



Desarrollo de un flujo de análisis de datos de secuenciación de bisulfito de genoma completo

**Maestría en Bioinformática y Biología de Sistemas
Universidad Nacional del Noroeste de la Provincia de
Buenos Aires**

Alumna: Bioq. Sofía Reynoso

Directora: Dra. Mg. Gabriela Alejandra Merino

Septiembre 2025

Agradecimientos

Estoy profundamente agradecida con todas las personas e instituciones que me han acompañado a lo largo de este recorrido de desarrollo profesional y personal. En especial quisiera agradecer:

A mi directora de tesis, Gabriela Merino, por confiar en mí para el desarrollo de este proyecto y acompañarme a lo largo de todo el proceso. Por sus ideas y su claridad para transmitirlos.

A la Dra. Gabriela Fernandez, por su constante acompañamiento y disposición para orientarme en cada etapa del proceso..

A Northwestern University, en especial a Luisa Cusick por su ayuda con la configuración del cluster.

A mis compañeros de la maestría, en especial a Anita, por los viajes a Pergamino.

A mis amigas, por las risas y el apoyo constante que me impulsan a seguir siempre hacia adelante.

A mis hermanos y cuñadas. A mis sobrinos, Meme y Pedro Pony. Por su amor. Por alegrarme cada día.

A mi madre, Liliana por su apoyo incondicional. Por sus sacrificios. Por incentivarme a que haga lo que me apasiona. Por su paciencia y, sobre todo, por su entendimiento. Por alegrarse conmigo, o hasta más que yo.

A Fer, mi compañero de vida. Por bancarme, entenderme y tenerme paciencia. Por estimularme y querer lo mejor para mí. Por hacerme ver las cosas de otra manera.

Resumen

La metilación del ADN constituye una de las marcas epigenéticas más estudiadas en eucariotas y cumple un rol esencial en la regulación de la expresión génica, el desarrollo embrionario y la determinación de fenotipos dependientes de tejido. Esta modificación, caracterizada por la adición de un grupo metilo al quinto carbono de la citosina para formar 5-metilcitosina (5mC), ocurre predominantemente en contextos CpG y se concentra en regiones genómicas conocidas como islas CpG (ICGs). Aproximadamente el 70% de los promotores génicos se solapan con ICGs y, en células, suelen encontrarse no metilados, favoreciendo la transcripción activa. Por el contrario, regiones intragénicas y cuerpos de genes suelen presentar altos niveles de metilación, contribuyendo al silenciamiento génico.

La secuenciación por bisulfito de genoma completo (WGBS) constituye la técnica de referencia para la caracterización de la metilación del ADN con resolución de nucleótido único. No obstante, el tratamiento con bisulfito implica una conversión de citosinas no metiladas a timinas que si bien reduce la complejidad del genoma, plantea desafíos técnicos en el alineamiento de las lecturas de secuenciación y en los análisis posteriores. Ante la gran diversidad de algoritmos y herramientas disponibles, la construcción de un flujo de análisis de principio a fin que asegure precisión técnica, eficiencia computacional y relevancia biológica representa un desafío no trivial.

En este trabajo se desarrollaron cuatro flujos de análisis integral para datos de WGBS, abarcando desde el control de calidad y el preprocesamiento de lecturas hasta el alineamiento, la cuantificación de niveles de metilación y la identificación e interpretación de regiones diferencialmente metiladas (DMRs). Para ello, se evaluaron y compararon dos herramientas ampliamente utilizadas para el alineamiento de lecturas y cuantificación de los perfiles de metilación, Bismark y BSMAP, junto con dos metodologías estadísticas para la detección de DMRs, DSS y methylKit.

Los resultados evidenciaron diferencias en el desempeño computacional de Bismark y BSMAP, aunque ambos mostraron porcentajes equivalentes de lecturas únicas y estimaciones globales de metilación. La comparación con datos de referencia de ENCODE reveló una alta correlación positiva ($r > 0,91$) entre los niveles de metilación predichos y los reportados experimentalmente, validando la robustez del pipeline desarrollado. El análisis funcional de los DMRs indicó que una proporción significativa de los cambios de metilación se localizan en elementos regulatorios como promotores y *enhancers*, confirmando la estrecha relación entre los perfiles epigenéticos y la regulación de la expresión génica dependiente de tejido. En particular, se observó que los promotores que cambian de estado de actividad entre corazón y riñón presentan mayores diferencias de metilación que aquellos que mantienen su estado.

En conjunto, en este trabajo se diseñaron cuatro flujos bioinformáticos flexibles, reproducibles y escalables para el análisis de datos de WGBS, capaces de integrar distintos alineadores y métodos estadísticos, y de proporcionar resultados consistentes con datos de referencia y biológicamente relevantes. Asimismo, su integración con anotaciones regulatorias de Ensembl abre la posibilidad de aplicar este enfoque en estudios comparativos de epigenomas de múltiples especies, en línea con iniciativas como ENCODE, GENE-SWitCH y AQUA-FAANG.

Abstract

DNA methylation is one of the best-characterized epigenetic marks in eukaryotes, with a fundamental role in gene expression regulation, embryonic development, and the establishment of tissue-specific phenotypes. This modification is defined by the addition of a methyl group to the fifth carbon of cytosine to generate 5-methylcytosine (5mC), occurs mainly in CpG contexts and tends to cluster in CpG islands (CGIs). Around 70% of gene promoters overlap with CGIs and are usually unmethylated in normal cells, which favors transcriptional activity. In contrast, intragenic regions and gene bodies are often methylated, contributing to transcriptional silencing.

Whole-genome bisulfite sequencing (WGBS) remains the gold standard for studying DNA methylation at single-nucleotide resolution. Although the chemical conversion of unmethylated cytosines to thymines reduces genome complexity, it introduces technical challenges for read alignment and downstream analyses. Given the large number of available tools and algorithms, designing an end-to-end workflow that balances technical accuracy, computational efficiency, and biological relevance is not straightforward.

In this work, four WGBS analysis workflows covering all steps from quality control and read preprocessing to alignment, methylation quantification, and the identification and interpretation of differentially methylated regions (DMRs) were developed. Two widely-used tools for mapping and methylation estimation, Bismark and BSMAP, together with two statistical approaches for DMR detection, DSS and methylKit, were systematically compared.

Although Bismark and BSMAP differed in computational performance, both produced comparable percentages of uniquely mapped reads and similar global methylation estimates. Importantly, a comparison with reference data from ENCODE revealed strong positive correlations ($r > 0.91$) between predicted methylation levels and experimental measurements, supporting the robustness of the developed pipelines. The functional analysis of DMRs further showed that many methylation changes occur in regulatory elements such as promoters and enhancers, highlighting the link between epigenetic variation and tissue-specific gene regulation. Notably, promoters that switched activity state between heart and kidney displayed larger methylation differences compared with those that maintained their status.

Together, these results demonstrate that the workflows developed here provide a flexible, reproducible, and scalable framework for WGBS data analysis. They integrate different alignment and statistical strategies, yield results consistent with reference data, and capture biologically meaningful patterns. In addition, the connection with regulatory annotations from Ensembl makes this approach suitable for comparative epigenomic studies across species, in line with international initiatives such as ENCODE, GENE-SWitCH, and AQUA-FAANG.

Abreviaturas

ADN: Ácido Desoxirribonucleico

ARN: Ácido Ribonucleico

CGI: CpG islands, islas CpG

NGS: Next Generation Sequencing, tecnologías de secuenciación masiva

WGBS: Whole Genome Bisulfite Sequencing, secuenciación de bisulfito de todo el genoma

RRBS: Reduced Representation Bisulfite, secuenciación de bisulfito de representación reducida

PCR: Polimerase Chain Reaction, reacción en cadena de la polimerasa

mC: 5-metilcitosina

OT: hebra superior original

CTOT: hebra complementaria a la hebra superior original

OB: hebra inferior original

CTOB: hebra complementaria a la hebra inferior original

DMRs: Differentially Methylated Regions, regiones diferencialmente metiladas

GEO: Gene Expression Omnibus

HMMs: Hidden Markov Models, modelos ocultos de Markov

AUC: area under curve, área bajo la curva

DMCs: Differentially Methylated Cytosines, citosinas diferencialmente metiladas

PE: paired-end, lecturas apareadas

HPC: high performance computing, entorno de cómputo de alto rendimiento

Índice general

Agradecimientos.....	2
Resumen.....	4
Abstract.....	5
Abreviaturas.....	8
Introducción y marco teórico.....	12
1.1 Epigenética.....	12
1.2 Metilación del ADN y WGBS.....	13
1.3 Análisis de datos de WGBS.....	16
1.4 Hipótesis.....	19
1.5 Objetivo.....	19
1.5.1 Objetivo General.....	19
1.5.2 Objetivos específicos.....	20
Materiales y Métodos.....	21
2.1 Revisión bibliográfica.....	21
2.3 Control de calidad y recorte.....	22
2.4 Alineamiento y post-procesamiento.....	22
2.5 Estimación de niveles de metilación.....	22
2.6 Análisis de metilación diferencial.....	22
2.7 Datos de prueba.....	23
2.8 Datos reales.....	23
2.9 Análisis de patrones de metilación en elementos regulatorios.....	24
2.10 Análisis computacional.....	24
Resultados.....	25
3.1 Revisión del Estado del Arte.....	25
3.1.1 Control de calidad y preprocesamiento de los datos.....	25
3.1.2 Herramientas de alineamiento y cuantificación para datos de secuenciación por bisulfito.....	26
3.1.3 Herramientas para la detección de Regiones Diferencialmente Metiladas (DMRs).....	29
3.1.4 Selección de herramientas para el flujo de análisis.....	33
3.2 Implementación y evaluación del flujo de análisis de datos de WGBS.....	36
3.2.1 Control de calidad inicial con FastQC.....	36
3.2.2 Recorte de adaptadores y limpieza de lecturas con Trim Galore!.....	37
3.2.3 Alineamiento de lecturas a la referencia.....	39
3.2.4 Post alineamiento: combinación de réplicas técnicas y de-duplicación.....	42
3.2.5 Estimación de los niveles de metilación.....	44

3.2.6 Análisis de metilación diferencial.....	45
3.2.7 Comparación con datos de metilación obtenidos de ENCODE.....	48
3.3 Análisis de los patrones de metilación del ADN de elementos regulatorios genómicos anotados en Ensembl, utilizando el flujo de análisis desarrollado.....	51
Discusión y Conclusiones.....	56
Bibliografía.....	58

Capítulo 1

Introducción y marco teórico

1.1 Epigenética

Históricamente, la palabra "epigenética" se utilizó para describir eventos que no podían explicarse mediante los principios genéticos. Conrad Waddington, a quien se le atribuye la creación del término, definió la epigenética como "la rama de la biología que estudia las interacciones causales entre los genes y sus productos, que dan lugar al fenotipo" (Waddington, 1942). Con el tiempo, esta definición ha evolucionado significativamente, reflejando un mejor entendimiento de los mecanismos moleculares que regulan la expresión génica sin alterar la secuencia del ADN. En un sentido amplio, la epigenética puede considerarse un puente entre el genotipo y el fenotipo, ya que modula la expresión de los genes sin modificar su secuencia subyacente. Este fenómeno explica cómo células con el mismo material genético pueden diferenciarse en tipos celulares con funciones específicas y perfiles de expresión génica estables a lo largo del desarrollo. Actualmente, la epigenética se define como el estudio de los cambios en la actividad o función de los genes que no implican modificaciones en la secuencia del ADN, donde tanto el ADN como las proteínas que interactúan con él pueden experimentar modificaciones químicas que alteran los niveles de expresión génica. (Dupont et al., 2009; Tolu et al., 2025).

La regulación epigenética juega un papel clave en la diferenciación celular y en la homeostasis de los organismos multicelulares, permitiendo que células especializadas mantengan perfiles de expresión génica específicos a lo largo del tiempo (Benincasa et al., 2021). A nivel biomolecular, la epigenética involucra macromoléculas estructuralmente cruciales: el ADN, las histonas, proteínas que se entrelazan en nucleosomas que luego se condensan para formar la cromatina, y el ARNnc (ARN no codificante) (Li, 2021).

Uno de los mecanismos epigenéticos más estudiados es la metilación del ADN, una modificación química que si bien no altera su secuencia, desempeña un papel fundamental en procesos como la regulación de la expresión génica, la inactivación del cromosoma X, la impronta genética y el silenciamiento de genes específicos de la línea germinal y de elementos repetitivos (Zhou et al., 2019). Otras modificaciones epigenéticas ampliamente estudiadas son la metilación, acetilación y fosforilación de las histonas que afectan la estructura de la cromatina. Al igual que la metilación del ADN, estas modificaciones alteran la accesibilidad de los genes, lo que repercute directamente en su silenciamiento y/o expresión (Tolu et al., 2025). Los ARNnc también regulan la expresión génica pero mediante la interacción con otras moléculas de ARNnc, ARNm, ADN y proteínas.

La metilación del ADN y las modificaciones en las histonas están reguladas por una compleja red de enzimas clasificadas en escritores, lectores y borradores epigenéticos, que coordinan la activación o represión génica. Los patrones de metilación del ADN y otras modificaciones epigenéticas se mantienen de manera estable a lo largo de la división celular somática y, en algunos casos, pueden heredarse entre generaciones. Sin embargo, alteraciones en estos patrones han sido asociadas con diversas patologías, incluyendo trastornos de impronta genómica y cáncer. Adicionalmente, factores ambientales como el tabaquismo, la actividad física, el consumo de alcohol y la dieta pueden inducir cambios en los patrones de metilación, afectando la expresión génica. Asimismo, la exposición a contaminantes ambientales, como el material particulado y el benceno, ha sido relacionada con cambios en los niveles de metilación del ADN, lo que podría aumentar el riesgo de enfermedades metabólicas y otros trastornos crónicos (Dor y Cedar, 2018; Wu et al., 2021; Zhang et al., 2024). Por ello, comprender cómo se establecen y mantienen estos patrones epigenéticos es fundamental para el estudio de la regulación génica y sus implicaciones en la salud y la enfermedad.

