flagC=83,163* para lecturas paired-end.
- Para establecer la calidad mínima de secuenciación en 20, se utiliza la opción *minQ=20*. Se ignoran las posiciones de las lecturas que tengan una calidad de secuenciación (*PHRED score*) menor de este valor.
- Para excluir las lecturas afectadas por fallo en la conversión por bisulfito, se utiliza la opción *methNonCpGs=0.9*. Se ignoran todas las lecturas que tengan un porcentaje

de citosinas metiladas en contexto no-CpGs mayor del 90%. Para desactivar está opción se debe fijar su valor en 0.

- Para eliminar posibles solapamientos entre los fragmentos forward y reverse de las lecturas paired-end, se utiliza la opción *peOverlap=Y*.
- Para que se reporten en los resultados todas las citosinas que estén cubiertas por alguna lectura, se utiliza la opción *minDepthMeth=1*.
- Las opciones utilizadas en la detección de variantes de secuencia son *minDepthSNV=1*, *varFraction=0.1*, *maxStrandBias=0.7* y *maxPval=0.05*. Así, una variante debe tener una frecuencia de al menos el 10%, una desviación entre hebras no superior al 70% y un valor-P menor o igual a 0.05 para que sea reportada en los resultados.

10.1.3 Obtención de perfiles de transcripción

Para la detección de los CpG-TLs caracterizados en esta Tesis Doctoral (véase sección 4.6), se utilizaron los perfiles de expresión de las muestras indicadas en la columna *CpG-TLs* de la Tabla 18. El perfil de expresión de cada muestra se obtuvo de **ENCODE PORTAL** (Sloan et al., 2016) en formato tabular, donde cada fila corresponde a un gen de **ENCODE 29** (Frankish et al., 2019) y cada columna una variable relacionada con el nivel de transcripción del gen.

Se procesaron estos ficheros y se transformaron en ficheros tabulares con las siguientes columnas:

- Cromosoma, coordenada de inicio, coordenada de fin y hebra. Estas columnas se obtuvieron a partir de la anotación génica de **ENCODE 29** para el identificador contenido en la primera columna del fichero original.
- Identificador de **ENCODE 29**, tasa de transcripción en TPM (Wagner et al., 2012) y tasa de transcripción en número de lecturas. Corresponden a las columnas primera, sexta y octava del fichero original.

10.2 Tablas suplementarias

10.2.1 Muestras e individuos

Tabla 18. Muestras de las que proceden las lecturas de WGBS

En la primera columna se indica el pseudónimo que recibe el individuo en el proyecto del que proceden los datos (véase **Tabla 19** para más información). En la segunda columna se indica el nombre común del órgano, tejido, tipo celular o tumor del que procede la muestra (véase Tabla 20 para más información). Las columnas tercera y cuarta indican, respectivamente, si la muestra se ha utilizado para calcular DMCs intra-individuales e inter-individuales (véase sección 4.5). Los DMCs de las muestras marcadas con el símbolo "P" están disponibles en la base de datos NGSmethDB (véase sección 5.4), pero no se han tenido en cuenta en el cálculo de los conjuntos estrictos de DMCs. La quinta columna indica si la muestra se ha utilizado para calcular CpG-TLs (véase sección 4.6).

Individuo	Muestra	intra-DMCs	inter-DMCs	CpG-TLs
1JKYN	Bazo	✓	✓	✗
1JKYN	Cuerpo del páncreas	✓	P	✗
1JKYN	Glándula tiroides	✓	P	✗
1JKYN	Lóbulo superior izquierdo del pulmón	✓	P	✗
1JKYN	Piel de la parte inferior de la pierna	✓	P	✗
1JKYN	Testículo	✓	P	✗
1K2DA	Cuerpo del páncreas	✓	P	✗
1K2DA	Estómago	✓	✓	✗
1K2DA	Glándula prostática	✓	✗	✗
1K2DA	Glándula suprarrenal	✓	✓	✗
1K2DA	Nervio tibial	✓	P	✗
1K2DA	Testículo	✓	P	✗
1LGRB	Glándula suprarrenal	P	✓	✗
1LGRB	Nervio tibial	P	P	✗
1LGRB	Ovario	P	P	✗
1LVAN	Bazo	✓	✓	✗
1LVAN	Estómago	✓	✓	✗
1LVAN	Glándula tiroides	✓	P	✗
1LVAN	Lóbulo superior izquierdo del pulmón	✓	P	✗
1LVAN	Ovario	✓	P	✗
1LVAN	Piel de la parte inferior de la pierna	✓	P	✗
A549	Adenocarcinoma alveolar	✗	✗	✗

GM12878	Células linfoblastoides	✗	✗	✓
GM23248	Fibroblastos del brazo	✗	✗	✓
H-22510	Corazón	✗	✗	✗
H-23769	Intestino delgado	P	✓	✓
H-23769	Intestino grueso	P	✗	✓
H-24720	Músculo esquelético del tronco	✗	✗	✗
H-24810	Estómago	✗	✓	✗
H-24943	Médula espinal	P	✗	✗
H-24943	Placenta	P	✗	✗
H-24943	Timo	P	P	✗
H-24996	Músculo de la pierna	✗	✗	✗
H-25008	Glándula suprarrenal	✗	✓	✗
H1	Células madre embrionarias	P	✗	✓
H1	Células madre mesenquimales	P	✗	✓
H9	Células musculares lisas	P	✗	✓
H9	Hepatocitos	P	✗	✓
HeLa-S3	Adenocarcinoma cervical	✗	✗	✓
HepG2	Carcinoma hepatocelular	✗	✗	✓
HSMM	Mioblasto del músculo esquelético	✗	✗	✓
HUES64	Células madre embrionarias	✓	✗	✓
HUES64	Ectodermo	✓	✗	✓
HUES64	Endodermo	✓	✗	✓
HUES64	Mesodermo	✓	✗	✓
IMR-90	Fibroblastos de pulmón fetal	✗	✗	✓
K562	Leucemia mielógena crónica	✗	✗	✓
OCI-LY7	Linfoma no-Hodgkin de células B	✗	✗	✗
RO-01549	Progenitor mieloide común CD34+	✗	✗	✓
RO-02035	Células NK	✓	✗	✗
RO-02035	Linfocitos B	✓	✗	✗
RO-02035	Linfocitos T	✓	✗	✗
RO-02035	Monocitos CD14+	✓	✗	✗
SK-N-SH	Neuroblastoma metastásico	✗	✗	✓
STL001	Bazo	✓	✓	✓
STL001	Colon sigmoide	✓	P	✓
STL001	Estómago	✓	✓	✓
STL001	Intestino delgado	✓	✓	✓
STL001	Músculo psoas	✓	P	✓

STL001	Pulmón	✓	P	✓
STL001	Tejido adiposo	✓	P	✓
STL001	Timo	✓	P	✓
STL001	Ventrículo derecho	✓	P	✓
STL001	Ventrículo izquierdo	✓	P	✓
STL002	Aorta	✓	✗	✓
STL002	Bazo	✓	✓	✓
STL002	Esófago	✓	P	✓
STL002	Estómago	✓	✓	✓
STL002	Glándula suprarrenal	✓	✓	✓
STL002	Intestino delgado	✓	✓	✓
STL002	Músculo psoas	✓	P	✓
STL002	Ovario	✓	P	✓
STL002	Páncreas	✓	P	✓
STL002	Pulmón	✓	P	✓
STL002	Tejido adiposo	✓	P	✓
STL003	Aurícula derecha	✓	✗	✓
STL003	Bazo	✓	✓	✓
STL003	Colon sigmoide	✓	P	✓
STL003	Esófago	✓	P	✓
STL003	Estómago	✓	✓	✓
STL003	Glándula suprarrenal	✓	✓	✓
STL003	Intestino delgado	✓	✓	✓
STL003	Páncreas	✓	P	✓
STL003	Tejido adiposo	✓	P	✓
STL003	Ventrículo derecho	✓	P	✓
STL003	Ventrículo izquierdo	✓	P	✓

Tabla 19. Información detallada sobre los individuos

En la primera columna se indica el pseudónimo que recibe el individuo en el proyecto del que proceden los datos. En la segunda columna se indica el identificador del individuo utilizado por **ENCODE PORTAL** (Sloan et al., 2016). La tercera columna indica la edad del individuo. Las abreviaturas "d.g." y "s.g." hacen alusión a días y semanas de gestación, respectivamente, en las muestras procedentes de fetos y embriones. Las columnas cuarta y quinta indican el sexo y la(s) etnia(s). Los datos que se desconocen se han indicado con el símbolo "D". En el caso de las etnias, "C" y "AA" hacen alusión a caucásica y afroamericana, respectivamente. Las columnas sexta y séptima indican el estado de saludo y el proyecto en el que ha participado el individuo, respectivamente. Todas las muestras de un mismo individuo pertenecen al mismo proyecto: Roadmap Epigenomics (Kundaje et al., 2015; Leung et al., 2015), ENCODE (Bernstein et al., 2012) o Enhancing GTEx (Van Wittenberghe et al., 2017).

Individuo	ID de ENCODE	Edad	Sexo	Etnia	Estado de salud	Proyecto
1JKYN	ENCDO845WKR	37 años	♂	D	Desconocido	GTEx
1K2DA	ENCDO451RUA	54 años	♂	D	Desconocido	GTEx
1LGRB	ENCDO793LXB	53 años	♀	D	Desconocido	GTEx
1LVAN	ENCDO271OUW	51 años	♀	D	Desconocido	GTEx
A549	ENCDO000AAZ	58 años	♂	C	Adenocarcinoma alveolar	ENCODE
GM12878	ENCDO000AAK	Adulto	♀	C	Desconocido	ENCODE
GM23248	ENCDO336AAA	53 años	♂	C	Desconocido	ENCODE
H-22510	ENCDO680EXU	101 d.g.	D	D	Sano	Roadmap
H-23769	ENCDO132ASK	108 d.g.	♂	D	Desconocido	Roadmap
H-24720	ENCDO585IIC	115 d.g.	♀	D	Sano	Roadmap
H-24810	ENCDO318GKQ	98 d.g.	♀	D	Sano	Roadmap
H-24943	ENCDO376MNT	115 d.g.	♀	D	Sano	Roadmap
H-24996	ENCDO022PVU	113 d.g.	♀	D	Desconocido	Roadmap
H-25008	ENCDO900GTZ	97 d.g.	♂	D	Desconocido	Roadmap
H1	ENCDO000AAW	Embrión	♂	D	Sano	Roadmap
H9	ENCDO222AAA	5 d.g.	♀	D	Sano	Roadmap
HeLa-S3	ENCDO000AAB	31 años	♀	AA	Adenocarcinoma cervical	ENCODE
HepG2	ENCDO000AAC	15 años	♂	C	Carcinoma hepatocelular	ENCODE
HSMM	ENCDO094AAA	22 años	♂	C	Desconocido	ENCODE
HUES64	ENCDO424FTP	Embrión	♂	D	Sano	Roadmap
IMR-90	ENCDO000AAX	16 s.g.	♀	C	Desconocido	ENCODE
K562	ENCDO000AAD	53 años	♀	D	Leucemia mielógena crónica	ENCODE
OCI-LY7	ENCDO351AAA	48 años	♂	D	Linfoma de células B	ENCODE
RO-01549	ENCDO412QUR	33 años	♀	C	Desconocido	Roadmap

RO-02035	ENCDO661BYS	37 años	♂	D	Sano	Roadmap
SK-N-SH	ENCDO000ABD	4 años	♀	D	Neuroblastoma metastático	ENCODE
STL001	ENCDO478OMA	3 años	♂	C y AA	Sano	Roadmap
STL002	ENCDO424HVB	30 años	♀	C	Trastorno bipolar	Roadmap
STL003	ENCDO058AAA	34 años	♂	C	Abuso de polisustancias	Roadmap

Tabla 20. Información detallada sobre las muestras

En la segunda columna se indica el nombre común del órgano, tejido, tipo celular o tumor del que procede la muestra. En la segunda columna se indica el identificador ontogenético de la muestra. Estos identificadores permiten clasificar las muestras de forma inequívoca. Dependiendo de si se trata de un órgano, un tejido, un tipo celular o un tumor, la ontología utilizada puede ser **UBERON** (Mungall, Torniai, Gkoutos, Lewis, & Haendel, 2012), **CL** (Lizio et al., 2015; Malladi et al., 2015) o **EFO** (Malone et al., 2010). El identificador ontogenético de cada muestra se ha tomado de **ENCODE PORTAL** (Sloan et al., 2016).

Muestra	ID de muestra	Método de obtención
Aorta	UBERON:0000947	Biopsia
Aurícula derecha	UBERON:0002078	Biopsia
Bazo	UBERON:0002106	Biopsia
Colon sigmoide	UBERON:0001159	Biopsia
Corazón	UBERON:0000948	Biopsia
Cuerpo del páncreas	UBERON:0001150	Biopsia
Esófago	UBERON:0001043	Biopsia
Estómago	UBERON:0000945	Biopsia
Glándula prostática	UBERON:0002367	Biopsia
Glándula suprarrenal	UBERON:0002369	Biopsia
Glándula tiroides	UBERON:0002046	Biopsia
Intestino delgado	UBERON:0002108	Biopsia
Intestino grueso	UBERON:0000059	Biopsia
Lóbulo superior izquierdo del pulmón	UBERON:0008952	Biopsia
Médula espinal	UBERON:0002240	Biopsia
Músculo de la pierna	UBERON:0001383	Biopsia
Músculo esquelético del tronco	UBERON:0001774	Biopsia
Músculo psoas	UBERON:0008450	Biopsia
Nervio tibial	UBERON:0001323	Biopsia
Ovario	UBERON:0000992	Biopsia
Páncreas	UBERON:0001264	Biopsia
Piel de la parte inferior de la pierna	UBERON:0004264	Biopsia
Placenta	UBERON:0001987	Biopsia

Pulmón	UBERON:0002048	Biopsia
Tejido adiposo	UBERON:0001013	Biopsia
Testículo	UBERON:0000473	Biopsia
Timo	UBERON:0002370	Biopsia
Ventrículo derecho	UBERON:0002080	Biopsia
Ventrículo izquierdo	UBERON:0002084	Biopsia
Células NK	CL:0000623	Biopsia y citometría de flujo
Linfocitos B	CL:0000236	Biopsia y citometría de flujo
Linfocitos T	CL:0000084	Biopsia y citometría de flujo
Monocitos CD14+	CL:0001054	Biopsia y citometría de flujo
Progenitor mieloide común CD34+	CL:0001059	Biopsia y citometría de flujo
Mioblasto del músculo esquelético	CL:0000515	Biopsia y cultivo primario
Células madre mesenquimales	CL:0000134	Células diferenciadas in vitro
Células musculares lisas	CL:0000192	Células diferenciadas in vitro
Ectodermo	CL:0000221	Células diferenciadas in vitro
Endodermo	CL:0000223	Células diferenciadas in vitro
Hepatocitos	CL:0000182	Células diferenciadas in vitro
Mesodermo	CL:0000222	Células diferenciadas in vitro
Adenocarcinoma alveolar	EFO:0001086	Línea celular A549
Células linfoblastoides	EFO:0002784	Línea celular GM12878
Fibroblastos del brazo	EFO:0005723	Línea celular GM23248
Células madre embrionarias	EFO:0003042	Línea celular H1
Adenocarcinoma cervical	EFO:0002791	Línea celular HeLa-S3
Carcinoma hepatocelular	EFO:0001187	Línea celular HepG2
Células madre embrionarias	EFO:0007089	Línea celular HUES64
Fibroblastos de pulmón fetal	EFO:0001196	Línea celular IMR-90
Leucemia mielógena crónica	EFO:0002067	Línea celular K562
Linfoma no-Hodgkin de células B	EFO:0006711	Línea celular OCI-LY7
Neuroblastoma metastásico	EFO:0003072	Línea celular SK-N-SH

10.2.2 Alineamiento en dos etapas

Tabla 21. Porcentajes de lecturas alineadas en cada tipo de alineamiento

La primera y la segunda columna indican el individuo y la muestra cuyos porcentajes de lecturas alineadas se muestran en la fila. Las columna tercera y cuarta son los porcentajes de lecturas con alineamiento único (% U) y con alineamiento ambiguo (% A) tras alinear frente al ensamblado primario de GRCh38/hg38. Las columnas quinta y sexta son los porcentajes de lecturas con alineamiento único y con alineamiento ambiguo tras alinear frente al ensamblado completo de GRCh38/hg38. Las columnas séptima y octava son los porcentajes de lecturas con alineamiento único y con alineamiento ambiguo tras alinear siguiendo la estrategia en dos etapas. La columna novena (% R) es el porcentaje de lecturas ambiguas recuperadas durante la segunda etapa de alineamiento. La columna décima (% L) es el porcentaje de lecturas recuperadas que se han asignado correctamente a las regiones correspondientes del ensamblado consenso.

Individuo	Muestra	Ensamblado primario		Ensamblado completo		Alineamiento en dos etapas			
		% U	% A	% U	% A	% U	% A	% R	% L
ENCDO058AAA	UBERON:0000945	87,40	8,85	85,96	10,31	87,43	8,84	14,21	98,71
ENCDO058AAA	UBERON:0001013	83,02	9,92	81,69	11,26	83,05	9,90	12,05	98,37
ENCDO058AAA	UBERON:0001043	87,93	8,39	86,48	9,85	87,96	8,37	15,03	98,66
ENCDO058AAA	UBERON:0001159	83,21	11,49	81,54	13,19	83,26	11,47	13,05	98,52
ENCDO058AAA	UBERON:0001264	45,70	9,34	44,85	10,19	45,78	9,26	9,15	90,61
ENCDO058AAA	UBERON:0002078	85,05	9,97	83,70	11,33	85,08	9,96	12,16	98,67
ENCDO058AAA	UBERON:0002080	84,36	9,65	83,03	11,00	84,40	9,64	12,40	98,67
ENCDO058AAA	UBERON:0002084	87,96	8,39	86,55	9,82	87,99	8,38	14,61	98,42
ENCDO058AAA	UBERON:0002106	82,29	11,96	80,58	13,69	82,34	11,93	12,81	98,39
ENCDO058AAA	UBERON:0002108	85,19	9,60	83,83	10,96	85,21	9,58	12,62	98,57
ENCDO058AAA	UBERON:0002369	84,30	9,17	82,99	10,50	84,33	9,16	12,78	98,68
ENCDO271OUW	UBERON:0000945	76,04	3,34	74,91	4,49	76,09	3,32	26,18	97,46
ENCDO271OUW	UBERON:0000992	69,16	3,27	68,14	4,31	69,20	3,24	24,74	97,19
ENCDO271OUW	UBERON:0002046	81,84	4,36	80,62	5,60	81,88	4,33	22,59	97,27
ENCDO271OUW	UBERON:0002106	79,06	3,60	77,88	4,80	79,10	3,58	25,52	97,62
ENCDO271OUW	UBERON:0004264	83,65	4,46	82,40	5,73	83,69	4,44	22,47	97,94
ENCDO271OUW	UBERON:0008952	77,18	3,86	76,00	5,06	77,23	3,84	24,14	98,23
ENCDO376MNT	UBERON:0001987	88,15	7,30	85,94	9,57	88,25	7,26	24,13	98,12
ENCDO376MNT	UBERON:0002240	87,93	6,75	85,30	9,46	88,01	6,75	28,67	98,62
ENCDO376MNT	UBERON:0002370	86,91	6,25	85,18	8,03	87,00	6,21	22,65	97,82
ENCDO424FTP	CL:0000221	84,65	7,41	82,52	9,58	84,69	7,40	22,71	99,03
ENCDO424FTP	CL:0000222	78,59	10,35	76,46	12,51	78,63	10,34	17,33	98,96
ENCDO424FTP	CL:0000223	75,55	9,51	73,71	11,38	75,60	9,49	16,65	98,63
ENCDO424FTP	EFO:0007089	80,99	6,99	79,43	8,57	81,04	6,96	18,74	98,66
ENCDO424HVB	UBERON:0000945	84,17	9,62	82,83	10,98	84,20	9,61	12,49	98,44
ENCDO424HVB	UBERON:0000947	83,94	9,44	82,58	10,81	83,97	9,43	12,81	98,54
ENCDO424HVB	UBERON:0000992	86,43	9,35	85,07	10,73	86,47	9,33	13,03	98,36

ENCDO424HVB	UBERON:0001013	84,86	9,22	83,52	10,57	84,90	9,20	12,99	98,53
ENCDO424HVB	UBERON:0001043	85,03	9,80	83,69	11,16	85,07	9,78	12,33	98,37
ENCDO424HVB	UBERON:0001264	86,34	9,48	84,97	10,87	86,38	9,47	12,92	98,41
ENCDO424HVB	UBERON:0002048	80,71	11,48	79,22	13,00	80,77	11,45	11,91	97,69
ENCDO424HVB	UBERON:0002106	88,01	8,59	86,61	10,01	88,04	8,58	14,29	98,60
ENCDO424HVB	UBERON:0002108	87,59	8,74	86,19	10,16	87,63	8,72	14,10	98,37
ENCDO424HVB	UBERON:0002369	85,99	9,27	84,67	10,61	86,03	9,26	12,76	98,50
ENCDO424HVB	UBERON:0008450	86,27	9,44	84,91	10,81	86,31	9,42	12,87	98,57
ENCDO451RUA	UBERON:0000473	77,99	4,39	76,81	5,60	78,04	4,36	22,05	97,40
ENCDO451RUA	UBERON:0000945	77,08	4,54	75,89	5,75	77,13	4,51	21,47	97,41
ENCDO451RUA	UBERON:0001150	71,96	3,08	70,93	4,13	72,01	3,05	26,09	97,20
ENCDO451RUA	UBERON:0001323	78,67	4,54	77,50	5,74	78,72	4,52	21,30	97,61
ENCDO451RUA	UBERON:0002367	80,31	4,67	79,05	5,96	80,36	4,65	22,00	97,93
ENCDO451RUA	UBERON:0002369	78,53	4,16	77,34	5,37	78,57	4,14	22,95	97,52
ENCDO478OMA	UBERON:0000945	80,11	10,64	78,83	11,93	80,14	10,62	10,97	98,41
ENCDO478OMA	UBERON:0001013	83,58	9,50	82,25	10,86	83,62	9,48	12,63	98,49
ENCDO478OMA	UBERON:0001159	87,90	8,28	86,49	9,71	87,93	8,27	14,87	98,42
ENCDO478OMA	UBERON:0002048	84,83	9,97	83,45	11,37	84,86	9,96	12,39	98,37
ENCDO478OMA	UBERON:0002080	85,94	9,45	84,54	10,85	85,97	9,43	13,15	98,52
ENCDO478OMA	UBERON:0002084	85,54	9,62	84,17	11,01	85,58	9,60	12,76	98,59
ENCDO478OMA	UBERON:0002106	87,30	8,55	85,89	9,98	87,33	8,54	14,47	98,53
ENCDO478OMA	UBERON:0002108	86,05	9,33	84,66	10,74	86,08	9,32	13,18	98,45
ENCDO478OMA	UBERON:0002370	86,67	9,03	85,26	10,45	86,69	9,01	13,76	98,49
ENCDO478OMA	UBERON:0008450	85,91	9,46	84,52	10,86	85,93	9,45	13,04	98,72
ENCDO661BYS	CL:0000084	87,34	8,42	85,21	10,59	87,43	8,37	20,95	98,14
ENCDO661BYS	CL:0000236	86,88	8,37	84,11	11,20	86,98	8,33	25,64	98,59
ENCDO661BYS	CL:0000623	86,88	8,80	84,83	10,89	86,97	8,75	19,63	98,01
ENCDO661BYS	CL:0001054	87,10	8,63	84,91	10,88	87,20	8,59	21,05	98,36
ENCDO793LXB	UBERON:0000992	78,03	3,73	76,88	4,91	78,08	3,71	24,54	97,84
ENCDO793LXB	UBERON:0001323	82,58	3,97	81,35	5,23	82,63	3,95	24,49	98,08
ENCDO793LXB	UBERON:0002369	82,09	4,44	80,89	5,68	82,14	4,42	22,08	97,97
ENCDO845WKR	UBERON:0000473	83,78	4,50	82,46	5,85	83,82	4,48	23,32	97,91
ENCDO845WKR	UBERON:0001150	61,59	3,11	60,65	4,08	61,64	3,09	24,33	97,03
ENCDO845WKR	UBERON:0002046	81,02	4,19	79,80	5,43	81,06	4,17	23,22	97,87
ENCDO845WKR	UBERON:0002106	82,39	4,34	81,16	5,60	82,44	4,33	22,71	97,84
ENCDO845WKR	UBERON:0004264	86,11	4,80	84,81	6,12	86,15	4,78	21,87	97,99
ENCDO845WKR	UBERON:0008952	82,00	4,71	80,79	5,94	82,05	4,68	21,23	97,71

10.2.3 DMCs intra- e inter-individuales

Tabla 22. DMCs intra-individuales detectar por pares de muestras

Individuo	Muestra 1 (ref.)	Muestra 2	Total de DMCs	DMCs hipermetilados	DMCs hipometilados
ENCDO000AAW	EFO:0003042	CL:0000134	2387171 (8,81%)	1855872 (6,85%)	531299 (1,96%)
ENCDO058AAA	UBERON:0000945	UBERON:0001013	1635774 (6,04%)	424820 (1,57%)	1210954 (4,47%)
ENCDO058AAA	UBERON:0000945	UBERON:0001043	1398514 (5,16%)	560177 (2,07%)	838337 (3,09%)
ENCDO058AAA	UBERON:0000945	UBERON:0001159	1841718 (6,80%)	911741 (3,37%)	929977 (3,43%)
ENCDO058AAA	UBERON:0000945	UBERON:0001264	3361802 (12,41%)	2805821 (10,36%)	555981 (2,05%)
ENCDO058AAA	UBERON:0000945	UBERON:0002080	1590450 (5,87%)	599206 (2,21%)	991244 (3,66%)
ENCDO058AAA	UBERON:0000945	UBERON:0002084	1697535 (6,27%)	647186 (2,39%)	1050349 (3,88%)
ENCDO058AAA	UBERON:0000945	UBERON:0002106	1971144 (7,28%)	619822 (2,29%)	1351322 (4,99%)
ENCDO058AAA	UBERON:0000945	UBERON:0002108	1311540 (4,84%)	898370 (3,32%)	413170 (1,53%)
ENCDO058AAA	UBERON:0000945	UBERON:0002369	1919689 (7,09%)	825936 (3,05%)	1093753 (4,04%)
ENCDO058AAA	UBERON:0001013	UBERON:0001043	1130005 (4,17%)	773386 (2,86%)	356619 (1,32%)
ENCDO058AAA	UBERON:0001013	UBERON:0001159	2464956 (9,10%)	1619982 (5,98%)	844974 (3,12%)
ENCDO058AAA	UBERON:0001013	UBERON:0001264	4219645 (15,58%)	3666219 (13,53%)	553426 (2,04%)
ENCDO058AAA	UBERON:0001013	UBERON:0002080	781840 (2,89%)	513297 (1,89%)	268543 (0,99%)
ENCDO058AAA	UBERON:0001013	UBERON:0002084	955883 (3,53%)	610803 (2,25%)	345080 (1,27%)
ENCDO058AAA	UBERON:0001013	UBERON:0002106	298421 (1,10%)	132457 (0,49%)	165964 (0,61%)
ENCDO058AAA	UBERON:0001013	UBERON:0002369	1102229 (4,07%)	777383 (2,87%)	324846 (1,20%)
ENCDO058AAA	UBERON:0001043	UBERON:0001159	2140242 (7,90%)	1216240 (4,49%)	924002 (3,41%)
ENCDO058AAA	UBERON:0001043	UBERON:0002080	1302987 (4,81%)	591859 (2,18%)	711128 (2,63%)
ENCDO058AAA	UBERON:0001043	UBERON:0002084	1401026 (5,17%)	630082 (2,33%)	770944 (2,85%)
ENCDO058AAA	UBERON:0001043	UBERON:0002106	1613011 (5,95%)	592772 (2,19%)	1020239 (3,77%)
ENCDO058AAA	UBERON:0001043	UBERON:0002369	1466309 (5,41%)	718848 (2,65%)	747461 (2,76%)
ENCDO058AAA	UBERON:0001159	UBERON:0002080	2670104 (9,86%)	1198894 (4,43%)	1471210 (5,43%)
ENCDO058AAA	UBERON:0001159	UBERON:0002084	2694560 (9,95%)	1206715 (4,45%)	1487845 (5,49%)
ENCDO058AAA	UBERON:0001159	UBERON:0002106	2446221 (9,03%)	871905 (3,22%)	1574316 (5,81%)
ENCDO058AAA	UBERON:0001159	UBERON:0002369	2605267 (9,62%)	1200187 (4,43%)	1405080 (5,19%)
ENCDO058AAA	UBERON:0001264	UBERON:0001043	3709833 (13,70%)	621639 (2,29%)	3088194 (11,40%)
ENCDO058AAA	UBERON:0001264	UBERON:0001159	4120051 (15,21%)	881189 (3,25%)	3238862 (11,96%)
ENCDO058AAA	UBERON:0001264	UBERON:0002080	3622148 (13,37%)	616238 (2,27%)	3005910 (11,10%)
ENCDO058AAA	UBERON:0001264	UBERON:0002084	3611862 (13,33%)	626008 (2,31%)	2985854 (11,02%)
ENCDO058AAA	UBERON:0001264	UBERON:0002106	4104052 (15,15%)	671462 (2,48%)	3432590 (12,67%)
ENCDO058AAA	UBERON:0001264	UBERON:0002369	2737698 (10,11%)	457349 (1,69%)	2280349 (8,42%)
ENCDO058AAA	UBERON:0002078	UBERON:0000945	1572414 (5,80%)	880687 (3,25%)	691727 (2,55%)
ENCDO058AAA	UBERON:0002078	UBERON:0001013	760975 (2,81%)	177019 (0,65%)	583956 (2,16%)

ENCDO058AAA	UBERON:0002078	UBERON:0001043	1260480 (4,65%)	592298 (2,19%)	668182 (2,47%)
ENCDO058AAA	UBERON:0002078	UBERON:0001159	2755862 (10,17%)	1398634 (5,16%)	1357228 (5,01%)
ENCDO058AAA	UBERON:0002078	UBERON:0001264	3563187 (13,15%)	2847800 (10,51%)	715387 (2,64%)
ENCDO058AAA	UBERON:0002078	UBERON:0002080	34480 (0,13%)	8092 (0,03%)	26388 (0,10%)
ENCDO058AAA	UBERON:0002078	UBERON:0002084	80500 (0,30%)	19387 (0,07%)	61113 (0,23%)
ENCDO058AAA	UBERON:0002078	UBERON:0002106	1524779 (5,63%)	487038 (1,80%)	1037741 (3,83%)
ENCDO058AAA	UBERON:0002078	UBERON:0002108	2111805 (7,80%)	1426386 (5,27%)	685419 (2,53%)
ENCDO058AAA	UBERON:0002078	UBERON:0002369	1201669 (4,44%)	527565 (1,95%)	674104 (2,49%)
ENCDO058AAA	UBERON:0002080	UBERON:0002106	1452579 (5,36%)	576073 (2,13%)	876506 (3,24%)
ENCDO058AAA	UBERON:0002080	UBERON:0002369	1134119 (4,19%)	581259 (2,15%)	552860 (2,04%)
ENCDO058AAA	UBERON:0002084	UBERON:0002080	264 (0,00%)	149 (0,00%)	115 (0,00%)
ENCDO058AAA	UBERON:0002084	UBERON:0002106	1520980 (5,61%)	621284 (2,29%)	899696 (3,32%)
ENCDO058AAA	UBERON:0002084	UBERON:0002369	1330024 (4,91%)	688175 (2,54%)	641849 (2,37%)
ENCDO058AAA	UBERON:0002106	UBERON:0002369	1411938 (5,21%)	867852 (3,20%)	544086 (2,01%)
ENCDO058AAA	UBERON:0002108	UBERON:0001013	2187797 (8,08%)	444750 (1,64%)	1743047 (6,43%)
ENCDO058AAA	UBERON:0002108	UBERON:0001043	1789010 (6,60%)	475594 (1,76%)	1313416 (4,85%)
ENCDO058AAA	UBERON:0002108	UBERON:0001159	904914 (3,34%)	263456 (0,97%)	641458 (2,37%)
ENCDO058AAA	UBERON:0002108	UBERON:0001264	2714129 (10,02%)	1868008 (6,90%)	846121 (3,12%)
ENCDO058AAA	UBERON:0002108	UBERON:0002080	2082630 (7,69%)	576262 (2,13%)	1506368 (5,56%)
ENCDO058AAA	UBERON:0002108	UBERON:0002084	2188833 (8,08%)	621081 (2,29%)	1567752 (5,79%)
ENCDO058AAA	UBERON:0002108	UBERON:0002106	2160411 (7,98%)	511323 (1,89%)	1649088 (6,09%)
ENCDO058AAA	UBERON:0002108	UBERON:0002369	1987283 (7,34%)	581305 (2,15%)	1405978 (5,19%)
ENCDO132ASK	UBERON:0000059	UBERON:0002108	11 (0,00%)	9 (0,00%)	2 (0,00%)
ENCDO222AAA	CL:0000182	CL:0000192	934855 (3,45%)	418481 (1,54%)	516374 (1,91%)
ENCDO271OUW	UBERON:0000945	UBERON:0000992	1737548 (6,41%)	724226 (2,67%)	1013322 (3,74%)
ENCDO271OUW	UBERON:0000945	UBERON:0002046	1863639 (6,88%)	552363 (2,04%)	1311276 (4,84%)
ENCDO271OUW	UBERON:0000945	UBERON:0002106	1539850 (5,68%)	642539 (2,37%)	897311 (3,31%)
ENCDO271OUW	UBERON:0000945	UBERON:0004264	1856730 (6,85%)	706311 (2,61%)	1150419 (4,25%)
ENCDO271OUW	UBERON:0000945	UBERON:0008952	1400302 (5,17%)	473432 (1,75%)	926870 (3,42%)
ENCDO271OUW	UBERON:0000992	UBERON:0002046	1750014 (6,46%)	617156 (2,28%)	1132858 (4,18%)
ENCDO271OUW	UBERON:0000992	UBERON:0004264	1555617 (5,74%)	712717 (2,63%)	842900 (3,11%)
ENCDO271OUW	UBERON:0000992	UBERON:0008952	1690159 (6,24%)	725920 (2,68%)	964239 (3,56%)
ENCDO271OUW	UBERON:0002046	UBERON:0004264	1711756 (6,32%)	1041033 (3,84%)	670723 (2,48%)
ENCDO271OUW	UBERON:0002106	UBERON:0000992	1811248 (6,69%)	893420 (3,30%)	917828 (3,39%)
ENCDO271OUW	UBERON:0002106	UBERON:0002046	1624776 (6,00%)	587399 (2,17%)	1037377 (3,83%)
ENCDO271OUW	UBERON:0002106	UBERON:0004264	1613085 (5,95%)	735080 (2,71%)	878005 (3,24%)
ENCDO271OUW	UBERON:0002106	UBERON:0008952	809351 (2,99%)	332580 (1,23%)	476771 (1,76%)
ENCDO271OUW	UBERON:0008952	UBERON:0002046	1369884 (5,06%)	552976 (2,04%)	816908 (3,02%)
ENCDO271OUW	UBERON:0008952	UBERON:0004264	1555793 (5,74%)	841573 (3,11%)	714220 (2,64%)
ENCDO376MNT	UBERON:0001987	UBERON:0002370	5538548 (20,45%)	245968 (0,91%)	5292580 (19,54%)
ENCDO376MNT	UBERON:0002240	UBERON:0001987	5129965 (18,94%)	4619975 (17,06%)	509990 (1,88%)
ENCDO376MNT	UBERON:0002240	UBERON:0002370	1269971 (4,69%)	273637 (1,01%)	996334 (3,68%)