1.2 Metilación del ADN y WGBS

La adición de grupos metilo a los residuos de carbono C5 (5mC) de las citosinas por las ADN metiltransferasas produce la metilación del ADN (Moore et al., 2013). El contexto de secuencia que rodea a una citosina es un factor clave en la regulación de su metilación. Las citosinas que forman parte de un dinucleótido C-G (CpG) suelen estar metiladas (~60-80% en mamíferos) y se regulan de manera diferente a las citosinas en otros contextos. La metilación de CpG generalmente ocurre en ambas cadenas de ADN para mantener los patrones de metilación durante la replicación del ADN. Las regiones de entre 500 y 2000 pares de bases (bp) con alta densidad de CpG se definen como islas CpG (CGI) (Smith et al., 2013; Rauluseviciute et al., 2019).

Una característica importante de las CGI es que exhiben diferentes patrones de metilación dependiendo de su ubicación en el genoma. La mayoría de las CGI que ocurren próximas a los sitios de inicio de transcripción de genes, es decir en los promotores, se encuentran hipometiladas lo que se ha demostrado está asociado con una reducción en la expresión de los transcritos correspondientes. Por el contrario, las CGI en el cuerpo del gen se encuentran generalmente metiladas, impactando positivamente en la expresión génica (Jones, 2012; Yang et al., 2016; Zhang et al., 2024). Al mismo tiempo, diferentes tipos celulares exhiben perfiles de metilación característicos relacionados con procesos celulares. Por ejemplo, los CpGs hipometilados asociados a promotores en diferentes tipos celulares generalmente están vinculados con procesos de mantenimiento celular, mientras que las regiones de promotores consistentemente hipermetiladas están relacionadas con el sistema reproductivo. Por otro lado, los perfiles de metilación específicos de cada tejido participan en funciones particulares del tejido y en la regulación de genes que determinan la identidad celular (Jones, 2012; Yang et al., 2016) (Figura 1.1).

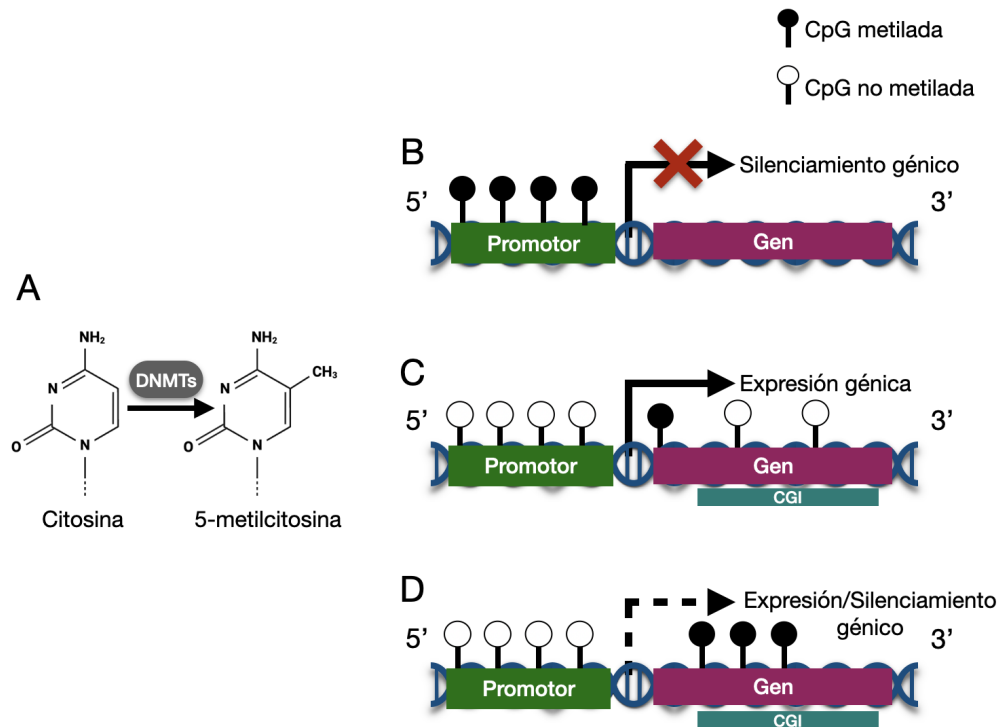


Figura 1.1. A) Representación esquemática simplificada del mecanismo molecular de metilación del ADN. B) La metilación en la región promotora de los genes silencia la expresión génica. C) La metilación en regiones fuera de las islas CpG (no-CpG) y en el cuerpo del gen activa la expresión de los genes. D) La metilación en islas CpG (CGI) en el cuerpo del gen puede activar o silenciar la expresión de los genes.

Existen varios métodos para estudiar el estado de metilación del ADN (Laird, 2010). Entre las primeras técnicas desarrolladas se destacan el enriquecimiento de fragmentos metilados (MeDIP) seguido de hibridación a microarreglos de ADN y MeDIP seguido de secuenciación (MeDIP-Seq). El desarrollo de las tecnologías de secuenciación masiva (o en inglés *Next Generation Sequencing*, NGS) ha permitido avances sin precedentes en la investigación epigenómica, particularmente en el estudio de la metilación del ADN en genomas completos.

Actualmente, los métodos más utilizados se basan en NGS y consisten en distinguir la citosina metilada de la no metilada mediante el tratamiento del ADN con bisulfito de sodio seguido de la secuenciación masiva. Esta estrategia permite estudiar los patrones de metilación de todo el genoma y se conoce como secuenciación de bisulfito de todo el genoma (del inglés *Whole Genome Bisulfite Sequencing*, WGBS). Una alternativa económica a WGBS, útil cuando sólo ciertas regiones del genoma son de interés, es la secuenciación de bisulfito de representación reducida (del inglés *Reduced Representation Bisulfite*, RRBS). RRBS incorpora un paso de digestión del ADN con enzimas de restricción previo al tratamiento de bisulfito que permite reducir la cantidad de ADN a secuenciar (Li et al., 2010; Beck et al., 2021). Si bien RRBS es útil para caracterizar un alto porcentaje de regiones promotoras, la cobertura que suele producir es baja principalmente en elementos distales que también participan en la regulación de la expresión de genes como los *enhancers*. A su vez, a diferencia de WGBS, RRBS no permite estudiar ni el contexto ni los niveles absolutos de metilación principalmente en regiones distales (Rauluseviciute et al., 2019).

El tratamiento del ADN con bisulfito de sodio convierte las citosinas no metiladas en uracilos, mientras que las citosinas metiladas permanecen sin modificaciones. Durante la secuenciación, estos uracilos se leen como timinas por la ADN polimerasa, por lo que al amplificar el ADN tratado con bisulfito mediante PCR, los productos resultantes tienen citosinas no metiladas que aparecen como timinas. Al comparar el ADN modificado con la secuencia original, se puede inferir el estado de metilación del ADN original. Para un locus genómico dado, el tratamiento con bisulfito y la amplificación por PCR generan hasta cuatro cadenas individuales de ADN que pueden terminar en un experimento de secuenciación (Figura 1.2).

Las bibliotecas de secuenciación de bisulfito pueden clasificarse en dos tipos distintos (Chen et al., 2010):

- Bibliotecas direccionales (o en inglés, *directional libraries*): La biblioteca de secuenciación se genera de manera direccional, lo que significa que las lecturas de secuenciación corresponderán a una versión convertida con bisulfito de la hebra original directa o inversa (Lister et al., 2009).
- Bibliotecas no direccionales (o en inglés, *non-directional libraries*): No se preserva la especificidad de la hebra, por lo que las cuatro hebras de ADN posibles pueden ser secuenciadas con una frecuencia aproximadamente igual.

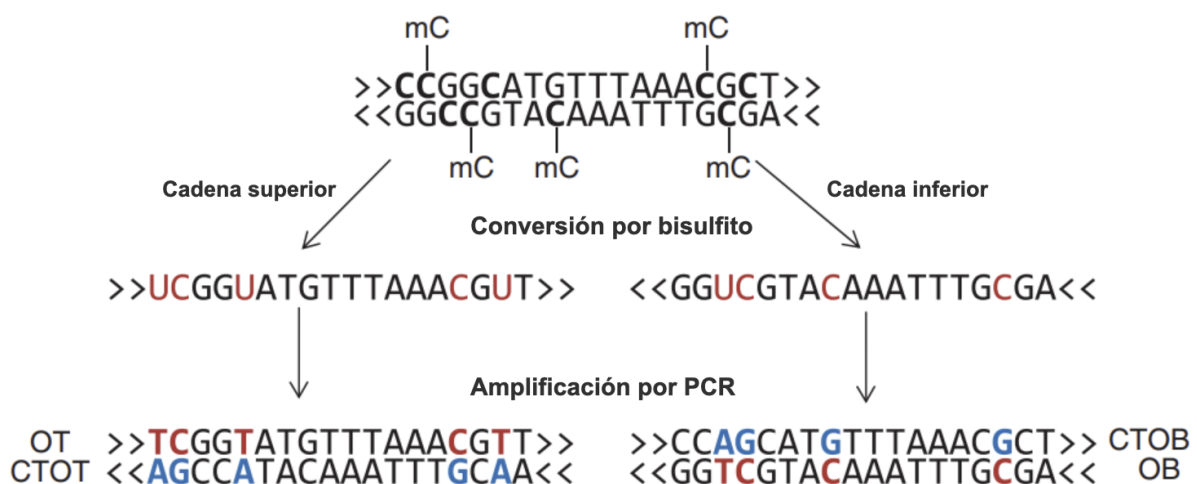


Figura 1.2. Efecto del tratamiento con bisulfito en el ADN. La conversión con bisulfito del ADN genómico y la posterior amplificación por PCR dan lugar a dos productos de PCR y hasta cuatro fragmentos de ADN potencialmente diferentes para cualquier locus dado. Las citosinas metiladas son resistentes a la conversión con bisulfito y pueden utilizarse como un indicador del estado de metilación del ADN. mC: 5-metilcitosina; OT: hebra superior original; CTOT: hebra complementaria a la hebra superior original; OB: hebra inferior original; y CTOB: hebra complementaria a la hebra inferior original. Ejemplo extraído de Krueger et al., 2012.

1.3 Análisis de datos de WGBS

Independientemente del tipo de biblioteca utilizada, como resultado de un experimento de WGBS, se obtienen millones de lecturas de secuencias de ADN tratadas con bisulfito. Sin embargo, estas lecturas por sí solas no son más que fragmentos de ADN repetidos y desordenados, ya que se desconoce su ubicación en el genoma. Por lo tanto, para poder interpretar y analizar los patrones de metilación del ADN, es fundamental llevar a cabo un procesamiento exhaustivo de estos datos.

Un flujo de trabajo completo (también conocido como *end-to-end pipeline*) para el análisis de datos de WGBS está compuesto por diversos pasos ilustrados en la Figura 1.3. En general, el procesamiento comienza con la evaluación de la calidad de los datos, a partir de la cual luego se realiza la eliminación de adaptadores de secuenciación y el recorte de bases pobremente identificadas con el objetivo de mejorar la calidad de las lecturas para los análisis posteriores, optimizando la fiabilidad de los resultados. Luego, se realiza el proceso de alineamiento contra el genoma de referencia, en el cual se asigna al menos una posición en el genoma a cada lectura de secuenciación mediante su comparación con la secuencia del genoma conocido. Posteriormente, si es necesario, se combinan los archivos de alineamiento de diferentes experimentos que provienen de la misma muestra biológica (réplicas técnicas) en un único archivo, aumentando así la cobertura. Además, se realiza un paso de de-duplicación, que elimina lecturas exactamente idénticas que mapean a la misma región que podrían haberse generado durante la amplificación por PCR. Luego, se realiza la cuantificación de los niveles de metilación, en la cual se estima el porcentaje de metilación en cada sitio CpG u otras regiones del genoma, permitiendo la identificación de los patrones de metilación en diversas regiones del ADN. Finalmente, se analizan los perfiles de metilación obtenidos en diferentes condiciones experimentales para detectar diferencias significativas entre ellos e identificar posibles cambios en la metilación asociados a procesos biológicos.

Se han desarrollado numerosas herramientas para abordar diversas etapas del proceso, las cuales requieren diferentes formatos de entrada y salida para los archivos de datos e involucran distintos lenguajes de programación (Rauluseviciute et al., 2019; Lin et al., 2023). Herramientas computacionales como TrimGalore! (Krueger, 2012), Trimmomatic (Bolger et al., 2014) y FastQC (Andrews, 2010) se utilizan para eliminar las secuencias de adaptadores y controlar la calidad de las lecturas. El alineamiento de las secuencias al genoma de referencia se realiza con herramientas de alineamiento específicas para WGBS, como Bismark (Krueger y Andrews, 2011), BS Seeker (Chen et al., 2010) y BSMAP (Xi y Li, 2009). Por lo general, estas herramientas también proveen funcionalidades para el filtrado de lecturas post-alineamiento y para generar los resultados del estado de metilación de cada citosina a lo largo del genoma.

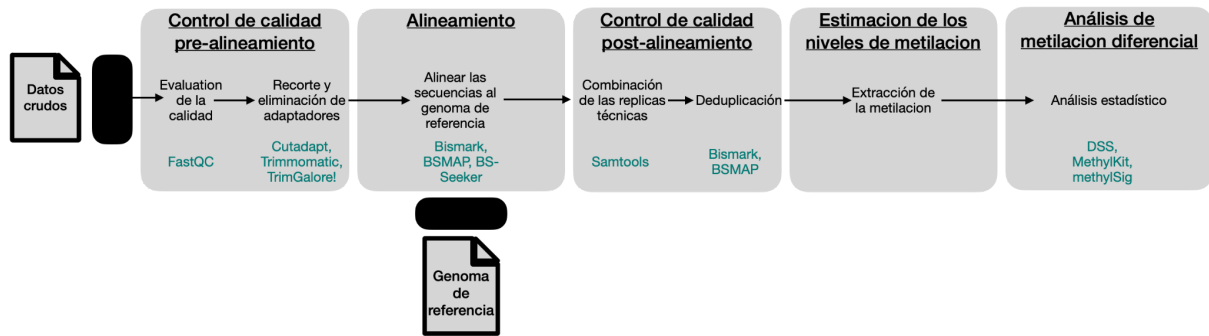


Figura 1.3. Flujo de trabajo para el análisis de datos de WGBS.