ENCDO424FTP	CL:0000222	CL:0000221	150657 (0,56%)	88317 (0,33%)	62340 (0,23%)
ENCDO424FTP	CL:0000223	CL:0000221	183345 (0,68%)	148761 (0,55%)	34584 (0,13%)
ENCDO424FTP	CL:0000223	CL:0000222	72707 (0,27%)	52420 (0,19%)	20287 (0,07%)
ENCDO424FTP	CL:0000223	EFO:0007089	111738 (0,41%)	91804 (0,34%)	19934 (0,07%)
ENCDO424FTP	EFO:0007089	CL:0000221	89757 (0,33%)	60390 (0,22%)	29367 (0,11%)
ENCDO424FTP	EFO:0007089	CL:0000222	86156 (0,32%)	36300 (0,13%)	49856 (0,18%)
ENCDO424HVB	UBERON:0000945	UBERON:0002106	1305278 (4,82%)	297293 (1,10%)	1007985 (3,72%)
ENCDO424HVB	UBERON:0000945	UBERON:0002108	550420 (2,03%)	281760 (1,04%)	268660 (0,99%)
ENCDO424HVB	UBERON:0000947	UBERON:0000945	1784950 (6,59%)	829820 (3,06%)	955130 (3,53%)
ENCDO424HVB	UBERON:0000947	UBERON:0000992	1876109 (6,93%)	966157 (3,57%)	909952 (3,36%)
ENCDO424HVB	UBERON:0000947	UBERON:0001013	1156597 (4,27%)	258120 (0,95%)	898477 (3,32%)
ENCDO424HVB	UBERON:0000947	UBERON:0001043	1156489 (4,27%)	353791 (1,31%)	802698 (2,96%)
ENCDO424HVB	UBERON:0000947	UBERON:0001264	1960424 (7,24%)	1062601 (3,92%)	897823 (3,31%)
ENCDO424HVB	UBERON:0000947	UBERON:0002048	1720373 (6,35%)	349690 (1,29%)	1370683 (5,06%)
ENCDO424HVB	UBERON:0000947	UBERON:0002106	1802545 (6,65%)	482724 (1,78%)	1319821 (4,87%)
ENCDO424HVB	UBERON:0000947	UBERON:0002108	1972176 (7,28%)	996866 (3,68%)	975310 (3,60%)
ENCDO424HVB	UBERON:0000947	UBERON:0002369	1268163 (4,68%)	313766 (1,16%)	954397 (3,52%)
ENCDO424HVB	UBERON:0000947	UBERON:0008450	1342820 (4,96%)	658896 (2,43%)	683924 (2,52%)
ENCDO424HVB	UBERON:0000992	UBERON:0000945	2228271 (8,23%)	1034448 (3,82%)	1193823 (4,41%)
ENCDO424HVB	UBERON:0000992	UBERON:0001013	1585753 (5,85%)	391164 (1,44%)	1194589 (4,41%)
ENCDO424HVB	UBERON:0000992	UBERON:0001043	1688596 (6,23%)	580872 (2,14%)	1107724 (4,09%)
ENCDO424HVB	UBERON:0000992	UBERON:0001264	2425854 (8,96%)	1326398 (4,90%)	1099456 (4,06%)
ENCDO424HVB	UBERON:0000992	UBERON:0002048	2187666 (8,08%)	437540 (1,62%)	1750126 (6,46%)
ENCDO424HVB	UBERON:0000992	UBERON:0002106	2341090 (8,64%)	645044 (2,38%)	1696046 (6,26%)
ENCDO424HVB	UBERON:0000992	UBERON:0002108	2428525 (8,97%)	1236392 (4,56%)	1192133 (4,40%)
ENCDO424HVB	UBERON:0000992	UBERON:0002369	1535928 (5,67%)	350664 (1,29%)	1185264 (4,38%)
ENCDO424HVB	UBERON:0000992	UBERON:0008450	2081558 (7,68%)	1056342 (3,90%)	1025216 (3,78%)
ENCDO424HVB	UBERON:0001013	UBERON:0000945	869900 (3,21%)	668407 (2,47%)	201493 (0,74%)
ENCDO424HVB	UBERON:0001013	UBERON:0002048	123599 (0,46%)	34398 (0,13%)	89201 (0,33%)
ENCDO424HVB	UBERON:0001013	UBERON:0002106	92280 (0,34%)	24634 (0,09%)	67646 (0,25%)
ENCDO424HVB	UBERON:0001013	UBERON:0002108	801490 (2,96%)	640123 (2,36%)	161367 (0,60%)
ENCDO424HVB	UBERON:0001043	UBERON:0000945	380620 (1,41%)	282137 (1,04%)	98483 (0,36%)
ENCDO424HVB	UBERON:0001043	UBERON:0001013	124427 (0,46%)	39823 (0,15%)	84604 (0,31%)
ENCDO424HVB	UBERON:0001043	UBERON:0001264	495083 (1,83%)	377427 (1,39%)	117656 (0,43%)
ENCDO424HVB	UBERON:0001043	UBERON:0002048	411856 (1,52%)	77608 (0,29%)	334248 (1,23%)
ENCDO424HVB	UBERON:0001043	UBERON:0002106	505092 (1,86%)	145134 (0,54%)	359958 (1,33%)
ENCDO424HVB	UBERON:0001043	UBERON:0002108	196998 (0,73%)	151111 (0,56%)	45887 (0,17%)
ENCDO424HVB	UBERON:0001264	UBERON:0000945	802076 (2,96%)	309305 (1,14%)	492771 (1,82%)
ENCDO424HVB	UBERON:0001264	UBERON:0001013	1329717 (4,91%)	281149 (1,04%)	1048568 (3,87%)

ENCDO424HVB	UBERON:0001264	UBERON:0002048	1968053 (7,27%)	301694 (1,11%)	1666359 (6,15%)
ENCDO424HVB	UBERON:0001264	UBERON:0002106	1805574 (6,67%)	397432 (1,47%)	1408142 (5,20%)
ENCDO424HVB	UBERON:0001264	UBERON:0002108	713949 (2,64%)	337662 (1,25%)	376287 (1,39%)
ENCDO424HVB	UBERON:0002048	UBERON:0000945	1153210 (4,26%)	986076 (3,64%)	167134 (0,62%)
ENCDO424HVB	UBERON:0002048	UBERON:0002106	273941 (1,01%)	136581 (0,50%)	137360 (0,51%)
ENCDO424HVB	UBERON:0002048	UBERON:0002108	1283909 (4,74%)	1110835 (4,10%)	173074 (0,64%)
ENCDO424HVB	UBERON:0002106	UBERON:0002108	979410 (3,62%)	805185 (2,97%)	174225 (0,64%)
ENCDO424HVB	UBERON:0002369	UBERON:0000945	1169370 (4,32%)	852451 (3,15%)	316919 (1,17%)
ENCDO424HVB	UBERON:0002369	UBERON:0001013	200522 (0,74%)	66948 (0,25%)	133574 (0,49%)
ENCDO424HVB	UBERON:0002369	UBERON:0001043	323747 (1,20%)	198694 (0,73%)	125053 (0,46%)
ENCDO424HVB	UBERON:0002369	UBERON:0001264	1122474 (4,14%)	896244 (3,31%)	226230 (0,84%)
ENCDO424HVB	UBERON:0002369	UBERON:0002048	681199 (2,51%)	207726 (0,77%)	473473 (1,75%)
ENCDO424HVB	UBERON:0002369	UBERON:0002106	525089 (1,94%)	206822 (0,76%)	318267 (1,17%)
ENCDO424HVB	UBERON:0002369	UBERON:0002108	1014008 (3,74%)	775835 (2,86%)	238173 (0,88%)
ENCDO424HVB	UBERON:0002369	UBERON:0008450	1104143 (4,08%)	855657 (3,16%)	248486 (0,92%)
ENCDO424HVB	UBERON:0008450	UBERON:0000945	1750880 (6,46%)	802993 (2,96%)	947887 (3,50%)
ENCDO424HVB	UBERON:0008450	UBERON:0001013	814003 (3,01%)	129060 (0,48%)	684943 (2,53%)
ENCDO424HVB	UBERON:0008450	UBERON:0001043	850707 (3,14%)	241282 (0,89%)	609425 (2,25%)
ENCDO424HVB	UBERON:0008450	UBERON:0001264	1830631 (6,76%)	1000323 (3,69%)	830308 (3,07%)
ENCDO424HVB	UBERON:0008450	UBERON:0002048	1683797 (6,22%)	268457 (0,99%)	1415340 (5,22%)
ENCDO424HVB	UBERON:0008450	UBERON:0002106	1528105 (5,64%)	326195 (1,20%)	1201910 (4,44%)
ENCDO424HVB	UBERON:0008450	UBERON:0002108	1614487 (5,96%)	791706 (2,92%)	822781 (3,04%)
ENCDO451RUA	UBERON:0000945	UBERON:0000473	1575915 (5,82%)	731155 (2,70%)	844760 (3,12%)
ENCDO451RUA	UBERON:0000945	UBERON:0001150	1984608 (7,33%)	1483347 (5,48%)	501261 (1,85%)
ENCDO451RUA	UBERON:0000945	UBERON:0001323	2063051 (7,62%)	820066 (3,03%)	1242985 (4,59%)
ENCDO451RUA	UBERON:0000945	UBERON:0002367	2190840 (8,09%)	998842 (3,69%)	1191998 (4,40%)
ENCDO451RUA	UBERON:0001150	UBERON:0000473	2228288 (8,23%)	541954 (2,00%)	1686334 (6,23%)
ENCDO451RUA	UBERON:0001150	UBERON:0001323	3130927 (11,56%)	728894 (2,69%)	2402033 (8,87%)
ENCDO451RUA	UBERON:0001323	UBERON:0000473	1374698 (5,07%)	794083 (2,93%)	580615 (2,14%)
ENCDO451RUA	UBERON:0002367	UBERON:0000473	1813961 (6,70%)	875785 (3,23%)	938176 (3,46%)
ENCDO451RUA	UBERON:0002367	UBERON:0001150	3099370 (11,44%)	2251767 (8,31%)	847603 (3,13%)
ENCDO451RUA	UBERON:0002367	UBERON:0001323	2087892 (7,71%)	936889 (3,46%)	1151003 (4,25%)
ENCDO451RUA	UBERON:0002369	UBERON:0000473	1095746 (4,05%)	458643 (1,69%)	637103 (2,35%)
ENCDO451RUA	UBERON:0002369	UBERON:0000945	1506202 (5,56%)	699278 (2,58%)	806924 (2,98%)
ENCDO451RUA	UBERON:0002369	UBERON:0001150	1497821 (5,53%)	1095675 (4,04%)	402146 (1,48%)
ENCDO451RUA	UBERON:0002369	UBERON:0001323	1883162 (6,95%)	679020 (2,51%)	1204142 (4,45%)
ENCDO451RUA	UBERON:0002369	UBERON:0002367	2309103 (8,52%)	1002865 (3,70%)	1306238 (4,82%)
ENCDO478OMA	UBERON:0000945	UBERON:0001013	252320 (0,93%)	191944 (0,71%)	60376 (0,22%)
ENCDO478OMA	UBERON:0000945	UBERON:0002048	668513 (2,47%)	215633 (0,80%)	452880 (1,67%)
ENCDO478OMA	UBERON:0000945	UBERON:0002080	641656 (2,37%)	263762 (0,97%)	377894 (1,40%)
ENCDO478OMA	UBERON:0000945	UBERON:0002084	934627 (3,45%)	298933 (1,10%)	635694 (2,35%)
ENCDO478OMA	UBERON:0000945	UBERON:0002106	874716 (3,23%)	151069 (0,56%)	723647 (2,67%)

ENCDO478OMA	UBERON:0000945	UBERON:0002108	140372 (0,52%)	16935 (0,06%)	123437 (0,46%)
ENCDO478OMA	UBERON:0000945	UBERON:0002370	2515783 (9,29%)	270581 (1,00%)	2245202 (8,29%)
ENCDO478OMA	UBERON:0001013	UBERON:0002048	1415506 (5,23%)	299273 (1,10%)	1116233 (4,12%)
ENCDO478OMA	UBERON:0001013	UBERON:0002080	963174 (3,56%)	258967 (0,96%)	704207 (2,60%)
ENCDO478OMA	UBERON:0001013	UBERON:0002084	1385538 (5,11%)	296231 (1,09%)	1089307 (4,02%)
ENCDO478OMA	UBERON:0001013	UBERON:0002106	1618324 (5,97%)	221816 (0,82%)	1396508 (5,16%)
ENCDO478OMA	UBERON:0001013	UBERON:0002108	910519 (3,36%)	147487 (0,54%)	763032 (2,82%)
ENCDO478OMA	UBERON:0001013	UBERON:0002370	3589663 (13,25%)	303105 (1,12%)	3286558 (12,13%)
ENCDO478OMA	UBERON:0001159	UBERON:0000945	615197 (2,27%)	564427 (2,08%)	50770 (0,19%)
ENCDO478OMA	UBERON:0001159	UBERON:0001013	1679281 (6,20%)	1528207 (5,64%)	151074 (0,56%)
ENCDO478OMA	UBERON:0001159	UBERON:0002048	313576 (1,16%)	243068 (0,90%)	70508 (0,26%)
ENCDO478OMA	UBERON:0001159	UBERON:0002080	1198724 (4,43%)	943780 (3,48%)	254944 (0,94%)
ENCDO478OMA	UBERON:0001159	UBERON:0002084	1191726 (4,40%)	884560 (3,27%)	307166 (1,13%)
ENCDO478OMA	UBERON:0001159	UBERON:0002106	157424 (0,58%)	63203 (0,23%)	94221 (0,35%)
ENCDO478OMA	UBERON:0001159	UBERON:0002108	219092 (0,81%)	190397 (0,70%)	28695 (0,11%)
ENCDO478OMA	UBERON:0001159	UBERON:0002370	1040448 (3,84%)	219948 (0,81%)	820500 (3,03%)
ENCDO478OMA	UBERON:0002080	UBERON:0002048	873789 (3,23%)	326907 (1,21%)	546882 (2,02%)
ENCDO478OMA	UBERON:0002084	UBERON:0002048	981369 (3,62%)	416368 (1,54%)	565001 (2,09%)
ENCDO478OMA	UBERON:0002084	UBERON:0002080	52 (0,00%)	48 (0,00%)	4 (0,00%)
ENCDO478OMA	UBERON:0002084	UBERON:0002106	998085 (3,68%)	299400 (1,11%)	698685 (2,58%)
ENCDO478OMA	UBERON:0002084	UBERON:0002108	992315 (3,66%)	485270 (1,79%)	507045 (1,87%)
ENCDO478OMA	UBERON:0002084	UBERON:0002370	2283701 (8,43%)	355664 (1,31%)	1928037 (7,12%)
ENCDO478OMA	UBERON:0002106	UBERON:0002048	76410 (0,28%)	63891 (0,24%)	12519 (0,05%)
ENCDO478OMA	UBERON:0002106	UBERON:0002080	927630 (3,42%)	692792 (2,56%)	234838 (0,87%)
ENCDO478OMA	UBERON:0002106	UBERON:0002108	660379 (2,44%)	492555 (1,82%)	167824 (0,62%)
ENCDO478OMA	UBERON:0002106	UBERON:0002370	431528 (1,59%)	59816 (0,22%)	371712 (1,37%)
ENCDO478OMA	UBERON:0002108	UBERON:0002048	608106 (2,24%)	286814 (1,06%)	321292 (1,19%)
ENCDO478OMA	UBERON:0002108	UBERON:0002080	849463 (3,14%)	515289 (1,90%)	334174 (1,23%)
ENCDO478OMA	UBERON:0002370	UBERON:0002048	733858 (2,71%)	652469 (2,41%)	81389 (0,30%)
ENCDO478OMA	UBERON:0002370	UBERON:0002080	2453114 (9,06%)	2117798 (7,82%)	335316 (1,24%)
ENCDO478OMA	UBERON:0002370	UBERON:0002108	2167923 (8,00%)	1876024 (6,93%)	291899 (1,08%)
ENCDO478OMA	UBERON:0008450	UBERON:0000945	749520 (2,77%)	369486 (1,36%)	380034 (1,40%)
ENCDO478OMA	UBERON:0008450	UBERON:0001013	975701 (3,60%)	638624 (2,36%)	337077 (1,24%)
ENCDO478OMA	UBERON:0008450	UBERON:0001159	1293047 (4,77%)	206495 (0,76%)	1086552 (4,01%)
ENCDO478OMA	UBERON:0008450	UBERON:0002048	1012853 (3,74%)	303154 (1,12%)	709699 (2,62%)
ENCDO478OMA	UBERON:0008450	UBERON:0002080	526102 (1,94%)	208417 (0,77%)	317685 (1,17%)
ENCDO478OMA	UBERON:0008450	UBERON:0002084	827898 (3,06%)	249398 (0,92%)	578500 (2,14%)
ENCDO478OMA	UBERON:0008450	UBERON:0002106	1078419 (3,98%)	214951 (0,79%)	863468 (3,19%)
ENCDO478OMA	UBERON:0008450	UBERON:0002108	944045 (3,49%)	290957 (1,07%)	653088 (2,41%)

ENCDO478OMA	UBERON:0008450	UBERON:0002370	2726746 (10,07%)	320721 (1,18%)	2406025 (8,88%)
ENCDO661BYS	CL:0000084	CL:0001054	1370482 (5,06%)	478534 (1,77%)	891948 (3,29%)
ENCDO661BYS	CL:0000236	CL:0000084	570192 (2,10%)	340999 (1,26%)	229193 (0,85%)
ENCDO661BYS	CL:0000236	CL:0001054	481847 (1,78%)	296650 (1,10%)	185197 (0,68%)
ENCDO661BYS	CL:0000623	CL:0000084	153834 (0,57%)	89808 (0,33%)	64026 (0,24%)
ENCDO661BYS	CL:0000623	CL:0000236	152692 (0,56%)	80763 (0,30%)	71929 (0,27%)
ENCDO661BYS	CL:0000623	CL:0001054	444790 (1,64%)	275945 (1,02%)	168845 (0,62%)
ENCDO793LXB	UBERON:0001323	UBERON:0000992	1732934 (6,40%)	1089680 (4,02%)	643254 (2,37%)
ENCDO793LXB	UBERON:0001323	UBERON:0002369	1542316 (5,69%)	866588 (3,20%)	675728 (2,49%)
ENCDO793LXB	UBERON:0002369	UBERON:0000992	1801990 (6,65%)	1047617 (3,87%)	754373 (2,78%)
ENCDO845WKR	UBERON:0000473	UBERON:0001150	2539332 (9,37%)	2162653 (7,98%)	376679 (1,39%)
ENCDO845WKR	UBERON:0000473	UBERON:0002046	1572290 (5,80%)	562770 (2,08%)	1009520 (3,73%)
ENCDO845WKR	UBERON:0000473	UBERON:0002106	2171977 (8,02%)	933281 (3,45%)	1238696 (4,57%)
ENCDO845WKR	UBERON:0000473	UBERON:0004264	2069786 (7,64%)	1088754 (4,02%)	981032 (3,62%)
ENCDO845WKR	UBERON:0000473	UBERON:0008952	1644543 (6,07%)	790394 (2,92%)	854149 (3,15%)
ENCDO845WKR	UBERON:0001150	UBERON:0002046	2241378 (8,27%)	219853 (0,81%)	2021525 (7,46%)
ENCDO845WKR	UBERON:0001150	UBERON:0002106	1965213 (7,25%)	231581 (0,85%)	1733632 (6,40%)
ENCDO845WKR	UBERON:0001150	UBERON:0008952	1409868 (5,20%)	134470 (0,50%)	1275398 (4,71%)
ENCDO845WKR	UBERON:0002106	UBERON:0002046	1215341 (4,49%)	552818 (2,04%)	662523 (2,45%)
ENCDO845WKR	UBERON:0002106	UBERON:0008952	656180 (2,42%)	409266 (1,51%)	246914 (0,91%)
ENCDO845WKR	UBERON:0004264	UBERON:0001150	1985171 (7,33%)	1642769 (6,06%)	342402 (1,26%)
ENCDO845WKR	UBERON:0004264	UBERON:0002046	1293932 (4,78%)	418299 (1,54%)	875633 (3,23%)
ENCDO845WKR	UBERON:0004264	UBERON:0002106	1530426 (5,65%)	558001 (2,06%)	972425 (3,59%)
ENCDO845WKR	UBERON:0004264	UBERON:0008952	991689 (3,66%)	425214 (1,57%)	566475 (2,09%)
ENCDO845WKR	UBERON:0008952	UBERON:0002046	752044 (2,78%)	266805 (0,98%)	485239 (1,79%)

Tabla 23. DMCs inter-individuales detectar por pares de muestras

Muestra	Individuo 1 (ref.)	Individuo 2	Total de DMCs	DMCs hipermetilados	DMCs hipometilados
UBERON:0000473	ENCDO845WKR	ENCDO451RUA	503336 (1,86%)	301093 (1,11%)	202243 (0,75%)
UBERON:0000945	ENCDO058AAA	ENCDO318GKQ	949626 (3,51%)	416064 (1,54%)	533562 (1,97%)
UBERON:0000945	ENCDO058AAA	ENCDO424HVB	386092 (1,43%)	210530 (0,78%)	175562 (0,65%)
UBERON:0000945	ENCDO058AAA	ENCDO451RUA	1344908 (4,96%)	811002 (2,99%)	533906 (1,97%)
UBERON:0000945	ENCDO058AAA	ENCDO478OMA	508570 (1,88%)	243161 (0,90%)	265409 (0,98%)
UBERON:0000945	ENCDO271OUW	ENCDO058AAA	1016285 (3,75%)	400450 (1,48%)	615835 (2,27%)
UBERON:0000945	ENCDO271OUW	ENCDO318GKQ	1136257 (4,19%)	434682 (1,60%)	701575 (2,59%)
UBERON:0000945	ENCDO271OUW	ENCDO424HVB	699200 (2,58%)	353382 (1,30%)	345818 (1,28%)
UBERON:0000945	ENCDO271OUW	ENCDO451RUA	1343258 (4,96%)	707546 (2,61%)	635712 (2,35%)
UBERON:0000945	ENCDO271OUW	ENCDO478OMA	649261 (2,40%)	264393 (0,98%)	384868 (1,42%)
UBERON:0000945	ENCDO318GKQ	ENCDO424HVB	739914 (2,73%)	489746 (1,81%)	250168 (0,92%)
UBERON:0000945	ENCDO318GKQ	ENCDO451RUA	1434966 (5,30%)	885265 (3,27%)	549701 (2,03%)
UBERON:0000945	ENCDO451RUA	ENCDO424HVB	881676 (3,25%)	449094 (1,66%)	432582 (1,60%)
UBERON:0000945	ENCDO478OMA	ENCDO318GKQ	470754 (1,74%)	215821 (0,80%)	254933 (0,94%)
UBERON:0000945	ENCDO478OMA	ENCDO424HVB	376207 (1,39%)	214928 (0,79%)	161279 (0,60%)
UBERON:0000945	ENCDO478OMA	ENCDO451RUA	737049 (2,72%)	454048 (1,68%)	283001 (1,04%)
UBERON:0000992	ENCDO271OUW	ENCDO793LXB	716978 (2,65%)	459294 (1,70%)	257684 (0,95%)
UBERON:0000992	ENCDO424HVB	ENCDO271OUW	833094 (3,08%)	276954 (1,02%)	556140 (2,05%)
UBERON:0000992	ENCDO424HVB	ENCDO793LXB	571377 (2,11%)	279226 (1,03%)	292151 (1,08%)
UBERON:0001013	ENCDO058AAA	ENCDO424HVB	403337 (1,49%)	222800 (0,82%)	180537 (0,67%)
UBERON:0001013	ENCDO478OMA	ENCDO058AAA	1944349 (7,18%)	465724 (1,72%)	1478625 (5,46%)
UBERON:0001013	ENCDO478OMA	ENCDO424HVB	1157211 (4,27%)	319338 (1,18%)	837873 (3,09%)
UBERON:0001043	ENCDO424HVB	ENCDO058AAA	351705 (1,30%)	151390 (0,56%)	200315 (0,74%)
UBERON:0001150	ENCDO845WKR	ENCDO451RUA	267502 (0,99%)	129310 (0,48%)	138192 (0,51%)
UBERON:0001159	ENCDO478OMA	ENCDO058AAA	2321150 (8,57%)	1760436 (6,50%)	560714 (2,07%)
UBERON:0001264	ENCDO424HVB	ENCDO058AAA	404131 (1,49%)	258440 (0,95%)	145691 (0,54%)
UBERON:0001323	ENCDO793LXB	ENCDO451RUA	724007 (2,67%)	371038 (1,37%)	352969 (1,30%)
UBERON:0002046	ENCDO271OUW	ENCDO845WKR	373467 (1,38%)	177471 (0,66%)	195996 (0,72%)
UBERON:0002048	ENCDO424HVB	ENCDO478OMA	652379 (2,41%)	355706 (1,31%)	296673 (1,10%)
UBERON:0002080	ENCDO478OMA	ENCDO058AAA	481421 (1,78%)	221289 (0,82%)	260132 (0,96%)
UBERON:0002084	ENCDO478OMA	ENCDO058AAA	565089 (2,09%)	274694 (1,01%)	290395 (1,07%)
UBERON:0002106	ENCDO271OUW	ENCDO058AAA	788129 (2,91%)	300335 (1,11%)	487794 (1,80%)
UBERON:0002106	ENCDO271OUW	ENCDO424HVB	489187 (1,81%)	188295 (0,70%)	300892 (1,11%)
UBERON:0002106	ENCDO271OUW	ENCDO478OMA	808510 (2,98%)	217933 (0,80%)	590577 (2,18%)
UBERON:0002106	ENCDO271OUW	ENCDO845WKR	778762 (2,87%)	276139 (1,02%)	502623 (1,86%)
UBERON:0002106	ENCDO424HVB	ENCDO058AAA	402423 (1,49%)	197713 (0,73%)	204710 (0,76%)
UBERON:0002106	ENCDO478OMA	ENCDO058AAA	603579 (2,23%)	355221 (1,31%)	248358 (0,92%)

UBERON:0002106	ENCDO478OMA	ENCDO424HVB	415448 (1,53%)	225825 (0,83%)	189623 (0,70%)
UBERON:0002106	ENCDO478OMA	ENCDO845WKR	540893 (2,00%)	323463 (1,19%)	217430 (0,80%)
UBERON:0002106	ENCDO845WKR	ENCDO058AAA	530612 (1,96%)	254030 (0,94%)	276582 (1,02%)
UBERON:0002106	ENCDO845WKR	ENCDO424HVB	408223 (1,51%)	205954 (0,76%)	202269 (0,75%)
UBERON:0002108	ENCDO058AAA	ENCDO132ASK	680128 (2,51%)	129186 (0,48%)	550942 (2,03%)
UBERON:0002108	ENCDO058AAA	ENCDO424HVB	135677 (0,50%)	62919 (0,23%)	72758 (0,27%)
UBERON:0002108	ENCDO058AAA	ENCDO478OMA	1519612 (5,61%)	303512 (1,12%)	1216100 (4,49%)
UBERON:0002108	ENCDO132ASK	ENCDO424HVB	514147 (1,90%)	398190 (1,47%)	115957 (0,43%)
UBERON:0002108	ENCDO478OMA	ENCDO132ASK	500097 (1,85%)	254637 (0,94%)	245460 (0,91%)
UBERON:0002108	ENCDO478OMA	ENCDO424HVB	765194 (2,82%)	569129 (2,10%)	196065 (0,72%)
UBERON:0002369	ENCDO424HVB	ENCDO058AAA	385006 (1,42%)	203510 (0,75%)	181496 (0,67%)
UBERON:0002369	ENCDO424HVB	ENCDO451RUA	434352 (1,60%)	282309 (1,04%)	152043 (0,56%)
UBERON:0002369	ENCDO424HVB	ENCDO793LXB	507985 (1,88%)	310118 (1,14%)	197867 (0,73%)
UBERON:0002369	ENCDO424HVB	ENCDO900GTZ	335604 (1,24%)	147401 (0,54%)	188203 (0,69%)
UBERON:0002369	ENCDO451RUA	ENCDO058AAA	532071 (1,96%)	199637 (0,74%)	332434 (1,23%)
UBERON:0002369	ENCDO451RUA	ENCDO793LXB	661748 (2,44%)	280441 (1,04%)	381307 (1,41%)
UBERON:0002369	ENCDO451RUA	ENCDO900GTZ	496659 (1,83%)	148771 (0,55%)	347888 (1,28%)
UBERON:0002369	ENCDO793LXB	ENCDO058AAA	689832 (2,55%)	292637 (1,08%)	397195 (1,47%)
UBERON:0002369	ENCDO793LXB	ENCDO900GTZ	531607 (1,96%)	190491 (0,70%)	341116 (1,26%)
UBERON:0002369	ENCDO900GTZ	ENCDO058AAA	575735 (2,13%)	325331 (1,20%)	250404 (0,92%)
UBERON:0002370	ENCDO478OMA	ENCDO376MNT	467707 (1,73%)	258096 (0,95%)	209611 (0,77%)
UBERON:0004264	ENCDO845WKR	ENCDO271OUW	804388 (2,97%)	402160 (1,48%)	402228 (1,48%)
UBERON:0008450	ENCDO478OMA	ENCDO424HVB	350982 (1,30%)	196042 (0,72%)	154940 (0,57%)
UBERON:0008952	ENCDO271OUW	ENCDO845WKR	727009 (2,68%)	350666 (1,29%)	376343 (1,39%)

10.2.4 Riqueza en DMCs y semáforos CpG de TFBSs

Tabla 24. Riqueza en intra-DMCs de los sitios de unión a factores de transcripción

FA: activador; FR: represor; M-M: Methyl-Minus; M-P: Methyl-Plus; R.O.: riqueza observada; R.E.: riqueza esperada; R: enriquecido; A: aleatorio; P: empobrecido

Nombre	FA	FR	M-M	M-P	R.O.	R.E. (media)	R. E. (SD)	Z-score	Valor-P	Estado
Arntl	N	N	S	N	1,039	0,999	0,000	102,612	0,001	R
Atoh1	S	N	N	N	1,361	1,001	0,004	91,858	0,001	R
BHLHE41	N	S	N	N	1,079	0,998	0,002	37,906	0,001	R
CREB1	N	N	S	N	1,241	0,999	0,005	48,027	0,001	R
CREB3	S	N	S	N	1,244	1,001	0,002	108,518	0,001	R
CREB3L1	N	S	N	N	1,170	0,999	0,001	244,238	0,001	R
CTCFL	S	N	N	N	0,405	1,003	0,004	-155,755	0,001	P
Dlx3	S	N	N	N	1,636	1,002	0,016	38,862	0,001	R
Dlx4	N	S	N	N	1,580	0,992	0,018	33,310	0,001	R
DUX4	S	N	N	N	1,550	1,005	0,005	105,651	0,001	R
E2F4	N	N	S	N	0,524	1,002	0,003	-180,009	0,001	P
E2F7	N	S	N	N	1,389	1,007	0,020	19,538	0,001	R
ELF1	N	N	S	N	1,097	0,997	0,004	24,906	0,001	R
ELF3	N	N	S	N	1,174	0,994	0,005	39,531	0,001	R
ELF4	N	N	S	N	0,829	1,000	0,004	-43,439	0,001	P
ELF5	N	N	S	N	1,151	1,001	0,002	72,926	0,001	R
ELK1	N	N	S	N	1,120	0,999	0,003	34,433	0,001	R
ELK3	N	S	S	N	1,196	1,000	0,001	164,517	0,001	R
EN1	N	S	N	N	1,589	0,987	0,011	53,653	0,001	R
ERF	N	S	N	N	1,078	1,000	0,003	24,684	0,001	R
ERG	N	N	S	N	1,168	1,000	0,002	97,082	0,001	R
ESR1	N	N	N	S	0,786	0,999	0,011	-18,781	0,001	P
Esrrg	S	N	N	N	0,925	0,999	0,002	-38,316	0,001	P
ETV1	N	N	S	N	1,128	0,999	0,002	70,737	0,001	R
ETV2	N	N	S	N	1,261	0,999	0,001	226,077	0,001	R
ETV3	N	N	S	N	1,228	0,998	0,001	181,257	0,001	R
ETV4	N	N	S	N	1,155	1,000	0,001	201,982	0,001	R
ETV5	N	N	S	N	1,065	0,999	0,002	35,832	0,001	R
EVX1	N	S	N	N	1,594	0,993	0,013	46,525	0,001	R
FEV	N	N	S	N	1,176	0,999	0,001	271,529	0,001	R
FLI1	N	N	S	N	1,170	1,000	0,001	283,308	0,001	R
FOXG1	N	S	N	N	1,754	0,993	0,011	69,178	0,001	R

FOXL1	S	N	N	N	1,617	0,998	0,012	50,151	0,001	R
FOXP3	N	N	S	N	1,713	0,994	0,006	128,439	0,001	R
Gabpa	N	N	S	N	0,865	1,000	0,003	-52,462	0,001	P
GBX2	S	N	N	N	1,590	0,994	0,008	70,895	0,001	R
GCM2	S	N	N	N	0,703	1,002	0,002	-182,148	0,001	P
Gfi1b	N	S	N	N	1,124	0,999	0,002	69,347	0,001	R
Gmeb1	N	N	S	N	0,968	0,996	0,003	-8,205	0,001	P
Hes1	N	N	S	N	0,733	1,000	0,001	-269,486	0,001	P
Hes2	N	N	S	N	1,031	0,999	0,001	28,111	0,001	R
HES5	N	N	S	N	0,701	1,004	0,004	-68,780	0,001	P
HES7	N	N	S	N	0,735	1,006	0,007	-40,298	0,001	P
HEY2	N	S	S	N	0,862	1,000	0,003	-52,839	0,001	P
HINFP	N	S	N	N	0,555	0,999	0,001	-543,599	0,001	P
HOXA10	N	N	N	S	1,680	0,998	0,004	160,660	0,001	R
Hoxa11	N	S	N	S	1,761	0,999	0,005	158,713	0,001	R
HOXA13	N	N	N	S	1,602	0,997	0,007	80,876	0,001	R
HOXB13	N	S	N	S	1,752	0,995	0,008	96,293	0,001	R
HOXC10	N	S	N	S	1,845	0,998	0,008	108,111	0,001	R
HOXC11	N	N	N	S	1,799	0,997	0,005	177,430	0,001	R
HOXC12	N	N	N	S	1,781	0,993	0,005	151,641	0,001	R
HOXC13	N	S	N	S	1,740	0,998	0,004	210,868	0,001	R
Hoxc9	N	S	N	S	1,756	1,003	0,006	122,704	0,001	R
HOXD11	N	N	N	S	1,826	0,997	0,005	160,719	0,001	R
HOXD13	N	N	N	S	1,745	0,993	0,004	207,788	0,001	R
Hoxd9	S	N	N	S	1,741	0,998	0,008	91,626	0,001	R
Klf12	N	S	N	N	0,589	0,999	0,003	-143,540	0,001	P
KLF14	N	S	N	N	0,499	0,999	0,002	-303,340	0,001	P
KLF16	N	S	N	N	0,497	1,000	0,002	-303,743	0,001	P
LHX2	S	N	N	N	1,687	0,993	0,010	71,208	0,001	R
Lhx4	S	N	N	N	1,826	1,000	0,016	52,263	0,001	R
MEIS2	N	N	N	S	1,145	1,001	0,006	22,664	0,001	R
MEIS3	N	N	N	S	1,167	0,999	0,008	19,855	0,001	R
MLX	N	N	S	N	1,057	1,000	0,002	24,320	0,001	R
MLXIPL	S	N	N	N	1,067	0,998	0,003	20,114	0,001	R
MYCN	N	N	S	N	0,818	1,000	0,002	-83,553	0,001	P
NFATC1	N	N	N	S	1,180	0,999	0,002	81,289	0,001	R
NFATC3	S	N	N	S	1,186	0,999	0,002	105,246	0,001	R
NR2F1	N	N	N	S	0,824	1,001	0,004	-48,576	0,001	P
Nr2f6	N	N	N	S	1,246	0,980	0,014	18,819	0,001	R
NR3C2	N	N	N	S	1,262	0,998	0,007	37,267	0,001	R
ONECUT2	N	N	S	N	1,822	1,000	0,007	123,013	0,001	R
OTX1	S	N	N	N	1,520	0,992	0,001	414,102	0,001	R

PAX9	N	N	N	S	0,916	0,994	0,002	-32,594	0,001	P
PBX1	N	N	N	S	1,609	1,006	0,013	45,401	0,001	R
PITX3	S	N	N	N	1,205	0,995	0,006	33,525	0,001	R
PKNX1	N	N	N	S	1,097	1,013	0,014	6,129	0,001	R
PKNX2	N	N	N	S	1,093	1,008	0,006	14,102	0,001	R
POU2F2	N	N	N	S	1,638	0,996	0,002	404,343	0,001	R
POU3F1	N	N	N	S	1,609	0,996	0,007	84,574	0,001	R
POU3F2	N	N	N	S	1,629	0,996	0,004	167,478	0,001	R
POU3F4	N	N	N	S	1,417	0,995	0,008	50,737	0,001	R
POU4F2	S	N	N	N	1,940	0,999	0,016	58,040	0,001	R
POU5F1	N	N	N	S	1,554	1,000	0,007	79,707	0,001	R
RARA	N	N	N	S	0,677	0,995	0,008	-39,988	0,001	P
Rarb	N	N	N	S	0,604	1,014	0,009	-46,118	0,001	P
Rarg	N	N	N	S	0,509	1,003	0,008	-62,650	0,001	P
RFX5	N	N	N	S	1,139	0,995	0,019	7,457	0,001	R
RORB	N	N	N	S	1,265	0,998	0,009	29,300	0,001	R
RUNX3	N	N	S	N	0,982	1,000	0,005	-3,855	0,001	P
RXRB	N	N	N	S	1,286	0,982	0,019	15,916	0,001	R
RXRG	N	N	N	S	1,264	0,991	0,014	19,642	0,001	R
SCRT1	N	N	N	S	1,090	1,002	0,006	13,596	0,001	R
SCRT2	N	N	N	S	1,116	1,001	0,004	26,934	0,001	R
SPDEF	N	N	S	N	1,090	1,001	0,002	50,471	0,001	R
SPIB	N	N	S	N	1,288	1,001	0,002	149,409	0,001	R
SPIC	S	N	N	N	1,528	1,000	0,008	67,704	0,001	R
TBX2	N	S	N	N	0,993	1,002	0,005	-1,854	0,001	A
TBX20	N	N	N	S	1,398	0,997	0,010	41,307	0,001	R
Tcf12	N	S	N	N	0,781	1,001	0,001	-168,043	0,001	P
Tcfl5	N	S	N	N	0,668	1,002	0,002	-151,475	0,001	P
TEAD4	N	S	N	N	1,321	1,002	0,005	61,126	0,001	R
TGIF1	N	S	N	N	1,066	1,014	0,017	2,953	0,001	R
ZBED1	N	N	S	N	1,132	1,002	0,003	40,050	0,001	R
ZBTB7A	N	S	N	N	0,688	1,000	0,003	-110,463	0,001	P
ZBTB7B	N	N	S	N	0,671	1,000	0,005	-70,124	0,001	P
ZIC1	S	N	N	N	0,529	1,000	0,004	-116,926	0,001	P

Tabla 25. Riqueza en inter-DMCs de los sitios de unión a factores de transcripción

FA: activador; FR: represor; M-M: Methyl-Minus; M-P: Methyl-Plus; R.O.: riqueza observada; R.E.: riqueza esperada; R: enriquecido; A: aleatorio; P: empobrecido