Una de las etapas críticas del análisis es el alineamiento de las lecturas al genoma. Como parte del protocolo de WGBS, el tratamiento del ADN con bisulfito convierte las citosinas no metiladas a timinas, lo cual reduce la complejidad de las lecturas de secuenciación, aumentando la posibilidad de mapeo ambiguo y volviendo más desafiantes el alineamiento contra referencia y los análisis posteriores. Además, el espacio de búsqueda se incrementa significativamente ya que las secuencias tratadas con bisulfito no son completamente complementarias entre las hebras de ADN. Dado que luego de la amplificación por PCR se generan cuatro cadenas distintas, el mapeo requiere hasta cuatro instancias de alineamiento de cadena diferentes para cada secuencia analizada (Xi y Li, 2009). Existen dos enfoques principales para mapear secuencias: el mapeo por "wild card" y el mapeo de "tres letras" (Figura 1.4). El primero reemplaza las citosinas (Cs) en la secuencia del ADN genómico con la letra "Y", que coincide tanto con Cs como con Ts en la secuencia de lectura, o modifica la matriz de puntuación de alineamiento de manera que las discrepancias entre Cs en la secuencia del ADN genómico y Ts en la secuencia de lectura no sean penalizadas. Este enfoque logra una mayor cobertura genómica, pero con un mayor riesgo de introducir sesgos hacia estimaciones más altas de metilación; sin embargo, mantiene las citosinas (Cs) en las lecturas convertidas por bisulfito y aumenta la complejidad de la secuencia hasta un nivel que garantiza alineamientos únicos con los genomas de referencia. Por otro lado, el mapeo de "tres letras" simplifica el alineamiento con bisulfito convirtiendo todas las citosinas (Cs) en timinas (Ts) en las lecturas y en ambas cadenas de la secuencia del ADN genómico, permitiendo realizar el alineamiento con un alfabeto de tres letras (A, G y T) y utilizar un alineador como Bowtie (Langmead et al., 2009). Esta reducción conduce a asignaciones falsas positivas en el genoma y a llamadas de metilación menos sensibles, pero elimina las Cs en las lecturas convertidas por bisulfito, disminuyendo la complejidad de la secuencia, lo que aumenta la posibilidad de posiciones de alineamiento ambiguas. (Bock, 2012; Grehl et al., 2020).

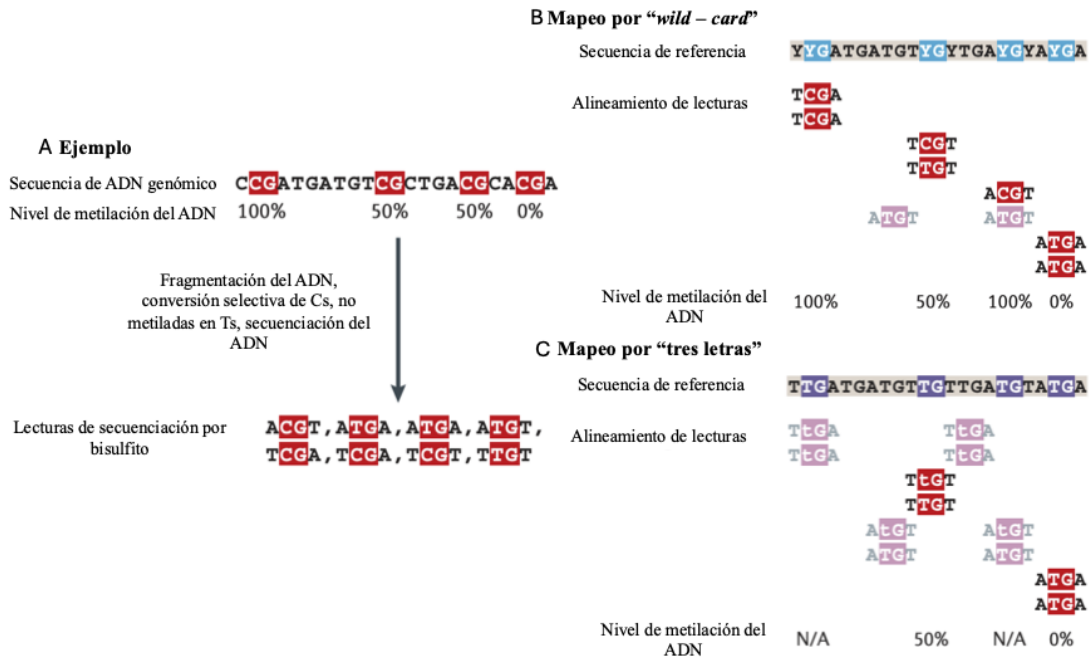


Figura 1.4. Dos estrategias alternativas para la alineación de bisulfito. A) Ejemplo de secuenciación por bisulfito de un fragmento de ADN con cuatro CpGs y ocho lecturas, usando fragmentos cortos para simplificar la visualización. B) Alineamiento con un alineador de “wild-card” que reemplaza Cs por Y. Las lecturas ambiguas se descartan, y la metilación se calcula como el porcentaje de Cs alineados en lecturas únicas. Esto puede causar asignaciones incorrectas de metilación, como en el tercer CpG. C) Alineamiento con un alineador de tres letras que reemplaza Cs por T. Reduce errores de metilación, pero muchas lecturas se descartan por mapear a múltiples posiciones, dejando algunos CpGs sin valores. La asignación aleatoria de lecturas ambiguas puede mejorar la precisión, pero con mayor incertidumbre en comparación con el alineador de comodín. Ejemplo extraído de Bock, 2012.

Los análisis posteriores al alineamiento y las herramientas a utilizar estarán delimitados por el objetivo biológico del experimento. Un objetivo frecuente en el análisis de datos de metilación del ADN es identificar los cambios en los patrones de metilación entre distintas condiciones experimentales, como diferentes tejidos, distintas etapas del desarrollo embrionario, e incluso entre un estado sano y un estado patológico. Las regiones diferencialmente metiladas (del inglés *Differentially Methylated Regions*, DMRs) pueden tener una longitud variable dependiendo del interés biológico y los métodos bioinformáticos utilizados para su identificación. Una característica del análisis de DMRs es que se basa en pruebas de hipótesis que analizan conteos de bases metiladas y no metiladas en un sitio determinado.

Un aspecto importante a tener en cuenta es que debido al costo relativamente alto de los experimentos de WGBS, el número de réplicas biológicas suele ser pequeño. Esto representa una limitación ya que puede provocar una estimación inestable de la varianza dentro de cada condición experimental y generar resultados no deseables en las pruebas de hipótesis. Como un primer acercamiento, la prueba de Fisher permite identificar metilación diferencial en muestras sin réplicas. Sin embargo, no considera la variabilidad biológica dentro de cada condición, lo que puede aumentar la tasa de falsos positivos. Debido a que la distribución de los datos de metilación es bimodal, los métodos que utilizan la distribución binomial o β -binomial tienden a mostrar un mejor desempeño (Singer, 2019). Dentro de las herramientas

para la identificación de DMRs se encuentran methylSig (Park et al., 2014), methylKit (Akalin et al., 2012), Metilene (Jühling et al., 2016), las cuales se basan en diferentes pruebas estadísticas como la prueba T, prueba de Fisher, prueba de Wilcoxon y modelos de regresión logística (Bock, 2012; Robinson et al., 2014; Yang X et al., 2016; Yang Y et al., 2023).

A pesar de la disponibilidad de numerosas herramientas bioinformáticas, la construcción de flujos de análisis integrados para WGBS sigue siendo un desafío. La diversidad de algoritmos disponibles, combinada con las especificidades técnicas de cada herramienta, hace que seleccionar y optimizar un flujo de trabajo sea una tarea compleja. Además, la falta de estandarización en la integración de datos de metilación con anotaciones regulatorias limita su aplicación en estudios funcionales.

En los últimos años, la comunidad científica ha adoptado un enfoque colaborativo a gran escala para recopilar y analizar grandes volúmenes de datos. Iniciativas como ENCODE han generado grandes volúmenes de datos de secuenciación de bisulfito del genoma completo (WGBS), proporcionando recursos esenciales para el estudio de la metilación del ADN en diferentes contextos biológicos. Estos esfuerzos colaborativos han permitido mejorar la comprensión de la regulación epigenética y su relación con diversas condiciones fisiológicas y patológicas.

Este proyecto aborda estas limitaciones mediante el diseño y la implementación de un flujo de análisis de datos de principio a fin, que no solo permita la cuantificación de niveles de metilación del ADN en genomas de vertebrados, sino que también facilite la integración de estos perfiles con anotaciones regulatorias de Ensembl. La aplicación de este flujo de trabajo a datos públicos provenientes de proyectos como ENCODE permitirá evaluar la utilidad de los perfiles de metilación como indicadores de actividad regulatoria en diferentes tipos celulares.

1.4 Hipótesis

La hipótesis que se plantea es que el desarrollo de un flujo de análisis de datos de secuenciación de bisulfito de genoma completo permitirá cuantificar con precisión los niveles de metilación del ADN en genomas de vertebrados y establecer relaciones significativas entre los patrones de metilación y los elementos regulatorios genómicos, como promotores, *enhancers* y regiones de cromatina abierta.

1.5 Objetivo

1.5.1 Objetivo General

El objetivo general de este proyecto es diseñar e implementar un flujo de análisis de datos de principio a fin (o *end-to-end*) para la cuantificación de los niveles de metilación del ADN en genomas vertebrados.

1.5.2 Objetivos específicos

Los objetivos específicos del proyecto se describen a continuación:

- Analizar el estado del arte de las distintas herramientas disponibles para el procesamiento de datos de WGBS;
- Diseñar, implementar y evaluar un flujo de análisis de datos de WGBS para obtener perfiles de metilación del ADN en regiones específicas de genomas de vertebrados;
- Utilizar el flujo de análisis desarrollado para el estudio de los patrones de metilación del ADN de elementos regulatorios genómicos anotados en Ensembl, como promotores, *enhancers* y regiones de cromatina abierta.

Capítulo 2

Materiales y Métodos

2.1 Revisión bibliográfica

Se realizó una revisión bibliográfica para seleccionar las herramientas y métodos más adecuados para el análisis de datos de WGBS. La búsqueda se llevó a cabo en bases de datos especializadas como PubMed¹ y Google Scholar², priorizando artículos de alto impacto publicados en revistas como Nature³, Nucleic Acids Research⁴ y Bioinformatics⁵. Se seleccionaron principalmente revisiones sistemáticas y estudios comparativos que evaluaran el rendimiento, disponibilidad y facilidad de uso de las herramientas de análisis de metilación. Se analizaron los resultados reportados y en base a los criterios listados anteriormente, se seleccionaron los métodos más adecuados para cada etapa.

2.2 Diseño e implementación del flujo de análisis

Se desarrollaron cuatro flujos de trabajo alternativos para el procesamiento y análisis de datos de WGBS, mediante la variación de las herramientas de alineamiento/cuantificación y detección de DMRs. Específicamente, se utilizaron Bismark y BSMAP para el alineamiento de lecturas y estimación de los niveles de metilación, mientras que DSS y methylKit se emplearon para la identificación de regiones diferencialmente metiladas (DMRs). Los flujos de trabajo evaluados fueron nombrados: Bismark+DSS, Bismark+methylKit, BSMAP+DSS y BSMAP+methylKit.

Todos los scripts desarrollados para la implementación de estos flujos de trabajo se encuentran disponibles en un repositorio público de GitHub: https://github.com/SofiaReynoso/WGBS_analysis_pipeline.git

¹ <https://pubmed.ncbi.nlm.nih.gov/>

² <https://scholar.google.com/>

³ <https://www.nature.com/>

⁴ <https://academic.oup.com/nar>

⁵ <https://academic.oup.com/bioinformatics>

2.3 Control de calidad y recorte.

Se seleccionó FastQC (v0.12.0)⁶ para la evaluación de calidad y Trim Galore! (v0.6.10)⁷ para el recorte de adaptadores y bases de baja calidad.

2.4 Alineamiento y post-procesamiento.

Se compararon dos herramientas de alineamiento: Bismark + Bowtie2 (con de-duplicación mediante deduplicate_bismark) y BSMAP (con de-duplicación integrada en methratio.py). Se utilizaron samtools merge para combinar réplicas técnicas en un único archivo BAM por muestra.

2.5 Estimación de niveles de metilación

Los niveles de metilación se estimaron con bismark_methylation_extractor (salida en .cov y .bedGraph) y methratio.py (archivo de texto tabulado).

2.6 Análisis de metilación diferencial

Se utilizaron los paquetes de R DSS y methylKit. DSS aplica un modelo beta-binomial jerárquico, mientras que methylKit emplea regresión logística con o sin sobredispersión. Se calcularon coeficientes de correlación de Pearson y se generaron gráficos con ggplot2.

Para adaptar las salidas de los alineadores al formato requerido por los paquetes de análisis, se realizaron transformaciones específicas. En el caso de Bismark, se utilizó el archivo de cobertura (*.bismark.cov), que contiene por cada CpG la información de cromosoma, posición inicial y final, porcentaje de metilación, número de lecturas metiladas y no metiladas. A partir de estos campos se generaron los archivos de entrada de DSS, que requieren las columnas: cromosoma, coordenada genómica, número total de lecturas y número de lecturas metiladas. Para methylKit, los mismos archivos de cobertura fueron leídos directamente mediante la función methRead, especificando el contexto CpG y un umbral mínimo de cobertura.