Nombre	FA	FR	M-M	M-P	R.O.	R.E. (media)	R. E. (SD)	Z-score	Valor-P	Estado
Arntl	N	N	S	N	0,961	1,006	0,004	-11,770	0,001	P
Atoh1	S	N	N	N	1,208	0,974	0,020	11,720	0,001	R
BHLHE41	N	S	N	N	1,032	1,003	0,010	3,017	0,001	R
CREB1	N	N	S	N	1,130	1,009	0,016	7,550	0,001	R
CREB3	S	N	S	N	1,098	1,005	0,016	5,922	0,001	R
CREB3L1	N	S	N	N	1,152	1,019	0,019	6,975	0,001	R
CTCFL	S	N	N	N	0,580	1,005	0,007	-58,179	0,001	P
Dlx3	S	N	N	N	1,292	0,972	0,049	6,588	0,001	R
Dlx4	N	S	N	N	1,353	0,977	0,068	5,574	0,001	R
DUX4	S	N	N	N	1,178	1,012	0,021	7,864	0,001	R
E2F4	N	N	S	N	0,707	0,998	0,005	-54,030	0,001	P
E2F7	N	S	N	N	1,258	0,982	0,091	3,022	0,001	R
ELF1	N	N	S	N	1,165	1,001	0,015	11,293	0,001	R
ELF3	N	N	S	N	1,206	1,005	0,008	25,045	0,001	R
ELF4	N	N	S	N	1,019	1,003	0,010	1,574	0,001	A
ELF5	N	N	S	N	1,192	1,002	0,006	29,220	0,001	R
ELK1	N	N	S	N	1,224	1,005	0,007	29,525	0,001	R
ELK3	N	S	S	N	1,247	1,005	0,009	26,778	0,001	R
EN1	N	S	N	N	1,355	0,982	0,058	6,426	0,001	R
ERF	N	S	N	N	1,150	0,996	0,003	44,054	0,001	R
ERG	N	N	S	N	1,209	1,004	0,006	36,353	0,001	R
ESR1	N	N	N	S	1,018	1,048	0,028	-1,075	0,16	A
Esrrg	S	N	N	N	0,943	1,000	0,009	-6,072	0,001	P
ETV1	N	N	S	N	1,208	0,999	0,002	115,728	0,001	R
ETV2	N	N	S	N	1,224	0,996	0,006	40,760	0,001	R
ETV3	N	N	S	N	1,194	1,002	0,006	33,185	0,001	R
ETV4	N	N	S	N	1,234	0,999	0,007	32,623	0,001	R
ETV5	N	N	S	N	1,172	0,998	0,007	26,658	0,001	R
EVX1	N	S	N	N	1,266	1,002	0,027	9,886	0,001	R
FEV	N	N	S	N	1,218	1,003	0,005	39,441	0,001	R
FLI1	N	N	S	N	1,213	1,004	0,006	35,326	0,001	R
FOXG1	N	S	N	N	1,208	1,003	0,032	6,390	0,001	R
FOXL1	S	N	N	N	1,382	1,036	0,029	11,850	0,001	R
FOXP3	N	N	S	N	1,197	1,002	0,023	8,361	0,001	R
Gabpa	N	N	S	N	1,013	1,004	0,003	2,476	0,001	A
GBX2	S	N	N	N	1,175	1,009	0,035	4,670	0,001	R

GCM2	S	N	N	N	0,882	1,000	0,005	-23,365	0,001	P
Gfi1b	N	S	N	N	1,062	0,996	0,006	11,557	0,001	R
Gmeb1	N	N	S	N	1,069	1,007	0,008	7,368	0,001	R
Hes1	N	N	S	N	0,876	1,003	0,006	-22,997	0,001	P
Hes2	N	N	S	N	0,938	0,999	0,005	-13,009	0,001	P
HES5	N	N	S	N	0,972	1,000	0,013	-2,215	0,001	A
HES7	N	N	S	N	0,977	1,007	0,017	-1,793	0,001	A
HEY2	N	S	S	N	0,900	1,001	0,004	-25,711	0,001	P
HINFP	N	S	N	N	0,785	1,007	0,008	-26,747	0,001	P
HOXA10	N	N	N	S	1,151	1,001	0,019	8,076	0,001	R
Hoxa11	N	S	N	S	1,129	1,003	0,006	21,208	0,001	R
HOXA13	N	N	N	S	1,069	0,998	0,019	3,821	0,001	R
HOXB13	N	S	N	S	1,164	0,996	0,027	6,285	0,001	R
HOXC10	N	S	N	S	1,168	1,002	0,012	14,345	0,001	R
HOXC11	N	N	N	S	1,138	1,004	0,005	26,236	0,001	R
HOXC12	N	N	N	S	1,157	1,004	0,009	16,734	0,001	R
HOXC13	N	S	N	S	1,156	0,997	0,012	12,947	0,001	R
Hoxc9	N	S	N	S	1,235	1,006	0,028	8,140	0,001	R
HOXD11	N	N	N	S	1,076	1,002	0,018	4,144	0,001	R
HOXD13	N	N	N	S	1,148	1,002	0,021	7,108	0,001	R
Hoxd9	S	N	N	S	1,168	1,014	0,025	6,119	0,001	R
Klf12	N	S	N	N	0,841	1,009	0,007	-23,744	0,001	P
KLF14	N	S	N	N	0,715	1,006	0,004	-65,107	0,001	P
KLF16	N	S	N	N	0,737	1,002	0,005	-57,302	0,001	P
LHX2	S	N	N	N	1,259	1,025	0,038	6,160	0,001	R
Lhx4	S	N	N	N	1,281	1,010	0,084	3,224	0,001	R
MEIS2	N	N	N	S	1,135	0,999	0,025	5,448	0,001	R
MEIS3	N	N	N	S	1,155	0,992	0,021	7,686	0,001	R
MLX	N	N	S	N	0,953	0,995	0,009	-4,692	0,001	P
MLXIPL	S	N	N	N	0,958	1,004	0,007	-6,509	0,001	P
MYCN	N	N	S	N	0,861	1,001	0,009	-14,784	0,001	P
NFATC1	N	N	N	S	1,034	1,001	0,004	7,596	0,001	R
NFATC3	S	N	N	S	1,036	1,002	0,005	7,136	0,001	R
NR2F1	N	N	N	S	0,907	0,996	0,009	-9,813	0,001	P
Nr2f6	N	N	N	S	1,086	0,994	0,045	2,067	0,001	A
NR3C2	N	N	N	S	1,121	1,012	0,015	7,122	0,001	R
ONECUT2	N	N	S	N	1,011	0,987	0,028	0,830	0,25	A
OTX1	S	N	N	N	1,153	1,003	0,013	11,419	0,001	R
PAX9	N	N	N	S	1,032	0,976	0,019	2,973	0,001	R

PBX1	N	N	N	S	1,219	1,006	0,023	9,223	0,001	R
PITX3	S	N	N	N	1,171	1,019	0,016	9,736	0,001	R
PKNX1	N	N	N	S	1,116	0,973	0,034	4,141	0,001	R
PKNX2	N	N	N	S	1,111	0,964	0,053	2,766	0,001	R
POU2F2	N	N	N	S	1,124	1,010	0,024	4,677	0,001	R
POU3F1	N	N	N	S	1,157	1,014	0,027	5,240	0,001	R
POU3F2	N	N	N	S	1,178	1,005	0,021	8,274	0,001	R
POU3F4	N	N	N	S	1,178	1,006	0,027	6,343	0,001	R
POU4F2	S	N	N	N	1,205	1,051	0,069	2,238	0,001	A
POU5F1	N	N	N	S	1,162	1,005	0,008	19,165	0,001	R
RARA	N	N	N	S	0,912	0,975	0,033	-1,906	0,001	A
Rarb	N	N	N	S	0,886	0,997	0,033	-3,339	0,001	P
Rarg	N	N	N	S	0,803	0,983	0,016	-11,541	0,001	P
RFX5	N	N	N	S	1,015	0,999	0,056	0,286	0,5	A
RORB	N	N	N	S	1,134	0,993	0,018	7,956	0,001	R
RUNX3	N	N	S	N	1,200	1,001	0,009	23,048	0,001	R
RXRB	N	N	N	S	1,114	1,019	0,070	1,345	0,001	A
RXRG	N	N	N	S	1,143	0,983	0,042	3,850	0,001	R
SCRT1	N	N	N	S	1,161	0,994	0,043	3,886	0,001	R
SCRT2	N	N	N	S	1,073	1,011	0,026	2,420	0,001	A
SPDEF	N	N	S	N	1,137	0,999	0,012	11,071	0,001	R
SPIB	N	N	S	N	1,262	1,002	0,009	29,213	0,001	R
SPIC	S	N	N	N	1,262	0,996	0,016	16,698	0,001	R
TBX2	N	S	N	N	0,989	1,003	0,012	-1,192	0,001	A
TBX20	N	N	N	S	1,134	1,000	0,021	6,481	0,001	R
Tcf12	N	S	N	N	0,981	1,009	0,018	-1,542	0,001	A
Tcfl5	N	S	N	N	0,734	0,999	0,008	-31,476	0,001	P
TEAD4	N	S	N	N	1,172	0,978	0,018	10,951	0,001	R
TGIF1	N	S	N	N	1,156	0,987	0,037	4,583	0,001	R
ZBED1	N	N	S	N	0,999	1,002	0,039	-0,087	0,5	A
ZBTB7A	N	S	N	N	0,882	1,005	0,009	-14,033	0,001	P
ZBTB7B	N	N	S	N	0,829	1,001	0,011	-15,154	0,001	P
ZIC1	S	N	N	N	0,671	1,007	0,006	-58,034	0,001	P

Tabla 26. Riqueza en CpG-TLs rojos de los sitios de unión a factores de transcripción

FA: activador; FR: represor; M-M: Methyl-Minus; M-P: Methyl-Plus; R.O.: riqueza observada; R.E.: riqueza esperada; R: enriquecido; A: aleatorio; P: empobrecido

Nombre	FA	FR	M-M	M-P	R.O.	R.E. (media)	R. E. (SD)	Z-score	Valor-P	Estado
Arntl	N	N	S	N	0,949	0,990	0,014	-2,870	0,001	P
Atoh1	S	N	N	N	0,932	0,998	0,031	-2,137	0,001	A
BHLHE41	N	S	N	N	0,997	0,996	0,018	0,073	0,5	A
CREB1	N	N	S	N	0,970	1,003	0,041	-0,791	0,25	A
CREB3	S	N	S	N	0,999	1,005	0,025	-0,281	0,25	A
CREB3L1	N	S	N	N	1,037	0,993	0,046	0,939	0,25	A
CTCFL	S	N	N	N	1,085	0,995	0,028	3,223	0,001	R
Dlx3	S	N	N	N	0,851	1,107	0,066	-3,865	0,001	P
Dlx4	N	S	N	N	1,124	1,124	0,085	0,000	0,5	A
DUX4	S	N	N	N	0,833	1,022	0,070	-2,710	0,001	P
E2F4	N	N	S	N	1,065	0,997	0,012	5,723	0,001	R
E2F7	N	S	N	N	1,178	1,073	0,088	1,195	0,25	A
ELF1	N	N	S	N	1,280	1,007	0,019	14,253	0,001	R
ELF3	N	N	S	N	1,179	0,994	0,035	5,327	0,001	R
ELF4	N	N	S	N	0,935	0,995	0,015	-4,090	0,001	P
ELF5	N	N	S	N	1,300	1,007	0,007	41,212	0,001	R
ELK1	N	N	S	N	1,285	1,010	0,018	15,214	0,001	R
ELK3	N	S	S	N	1,274	1,007	0,014	19,567	0,001	R
EN1	N	S	N	N	0,975	1,091	0,048	-2,428	0,001	A
ERF	N	S	N	N	1,202	1,016	0,021	8,688	0,001	R
ERG	N	N	S	N	1,293	0,996	0,007	41,477	0,001	R
ESR1	N	N	N	S	0,864	0,940	0,101	-0,757	0,25	A
Esrrg	S	N	N	N	0,875	0,994	0,022	-5,334	0,001	P
ETV1	N	N	S	N	1,297	0,999	0,010	30,507	0,001	R
ETV2	N	N	S	N	1,306	0,986	0,019	16,799	0,001	R
ETV3	N	N	S	N	1,244	1,013	0,027	8,439	0,001	R
ETV4	N	N	S	N	1,287	0,998	0,006	46,531	0,001	R
ETV5	N	N	S	N	1,276	1,001	0,016	17,435	0,001	R
EVX1	N	S	N	N	0,980	1,056	0,041	-1,845	0,001	A
FEV	N	N	S	N	1,269	0,999	0,005	49,942	0,001	R
FLI1	N	N	S	N	1,293	0,996	0,008	36,750	0,001	R
FOXG1	N	S	N	N	0,906	1,003	0,038	-2,550	0,001	A
FOXL1	S	N	N	N	0,901	1,022	0,056	-2,175	0,001	A
FOXP3	N	N	S	N	0,846	0,988	0,020	-7,120	0,001	P
Gabpa	N	N	S	N	1,287	1,007	0,005	55,071	0,001	R

GBX2	S	N	N	N	0,962	1,042	0,083	-0,970	0,25	A
GCM2	S	N	N	N	1,061	0,988	0,016	4,459	0,001	R
Gfi1b	N	S	N	N	0,907	1,007	0,042	-2,394	0,001	A
Gmeb1	N	N	S	N	1,111	1,004	0,019	5,640	0,001	R
Hes1	N	N	S	N	1,059	0,997	0,014	4,338	0,001	R
Hes2	N	N	S	N	0,990	0,995	0,019	-0,275	0,25	A
HES5	N	N	S	N	0,933	0,992	0,010	-5,682	0,001	P
HES7	N	N	S	N	0,934	0,970	0,032	-1,115	0,13	A
HEY2	N	S	S	N	1,043	1,007	0,013	2,830	0,001	R
HINFP	N	S	N	N	1,074	1,008	0,019	3,402	0,001	R
HOXA10	N	N	N	S	0,774	1,016	0,076	-3,197	0,001	P
Hoxa11	N	S	N	S	0,811	1,002	0,016	-11,987	0,001	P
HOXA13	N	N	N	S	0,845	1,037	0,091	-2,118	0,001	A
HOXB13	N	S	N	S	0,745	0,956	0,092	-2,292	0,001	A
HOXC10	N	S	N	S	0,845	0,993	0,023	-6,362	0,001	P
HOXC11	N	N	N	S	0,804	1,006	0,010	-19,518	0,001	P
HOXC12	N	N	N	S	0,831	0,993	0,014	-11,584	0,001	P
HOXC13	N	S	N	S	0,854	0,995	0,031	-4,591	0,001	P
Hoxc9	N	S	N	S	0,945	1,025	0,034	-2,384	0,001	A
HOXD11	N	N	N	S	0,780	0,998	0,017	-13,137	0,001	P
HOXD13	N	N	N	S	0,737	1,023	0,104	-2,750	0,001	P
Hoxd9	S	N	N	S	0,773	1,036	0,069	-3,823	0,001	P
Klf12	N	S	N	N	0,886	0,993	0,013	-8,457	0,001	P
KLF14	N	S	N	N	0,853	0,986	0,007	-17,849	0,001	P
KLF16	N	S	N	N	0,929	0,993	0,015	-4,404	0,001	P
LHX2	S	N	N	N	0,988	1,000	0,065	-0,192	0,25	A
Lhx4	S	N	N	N	0,957	1,119	0,211	-0,768	0,25	A
MEIS2	N	N	N	S	1,133	0,961	0,027	6,449	0,001	R
MEIS3	N	N	N	S	1,119	1,012	0,042	2,554	0,001	A
MLX	N	N	S	N	0,843	0,985	0,008	-18,434	0,001	P
MLXIPL	S	N	N	N	0,854	0,997	0,011	-12,768	0,001	P
MYCN	N	N	S	N	0,984	0,988	0,020	-0,186	0,5	A
NFATC1	N	N	N	S	1,074	0,989	0,014	6,147	0,001	R
NFATC3	S	N	N	S	1,081	0,984	0,015	6,626	0,001	R
NR2F1	N	N	N	S	0,852	0,993	0,020	-7,179	0,001	P
Nr2f6	N	N	N	S	1,231	0,994	0,142	1,666	0,001	A
NR3C2	N	N	N	S	1,014	0,980	0,056	0,611	0,25	A
ONECUT2	N	N	S	N	0,681	1,005	0,022	-14,678	0,001	P
OTX1	S	N	N	N	0,894	1,032	0,019	-7,269	0,001	P
PAX9	N	N	N	S	0,969	0,981	0,033	-0,376	0,38	A
PBX1	N	N	N	S	0,930	1,055	0,032	-3,851	0,001	P
PITX3	S	N	N	N	0,991	1,004	0,021	-0,613	0,5	A

PKNX1	N	N	N	S	1,094	1,026	0,071	0,966	0,25	A
PKNX2	N	N	N	S	0,965	0,989	0,068	-0,351	0,25	A
POU2F2	N	N	N	S	0,769	0,998	0,048	-4,822	0,001	P
POU3F1	N	N	N	S	0,805	1,005	0,017	-12,124	0,001	P
POU3F2	N	N	N	S	0,819	1,027	0,018	-11,406	0,001	P
POU3F4	N	N	N	S	0,842	1,002	0,041	-3,918	0,001	P
POU4F2	S	N	N	N	0,885	0,980	0,142	-0,666	0,25	A
POU5F1	N	N	N	S	0,780	0,997	0,027	-8,151	0,001	P
RARA	N	N	N	S	0,472	1,015	0,084	-6,458	0,001	P
Rarb	N	N	N	S	0,616	1,030	0,078	-5,312	0,001	P
Rarg	N	N	N	S	0,520	1,006	0,026	-18,876	0,001	P
RFX5	N	N	N	S	0,965	0,991	0,095	-0,265	0,25	A
RORB	N	N	N	S	0,995	1,016	0,063	-0,340	0,5	A
RUNX3	N	N	S	N	1,232	0,996	0,013	17,735	0,001	R
RXRB	N	N	N	S	1,311	0,987	0,129	2,519	0,001	A
RXRG	N	N	N	S	1,102	1,035	0,140	0,480	0,25	A
SCRT1	N	N	N	S	1,227	0,986	0,087	2,763	0,001	R
SCRT2	N	N	N	S	1,214	1,016	0,052	3,792	0,001	R
SPDEF	N	N	S	N	1,124	0,999	0,009	13,616	0,001	R
SPIB	N	N	S	N	1,200	1,012	0,017	11,288	0,001	R
SPIC	S	N	N	N	1,096	1,016	0,036	2,244	0,001	A
TBX2	N	S	N	N	0,826	0,985	0,008	-19,029	0,001	P
TBX20	N	N	N	S	0,948	0,990	0,028	-1,513	0,001	A
Tcf12	N	S	N	N	1,207	1,011	0,020	9,585	0,001	R
Tcfl5	N	S	N	N	1,067	0,997	0,011	6,368	0,001	R
TEAD4	N	S	N	N	0,965	1,035	0,033	-2,114	0,001	A
TGIF1	N	S	N	N	1,007	1,007	0,061	0,000	0,37	A
ZBED1	N	N	S	N	0,802	1,014	0,021	-10,323	0,001	P
ZBTB7A	N	S	N	N	1,121	1,004	0,005	25,065	0,001	R
ZBTB7B	N	N	S	N	0,972	0,991	0,022	-0,856	0,25	A
ZIC1	S	N	N	N	1,277	1,003	0,018	14,897	0,001	R

Tabla 27. Riqueza en CpG-TLs verdes de los sitios de unión a factores de transcripción

FA: activador; FR: represor; M-M: Methyl-Minus; M-P: Methyl-Plus; R.O.: riqueza observada; R.E.: riqueza esperada; R: enriquecido; A: aleatorio; P: empobrecido

Nombre	FA	FR	M-M	M-P	R.O.	R.E. (media)	R. E. (SD)	Z-score	Valor-P	Estado
Arntl	N	N	S	N	1,462	0,996	0,020	22,880	0,001	R
Atoh1	S	N	N	N	0,948	1,023	0,038	-1,983	0,001	A
BHLHE41	N	S	N	N	1,371	1,005	0,022	16,501	0,001	R
CREB1	N	N	S	N	1,153	1,002	0,031	4,927	0,001	R
CREB3	S	N	S	N	1,444	1,022	0,047	8,960	0,001	R
CREB3L1	N	S	N	N	1,444	1,025	0,040	10,469	0,001	R
CTCFL	S	N	N	N	0,925	0,993	0,017	-4,078	0,001	P
Dlx3	S	N	N	N	0,786	1,043	0,069	-3,689	0,001	P
Dlx4	N	S	N	N	0,954	1,042	0,134	-0,651	0,25	A
DUX4	S	N	N	N	1,162	0,985	0,014	12,972	0,001	R
E2F4	N	N	S	N	0,848	0,994	0,012	-12,585	0,001	P
E2F7	N	S	N	N	0,796	1,092	0,100	-2,959	0,001	P
ELF1	N	N	S	N	1,166	0,987	0,015	11,886	0,001	R
ELF3	N	N	S	N	1,156	0,980	0,017	10,298	0,001	R
ELF4	N	N	S	N	0,842	0,990	0,014	-10,463	0,001	P
ELF5	N	N	S	N	1,120	1,001	0,011	10,753	0,001	R
ELK1	N	N	S	N	1,020	1,011	0,016	0,617	0,25	A
ELK3	N	S	S	N	1,027	1,004	0,016	1,482	0,25	A
EN1	N	S	N	N	1,096	0,991	0,089	1,170	0,25	A
ERF	N	S	N	N	0,992	1,009	0,017	-0,971	0,001	A
ERG	N	N	S	N	1,059	0,993	0,019	3,508	0,001	R
ESR1	N	N	N	S	1,036	0,969	0,068	0,992	0,25	A
Esrrg	S	N	N	N	1,171	1,002	0,017	9,769	0,001	R
ETV1	N	N	S	N	1,014	0,998	0,021	0,749	0,25	A
ETV2	N	N	S	N	1,110	0,992	0,008	14,156	0,001	R
ETV3	N	N	S	N	1,057	1,008	0,020	2,479	0,001	A
ETV4	N	N	S	N	1,016	1,007	0,019	0,461	0,25	A
ETV5	N	N	S	N	0,980	1,013	0,021	-1,559	0,001	A
EVX1	N	S	N	N	1,012	1,015	0,016	-0,159	0,5	A
FEV	N	N	S	N	1,042	1,005	0,018	1,971	0,001	A
FLI1	N	N	S	N	1,056	0,991	0,022	3,012	0,001	R
FOXG1	N	S	N	N	1,110	0,971	0,030	4,617	0,001	R
FOXL1	S	N	N	N	0,751	0,961	0,094	-2,242	0,001	A
FOXP3	N	N	S	N	1,172	0,938	0,027	8,707	0,001	R
Gabpa	N	N	S	N	1,059	1,010	0,010	5,019	0,001	R
GBX2	S	N	N	N	0,842	1,015	0,044	-3,957	0,001	P

GCM2	S	N	N	N	0,856	1,007	0,005	-27,779	0,001	P
Gfi1b	N	S	N	N	0,931	0,994	0,011	-5,912	0,001	P
Gmeb1	N	N	S	N	0,974	0,996	0,020	-1,071	0,25	A
Hes1	N	N	S	N	1,210	0,998	0,002	86,113	0,001	R
Hes2	N	N	S	N	1,379	1,000	0,008	49,007	0,001	R
HES5	N	N	S	N	1,071	1,005	0,022	2,965	0,001	R
HES7	N	N	S	N	1,191	0,993	0,024	8,296	0,001	R
HEY2	N	S	S	N	1,286	0,999	0,010	27,960	0,001	R
HINFP	N	S	N	N	0,933	1,004	0,019	-3,705	0,001	P
HOXA10	N	N	N	S	0,932	1,021	0,020	-4,495	0,001	P
Hoxa11	N	S	N	S	1,098	1,014	0,020	4,274	0,001	R
HOXA13	N	N	N	S	1,033	0,992	0,023	1,767	0,001	A
HOXB13	N	S	N	S	1,099	1,012	0,061	1,410	0,001	A
HOXC10	N	S	N	S	1,111	0,996	0,022	5,189	0,001	R
HOXC11	N	N	N	S	1,099	1,019	0,014	5,780	0,001	R
HOXC12	N	N	N	S	1,051	1,004	0,011	4,387	0,001	R
HOXC13	N	S	N	S	1,181	1,003	0,021	8,401	0,001	R
Hoxc9	N	S	N	S	1,062	1,015	0,029	1,645	0,001	A
HOXD11	N	N	N	S	1,145	1,001	0,008	17,097	0,001	R
HOXD13	N	N	N	S	1,077	1,020	0,040	1,457	0,001	A
Hoxd9	S	N	N	S	0,866	1,032	0,031	-5,345	0,001	P
Klf12	N	S	N	N	0,986	1,010	0,006	-4,187	0,001	P
KLF14	N	S	N	N	0,784	1,001	0,009	-23,829	0,001	P
KLF16	N	S	N	N	0,799	1,010	0,015	-14,521	0,001	P
LHX2	S	N	N	N	1,013	1,001	0,027	0,436	0,37	A
Lhx4	S	N	N	N	0,974	1,051	0,048	-1,612	0,001	A
MEIS2	N	N	N	S	1,183	0,992	0,038	5,043	0,001	R
MEIS3	N	N	N	S	1,214	1,008	0,018	11,606	0,001	R
MLX	N	N	S	N	1,314	1,009	0,020	15,578	0,001	R
MLXIPL	S	N	N	N	1,423	1,001	0,019	22,348	0,001	R
MYCN	N	N	S	N	1,496	1,004	0,011	42,760	0,001	R
NFATC1	N	N	N	S	1,043	0,999	0,019	2,338	0,001	A
NFATC3	S	N	N	S	1,050	0,998	0,023	2,282	0,001	A
NR2F1	N	N	N	S	1,123	1,016	0,012	8,899	0,001	R
Nr2f6	N	N	N	S	1,438	1,056	0,065	5,854	0,001	R
NR3C2	N	N	N	S	1,194	0,929	0,028	9,608	0,001	R
ONECUT2	N	N	S	N	1,002	1,016	0,031	-0,444	0,25	A
OTX1	S	N	N	N	0,867	1,011	0,028	-5,184	0,001	P
PAX9	N	N	N	S	0,858	0,972	0,043	-2,644	0,001	P

PBX1	N	N	N	S	0,910	1,083	0,025	-6,813	0,001	P
PITX3	S	N	N	N	0,927	0,997	0,037	-1,864	0,001	A
PKNX1	N	N	N	S	1,056	1,031	0,091	0,275	0,5	A
PKNX2	N	N	N	S	1,098	1,060	0,085	0,441	0,5	A
POU2F2	N	N	N	S	1,131	1,019	0,025	4,492	0,001	R
POU3F1	N	N	N	S	0,938	1,020	0,023	-3,513	0,001	P
POU3F2	N	N	N	S	0,944	1,006	0,034	-1,814	0,001	A
POU3F4	N	N	N	S	0,792	1,019	0,029	-7,774	0,001	P
POU4F2	S	N	N	N	0,997	1,074	0,181	-0,423	0,25	A
POU5F1	N	N	N	S	1,133	0,982	0,016	9,210	0,001	R
RARA	N	N	N	S	1,263	0,985	0,057	4,840	0,001	R
Rarb	N	N	N	S	1,128	0,985	0,037	3,900	0,001	R
Rarg	N	N	N	S	0,807	1,003	0,013	-15,222	0,001	P
RFX5	N	N	N	S	1,181	1,011	0,081	2,091	0,001	A
RORB	N	N	N	S	1,264	1,007	0,048	5,352	0,001	R
RUNX3	N	N	S	N	0,847	0,996	0,034	-4,376	0,001	P
RXRB	N	N	N	S	1,480	1,025	0,063	7,251	0,001	R
RXRG	N	N	N	S	1,363	1,010	0,018	19,448	0,001	R
SCRT1	N	N	N	S	1,193	0,976	0,015	14,205	0,001	R
SCRT2	N	N	N	S	1,282	0,984	0,026	11,527	0,001	R
SPDEF	N	N	S	N	1,070	0,999	0,024	3,007	0,001	R
SPIB	N	N	S	N	0,998	1,000	0,014	-0,140	0,25	A
SPIC	S	N	N	N	1,177	0,983	0,029	6,754	0,001	R
TBX2	N	S	N	N	0,897	1,003	0,006	-17,207	0,001	P
TBX20	N	N	N	S	1,193	1,038	0,024	6,410	0,001	R
Tcf12	N	S	N	N	0,965	1,000	0,004	-9,700	0,001	P
Tcfl5	N	S	N	N	1,316	0,996	0,011	28,005	0,001	R
TEAD4	N	S	N	N	1,008	0,993	0,025	0,596	0,25	A
TGIF1	N	S	N	N	0,990	1,042	0,093	-0,549	0,5	A
ZBED1	N	N	S	N	0,946	1,027	0,029	-2,811	0,001	P
ZBTB7A	N	S	N	N	0,952	0,999	0,002	-26,252	0,001	P
ZBTB7B	N	N	S	N	0,960	1,005	0,018	-2,535	0,001	A
ZIC1	S	N	N	N	1,020	1,002	0,010	1,751	0,001	A

Glosario

BS	*Secuenciación de ADN tratado con bisulfito*
RRBS	*Reduced representation BS*
WGBS	*Whole-Genome BS*
CpG	*Dinucleótido citosina-guanina*
CGI	*Isla CpG*
DMC	*CpG diferencialmente metilado*
Inter-DMC	*DMC inter-individual*
Intra-DMC	*DMC intra-individual*
CpG-TL	*Semáforo CpG*
TSS	*Sitio de inicio de la transcripción*
TES	*Sitio de fin de la transcripción*
TF	*Factor de transcripción*
TFBS	*Sitio de unión a factor de transcripción*
Methyl-Plus	*TF con preferencia por sitios metilados*
Methyl-Minus	*TF con preferencia por sitios no-metilados*
DHS	*Sitio de hipersensibilidad a la ADNasa I*
CTCFBS	*Sitio de unión a CTCF*
SNP	*Polimorfismo de un solo nucleótido*
Indel	*Inserción o deleción*
M	*Sitio CpG metilado*
I	*Sitio CpG con metilación intermedia*
U	*Sitio CpG no-metilado*

Referencias

Adams, D., Altucci, L., Antonarakis, S. E., Ballesteros, J., Beck, S., Bird, A., … Willcocks, S. (2012). BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotechnology*. https://doi.org/10.1038/nbt.2153

Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Ech, M., … Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, *46*(W1), W537–W544. https://doi.org/10.1093/nar/gky379

Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., … Zhu, J. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204. https://doi.org/10.1038/nature24277

Amoreira, C., Hindermann, W., & Grunau, C. (2003). An improved version of the DNA methylation database (MethDB). *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkg093

Andrews, S. (2018). FastQC: a quality control tool for high throughput sequence data. Retrieved from https://www.bioinformatics.babraham.ac.uk/projects/fastqc

Antequera, F. (2003). Structure, function and evolution of CpG island promoters. *Cellular and Molecular Life Sciences*. https://doi.org/10.1007/s00018-003-3088-6

Aran, D., Sabato, S., & Hellman, A. (2013). DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biology*. https://doi.org/10.1186/gb-2013-14-3-r21

Aran, D., Toperoff, G., Rosenberg, M., & Hellman, A. (2011). Replication timing-related and gene body-specific methylation of active human genes. *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddq513

Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y., & Shirakawa, M. (2008). Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature*, *455*(7214), 818–821. https://doi.org/10.1038/nature07249

Atlasi, Y., & Stunnenberg, H. G. (2017). The interplay of epigenetic marks during stem cell differentiation and development. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg.2017.57

Bachman, M., Uribe-Lewis, S., Yang, X., Williams, M., Murrell, A., & Balasubramanian, S. (2014). 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nature Chemistry*. https://doi.org/10.1038/nchem.2064

Bahar Halpern, K., Vana, T., & Walker, M. D. (2014). Paradoxical role of DNA methylation in activation of FoxA2 gene expression during endoderm development. *Journal of Biological Chemistry*, *289*(34), 23882–23892. https://doi.org/10.1074/jbc.M114.573469

Baker, M. (2016). 1,500 Scientists Lift the Lid on Reproducibility. *Nature*, *533*(7604), 452–454.

https://doi.org/10.1038/533452a

Banks, E., Lunter, G., Albers, C. A., Durbin, R., Danecek, P., Auton, A., … DePristo, M. A. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Barau, J., Teissandier, A., Zamudio, N., Roy, S., Nalesso, V., Herault, Y., … Bourc'his, D. (2016). The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science (New York, N.Y.)*, *354*(6314), 909–912. https://doi.org/10.1126/science.aah5143

Barturen, G., Oliver, J. L., & Hackenberg, M. (2017). Error correction in methylation profiling from NGS bisulfite protocols. In M. Elloumi (Ed.), *Algorithms for Next-Generation Sequencing Data: Techniques, Approaches, and Applications* (pp. 167–183). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-59826-0_8

Barturen, G., Rueda, A., Oliver, J. L., & Hackenberg, M. (2014). MethylExtract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000Research*. https://doi.org/10.12688/f1000research.2-217.v2

Beck, C. R., Collier, P., Macfarlane, C., Malig, M., Kidd, J. M., Eichler, E. E., … Moran, J. V. (2010). LINE-1 retrotransposition activity in human genomes. *Cell*. https://doi.org/10.1016/j.cell.2010.05.021

Bell, A. C., & Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*. https://doi.org/10.1038/35013100

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*. https://doi.org/10.1017/CBO9781107415324.004

Bennett, E. A., Keller, H., Mills, R. E., Schmidt, S., Moran, J. V., Weichenrieder, O., & Devine, S. E. (2008). Active Alu retrotransposons in the human genome. *Genome Research*. https://doi.org/10.1101/gr.081737.108

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*. https://doi.org/10.1038/nature11247

Bestor, T. H., & Ingram, V. M. (1983). Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.80.18.5559

Bird, A. P. (1978). Use of restriction enzymes to study eukaryotic DNA methylation. II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. *Journal of Molecular Biology*. https://doi.org/10.1016/0022-2836(78)90243-7

Bird, A. P. (1984). DNA methylation versus gene expression. *Journal of Embryology and Experimental Morphology*, *83 Suppl*, 31–40. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6598198

Bird, A. P. (1993). Functions for DNA methylation in vertebrates. *Cold Spring Harbor Symposia on Quantitative Biology*, *58*, 281–285. https://doi.org/10.1101/SQB.1993.058.01.033

Bird, A. P., & Southern, E. M. (1978). Use of restriction enzymes to study eukaryotic DNA methylation. I. The methylation pattern in ribosomal DNA from Xenopus laevis. *Journal of Molecular Biology*. https://doi.org/10.1016/0022-2836(78)90242-5

Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z. D., … Meissner, A. (2011). Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell*, *144*(3), 439–452. https://doi.org/10.1016/j.cell.2010.12.032

Bock, C., Tomazou, E. M., Brinkman, A. B., Müller, F., Simmer, F., Gu, H., … Meissner, A. (2010).

Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature Biotechnology*. https://doi.org/10.1038/nbt.1681

Booth, M. J., Ost, T. W. B., Beraldi, D., Bell, N. M., Branco, M. R., Reik, W., & Balasubramanian, S. (2013). Oxidative bisulfite sequencing of 5-methylcytosine and 5- hydroxymethylcytosine. *Nature Protocols*. https://doi.org/10.1038/nprot.2013.115

Bostick, M., Kim, J. K., Esteve, P.-O., Clark, A., Pradhan, S., & Jacobsen, S. E. (2007). UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science (New York, N.Y.)*, *317*(5845), 1760–1764. https://doi.org/10.1126/science.1147939

Boyes, J., & Bird, A. (1992). Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. *The EMBO Journal*, *11*(1), 327–333. https://doi.org/10.1002/j.1460-2075.1992.tb05055.x

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., … Crawford, G. E. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*. https://doi.org/10.1016/j.cell.2007.12.014

Brinkman, A. B., Simmer, F., Ma, K., Kaan, A., Zhu, J., & Stunnenberg, H. G. (2010). Whole-genome DNA methylation profiling using MethylCap-seq. *Methods*. https://doi.org/10.1016/j.ymeth.2010.06.012

Broad Institute. (2017). GATK3: variant Discovery in High-Throughput Sequencing Data.

Broad Institute. (2019). Picard: a set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. Retrieved from https://broadinstitute.github.io/picard

Bürglin, T. R. (2011). Homeodomain subtypes and functional diversity. *Sub-Cellular Biochemistry*, *52*, 95–122. https://doi.org/10.1007/978-90-481-9069-0_5

Campbell, C. L., Scheller, C., Horn, H., Kidd, J. M., Doddapaneni, H., Underhill, P. A., … Fitzgerald, T. W. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., … Kent, W. J. (2018). The UCSC Genome Browser database: 2018 update. *Nucleic Acids Research*, *46*(D1), D762–D769. https://doi.org/10.1093/nar/gkx1020

Cattanach, B. M., & Kirk, M. (1985). Differential activity of maternally and paternally derived chromosome regions in mice. *Nature*. https://doi.org/10.1038/315496a0

Cheedipudi, S., Genolet, O., & Dobreva, G. (2014). Epigenetic inheritance of cell fates during embryonic development. *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2014.00019

Chen, P. Y., Cokus, S. J., & Pellegrini, M. (2010). BS Seeker: Precise mapping for bisulfite sequencing. *BMC Bioinformatics*, *11*. https://doi.org/10.1186/1471-2105-11-203

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., … Hubbard, T. (2011). Modernizing reference genome assemblies. *PLoS Biology*, *9*(7). https://doi.org/10.1371/journal.pbio.1001091

Clark, C., Palta, P., Joyce, C. J., Scott, C., Grundberg, E., Deloukas, P., … Coffey, A. J. (2012). A Comparison of the Whole Genome Approach of MeDIP-Seq to the Targeted Approach of the Infinium HumanMethylation450 BeadChip® for Methylome Profiling. *PLoS ONE*.

https://doi.org/10.1371/journal.pone.0050233

Clevers, H., Rafelski, S., Elowitz, M., Klein, A., Shendure, J., Trapnell, C., … Love, J. C. (2017). What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism? *Cell Systems*, *4*(3), 255–259. https://doi.org/10.1016/j.cels.2017.03.006

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., Rice, P. M., Institute, T. J. H., … Laboratory, E. M. B. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*(6), 1767–1771. https://doi.org/10.1093/nar/gkp1137

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., … Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, *452*(7184), 215–219. https://doi.org/10.1038/nature06745

Collins, F. S., Lander, E. S., Rogers, J., & Waterson, R. H. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

Colwell, M., Drown, M., Showel, K., Drown, C., Palowski, A., & Faulk, C. (2018). Evolutionary conservation of DNA methylation in CpG sites within ultraconserved noncoding elements. *Epigenetics*. https://doi.org/10.1080/15592294.2017.1411447

Compere, S. J., & Palmiter, R. D. (1981). DNA methylation controls the inducibility of the mouse metallothionein-I gene in lymphoid cells. *Cell*. https://doi.org/10.1016/0092-8674(81)90248-8