En el caso de BSMAP, el archivo de salida de methratio.py incluye coordenada, contexto de metilación, conteos de C y T, y proporciones de metilación. Para generar las entradas de methylKit, dichos campos se reordenaron y transformaron para obtener un archivo tabulado con identificador de posición, cromosoma, coordenada, hebra, cobertura, porcentaje de C y

⁶ <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

⁷ https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

porcentaje de T, siguiendo el formato requerido para la función `methRead`. Para DSS, se construyeron archivos simplificados que contienen únicamente cromosoma, posición, número total de lecturas y número de lecturas metiladas. En ambos casos, las modificaciones se realizaron con comandos basados en `awk`.

2.7 Datos de prueba

Para validar cada módulo durante la implementación de los flujos de análisis, se utilizaron datos de prueba disponibles públicamente, incluyendo un subconjunto de lecturas *paired-end* de un experimento de referencia (GSM1027571) y archivos de ejemplo proporcionados por los desarrolladores de las diferentes herramientas.

2.8 Datos reales

Los datos de secuenciación de bisulfito de genoma completo (WGBS) fueron descargados de la base de datos ENCODE y están disponibles en el Gene Expression Omnibus (GEO) bajo los identificadores de acceso GSE82658 (corazón) y GSE82451 (riñón). Se utilizaron muestras de tejido de riñón y corazón de ratón (*Mus musculus*) de la cepa *B6NTac* y *B6NCrl/Lap* en el día postnatal 0. Cada muestra de tejido contiene dos réplicas isogénicas (o réplicas biológicas) con dos réplicas técnicas. Todos los experimentos fueron realizados utilizando secuenciación single-end de 100 pb (SE100). Los datos crudos (archivos FASTQ) fueron procesados en esta tesis. Los datos de referencia para el alineamiento y análisis de metilación se basaron en el ensamblaje del genoma de ratón *GRCm39*, descargado de Ensembl.

Para los datos de corazón, se emplearon las muestras de la serie GSE82658, correspondientes al experimento ENCSR397YEG⁸:

- Réplica 1: ENCFF001ZVM (718506938 lecturas; 71.8 Gbp), ENCFF001ZVP (652720883 lecturas; 65.2 Gbp).
- Réplica 2: ENCFF001ZVR (641332810 lecturas; 64.1 Gbp), ENCFF001ZVT (661804193 lecturas; 66.1 Gbp)

Para los datos de riñón, se emplearon las muestras de la serie GSE82451, correspondientes al experimento ENCSR128HOP⁹:

- Réplica 1: ENCFF001ZVU (643968828 lecturas; 64.3 Gbp), ENCFF001ZVX (706610024 lecturas; 70.6 Gbp)
- Réplica 2: ENCFF001ZWA (635147721 lecturas; 63.5 Gbp), ENCFF001ZWD (654339599 lecturas; 65.4 Gbp)

⁸ <https://www.encodeproject.org/ENCSR397YEG/>

⁹ <https://www.encodeproject.org/ENCSR128HOP/>

2.9 Análisis de patrones de metilación en elementos regulatorios

Se utilizaron anotaciones regulatorias de Ensembl para ratón (ensamblado *GRCm39*¹⁰), incluyendo promotores, enhancers y regiones de cromatina abierta. En total, se encuentran anotados 31.408 promotores, 161.385 enhancers y 13.677 regiones de cromatina abierta.

El estado de actividad de los promotores reportados para las muestras de riñón y corazón se integró a los perfiles de metilación. En R se usaron los paquetes GenomicRanges, rtracklayer, ggplot2 y patchwork. En Python se utilizaron las librerías pandas, matplotlib y seaborn para la integración y visualización de datos.

2.10 Análisis computacional

El análisis computacional se llevó a cabo utilizando los nodos "Genomics Nodes" y "Analytics Nodes" en el clúster de computación de alto rendimiento Quest de Northwestern University. Para las tareas de control de calidad y cuantificación requirieron entre 4 a 6 hilos y hasta 100 GB de RAM. El análisis de metilación diferencial requirió de 14 nodos con 24 Gb de RAM.

¹⁰https://ftp.ensembl.org/pub/current_regulation/mus_musculus/GRCm39/annotation/Mus_musculus.GRCm39.regulatory_features.v114.gff3.gz

Capítulo 3

Resultados

3.1 Revisión del Estado del Arte

Como se ha descrito en la Introducción, por lo general, un flujo de trabajo completo para el análisis de datos de WGBS está compuesto por diversas etapas que pueden agruparse en: pre-procesamiento de las lecturas de secuenciación, su alineamiento contra el genoma de referencia, el post-procesamiento de los alineamientos, la cuantificación de los perfiles de metilación y la detección de regiones diferencialmente metiladas (DMRs). Dada la vasta disponibilidad de herramientas actualmente disponibles para cada una de estas etapas, en una primera instancia se realizó una revisión exhaustiva del estado del arte con el fin de determinar cuáles son las más adecuadas teniendo en cuenta los siguientes criterios:

- Herramientas de código abierto, ampliamente aceptadas en la comunidad y con mantenimiento actual
- Accesibilidad y facilidad de uso
- Tiempos de ejecución y requerimientos computacionales
- Aplicabilidad en vertebrados, especialmente en humano, ratón, cerdo, pollo, vaca, que son las especies de interés del grupo Ensembl Regulation.

Los resultados detallados obtenidos del análisis bibliográfico se presentan a continuación.

3.1.1 Control de calidad y preprocesamiento de los datos

El preprocesamiento constituye una etapa fundamental en el análisis de datos de WGBS, ya que condiciona la calidad del alineamiento y, en consecuencia, la precisión en la cuantificación de los niveles de metilación del ADN. Esta etapa incluye el control de calidad inicial, la remoción de adaptadores de secuenciación, el recorte y/o eliminación de las lecturas que no cumplan con criterios mínimos de calidad.

En el contexto de la cuantificación de perfiles de metilación, la calidad de las lecturas antes del alineamiento juega un papel clave. Un error frecuente es confundir bases mal llamadas con conversiones reales de C a T, lo que puede generar falsos positivos o negativos en la detección de sitios metilados. Por ello, es crucial evaluar la calidad de las bases (del inglés, *Phred quality score*), ya que errores sistemáticos en el *base-calling* (proceso de asignar una base a cada señal detectada durante la secuenciación) pueden comprometer la fidelidad del análisis.

Durante la secuenciación es posible además que se lean parcial o completamente algunos adaptadores que pueden interferir con el alineamiento (ya que formarán parte de las lecturas) o introducir conversiones erróneas de C a T. Otro problema habitual es la presencia de bases con baja calidad en los extremos de las lecturas que pueden ser confundidas con conversiones

reales durante el análisis de metilación. Por lo tanto, la correcta eliminación de adaptadores y el recorte de regiones de baja calidad no solo mejoran las tasas de alineamiento, sino que también reducen la probabilidad de errores en la cuantificación de metilación, al evitar la interpretación errónea de bases mal llamadas como conversiones genuinas de citosina a timina.

Para abordar estos problemas, existen múltiples herramientas, muchas de las cuales son ampliamente utilizadas en otros flujos de análisis de datos de NGS, como RNA-seq, ATAC-seq o ChIP-seq, lo que demuestra su robustez y su aceptación en la comunidad bioinformática.

Una de las principales herramientas útil para evaluar la calidad inicial de las lecturas es FastQC¹¹, la cual permite identificar problemas como pérdida de calidad en los extremos de las lecturas, presencia de adaptadores, sesgos en la composición de bases, y otras anomalías que podrían afectar los resultados del análisis. Para ello, dado un archivo (o un par de archivos, en el caso de experimentos *pair-end*) de lecturas de secuenciación, FastQC genera un reporte HTML con distintas medidas estadísticas y gráficos de análisis exploratorio. Ahora bien, si se detecta algún problema con las lecturas analizadas, es necesario usar otras herramientas para tratar de solucionarlo ya que FastQC no provee herramientas para manipular y/o editar los archivos y/o las lecturas que éstos contiene

Entre las alternativas disponibles para editar lecturas de secuenciación, las herramientas que más se destacan son:

- Trim Galore!¹²: Es un *wrapper* que combina Cutadapt con FastQC. Automatiza la eliminación de adaptadores y el recorte por calidad, y genera informes antes y después del recorte de bases. Está optimizado para datos de bisulfito, aunque su uso se ha extendido a otros tipos de datos.
- Cutadapt: Programa independiente que detecta y recorta adaptadores, además de permitir el recorte de bases con baja calidad. Es el motor utilizado por Trim Galore! (Martin, 2011).
- Trimmomatic: Permite una configuración altamente flexible para recorte por calidad, longitud mínima y eliminación de adaptadores. Se emplea frecuentemente en estudios de RNA-seq y otros análisis de NGS (Bolger et al., 2014).

3.1.2 Herramientas de alineamiento y cuantificación para datos de secuenciación por bisulfito

El alineamiento preciso de las lecturas de secuencias tratadas con bisulfito es esencial para la posterior interpretación de los patrones de metilación del ADN. Tanto la correcta identificación de sitios metilados como la precisa determinación de la ubicación de cada lectura en el genoma son claves para caracterizar el estado epigenético de una muestra. Dado que el alineamiento contra el genoma de referencia es uno de los pasos más críticos en el análisis de datos de WGBS, el análisis bibliográfico se ha centrado en la exploración de herramientas de alineamiento, con el objetivo de determinar cuál de ellas ofrece los mejores resultados en términos de precisión, rendimiento y facilidad de uso.

¹¹ <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

¹² https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

En la literatura científica existen numerosos estudios que abordan la evaluación y comparación de herramientas de alineamiento para datos de secuenciación de bisulfito. Dichos trabajos analizan el desempeño de tales métodos considerando múltiples criterios: la sensibilidad y especificidad en la detección de sitios metilados, la exactitud en la estimación de niveles de metilación, cantidad de lecturas alineadas de manera única, así como también métricas como el tiempo de ejecución, el consumo de memoria RAM y CPU, y la robustez frente a variaciones en la calidad o longitud de las lecturas. Gong et al. (2022) compararon 14 algoritmos de alineamiento ampliamente utilizados en estudios de metilación en mamíferos, incluyendo humanos, bovinos y cerdos. Para el análisis utilizaron datos tanto simulados como reales. El objetivo fue evaluar en forma integral el desempeño de estos algoritmos en términos de tiempo de ejecución, uso de memoria, lecturas alineadas de manera única, lecturas alineadas de forma insatisfactoria, precisión de mapeo, sensibilidad y F1-score. En cuanto a la velocidad y el consumo de memoria, variables que suelen estar inversamente correlacionadas, Walt (Chen et al., 2016) y BSMAP fueron los alineadores más rápidos, mientras que BSSeeker2-bwt2-local fue el más lento. Los autores reportaron además que BSMAP, aunque rápido, consume más memoria en comparación con Bismark-bwt2-e2e, Bwa-meth y BSSeeker2-bwt.

En este estudio, Bismark-bwt2-e2e presentó una mayor cantidad de lecturas alineadas de manera única en comparación con BSSeeker2. Además, aunque bwt2-e2e fue más lento que his2, reportó un mayor número de lecturas alineadas de manera única manteniendo una precisión de mapeo, sensibilidad y F1-score similares. En términos generales, los alineadores Bwa-meth, BSBolt (Farrell et al., 2021), BSMAP, Bismark-bwt2-e2e y Walt lograron una precisión del mapeo, sensibilidad y F1-score superiores al 90 % en humanos, bovinos y porcinos, lo que los posiciona como herramientas robustas para estudios de metilación en mamíferos.

Además de estos estudios en mamíferos, se han realizado comparaciones en plantas para evaluar la variabilidad interespecífica en el proceso de alineamiento del ADN metilado. Grehl et al. (2020) reportaron una evaluación similar en plantas mostrando la variabilidad en el rendimiento de herramientas de alineamiento de datos de WGBS entre especies y tamaños genómicos. En este estudio, realizado con datos simulados de *Arabidopsis thaliana*, *Brassica napus* (colza), *Glycine max* (soja), *Solanum tuberosum* (papa) y *Zea mays* (maíz), se evaluó el impacto de diferentes tasas de conversión de bisulfito, tasa de error de secuenciación y el número máximo de desajustes permitidos en el rendimiento de los alineadores. Este estudio puso en evidencia que tanto el tiempo de ejecución como el consumo de memoria varían ampliamente entre las herramientas, especialmente en genomas grandes como el de maíz. Por ejemplo, Segemehl resultó inviable para genomas grandes debido a su elevado requerimiento de memoria (mayor a 40 GB) y tiempo de procesamiento (al menos 80 horas).

Herramientas como BSMAP, GEM3 y GSNAP mostraron un bajo incremento en el tiempo de ejecución con el aumento del tamaño genómico, pero a costa de un mayor uso de memoria. Por otro lado, Bismark demostró ser una herramienta robusta, con bajo consumo de memoria, buen rendimiento en tiempo de ejecución y alta precisión en todos los genomas evaluados, aunque con una leve disminución en precisión en genomas altamente repetitivos como el de maíz. Esta pérdida de precisión puede explicarse por la elevada proporción de secuencias repetidas en dicho genoma (~75%), las cuales representan un desafío para la asignación única de lecturas. Además, se observó que algunos alineadores tienden a generar alineamientos

falsos positivos en regiones ricas en repeticiones (como GEM3 y GSNAP), mientras que otros excluyen dichas regiones, afectando la cobertura. En cuanto a la detección de sitios CpG, herramientas como Bwa-meth y BS-Seeker2, presentaron en este trabajo tasas de error elevadas en genomas vegetales grandes y ricos en repeticiones, lo que contrasta con su buen desempeño reportado en datos humanos.

En un estudio reciente, Qun-ting Lin et al. (2023) evaluaron cinco herramientas ampliamente utilizadas para el alineamiento de datos de secuenciación de bisulfito en mamíferos: BSMAP, Bismark, BatMeth2, BSBolt y BS-Seeker2. El estudio generó un total de ocho metilomas (cuatro humanos de la línea celular HEK293 y cuatro de hígado de ratón adulto), utilizando dos plataformas de secuenciación de nueva generación: GenoLab M (GeneMind Biosciences) y NovaSeq 6000 (Illumina). Esta configuración experimental permitió una comparación cruzada entre plataformas y algoritmos, bajo condiciones controladas.