Cortellino, S., Xu, J., Sannai, M., Moore, R., Caretti, E., Cigliano, A., … Bellacosa, A. (2011). Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell*. https://doi.org/10.1016/j.cell.2011.06.020

Creighton, C. J., Morgan, M., Gunaratne, P. H., Wheeler, D. A., Gibbs, R. A., Robertson, G., … Sofia, H. J. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, *499*(7456), 43–49. https://doi.org/10.1038/nature12222

D'Urso, A., & Brickner, J. H. (2014). Mechanisms of epigenetic memory. *Trends in Genetics*. https://doi.org/10.1016/j.tig.2014.04.004

De Mendoza, A., Bonnet, A., Vargas-Landin, D. B., Ji, N., Hong, F., Yang, F., … Lister, R. (2018). Recurrent acquisition of cytosine methyltransferases into eukaryotic retrotransposons. *Nature Communications*. https://doi.org/10.1038/s41467-018-03724-9

Dean, W. (2014). DNA methylation and demethylation: A pathway to gametogenesis and development. *Molecular Reproduction and Development*. https://doi.org/10.1002/mrd.22280

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes and Development*, *25*(10), 1010–1022. https://doi.org/10.1101/gad.2037511

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., & Fuks, F. (2011). Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, *3*(6), 771–784. https://doi.org/10.2217/epi.11.105

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Dimitrieva, S., & Bucher, P. (2013). UCNEbase - A database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gks1092

Donaghey, J., Thakurela, S., Charlton, J., Chen, J. S., Smith, Z. D., Gu, H., … Meissner, A. (2018). Genetic determinants and epigenetic effects of pioneer-factor occupancy. *Nature Genetics*. https://doi.org/10.1038/s41588-017-0034-3

Doskočil, J., & Šorm, F. (1962). Distribution of 5-methylcytosine in pyrimidine sequences of deoxyribonucleic

acids. *BBA Specialized Section on Nucleic Acids and Related Subjects*. https://doi.org/10.1016/0926-6550(62)90353-5

Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E., … Beck, S. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotechnology*. https://doi.org/10.1038/nbt1414

Dreos, R., Ambrosini, G., Périer, R. C., & Bucher, P. (2013). EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Research*, *41*(D1), D157–D164. https://doi.org/10.1093/nar/gks1233

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., … Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, *38*(12), 1378–1385. https://doi.org/10.1038/ng1909

ECMA International. (2013). The JSON Data Interchange Format. *Standard ECMA-404*. https://doi.org/10.17487/rfc7158

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Ewels, P. (2019). nf-core/methylseq: a bioinformatics best-practice analysis pipeline used for Methylation (BS-Seq) data analysis. Retrieved from https://github.com/nf-core/methylseq

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., & Bock, C. (2015). Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Reports*. https://doi.org/10.1016/j.celrep.2015.02.001

Fielding, R. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. *Building*. https://doi.org/10.1.1.91.2433

Fisher, R. A. (2006). On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87. https://doi.org/10.2307/2340521

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., … Cohen, D. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database : The Journal of Biological Databases and Curation*, *2017*, bax028-bax028. https://doi.org/10.1093/database/bax028

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., … Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, *47*(D1), D766–D773. https://doi.org/10.1093/nar/gky955

Frommer, M., LE, M., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., … Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual {DNA} strands. *Proceedings of the National Academy of Sciences of the United States of America*.

Fuller, C. W., Middendorf, L. R., Benner, S. A., Church, G. M., Harris, T., Huang, X., … Vezenov, D. V. (2009). The challenges of sequencing by synthesis. *Nature Biotechnology*. https://doi.org/10.1038/nbt.1585

Fürst, R. W., Kliem, H., Meyer, H. H. D., & Ulbrich, S. E. (2012). A differentially methylated single CpG-site is correlated with estrogen receptor alpha transcription. *Journal of Steroid Biochemistry and Molecular Biology*, *130*(1–2), 96–104. https://doi.org/10.1016/j.jsbmb.2012.01.009

Gardiner-Garden, M., & Frommer, M. (1987). CpG Islands in vertebrate genomes. *Journal of Molecular Biology*. https://doi.org/10.1016/0022-2836(87)90689-9

Garg, P., Joshi, R. S., Watson, C., & Sharp, A. J. (2018). A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. *PLoS Genetics*, *14*(10). https://doi.org/10.1371/journal.pgen.1007707

Geisen, S., Barturen, G., Alganza, Á. M., Hackenberg, M., & Oliver, J. L. (2014). NGSmethDB: An updated genome resource for high quality, single-cytosine resolution methylomes. *Nucleic Acids Research*, *42*(D1). https://doi.org/10.1093/nar/gkt1202

Gibbs, R. A. (1990). DNA Amplification by the Polymerase Chain Reaction. *Analytical Chemistry*. https://doi.org/10.1021/ac00212a004

Gómez-Martín, C., Lebrón, R., Oliver, J. L., & Hackenberg, M. (2018). Prediction of CpG Islands as an intrinsic clustering property found in many Eukaryotic DNA sequences and its relation to DNA methylation. In T. Vavouri & M. A. Peinado (Eds.), *Methods in Molecular Biology* (Vol. 1766, pp. 31–47). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-7768-0_3

Green, B. B., Houseman, E. A., Johnson, K. C., Guerin, D. J., Armstrong, D. A., Christensen, B. C., & Marsit, C. J. (2016). Hydroxymethylation is uniquely distributed within term placenta, and is associated with gene expression. *FASEB Journal*. https://doi.org/10.1096/fj.201600310R

Grunau, C. (2002). MethDB--a public database for DNA methylation data. *Nucleic Acids Research*. https://doi.org/10.1093/nar/29.1.270

Gu, F., Doderer, M. S., Huang, Y. W., Roa, J. C., Goodfellow, P. J., Kizer, E. L., … Chen, Y. (2013). CMS: A Web-Based System for Visualization and Analysis of Genome-Wide Methylation Data of Human Cancers. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0060980

Guo, H., Zhu, P., Guo, F., Li, X., Wu, X., Fan, X., … Tang, F. (2015). Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nature Protocols*. https://doi.org/10.1038/nprot.2015.039

Guo, J. U., Su, Y., Zhong, C., Ming, G. L., & Song, H. (2011). Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell*, *145*(3), 423–434. https://doi.org/10.1016/j.cell.2011.03.022

Guo, W., Fiziev, P., Yan, W., Cokus, S., Sun, X., Zhang, M. Q., … Pellegrini, M. (2013). BS-Seeker2: A versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics*, *14*(1). https://doi.org/10.1186/1471-2164-14-774

Hackenberg, M., Barturen, G., Carpena, P., Luque-Escamilla, P. L., Previti, C., & Oliver, J. L. (2010). Prediction of CpG-island function: CpG clustering vs. sliding-window methods. *BMC Genomics*, *11*(1), 327. https://doi.org/10.1186/1471-2164-11-327

Hackenberg, M., Barturen, G., & L., J. (2012). DNA Methylation Profiling from High-Throughput Sequencing Data. In G. Barturen (Ed.), *DNA Methylation - From Genomics to Technology* (p. Ch. 2). Rijeka: IntechOpen. https://doi.org/10.5772/34825

Hackenberg, M., Barturen, G., & Oliver, J. L. (2011). NGSmethDB: A database for next-generation sequencing single-cytosine- resolution DNAmethylation data. *Nucleic Acids Research*, *39*(SUPPL. 1). https://doi.org/10.1093/nar/gkq942

Hackenberg, M., Previti, C., Luque-Escamilla, P. L., Carpena, P., Martínez-Aroza, J., & Oliver, J. L. (2006). CpGcluster: A distance-based algorithm for CpG-island detection. *BMC Bioinformatics*, *7*. https://doi.org/10.1186/1471-2105-7-446

Halley-Stott, R. P., & Gurdon, J. B. (2013). Epigenetic memory in the context of nuclear reprogramming and cancer. *Briefings in Functional Genomics*, *12*(3), 164–173. https://doi.org/10.1093/bfgp/elt011

Hammerman, P. S., Voet, D., Lawrence, M. S., Voet, D., Jing, R., Cibulskis, K., … Shen, R. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. https://doi.org/10.1038/nature11404

Han, H., Cho, J. W., Lee, S., Yun, A., Kim, H., Bae, D., … Lee, I. (2018). TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, *46*(D1), D380–D386. https://doi.org/10.1093/nar/gkx1013

Hannon, E., Knox, O., Sugden, K., Burrage, J., Wong, C. C. Y., Belsky, D. W., … Mill, J. (2018). Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS Genetics*, *14*(8), e1007544. https://doi.org/10.1371/journal.pgen.1007544

Hansen, K. D., Langmead, B., & Irizarry, R. A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, *13*(10). https://doi.org/10.1186/gb-2012-13-10-R83

Harbers, M., Medvedeva, Y. A., Lassmann, T., Bajic, V. B., Bhuyan, M. S. I., Ba-Alawi, W., … Kawaji, H. (2014). Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics*, *15*(1), 119. https://doi.org/10.1186/1471-2164-15-119

Hark, A. T., Schoenherr, C. J., Katz, D. J., Ingram, R. S., Levorse, J. M., & Tilghman, S. M. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*. https://doi.org/10.1038/35013106

He, X.-J., Chen, T., & Zhu, J.-K. (2011). Regulation and function of DNA methylation in plants and animals. *Cell Research*, *21*(3), 442–465. https://doi.org/10.1038/cr.2011.23

He, Y. F., Li, B. Z., Li, Z., Liu, P., Wang, Y., Tang, Q., … Xu, G. L. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science*. https://doi.org/10.1126/science.1210944

Hellman, A., & Chess, A. (2007). Gene body-specific methylation on the active X chromosome. *Science*. https://doi.org/10.1126/science.1136352

Henikoff, S., & Greally, J. M. (2016). Epigenetics, cellular memory and gene regulation. *Current Biology*, *26*(14), R644–R648. https://doi.org/10.1016/j.cub.2016.06.011

Hermann, A., Goyal, R., & Jeltsch, A. (2004). DNA Processively with High Preference for Hemimethylated. *Journal of Biological Chemistry*, *279*(46), 48350–48359. https://doi.org/10.1074/jbc.M403427200

Hinrichs, A. S., Raney, B. J., Speir, M. L., Rhead, B., Casper, J., Karolchik, D., … Kent, W. J. (2016). UCSC Data Integrator and Variant Annotation Integrator. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv766

Holliday, R., & Pugh, J. (1975). DNA modification mechanisms and gene activity during development. *Science*. https://doi.org/10.1126/science.1111098

Hotchkiss, R. D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *The Journal of Biological Chemistry*. https://doi.org/10.1038/nrg2063

Huang, Y., Pastor, W. A., Shen, Y., Tahiliani, M., Liu, D. R., & Rao, A. (2010). The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE*, *5*(1).

https://doi.org/10.1371/journal.pone.0008888

Illingworth, R. S., Gruenewald-Schneider, U., Webb, S., Kerr, A. R. W., James, K. D., Turner, D. J., … Bird, A. P. (2010). Orphan CpG Islands Identify numerous conserved promoters in the mammalian genome. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1001134

Irizarry, R. A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S. A., Jeddeloh, J. A., … Feinberg, A. P. (2008). Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Research*. https://doi.org/10.1101/gr.7301508

Ito, S., Dalessio, A. C., Taranova, O. V., Hong, K., Sowers, L. C., & Zhang, Y. (2010). Role of tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*. https://doi.org/10.1038/nature09303

Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. A., … Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*. https://doi.org/10.1126/science.1210597

Iwafuchi-Doi, M. (2019). The mechanistic basis for chromatin regulation by pioneer transcription factors. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *11*(1), e1427. https://doi.org/10.1002/wsbm.1427

Iwafuchi-Doi, M., & Zaret, K. S. (2014). Pioneer transcription factors in cell reprogramming. *Genes and Development*. https://doi.org/10.1101/gad.253443.114

Iwafuchi-Doi, M., & Zaret, K. S. (2016). Cell fate control by pioneer transcription factors. *Development*, *143*(11), 1833–1837. https://doi.org/10.1242/dev.133900

Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nature Genetics*, *33*(3S), 245–254. https://doi.org/10.1038/ng1089

James Kent, W., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*. https://doi.org/10.1101/gr.229102. Article published online before print in May 2002

Jeziorska, D. M., Murray, R. J. S., De Gobbi, M., Gaentzsch, R., Garrick, D., Ayyub, H., … Tufarelli, C. (2017). DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proceedings of the National Academy of Sciences*, *114*(36), E7526–E7535. https://doi.org/10.1073/pnas.1703087114

Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A., & Cheng, X. (2007). Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature*, *449*(7159), 248–251. https://doi.org/10.1038/nature06146

Johnson, T. B., & Coghill, R. D. (1925). Researches on pyrimidines. C111. The discovery of 5-methyl-cytosine in tuberculinic acid, the nucleic acid of the tubercle bacillus. *Journal of the American Chemical Society*, *47*(11), 2838–2844. https://doi.org/10.1021/ja01688a030

Jones, P. A. (1999). The DNA methylation paradox. *Trends in Genetics*. https://doi.org/10.1016/S0168-9525(98)01636-9

Jones, P. A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, *13*(7), 484–492. https://doi.org/10.1038/nrg3230

Kanduri, C., Pant, V., Loukinov, D., Pugacheva, E., Qi, C. F., Wolffe, A., … Lobanenkov, V. V. (2000). Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive. *Current Biology*. https://doi.org/10.1016/S0960-9822(00)00597-2

Karolchik, D. (2003). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, *32*(90001), 493D–

496. https://doi.org/10.1093/nar/gkh103

Karsch-Mizrachi, I., Takagi, T., & Cochrane, G. (2017). The international nucleotide sequence database collaboration. *Nucleic Acids Research*, *46*(D1), D48–D51. https://doi.org/10.1093/nar/gkx1097

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., Van Der Lee, R., … Mathelier, A. (2018). JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, *46*(D1), D260–D266. https://doi.org/10.1093/nar/gkx1126

Khan, A., & Zhang, X. (2016). DbSUPER: A database of Super-enhancers in mouse and human genome. *Nucleic Acids Research*, *44*(D1), D164–D171. https://doi.org/10.1093/nar/gkv1002

Kim, M., & Costello, J. (2017). DNA methylation: An epigenetic mark of cellular memory. *Experimental and Molecular Medicine*, *49*(4), e322. https://doi.org/10.1038/emm.2017.10

Kim, M., Park, Y.-K., Kang, T.-W., Lee, S.-H., Rhee, Y.-H., Park, J.-L., … Kim, Y. S. (2014). Dynamic changes in DNA methylation and hydroxymethylation when hES cells undergo differentiation toward a neuronal lineage. *Human Molecular Genetics*, *23*(3), 657–667. https://doi.org/10.1093/hmg/ddt453

Kim, S., Yu, N. K., & Kaang, B. K. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. *Experimental & Molecular Medicine*. https://doi.org/10.1038/emm.2015.33

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., … Ding, L. (2009). VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, *25*(17), 2283–2285. https://doi.org/10.1093/bioinformatics/btp373

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., & Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods*, *6*(4), 291–295. https://doi.org/10.1038/nmeth.1311

Krueger, F. (2018). Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries.

Krueger, F., & Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, *27*(11), 1571–1572. https://doi.org/10.1093/bioinformatics/btr167

Krueger, F., Kreck, B., Franke, A., & Andrews, S. R. (2012). DNA methylome analysis using short bisulfite sequencing data. *Nature Methods*. https://doi.org/10.1038/nmeth.1828

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1952.10483441

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–329. https://doi.org/10.1038/nature14248

Laird, P. W. (2010). Principles and challenges of genomewide DNA methylation analysis. *Nature Reviews Genetics*, *11*(3), 191–203. https://doi.org/10.1038/nrg2732

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., … Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, *172*(4), 650–665. https://doi.org/10.1016/j.cell.2018.01.029

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of

short DNA sequences to the human genome. *Genome Biology*, *10*(3). https://doi.org/10.1186/gb-2009-10-3-r25

Larsen, F., Solheim, J., & Prydz, H. (1993). A methylated CpG Island 3′ in the apolipoprotein-E gene does not repress its transcription. *Human Molecular Genetics*. https://doi.org/10.1093/hmg/2.6.775

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., … Wei, C. L. (2010). Dynamic changes in the human methylome during differentiation. *Genome Research*, *20*(3), 320–331. https://doi.org/10.1101/gr.101907.109

Lebrón, R., Barturen, G., Gómez-Martín, C., Oliver, J. L., & Hackenberg, M. (2016). MethFlowVM: a virtual machine for the integral analysis of bisulfite sequencing data. *BioRxiv: Http://Biorxiv.Org/Content/Early/2016/07/31/066795*. https://doi.org/10.1101/066795

Lebrón, R., Gómez-Martín, C., Carpena, P., Bernaola-Galván, P., Barturen, G., Hackenberg, M., & Oliver, J. L. (2017). NGSmethDB 2017: Enhanced methylomes and differential methylation. *Nucleic Acids Research*, *45*(D1), D97–D103. https://doi.org/10.1093/nar/gkw996

Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., … Ren, B. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, *518*(7539), 350–354. https://doi.org/10.1038/nature14217

Levy, S. E., & Myers, R. M. (2016). Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics*. https://doi.org/10.1146/annurev-genom-083115-022413

Li, D., Zhang, B., Xing, X., & Wang, T. (2015). Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. *Methods*. https://doi.org/10.1016/j.ymeth.2014.10.032

Li, E., Bestor, T. H., & Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*. https://doi.org/10.1016/0092-8674(92)90611-F

Li, H. (2011). Tabix: Fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, *27*(5), 718–719. https://doi.org/10.1093/bioinformatics/btq671

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, S., Figueroa, M. E., Kormaksson, M., Melnick, A., Mason, C. E., Garrett-Bakelman, F. E., & Akalin, A. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, *13*(10), R87. https://doi.org/10.1186/gb-2012-13-10-r87

Lin, X., Sun, D., Rodriguez, B., Zhao, Q., Sun, H., Zhang, Y., … Bishop, M. (2013). BSeQC: Quality control of bisulfite sequencing experiments. *Bioinformatics*, *29*(24), 3227–3229. https://doi.org/10.1093/bioinformatics/btt548

Lioznova, A. V., Khamis, A. M., Artemov, A. V., Besedina, E., Ramensky, V., Bajic, V. B., … Medvedeva, Y. A. (2019). CpG traffic lights are markers of regulatory regions in human genome. *BMC Genomics*, *20*(1), 102. https://doi.org/10.1186/s12864-018-5387-1

Lisanti, S., Omar, W. A. W., Tomaszewski, B., De Prins, S., Jacobs, G., Koppen, G., … Langie, S. A. S. (2013). Comparison of methods for quantification of global DNA methylation in human cells and tissues. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0079044

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., … Ecker, J. R. (2013). Global epigenomic reconfiguration during mammalian brain development. *Science*. https://doi.org/10.1126/science.1237905

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008).

Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, *133*(3), 523–536. https://doi.org/10.1016/j.cell.2008.03.029

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., … Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*(7271), 315–322. https://doi.org/10.1038/nature08514

Liu, Y., Siegmund, K. D., Laird, P. W., & Berman, B. P. (2012). Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biology*, *13*(7). https://doi.org/10.1186/gb-2012-13-7-r61

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., … Kawaji, H. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology*, *16*(1), 22. https://doi.org/10.1186/s13059-014-0560-6

Luo, H., Xi, Y., Li, W., Li, J., Li, Y., Dong, S., … Yu, W. (2017). Cell identity bookmarking through heterogeneous chromatin landscape maintenance during the cell cycle. *Human Molecular Genetics*, *26*(21), 4231–4243. https://doi.org/10.1093/hmg/ddx312

Luo, Y., Lu, X., & Xie, H. (2014). Dynamic Alu Methylation during Normal Development, Aging, and Tumorigenesis. *BioMed Research International*, *2014*. https://doi.org/10.1155/2014/784706

Lv, J., Liu, H., Su, J., Wu, X., Liu, H., Li, B., … Zhang, Y. (2012). DiseaseMeth: A human disease methylation database. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkr1169

Magdinier, F., & Wolffe, A. P. (2002). Selective association of the methyl-CpG binding protein MBD2 with the silent p14/p16 locus in human neoplasia. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.101617298

Maienschein, J. (2012). Epigenesis and Preformationism. Retrieved from http://stanford.library.usyd.edu.au/entries/epigenesis/

Maiti, A., & Drohat, A. C. (2011). Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: Potential implications for active demethylation of CpG sites. *Journal of Biological Chemistry*, *286*(41), 35334–35338. https://doi.org/10.1074/jbc.C111.284620

Malladi, V. S., Erickson, D. T., Podduturi, N. R., Rowe, L. D., Chan, E. T., Davidson, J. M., … Hong, E. L. (2015). Ontology application and use at the ENCODE DCC. *Database*, *2015*. https://doi.org/10.1093/database/bav010

Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., … Parkinson, H. (2010). Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq099

Marques-Rocha, J. L., Milagro, F. I., Mansego, M. L., Mourão, D. M., Martínez, J. A., & Bressan, J. (2016). LINE-1 methylation is positively associated with healthier lifestyle but inversely related to body fat mass in healthy young individuals. *Epigenetics*. https://doi.org/10.1080/15592294.2015.1135286

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Maunakea, A. K., Chepelev, I., Cui, K., & Zhao, K. (2013). Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Research*. https://doi.org/10.1038/cr.2013.110

Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., Dsouza, C., Fouse, S. D., … Costello, J. F.

(2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*. https://doi.org/10.1038/nature09165

Mayran, A., Khetchoumian, K., Hariri, F., Pastinen, T., Gauthier, Y., Balsalobre, A., & Drouin, J. (2018). Pioneer factor Pax7 deploys a stable enhancer repertoire for specification of cell fate. *Nature Genetics*. https://doi.org/10.1038/s41588-017-0035-2

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., … Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*. https://doi.org/10.1038/nbt.1630

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., … Lander, E. S. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, *454*(7205), 766–770. https://doi.org/10.1038/nature07107

Mendoza, M. L., Steinberg, K., Church, D. M., Kitts, P. A., Durbin, R., Flicek, P., … Hoffman, M. M. (2015). Extending reference assembly models. *Genome Biology*, *16*(1), 13. https://doi.org/10.1186/s13059-015-0587-3

Miura, F., Enomoto, Y., Dairiki, R., & Ito, T. (2012). Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Research*, *40*(17). https://doi.org/10.1093/nar/gks454

Moarefi, A. H., & Chédin, F. (2011). ICF syndrome mutations cause a broad spectrum of biochemical defects in DNMT3B-mediated de novo DNA methylation. *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2011.04.050

Müller, F., Scherer, M., Assenov, Y., Lutsik, P., Walter, J., Lengauer, T., & Bock, C. (2019). RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biology*, *20*(1), 55. https://doi.org/10.1186/s13059-019-1664-9

Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E., & Haendel, M. A. (2012). Uberon, an integrative multi-species anatomy ontology. *Genome Biology*. https://doi.org/10.1186/gb-2012-13-1-r5

Negre, V., & Grunau, C. (2006). The MethDB DAS Server: Adding an epigenetic information layer to the human genome. *Epigenetics*. https://doi.org/10.4161/epi.1.2.2765

Nestor, C. E., & Meehan, R. R. (2014). Hydroxymethylated DNA immunoprecipitation (hmeDIP). *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-62703-706-8_20

Okano, M., Bell, D. W., Haber, D. A., & Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, *99*(3), 247–257.

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv566

Ollikainen, M., Smith, K. R., Joo, E. J. H., Ng, H. K., Andronikos, R., Novakovic, B., … Craig, J. M. (2010). DNA methylation analysis of multiple tissues from newborn twins reveals both genetic and intrauterine components to variation in the human neonatal epigenome. *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddq336

Olova, N., Krueger, F., Andrews, S., Oxley, D., Berrens, R. V., Branco, M. R., & Reik, W. (2018). Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biology*, *19*(1), 33. https://doi.org/10.1186/s13059-018-1408-2

Ooi, S. K. T., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., … Bestor, T. H. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, *448*(7154), 714–717. https://doi.org/10.1038/nature05987

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., … Clark, S. J. (2016). Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, *17*(1), 208. https://doi.org/10.1186/s13059-016-1066-1

Probst, A. V, Dunleavy, E., & Almouzni, G. (2009). Epigenetic inheritance during the cell cycle. *Nature Reviews. Molecular Cell Biology*, *10*(3), 192–206. https://doi.org/10.1038/nrm2640

Ptashne, M. (2007). On the use of the word "epigenetic." *Current Biology*. https://doi.org/10.1016/j.cub.2007.02.030

Qu, K., Garamszegi, S., Wu, F., Thorvaldsdottir, H., Liefeld, T., Ocana, M., … Mesirov, J. P. (2016). Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace. *Nature Methods*, *13*(3), 245–247. https://doi.org/10.1038/nmeth.3732

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rackham, O., Itoh, M., Suzuki, H., Lilje, B., Hayashizaki, Y., Furuhata, E., … Bertin, N. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493), 455–461. https://doi.org/10.1038/nature12787

Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P., & Jaenisch, R. (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences*, *97*(10), 5237–5242. https://doi.org/10.1073/pnas.97.10.5237

Raney, B. J., Dreszer, T. R., Barber, G. P., Clawson, H., Fujita, P. A., Wang, T., … Kent, W. J. (2014). Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btt637

Riggs, A. D. (1975). X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*, *14*(1), 9–25. https://doi.org/10.1159/000130315

Rinaldi, L., Datta, D., Serrat, J., Morey, L., Solanas, G., Avgustinova, A., … Benitah, S. A. (2016). Dnmt3a and Dnmt3b Associate with Enhancers to Regulate Human Epidermal Stem Cell Homeostasis. *Cell Stem Cell*, *19*(4), 491–501. https://doi.org/10.1016/j.stem.2016.06.020

Rishi, V., Bhattacharya, P., Chatterjee, R., Rozenberg, J., Zhao, J., Glass, K., … Vinson, C. (2010). CpG methylation of half-CRE sequences creates C/EBP binding sites that activate some tissue-specific genes. *Proceedings of the National Academy of Sciences*, *107*(47), 20311–20316. https://doi.org/10.1073/pnas.1008688107

Rothbart, S. B., Krajewski, K., Nady, N., Tempel, W., Xue, S., Badeaux, A. I., … Strahl, B. D. (2012). Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nature Structural and Molecular Biology*, *19*(11), 1155–1160. https://doi.org/10.1038/nsmb.2391

Rulands, S., Lee, H. J., Clark, S. J., Angermueller, C., Smallwood, S. A., Krueger, F., … Reik, W. (2018). Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency. *Cell Systems*, *7*(1), 63–76.e12. https://doi.org/10.1016/j.cels.2018.06.012

Sandoval, J., Heyn, H. A., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., & Esteller, M. (2011). Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*. https://doi.org/10.4161/epi.6.6.16196

Saxonov, S., Berg, P., & Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, *103*(5), 1412–1417. https://doi.org/10.1073/pnas.0510310103

Schon, U., Diem, O., Leitner, L., Gunzburg, W. H., Mager, D. L., Salmons, B., & Leib-Mosch, C. (2009). Human Endogenous Retroviral Long Terminal Repeat Sequences as Cell Type-Specific Promoters in Retroviral Vectors. *Journal of Virology*. https://doi.org/10.1128/JVI.00858-09

Schumacher, T. N., Wold, B., Lein, E., Sharma, P., Netea, M., Eberwine, J., … Linnarsson, S. (2017). The Human Cell Atlas. *ELife*, *6*. https://doi.org/10.7554/elife.27041

Schwartzman, O., & Tanay, A. (2015). Single-cell epigenomics: Techniques and emerging applications. *Nature Reviews Genetics*, *16*(12), 716–726. https://doi.org/10.1038/nrg3980

Sender, R., Fuchs, S., & Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology*, *14*(8), e1002533. https://doi.org/10.1371/journal.pbio.1002533

Sharifi-Zarchi, A., Gerovska, D., Adachi, K., Totonchi, M., Pezeshk, H., Taft, R. J., … Araúzo-Bravo, M. J. (2017). DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism. *BMC Genomics*. https://doi.org/10.1186/s12864-017-4353-7

Shayevitch, R., Askayo, D., Keydar, I., & Ast, G. (2018). The importance of DNA methylation of exons on alternative splicing. *RNA*. https://doi.org/10.1261/rna.064865.117

Sherry, S. T. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Shipony, Z., Mukamel, Z., Cohen, N. M., Landan, G., Chomsky, E., Zeliger, S. R., … Tanay, A. (2014). Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature*, *513*(7516), 115–119. https://doi.org/10.1038/nature13458

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., … Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Siggens, L., & Ekwall, K. (2014). Epigenetics, chromatin and genome organization: recent advances from the ENCODE project. *Journal of Internal Medicine*, *276*(3), 201–214. https://doi.org/10.1111/joim.12231

Sloan, C. A., Chan, E. T., Davidson, J. M., Malladi, V. S., Strattan, J. S., Hitz, B. C., … Cherry, J. M. (2016). ENCODE data at the ENCODE portal. *Nucleic Acids Research*, *44*(D1), D726–D732. https://doi.org/10.1093/nar/gkv1160

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., … Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, *11*(8), 817–820. https://doi.org/10.1038/nmeth.3035

Smit, A. (2018). RepeatMasker: a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. Retrieved from http://www.repeatmasker.org

Smith, Z. D., Gu, H., Bock, C., Gnirke, A., & Meissner, A. (2009). High-throughput bisulfite sequencing in mammalian genomes. *Methods*. https://doi.org/10.1016/j.ymeth.2009.05.003

Smith, Z. D., & Meissner, A. (2013). DNA methylation: Roles in mammalian development. *Nature Reviews Genetics*, *14*(3), 204–220. https://doi.org/10.1038/nrg3354

Song, Q., Decato, B., Hong, E. E., Zhou, M., Fang, F., Qu, J., … Smith, A. D. (2013). A reference methylome

database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0081148

Spearman, C. (1987). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*. https://doi.org/10.2307/1422689

Stamp, M. (2004). A revealing introduction to hidden Markov models. *Department of Computer Science San Jose State University*. https://doi.org/10.1.1.136.137

Stern, C. D. (2000). Conrad H. Waddington's contributions to avian and mammalian development, 1930-1940. *International Journal of Developmental Biology*, *44*(1), 15–22.

Stirzaker, C., Taberlay, P. C., Statham, A. L., & Clark, S. J. (2014). Mining cancer methylomes: prospects and challenges. *Trends in Genetics : TIG*, *30*(2), 75–84. https://doi.org/10.1016/j.tig.2013.11.004

Stocks, B. J., Lynham, T. J., Lawson, B. D., Alexander, M. E., Wagner, C. E. Van, McAlpine, R. S., & Dubé, D. E. (1989). Canadian Forest Fire Danger Rating System: An Overview. *The Forestry Chronicle*, *65*(4), 258–265. https://doi.org/10.5558/tfc65258-4

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

Sutherland, E., Coe, L., & Raleigh, E. A. (1992). McrBC: a multisubunit GTP-dependent restriction endonuclease. *Journal of Molecular Biology*. https://doi.org/10.1016/0022-2836(92)90925-A

Sved, J., & Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the National Academy of Sciences*, *87*(12), 4692–4696. https://doi.org/10.1073/pnas.87.12.4692

Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., … Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*. https://doi.org/10.1126/science.1170116

Taiwo, O., Wilson, G. A., Morris, T., Seisenberger, S., Reik, W., Pearce, D., … Butcher, L. M. (2012). Methylome analysis using MeDIP-seq with low DNA concentrations. *Nature Protocols*. https://doi.org/10.1038/nprot.2012.012

Taub, M. A., Corrada Bravo, H., & Irizarry, R. A. (2010). Overcoming bias and systematic errors in next generation sequencing data. *Genome Medicine*, *2*(12). https://doi.org/10.1186/gm208

Tomizawa, S. -i. S., Kobayashi, H., Watanabe, T., Andrews, S., Hata, K., Kelsey, G., & Sasaki, H. (2011). Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes. *Development*. https://doi.org/10.1242/dev.061416

Tomso, D. J., & 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. *Journal of Molecular Biology*, *327*(2), 303–308. https://doi.org/10.1016/S0022-2836(03)00120-7

Torres-Padilla, M.-E., & Chambers, I. (2014). Transcription factor heterogeneity in pluripotent stem cells: a stochastic advantage. *Development*, *141*(11), 2173–2181. https://doi.org/10.1242/dev.102624

Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J., & Ecker, J. R. (2015). MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nature Protocols*.

https://doi.org/10.1038/nprot.2014.114

Van Wittenberghe, N., Roe, B., Johnson, M., Neri, F. J., Smith, K. S., Tsang, E. K., … Wheeler, J. (2017). Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nature Genetics*, *49*(12), 1664–1670. https://doi.org/10.1038/ng.3969

Vanhove, J. (2016). What data patterns can lie behind a correlation coefficient?

Vickaryous, M. K., & Hall, B. K. (2006). Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biological Reviews of the Cambridge Philosophical Society*. https://doi.org/10.1017/S1464793106007068

Visel, A., Minovitsky, S., Dubchak, I., & Pennacchio, L. A. (2007). VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Research*, *35*(SUPPL. 1), D88–D92. https://doi.org/10.1093/nar/gkl822

Wagner, G. P., Kin, K., & Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, *131*(4), 281–285. https://doi.org/10.1007/s12064-012-0162-3

Wang, Y. M., Zhou, P., Wang, L. Y., Li, Z. H., Zhang, Y. N., & Zhang, Y. X. (2012). Correlation between DNase I hypersensitive site distribution and gene expression in HeLa S3 cells. *PLoS ONE*, *7*(8). https://doi.org/10.1371/journal.pone.0042414

Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Pääbo, S., Rebhan, M., & Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*, *39*(4), 457–466. https://doi.org/10.1038/ng1990

Weng, Y.-I., Huang, T. H.-M., & Yan, P. S. (2009). Methylated DNA Immunoprecipitation and Microarray-Based Analysis: Detection of DNA Methylation in Breast Cancer Cell Lines. https://doi.org/10.1007/978-1-60327-378-7_10

Wolf, S. F., Jollyt, D. J., Lunnen, K. D., Friedmannt, T., & Migeon, B. R. (1984). Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome : Implications for X-chromosome inactivation. *Proceedings of the National Academy of Sciences of the United States of America*.

Wu, H., & Zhang, Y. (2015). Charting oxidized methylcytosines at base resolution. *Nature Structural & Molecular Biology*. https://doi.org/10.1038/nsmb.3071

Wyatt, G. R. (1950). Occurrence of 5-methyl-cytosine in nucleic acids. *Nature*. https://doi.org/10.1038/166237b0

Wyatt, G. R. (1951). Recognition and estimation of 5-methylcytosine in nucleic acids. *Biochemical Journal*, *48*(5), 581–584. https://doi.org/10.1042/bj0480581

Xie, W., Barr, C. L., Kim, A., Yue, F., Lee, A. Y., Eubanks, J., … Ren, B. (2012). Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell*. https://doi.org/10.1016/j.cell.2011.12.035

Xiong, Y., Wei, Y., Gu, Y., Zhang, S., Lyu, J., Zhang, B., … Zhang, Y. (2017). DiseaseMeth version 2.0: A major expansion and update of the human disease methylation database. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw1123

Xuan Lin, Q. X., Sian, S., An, O., Thieffry, D., Jha, S., & Benoukraf, T. (2019). MethMotif: an integrative cell specific database of transcription factor binding motifs coupled with DNA methylation profiles. *Nucleic Acids Research*, *47*(D1), D145–D154. https://doi.org/10.1093/nar/gky1005

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., … Taipale, J. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, *356*(6337),

eaaj2239. https://doi.org/10.1126/science.aaj2239

Yoder, J. A., Walsh, C. P., & Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*. https://doi.org/10.1016/S0168-9525(97)01181-5

Yong, W. S., Hsu, F. M., & Chen, P. Y. (2016). Profiling genome-wide DNA methylation. *Epigenetics and Chromatin*, *9*(1), 26. https://doi.org/10.1186/s13072-016-0075-3

Yu, D.-H., Ware, C., Waterland, R. A., Zhang, J., Chen, M.-H., Gadkari, M., … Shen, L. (2013). Developmentally Programmed 3′ CpG Island Methylation Confers Tissue- and Cell-Type-Specific Transcriptional Activation. *Molecular and Cellular Biology*, *33*(9), 1845–1858. https://doi.org/10.1128/MCB.01124-12

Yu, M., Hon, G. C., Szulwach, K. E., Song, C. X., Jin, P., Ren, B., & He, C. (2012). Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nature Protocols*. https://doi.org/10.1038/nprot.2012.137

Yu, M., Hon, G. C., Szulwach, K. E., Song, C. X., Zhang, L., Kim, A., … He, C. (2012). Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*. https://doi.org/10.1016/j.cell.2012.04.027

Zaret, K. S., & Mango, S. E. (2016). Pioneer transcription factors, chromatin dynamics, and cell fate control. *Current Opinion in Genetics and Development*. https://doi.org/10.1016/j.gde.2015.12.003

Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T., & Flicek, P. R. (2015). The Ensembl Regulatory Build. *Genome Biology*, *16*(1), 56. https://doi.org/10.1186/s13059-015-0621-5

Zhang, Y., Zhang, D., Li, Q., Liang, J., Sun, L., Yi, X., … Shang, Y. (2016). Nucleation of DNA repair factors by FOXA1 links DNA demethylation to transcriptional pioneering. *Nature Genetics*. https://doi.org/10.1038/ng.3635

Zhao, M.-T., Whyte, J. J., Hopkins, G. M., Kirk, M. D., & Prather, R. S. (2014). Methylated DNA Immunoprecipitation and High-Throughput Sequencing (MeDIP-seq) Using Low Amounts of Genomic DNA. *Cellular Reprogramming*. https://doi.org/10.1089/cell.2014.0002

Zhou, J., Sears, R. L., Xing, X., Zhang, B., Li, D., Rockweiler, N. B., … Wang, T. (2017). Tissue-specific DNA methylation is conserved across human, mouse, and rat, and driven by primary sequence conservation. *BMC Genomics*. https://doi.org/10.1186/s12864-017-4115-6

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T. Y., Kohlbacher, O., … Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature*. https://doi.org/10.1038/nature12433

Ziller, M. J., Muller, F., Liao, J., Zhang, Y., Gu, H., Bock, C., … Meissner, A. (2011). Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS Genetics*, *7*(12), e1002389. https://doi.org/10.1371/journal.pgen.1002389