El análisis incluyó múltiples métricas: tiempo computacional, porcentaje de alineamiento único, porcentaje de lecturas duplicadas, profundidad y cobertura genómica, y porcentaje de citosinas metiladas detectadas por cada herramienta. En términos de desempeño, BatMeth2 y BSBolt mostraron los mayores porcentajes de alineamiento (hasta un 91%), mientras que BSMAP destacó por su rapidez computacional y su alta concordancia entre réplicas y plataformas. Bismark, por su parte, aunque detectó un porcentaje menor de sitios metilados (~65%), se caracterizó por su estricto filtrado de lecturas (requiriendo mapeo único en ambos extremos de los pares), baja tasa de duplicación (~2%), y su capacidad para generar informes integrales del análisis, lo cual lo convierte en una herramienta especialmente amigable para el usuario. Aunque BS-Seeker2 fue incluido en la comparación, su rendimiento fue menos destacado en los resultados cuantitativos presentados.

A modo de resumen y en base a los estudios previamente descritos, la Tabla 3.1 presenta una comparación de cinco herramientas ampliamente utilizadas para el alineamiento de datos de WGBS. Entre ellas, Bismark, introducida en 2011, se encuentra dentro de las más citadas, con más de 4400 citas y actualizaciones recientes hasta 2024, lo cual refleja un desarrollo activo y una comunidad de usuarios consolidada. Esta herramienta emplea la estrategia denominada “tres letras”, en la que se generan versiones modificadas del genoma de referencia para permitir una alineación eficiente mediante Bowtie2. Su rendimiento puede considerarse intermedio en términos de velocidad de ejecución, y presenta una demanda moderada de memoria, lo que la hace adecuada para análisis de mediana a gran escala. Tiene la ventaja de presentar herramientas para el control de calidad post-alineamiento como por ejemplo, el sesgo introducido por la PCR.

Por otro lado, BSMAP, publicado en 2009, utiliza una estrategia tipo “*wild-card*”, permitiendo una alineación rápida a costa de un mayor consumo de memoria RAM. Cuenta con un menor número de citas en comparación con Bismark y su última actualización registrada fue en 2020. Al igual que Bismark, durante el proceso de estimación de los niveles de metilación, tiene la posibilidad de realizar un paso de de-duplicación.

Por su parte, BS-Seeker2 y Bwa-meth emplean estrategias similares (“tres letras”) basadas en Bowtie2 y BWA-MEM, respectivamente. No obstante, BS-Seeker2 ha mostrado limitaciones en cuanto a velocidad, mientras que Bwa-meth ofrece un rendimiento más equilibrado con velocidad media y bajo consumo de memoria, y se ha mantenido activo hasta 2023. En el

caso de BSBolt, mostró una precisión significativamente menor en la identificación de sitios CpG en comparación con otros algoritmos, afectando la confiabilidad de los resultados.

Tabla 3.1. Comparación de herramientas de alineamiento y cuantificación para datos de secuenciación por bisulfito. Se listan cinco alineadores ampliamente utilizados en análisis de metilación de ADN: BSMAP, Bismark, BSBolt, BS-Seeker2 y Bwa-meth.

Herramienta	Año	Citas	Última actualización	Herramienta	Estrategia	Velocidad (tiempo ejecución)	Consumo de RAM
BSMAP	2009	687	2020	SOAP	Wild-card	Alta	Alta
Bismark	2011	4475	2024	Bowtie2	tres letras	Media	Media
BSBolt	2021	56	2023	BWA-MEM	tres letras versión modificada de BWA-MEM y HTSLIB	Alta	Media
BS-Seeker2	2013	248	2018	Bowtie2	tres letras	Baja	Baja
Bwa-meth	2014	186	2023	BWA-MEM	tres letras	Media	Baja

3.1.3 Herramientas para la detección de Regiones Diferencialmente Metiladas (DMRs)

Una vez que se obtienen los niveles de metilación por base, generalmente se procede a estudiar las alteraciones en los perfiles de metilación. Cuando se dispone de múltiples grupos de muestras, es común que este análisis se centre en identificar bases o regiones con diferentes proporciones de metilación entre las muestras, conocidas como regiones diferencialmente metiladas (DMRs, del inglés *differentially methylated regions*).

La detección de metilación diferencial se realiza habitualmente comparando la proporción de citosinas metiladas entre un grupo de prueba y un grupo control. Las herramientas más utilizadas actualmente y sus principales características se resumen en la Tabla 3.2. En el caso más simple, donde no se cuenta con réplicas biológicas, es común aplicar pruebas estadísticas como la prueba exacta de Fisher, implementada en herramientas como methylKit (Akalin et al., 2012) y RnBeads (Assenov et al., 2014). Alternativamente, es posible utilizar modelos ocultos de Markov (HMMs, del inglés *Hidden Markov Models*) como MethPipe (Song et al., 2013). No obstante, al no considerar la variabilidad biológica entre muestras, estos enfoques presentan limitaciones en estudios con mayor complejidad experimental (Robinson et al., 2014).

Cuando se dispone de réplicas, los métodos basados en regresión resultan más apropiados, ya que permiten modelar la relación entre los niveles de metilación y las condiciones experimentales, incorporando además covariables técnicas o biológicas (por ejemplo, edad, sexo, tipo celular). Entre los enfoques más simples se encuentra la regresión lineal,

implementada en el paquete R limma (Ritchie et al., 2015), originalmente diseñado para análisis de expresión génica en microarreglos. Limma ajusta modelos lineales para cada CpG y aplica una moderación bayesiana empírica de los errores estándar, mejorando la robustez del análisis.

Otro enfoque lineal es el implementado en BSmooth (Hansen et al., 2012), que aplica un suavizado local de los niveles de metilación a lo largo del genoma, asumiendo una distribución binomial para los datos. Sin embargo, uno de los inconvenientes de estos modelos es que los valores ajustados pueden salirse del intervalo $[0,1]$, propio de las medidas de proporción, lo cual puede distorsionar las inferencias si no se aplican transformaciones adecuadas.

Por otro lado, se han desarrollado métodos basados en regresión logística, que respetan los límites naturales de los valores de metilación y permiten ajustar la varianza mediante factores de sobre/infra-dispersión. methylKit permite aplicar regresión logística con o sin corrección de sobre-dispersión, utilizando pruebas como la Chi-cuadrado o la prueba F para contrastar la hipótesis nula de ausencia de efecto.

Una mejora adicional a estos enfoques se da mediante el uso de modelos beta-binomiales, que modelan la variabilidad observada en la proporción de metilación mediante una distribución beta, y la observada en la cobertura mediante una distribución binomial. Estos modelos permiten una estimación más realista de la varianza y tienen mejor desempeño en presencia de sobre-dispersión. Herramientas como DSS (Feng et al., 2014), MOABS (Sun et al., 2014), RADMeth, BiSeq y methylSig emplean esta estrategia, combinándola en algunos casos con métodos bayesianos empíricos para compartir información entre loci y aumentar la potencia estadística (Zhang et al., 2018; Singer, 2019).

Tabla 3.2. Métodos para la detección de metilación diferencial basados en el tipo de modelo estadístico.

Herramienta	Modelo asumido	Prueba de metilación diferencial	Segmentación	Lenguaje	Cita
methylKit (sin réplicas)	Ninguno	Prueba exacta de Fisher	Ventana de desplazamiento (tiling window)	R	Akalin et al., 2012
methylKit (con réplicas)	Regresión logística (con o sin sobredispersión)	Prueba Chi-cuadrado o F	Ventana de desplazamiento o regiones predefinidas	R	Akalin et al., 2012
BSmooth	Distribución binomial	Prueba t modificada	Unión de CpGs consecutivos	R	Hansen et al., 2012
DSS	Modelo beta-binomial jerárquico bayesiano	Prueba de Wald	Unión de CpGs basada en valor p	R	Feng et al., 2014
MOABS	Modelo beta-binomial	Prueba de razón de verosimilitud	No aplica segmentación explícita	C++	Sun et al., 2014

RADMeth	Regresión beta-binomial	Prueba de razón de verosimilitud logarítmica	Correlación entre pares de valores p	C++	Dolzhenko and Smith, 2014
BiSeq	Modelo de regresión beta-binomial	Prueba de Wald	Unión de CpGs consecutivos	R	Hebestreit et al., 2013
methylSig	Modelo beta-binomial	Prueba de razón de verosimilitud	Ventana de desplazamiento (tiling window)	R	Park et al., 2014

En un estudio realizado por Liu et al. (2020), se evaluaron siete herramientas para la detección de DMRs mediante la simulación de datos RRBS. Se analizó su rendimiento en términos de tasa de error tipo I, precisión/sensibilidad, área bajo la curva ROC (AUC), cobertura de secuenciación, longitud de lectura, longitud de DMRs y tamaño de muestra.

En una primera instancia, evaluaron la tasa de error tipo I de diferentes herramientas bajo la hipótesis nula. Todas las herramientas fueron capaces de controlar los falsos positivos por debajo del nivel deseado (tasa de error tipo I inferior al 5%). BiSeq y HMM-DM fueron demasiado conservadoras y tuvieron las tasas de error tipo I más bajas. En contraste, DMRfinder presentó una tasa de error tipo I algo más alta en comparación con las otras herramientas, pero aún por debajo del 5%.

Luego evaluaron el rendimiento bajo la hipótesis alternativa, enfocándose en cómo la diferencia de metilación en los DMRs, la longitud de los DMRs, la profundidad de secuenciación, la longitud de lectura y el tamaño de muestra afectan el rendimiento de estas herramientas para detectar DMRs. Se utilizaron dos criterios, las curvas de precisión-sensibilidad y el área bajo la curva (ABC). Todas las herramientas lograron valores de ABC más altos al aumentar la diferencia de metilación en los DMRs. Cuando la diferencia de metilación para un DMR era pequeña ($<0,2$), methylSig alcanzó valores de ABC mayores a 0,85. DMRfinder, methylKit y BiSeq también tuvieron un rendimiento bueno con valores de ABC superiores a 0,7.

Cuando la diferencia de metilación en los DMRs era pequeña, DMRfinder tuvo una mayor sensibilidad, pero una precisión relativamente baja, mientras que BiSeq tuvo la mayor precisión, pero una sensibilidad más baja. HMM-DM mostró el peor rendimiento, ya que no detectó ningún DMR cuando la diferencia de metilación era pequeña.

También evaluaron el rendimiento bajo diferentes profundidades de cobertura de secuenciación (10×, 30× y 50×). Bajo una profundidad de 10×, DMRfinder superó a las demás herramientas. MethylSig tuvo un ABC casi igual al de DMRfinder, y methylKit ocupó el tercer lugar. DSS no tuvo un buen rendimiento, cuando la profundidad fue 10×, su valor de ABC fue aproximadamente 0,5.

En relación con la longitud de los DMRs, la sensibilidad aumenta con la longitud de los DMRs. Para los DMRs largos (>1000 pb), DMRfinder y methylKit pudieron encontrar casi todos, mientras que DSS y HMM-DM encontraron pocos DMRs. methylSig y metilene,

tuvieron una sensibilidad ligeramente más baja para los DMRs largos que para los DMRs medianos (500 pb ~ 1000 pb).

Por otro lado, todas las herramientas mejoraron su rendimiento a medida que aumentaba el tamaño de la muestra. El valor ABC generalmente fue inferior a 0,75 cuando el tamaño fue de dos muestras por grupo. DMRfinder tuvo un rendimiento mucho mejor cuando el tamaño de la muestra era pequeño, mientras que BiSeq no funcionó normalmente cuando había solo 2 muestras en cada grupo.

Con respecto al tiempo de ejecución y el uso de memoria, las herramientas más rápidas fueron methylKit, methylSig y metilene. Mientras que BiSeq y HMM-DM fueron mucho más lentas. En cuanto al uso de memoria, BiSeq ocupó una gran cantidad de memoria en comparación con las demás herramientas, mientras que DMRfinder y metilene gestionaron el uso de memoria de manera más eficiente.

Piao et. al (2021) evaluaron ocho métodos utilizados para el análisis de metilación diferencial. En una primera instancia evaluaron el desempeño en la detección de citosinas diferencialmente metiladas (DMCs, del inglés Differentially Methylated Cytosines), estudiando la tasa de verdaderos positivos de cada método con un valor de p ajustado por Benjamini-Hochberg del 5%, variando la profundidad media de secuenciación y el número de réplicas, observando que methylSig, BSmooth, Biseq y metilene mostraron un desempeño inferior en comparación con methylKit, la prueba exacta de Fisher, DSS y RADMeth. Por otro lado, methylKit, DSS y RADMeth identificaron DMCs con precisión y mostraron un desempeño similar cuando la profundidad de secuenciación fue $\geq 15\times$ o cuando había ≥ 3 réplicas por condición. Además, DSS y methylKit tuvieron tasas más altas de verdaderos positivos cuando el número de réplicas era pequeño. En cuanto al número de falsos positivos detectados, methylSig y el test exacto de Fisher rechazaron correctamente todos los no-DMCs a un valor de $p < 0,05$, mientras que Biseq mostró un gran número de predicciones falsas. MethylKit, DSS y RADMeth presentaron un número reducido de falsos positivos.

Además, en este trabajo estudiaron el desempeño en la detección de regiones diferencialmente metiladas. methylKit y la prueba exacta de Fisher detectaron todas las DMRs verdaderas. RADmeth y methylSig tuvieron buen desempeño con cobertura $\geq 15\times$ o ≥ 3 réplicas, pero su rendimiento disminuyó con menor cobertura. Metilene mostró mejor desempeño en la identificación precisa de los límites de las DMRs. BSmooth también presentó buen poder de detección, salvo por errores en coberturas específicas. Por último, DSS y Biseq se destacaron como más apropiados para el análisis de citosinas individuales.

Por otro lado, probaron los métodos en datos biológicos reales, utilizando 32 DMRs verificadas biológicamente como estándar de referencia. MethylKit y la prueba exacta de Fisher cubrieron las 32 DMRs con un solapamiento mayor al 60%. RADmeth también cubrió todos los verdaderos positivos, pero con una fracción de superposición relativamente menor. DSS y metilene mostraron un desempeño general similar, en el cual metilene logró más superposiciones ($>60\%$) que DSS, mientras que DSS cubrió un mayor número de DMRs que metilene. Por otro lado, Biseq no logró detectar la mayoría de las DMRs.

El análisis bibliográfico y del estado del arte han revelado que los estudios comparativos recientes evidencian que no existe una única herramienta óptima en todos los contextos, ya que el desempeño puede variar según el organismo, el tamaño del genoma, la complejidad de

las regiones repetitivas y la plataforma de secuenciación utilizada. Esta diversidad de enfoques y herramientas pone de manifiesto la necesidad de un análisis cuidadoso a la hora de diseñar un flujo de trabajo bioinformático para el análisis de metilación del ADN.

3.1.4 Selección de herramientas para el flujo de análisis

Para el control de calidad y preprocesamiento de las lecturas, se optó por FastQC y Trim Galore!, debido a su capacidad para evaluar la calidad de las lecturas, eliminar adaptadores y recortar bases de baja calidad, garantizando la integridad de los datos previos al alineamiento.

En la etapa de alineamiento, se seleccionaron Bismark y BSMAP, dos herramientas con desempeño comprobado en términos de precisión, flexibilidad y eficiencia computacional, ampliamente adoptadas en la comunidad científica y que cuentan con documentación y soporte.

En cuanto al análisis diferencial de metilación, las herramientas más destacadas en la literatura son DSS y methylKit, por su capacidad para modelar de forma adecuada la variabilidad biológica y técnica de los datos, y por ofrecer enfoques estadísticos robustos para la detección de diferencias significativas en los niveles de metilación entre condiciones experimentales. Estas herramientas han demostrado un buen equilibrio entre sensibilidad y especificidad, lo que las posiciona como opciones confiables y ampliamente utilizadas en estudios epigenómicos con datos de WGBS.

3.1.5 Construcción de un flujo de procesamiento de datos de secuenciación por bisulfito

Teniendo en cuenta las diferentes opciones, se diseñaron e implementaron cuatro flujos alternativos de procesamiento para datos de WGBS. Para el desarrollo de cada etapa se realizó un análisis de manuales, tutoriales, foros comunitarios y repositorios como GitHub, lo cual permitió evaluar la capacidad analítica de las herramientas, así como también aspectos como la facilidad de instalación, compatibilidad con distintas plataformas y soporte activo por parte de la comunidad científica.

La Figura 3.1 ilustra el flujo de análisis propuesto con las distintas herramientas seleccionadas para cada una de las etapas de análisis presentadas anteriormente en la Figura 1.3.

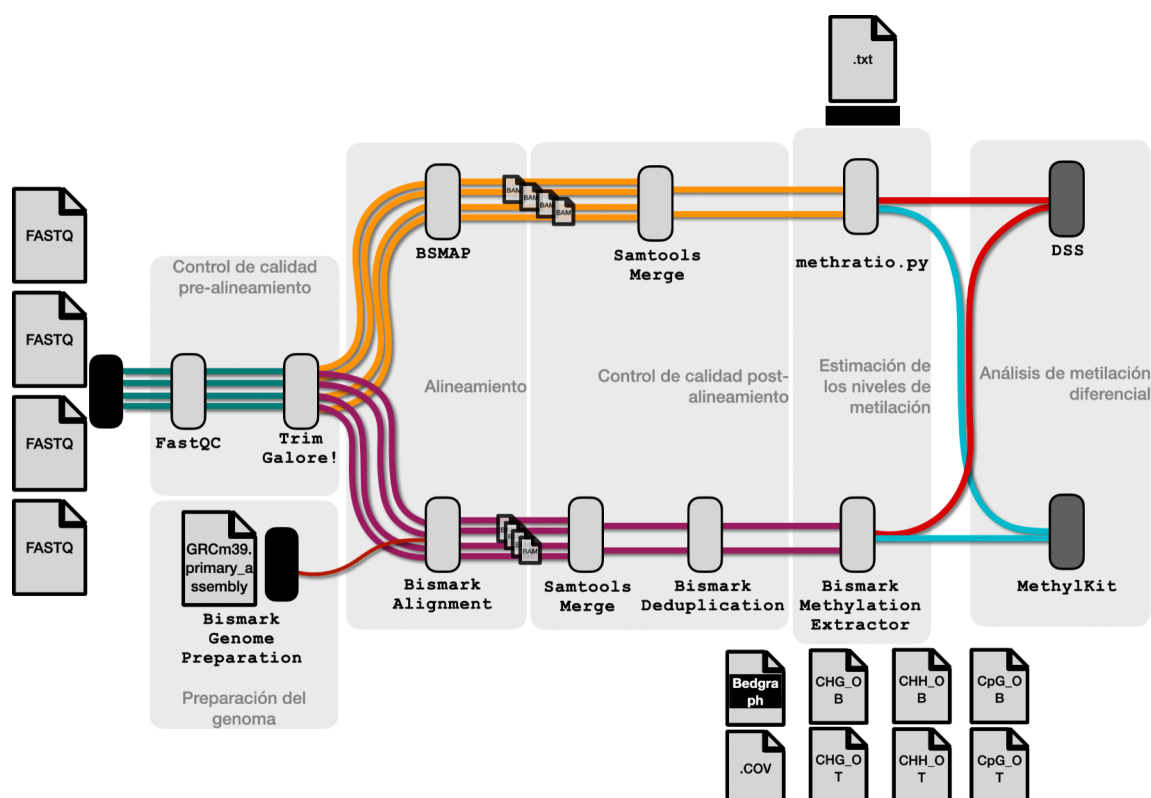


Figura 3.1. Flujo de análisis propuesto.

Inicialmente, cada herramienta fue puesta a prueba utilizando archivos de ejemplo provenientes de recursos educativos disponibles en línea, como los proporcionados en el curso “*Analysing bisulfite methylation sequencing data (One day)*”, desarrollado por el Babraham Institute (Krueger & Andrews, 2021). Esto permitió familiarizarse con el funcionamiento de cada etapa del flujo y validar su correcta ejecución antes de aplicar el análisis a datos reales. Los datos empleados en este ejercicio correspondían a un subconjunto de 200.000 lecturas de un experimento de secuenciación *paired-end* (PE) (GSM1027571). Además, se utilizaron los archivos de prueba oficialmente provistos por los desarrolladores de las distintas herramientas, principalmente para verificar que las instalaciones se hayan realizado correctamente y que los programas se comporten como se espera. Si bien la cantidad de lecturas de los archivos de prueba es sustancialmente menor que la generada en experimentos reales, esta estrategia permitió evaluar el comportamiento de cada herramienta del flujo sin requerir grandes tiempos de cómputo.

El proceso de análisis inicia con archivos crudos en formato FASTQ, los cuales fueron sometidos a un control de calidad inicial mediante la herramienta FastQC, con el objetivo de identificar lecturas de baja calidad o presencia de adaptadores. Posteriormente, de ser necesario, se utiliza Trim Galore! para el recorte de regiones no deseadas y la generación de archivos limpios que se utilizan como entrada para la etapa de alineamiento.

Luego, con las lecturas preprocesadas y el genoma de referencia, se continúa con el alineamiento. Para esto, el flujo propuesto provee dos alternativas: Bismark y BSMAP. En el caso de Bismark, previo al alineamiento, es necesario preparar el genoma de referencia con el módulo Bismark_Genome_Preparation que construye versiones bisulfito del genoma y las indexa para su posterior uso. Cabe destacar que si bien el alineamiento se debe ejecutar para

cada muestra en forma independiente, la preparación del genoma se realiza solo una vez. La herramienta de alineamiento elegida para Bismark es Bowtie2. Bismark requiere que se especifiquen las secuencias a analizar en formato FASTQ o FASTA así como también la carpeta que contiene el genoma de referencia en formato fa o FASTA, junto con el resultado de Bismark_Genome_Preparation. Por otro lado, BSMAP no requiere del paso previo de preparación de genoma y requiere archivos de entrada con las lecturas en formato FASTA, FASTQ o BAM, y la secuencia de referencia en formato FASTA. En ambos casos, por defecto, el flujo de análisis emplea los parámetros predeterminados por los desarrolladores. En cuanto al número de procesadores a utilizar, BSMAP utiliza por defecto el número de procesadores detectados, mientras que en Bismark se especificó el número de procesadores a utilizar (4 u 8). Sin embargo, si bien ambas herramientas permiten especificar el número de procesadores a utilizar, los desarrolladores coinciden en que el rendimiento en paralelo escala bien con 12 hilos o menos y que no hay una ganancia significativa de velocidad con más de 12 hilos. Como resultado del alineamiento, cada herramienta genera archivos BAM correspondientes a los alineamientos individuales por muestra.

El post-procesamiento de los archivos de alineamientos utiliza Samtools Merge para la combinación de alineamientos de réplicas técnicas de una misma réplica biológica, si las hubiese. Una etapa clave en los flujos es la remoción de lecturas idénticas, potencialmente introducidas por la amplificación por PCR, denominada de-duplicación. Tanto en el flujo de trabajo basado en Bismark como en BSMAP, este paso se realiza de manera serial, ya que las posiciones mapeadas deben almacenarse en memoria dentro de una única estructura de datos. Por este motivo, el tiempo requerido para este paso está determinado principalmente por el tiempo necesario para leer el archivo BAM completo una vez. Esta característica limita la posibilidad de paralelización de esta etapa y puede representar un cuello de botella en flujos de trabajo con grandes volúmenes de datos. En el flujo de Bismark, el proceso de de-duplicación se ejecuta como un paso explícito posterior al alineamiento, a través del módulo `deduplicate_bismark`. En cambio, en BSMAP, dicha funcionalidad se encuentra integrada en el script `methratio.py`, que además calcula la proporción de metilación. En este caso, el uso del parámetro `-r` permite activar la remoción de duplicados.

Posteriormente se estima el nivel de metilación en cada posición CpG, para lo cual el *pipeline* propone dos métodos alternativos. Por un lado, `methratio.py` que permite calcular proporciones de metilación a partir de los alineamientos generados por BSMAP, produciendo archivos de texto que resumen el estado de metilación en cada sitio genómico: cromosoma, coordenada, hebra, contexto CG/CHG/CHH, ratio de metilación, cobertura efectiva C + T, número de C, número de C + T bruto, G y G + A en la hebra inversa, e intervalo de confianza 95 % inferior/superior. Por otro lado, Bismark_Methylation_Extractor genera archivos en múltiples formatos (.bedGraph, .cov, CHG/CHH/CpG con orientación OT/OB), compatibles con análisis posteriores en R.

Finalmente, el flujo de análisis concluye con los métodos de análisis de metilación diferencial. Al igual que en otras etapas, el flujo provee dos alternativas: DSS y MethylKit. Tanto DSS como MethylKit son paquetes de R, compatibles con los formatos de salida generados por Bismark y `methratio.py` (BSMAP), lo que facilita su integración fluida con las etapas anteriores. Además, para ambas herramientas se propone utilizar los parámetros predeterminados por los desarrolladores.

3.2 Implementación y evaluación del flujo de análisis de datos de WGBS

Luego de llevar a cabo la validación de cada módulo por separado utilizando los archivos de prueba, se continuó con la evaluación de las distintas alternativas del flujo propuesto en el contexto de un experimento real. Para ello, se utilizaron datos provenientes de experimentos depositados en la base de datos de ENCODE¹³. En particular, se seleccionaron experimentos correspondientes a muestras de distintos tejidos pero tomadas de los mismos individuos de manera de asegurar que las diferencias observadas a la hora de realizar análisis comparativos se deben a diferencias entre los correspondientes tejidos y no a otras fuentes de variabilidad. Más precisamente, los tejidos seleccionados son cardíaco y renal en ratones de sexo masculino recién nacidos. Los identificadores de los experimentos seleccionados son ENCSR397YEG (corazón) y ENCSR128HOP (riñón). En ambos casos, cada experimento involucra dos réplicas biológicas, las cuales a su vez tienen dos réplicas técnicas asociadas.

3.2.1 Control de calidad inicial con FastQC

La primera etapa del flujo de análisis consistió en la evaluación de la calidad de los datos crudos mediante la herramienta FastQC (v0.12.0).

FastQC se ejecutó en un entorno de cómputo de alto rendimiento (HPC), asignando 4 núcleos de CPU y 24 GB de memoria RAM. El mismo se aplicó a todos los archivos FASTQ comprimidos (.fastq.gz) correspondientes a las lecturas pareadas de ambas muestras seleccionadas.

El tiempo total de ejecución para el conjunto de archivos de cada experimento fue de aproximadamente 1,24 horas, con un uso total de memoria de 1,02 GB, sin variaciones entre los datos de riñón y corazón. La Tabla 3.3 resume las métricas principales obtenidas con FastQC para cada archivo FASTQ crudo, incluyendo el número total de lecturas, el total de bases secuenciadas y el rango de longitud de las lecturas.

Tabla 3.3. Métricas de secuencias de los datos crudos según el análisis con FastQC.

Muestra	Accession	Lecturas totales	Bases totales (Gbp)	Longitud mínima - máxima (pb)
Riñón	ENCFF001ZVX	706.610.024	70,6	100
	ENCFF001ZVU	643.968.828	64,3	100
	ENCFF001ZWD	654.339.599	65,4	100
	ENCFF001ZWA	635.147.721	63,5	100

¹³ <https://www.encodeproject.org/>

Corazón	ENCFF001ZVM	718.506.938	71,8	100
	ENCFF001ZVP	652.720.883	65,2	100
	ENCFF001ZVR	641.332.810	64,1	100
	ENCFF001ZVT	661.804.193	66,1	100

3.2.2 Recorte de adaptadores y limpieza de lecturas con Trim Galore!

Una vez realizado el control de calidad inicial, se realizó el recorte de adaptadores y la eliminación de secuencias de baja calidad utilizando la herramienta Trim Galore! (v0.6.10).

El análisis se realizó en un entorno HPC, en un entorno virtual de Conda con cutadapt instalado. Se determinó un máximo de 6 núcleos de CPU y 100 GB de memoria RAM por tarea, aunque el consumo real fue considerablemente menor. Cada archivo fue procesado de forma individual.

En cuanto al desempeño computacional, el tiempo de ejecución promedio fue de aproximadamente 2 horas por archivo, y el uso de memoria no superó los 0,47 GB, como se muestra en la Figura 3.2.

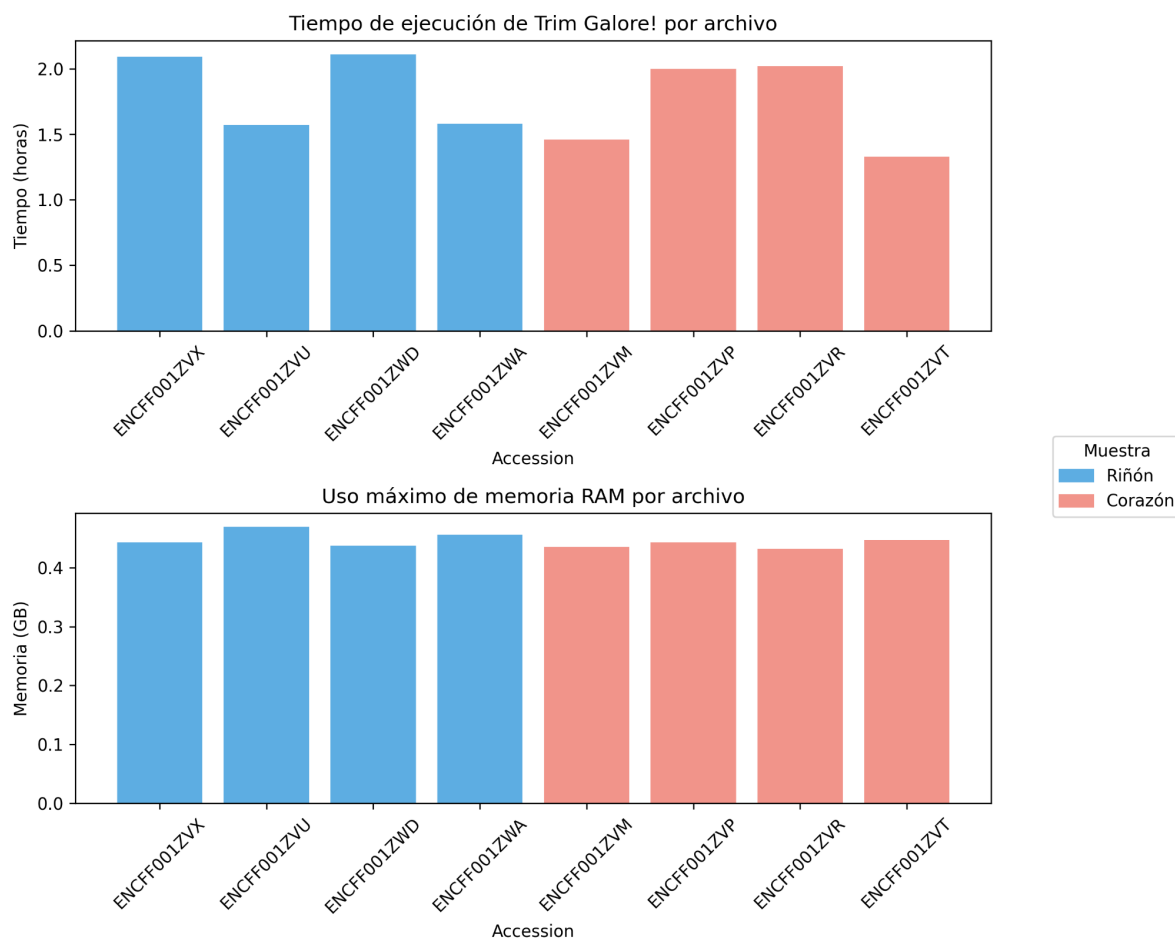


Figura 3.2. Tiempo de ejecución (en horas) y uso máximo de memoria RAM por archivo (en GB) de Trim Galore! para las muestras de riñón (azul) y corazón (rojo).

Los datos ya procesados con TrimGalore! se volvieron a analizar con FastQC, obteniendo los resultados listados en la Tabla 3.4. Luego de esta etapa de preprocesamiento, el número total de lecturas conservadas es entre 600,5 y 672,3 millones de secuencias por archivo, con longitudes comprendidas entre 20 y 100 pares de bases

Tabla 3.4. Métricas de secuencias post-procesamiento con Trim Galore! según análisis con FastQC.

Muestra	Accession	Lecturas totales	Bases totales (Gbp)	Longitud mínima - máxima (pb)
Riñón	ENCFF001ZVX	668.174.519	63,5	20–100
	ENCFF001ZVU	611.770.138	58,2	20–100
	ENCFF001ZWD	620.480.827	59,0	20–100
	ENCFF001ZWA	606.805.186	57,9	20–100

Corazón	ENCFF001ZVM	609.707.563	57,2	20–100
	ENCFF001ZVP	672.368.669	63,1	20–100
	ENCFF001ZVR	605.938.196	57,2	20–100
	ENCFF001ZVT	600.586.621	55,1	20–100

La Figura 3.3 A) ilustra el resultado del análisis de calidad inicial de uno de los archivos con FastQC, donde se puede observar una disminución progresiva en la calidad de las bases hacia el extremo 3' de las secuencias. El panel B) muestra el resultado de FastQC para el mismo archivo luego del procesamiento de las lecturas con TrimGalore!, evidenciando la mejora lograda en la calidad de las secuencias.

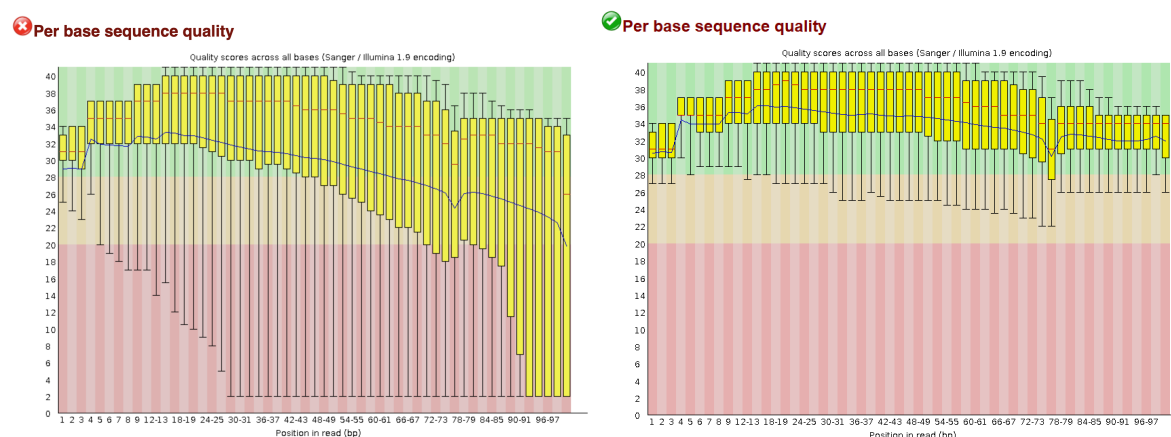


Figura 3.3. Reporte de calidad de las secuencias de FastQC para una de las muestras de corazón A) antes y B) después de realizar el recorte con TrimGalore!(Accession ENCFF001ZVT)

3.2.3 Alineamiento de lecturas a la referencia

La segunda etapa del flujo consistió en el alineamiento de las lecturas a la secuencia de referencia. Se evaluaron dos alineadores, Bismark y BSMAP, sobre las mismas muestras de riñón y corazón, utilizando distintos parámetros de paralelización para evaluar el impacto en el desempeño computacional.

Para Bismark, las muestras de riñón se procesaron utilizando 4 hilos y un máximo de 150 GB de memoria disponible. Por otro lado, para las muestras de corazón, se emplearon 8 hilos manteniendo la misma disponibilidad de memoria.

El análisis de la Figura 3.4 revela que el uso de mayor paralelización (8 hilos) en las muestras de corazón permitió reducir el tiempo de ejecución en comparación con las muestras de riñón (4 hilos), aunque con un aumento considerable en el uso de memoria RAM (aproximadamente el doble). Esto coincide con lo reportado por los desarrolladores, un incremento en el número de hilos disponibles permite la aceleración casi lineal del proceso

pero con un incremento proporcional en el consumo de CPU y memoria. A su vez, evidencia la necesidad de balancear el uso de recursos computacionales dependiendo del entorno de ejecución disponible y los objetivos del análisis.

Alternativamente, se implementó un alineamiento utilizando BSMAP en el mismo entorno HPC, pero ahora empleando una asignación de 10 hilos de procesamiento y un máximo de 150 GB de memoria RAM disponible por trabajo. Cada muestra fue procesada de forma independiente, utilizando un máximo de 8 hilos efectivos por muestra.

Los resultados (Figura 3.4) mostraron que el tiempo de ejecución de BSMAP fue considerablemente menor en comparación con Bismark, al comparar la misma muestra. En promedio, cada muestra fue procesada en un intervalo de 2,54 a 3,48 horas, con un uso máximo de memoria RAM de aproximadamente 7,5 GB por archivo.

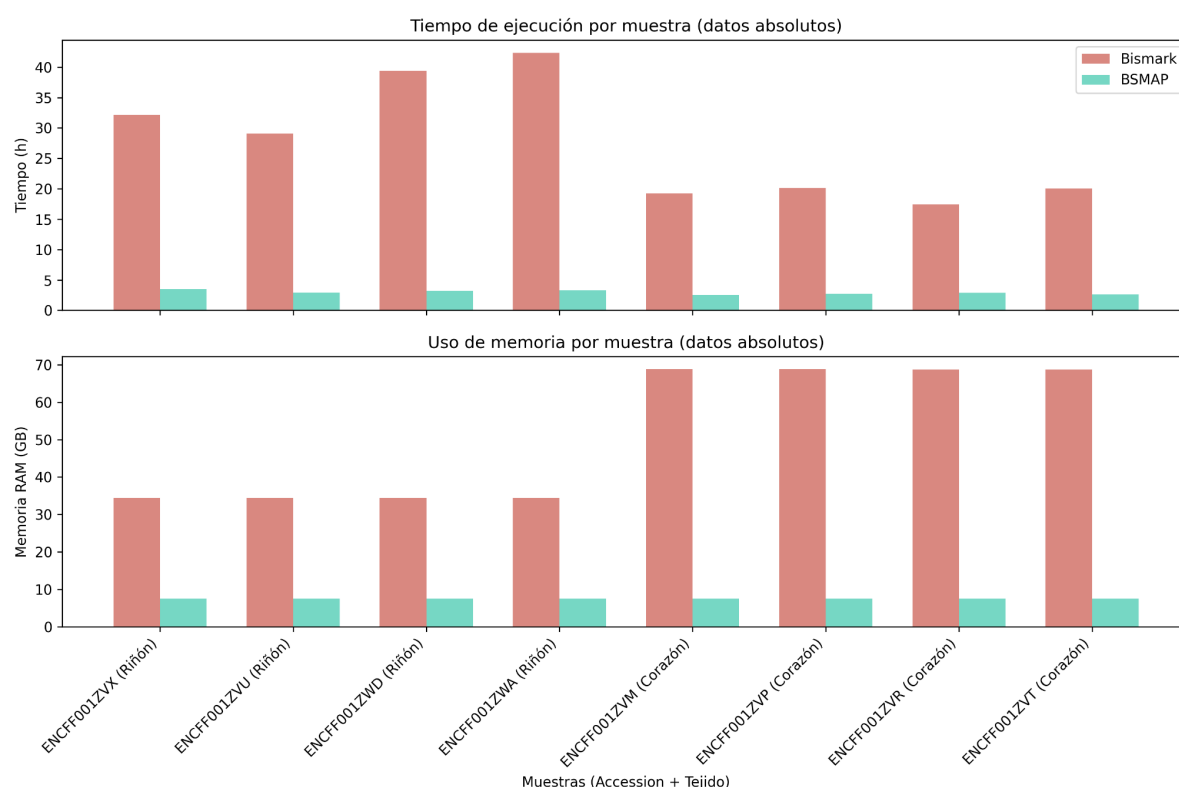


Figura 3.4. Tiempo de ejecución (en horas) y uso máximo de memoria RAM por archivo (en GB) de Bismark (rojo) y BSMAP (verde) para las muestras de riñón y corazón durante el proceso de alineamiento.

Dado que las muestras presentan cantidades de lecturas distintas, se generó una versión normalizada de los resultados (Figura 3.5), en la que el tiempo de ejecución y la memoria RAM se escalaron al mismo número de lecturas para todas las muestras, permitiendo comparar la eficacia intrínseca de cada alineador sin el sesgo del tamaño de los datos. Independientemente de la cantidad de lecturas, BSMAP demostró ser más rápido y consumir menos memoria RAM que Bismark.

En conjunto, estos resultados sugieren que BSMAP ofrece ventajas en eficiencia de tiempo y uso de memoria, especialmente en escenarios con recursos limitados o en los que se requiere procesar un gran número de muestras en paralelo.

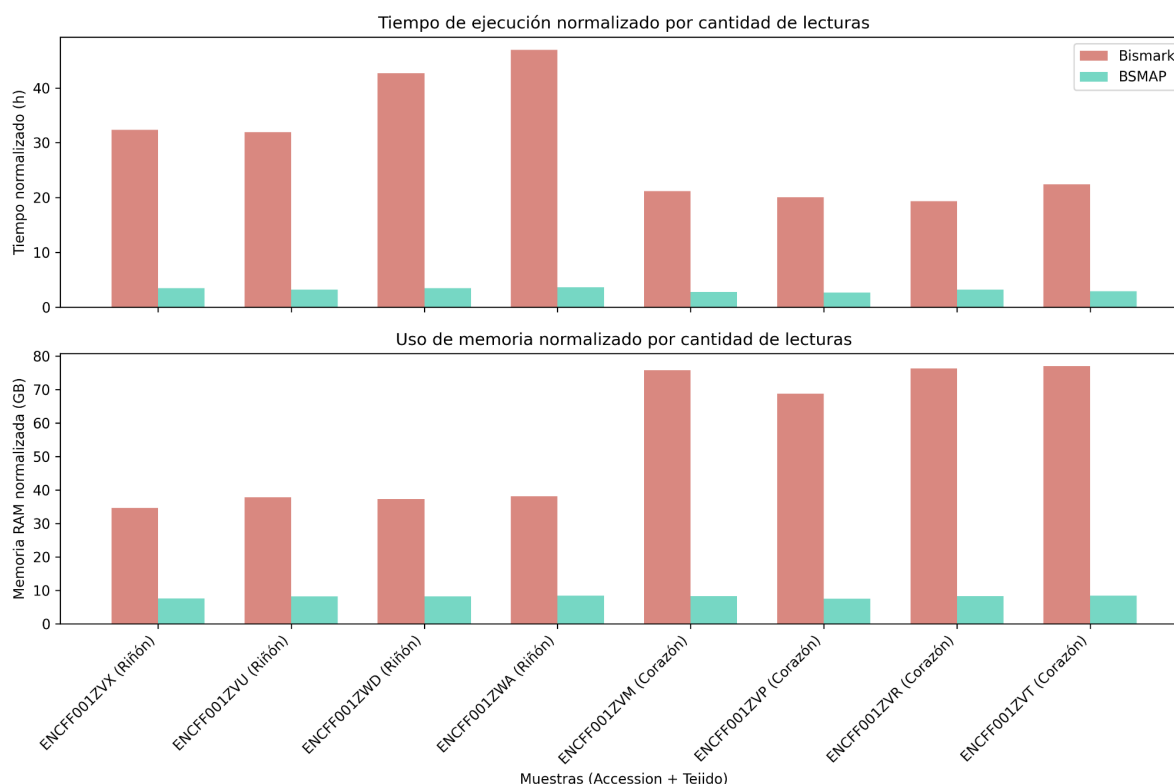


Figura 3.5. Tiempo de ejecución (en horas) y uso máximo de memoria RAM por archivo (en GB) normalizado de Bismark (rojo) y BSMAP (verde) para las muestras de riñón y corazón durante el proceso de alineamiento.

Respecto del número de alineamientos únicos, ambas herramientas produjeron porcentajes de alineamientos únicos muy similares en todas las muestras analizadas (Tabla 3.5). Bismark reportó una eficiencia de alineamiento único entre 77,8% y 80,4%, mientras que BSMAP mostró porcentajes equivalentes, entre 79,8% y 82,0% . Estas diferencias indican que, en términos de lecturas útiles para el análisis de metilación, ninguna de las dos herramientas presenta una ventaja en este conjunto de datos.

Tabla 3.5. Porcentaje de alineamientos únicos obtenidos con Bismark y BSMAP

Muestra	Accession	Unique Bismark (%)	Unique BSMAP (%)
Riñón	ENCFF001ZVX	80,4	82,0
	ENCFF001ZVU	80,4	79,8
	ENCFF001ZWD	79,4	81,1
	ENCFF001ZWA	79,7	81,2
Corazón	ENCFF001ZVM	79,7	81,8

	ENCFF001ZVP	79,8	81,9
	ENCFF001ZVR	79,5	81,3
	ENCFF001ZVT	77,8	80,4

Esta comparación sugiere que la proporción de lecturas únicas alineadas es altamente consistente entre Bismark y BSMAP, lo que permite cierta flexibilidad en la elección de la herramienta para el alineamiento dependiendo de otros factores como el tiempo de ejecución, la compatibilidad con formatos o la facilidad de integración en flujos de trabajo personalizados.

3.2.4 Post alineamiento: combinación de réplicas técnicas y de-duplicación.

Una vez finalizado el alineamiento, se continuó con la etapa de combinar los archivos BAM resultantes para las distintas réplicas técnicas mediante la herramienta samtools merge. Este paso permite obtener los alineamientos de cada réplica biológica en un único archivo, facilitando el análisis cuantitativo de la metilación a nivel de muestra biológica.

Esta tarea se realizó en un entorno HPC, asignando 4 hilos de procesamiento y hasta 50 GB de memoria RAM por tarea. Cada operación de *merge* se llevó a cabo por separado para las muestras procesadas con Bismark y BSMAP.

En todos los casos, el consumo de memoria fue bajo en comparación con otras etapas del proceso (<0.02 GB), sin diferencias entre los datos de salida de cada herramienta de alineamiento. En cuanto al tiempo de ejecución, se observaron valores promedio de $2,26 \pm 0,08$ h para Bismark y $1,34 \pm 0,05$ h para BSMAP, lo que implica una reducción de aproximadamente 30% en BSMAP respecto a Bismark. Esta diferencia se encuentra asociada a la complejidad de los archivos BAM generados por cada alineador. Los archivos Bismark resultaron ser de mayor tamaño debido a que incluyen cinco campos adicionales, mientras que los de BSMAP contienen solo dos campos adicionales, incrementando el volumen de información por lectura. (Figura 3.6).

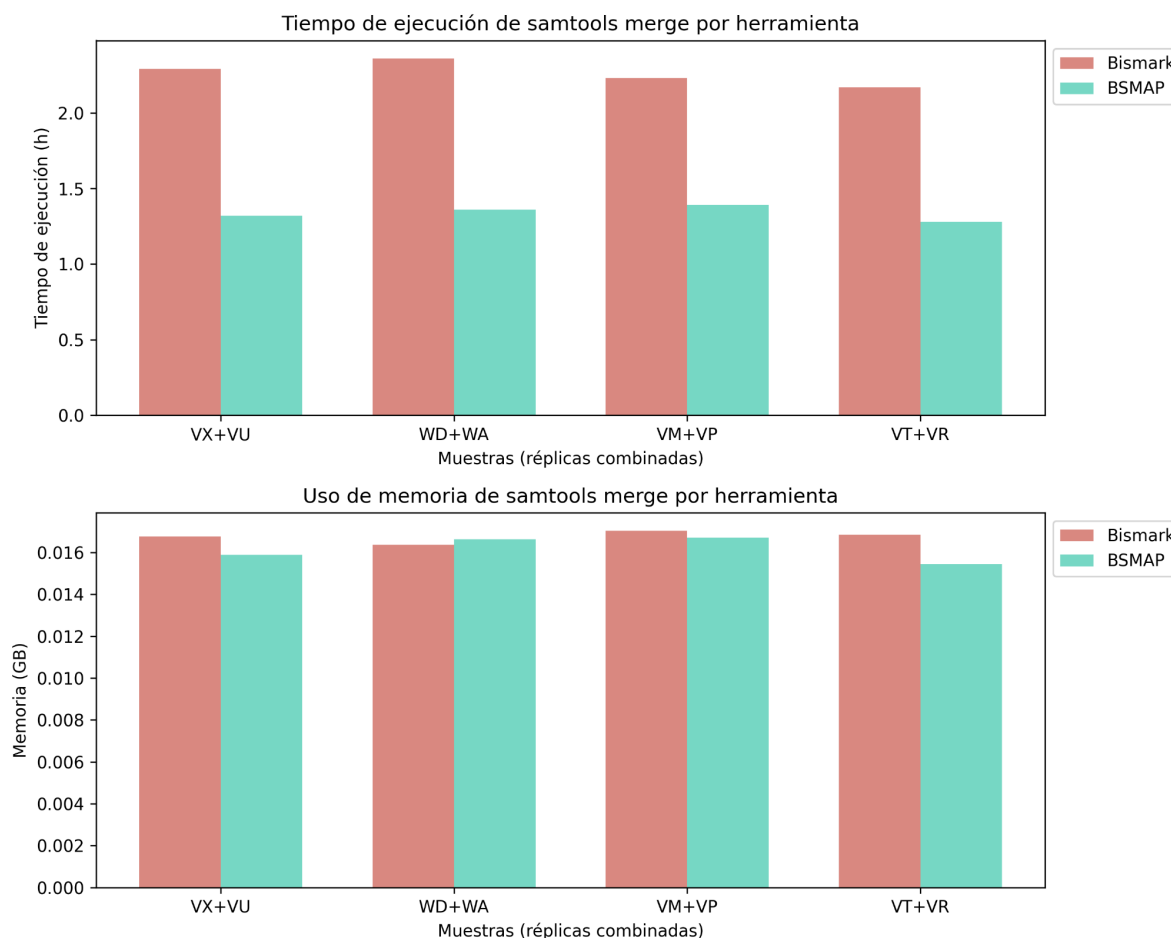


Figura 3.6. Tiempo de ejecución (en horas) y uso máximo de memoria RAM por archivo (en GB) de samtools merge para los datos de salida de Bismark (rojo) y BSMAP (verde).

Luego se procedió a la eliminación de lecturas duplicadas. En el caso del flujo que utiliza Bismark, este paso se realiza de manera explícita mediante el módulo `deduplicate_bismark`, que identifica duplicados a partir de coordenadas de mapeo idénticas entre lecturas.

En cambio, para el flujo que utiliza BSMAP, la de-duplicación no se realiza como un paso separado. En su lugar, el script `methratio.py` que se utiliza para la estimación de los niveles de metilación proporciona un parámetro que permite excluir lecturas duplicadas antes de realizar la cuantificación.

La de-duplicación con Bismark se llevó a cabo en el entorno HPC, utilizando 4 hilos de CPU por trabajo y un máximo de 150 GB de memoria RAM disponible. En la Tabla 3.6, se resumen los recursos computacionales utilizados para esta etapa. El requerimiento de memoria durante esta etapa es uno de los más elevados de todo el flujo de análisis. Esto se debe a que durante este proceso Bismark almacena todas las posiciones mapeadas de las lecturas en memoria en una única estructura de datos que luego recorre en forma serial para poder identificar duplicados. En consonancia con esto, la literatura muestra que durante el alineamiento con Bismark, incluyendo la de-duplicación, se utilizan aproximadamente 90 GB de memoria RAM.

Tabla 3.6. Tiempo de ejecución y uso máximo de memoria RAM durante el proceso de de-duplicación utilizando Bismark como herramienta de alineamiento.

Herramienta	Accession	Tiempo (h)	Memoria (GB)
Bismark	ENCFF001ZVX ENCFF001ZVU	3,56	102,44
Bismark	ENCFF001ZWD ENCFF001ZWA	3,56	97,56
Bismark	ENCFF001ZVM ENCFF001ZVP	5,02	102,01
Bismark	ENCFF001ZVT ENCFF001ZVR	5,02	101,32

3.2.5 Estimación de los niveles de metilación

La siguiente etapa comprende la estimación de los niveles de metilación a partir de los archivos de alineamiento previamente procesados. Para el caso de Bismark, esta tarea se realizó utilizando la función `bismark_methylation_extractor`, que genera archivos comprimidos (gzip) en formato `.bedGraph`¹⁴ con los niveles de metilación por posición genómica. Esta herramienta permite el uso de múltiples hilos en paralelo, lo cual optimiza significativamente los tiempos de ejecución. En este análisis se utilizaron 8 hilos.

Por otro lado, el flujo basado en BSMAP continuó con la ejecución del script `methratio.py`, que calcula las proporciones de metilación por posición, integrando de forma simultánea la función de de-duplicación. Este enfoque, si bien es funcionalmente compacto, presenta una mayor demanda computacional, principalmente en el uso de memoria RAM.

En la Tabla 3.7 se presentan los valores de tiempo de ejecución y uso máximo de memoria para ambos enfoques.

Tabla 3.7. Tiempo de ejecución y uso máximo de memoria RAM durante la estimación de los niveles de metilación para cada flujo de análisis.

Herramienta	Accession	Tiempo (h)	Memoria (GB)
Bismark	ENCFF001ZVX ENCFF001ZVU	4,31	2,02
Bismark	ENCFF001ZWD ENCFF001ZWA	4,35	2,02
Bismark	ENCFF001ZVM ENCFF001ZVP	4,51	2,02
Bismark	ENCFF001ZVT ENCFF001ZVR	4,48	2,02

¹⁴ <https://genome.ucsc.edu/goldenpath/help/bedgraph.html>

BSMAP	ENCFF001ZVX ENCFF001ZVU	11,39	25,42
BSMAP	ENCFF001ZWD ENCFF001ZWA	10,16	25,42
BSMAP	ENCFF001ZVM ENCFF001ZVP	11,36	25,42
BSMAP	ENCFF001ZVT ENCFF001ZVR	10,59	25,42

Como se observa, Bismark presenta un rendimiento computacional más eficiente en términos de tiempo y uso de memoria. En todos los casos, el proceso con Bismark se completó en aproximadamente 4,3–4,5 horas, con un uso estable de memoria de 2,02 GB.

Por otro lado, el script methratio.py de BSMAP necesitó entre 10 y 11 horas, utilizando alrededor de 25 GB de memoria por archivo. Este mayor consumo de tiempo y memoria observado en BSMAP podría deberse a que se combina la estimación de niveles de metilación con un paso de de-duplicación interna, lo que implica una mayor carga computacional durante su ejecución.

Cabe destacar que si se consideran la etapa de de-duplicación y la de estimación de los niveles de metilación, ambas herramientas necesitaron tiempos similares de ejecución, pero Bismark requirió más memoria RAM, por lo que se puede afirmar que BSMAP es más eficiente.

3.2.6 Análisis de metilación diferencial

El análisis de DMRs fue realizado en un entorno local, con 14 núcleos de CPU y 24 GB de memoria unificada para la ejecución de las diferentes alternativas: Bismark+DSS, Bismark+methylKit, BSMAP+DSS y BSMAP+methylKit.

La Figura 3.7 ilustra el número de DMRs identificados con las cuatro alternativas de análisis. DSS evidenció resultados similares para BSMAP y Bismark. Por el contrario, methylKit reportó una diferencia más marcada, identificando aproximadamente tres veces más DMRs cuando se utilizó BSMAP en lugar de Bismark.

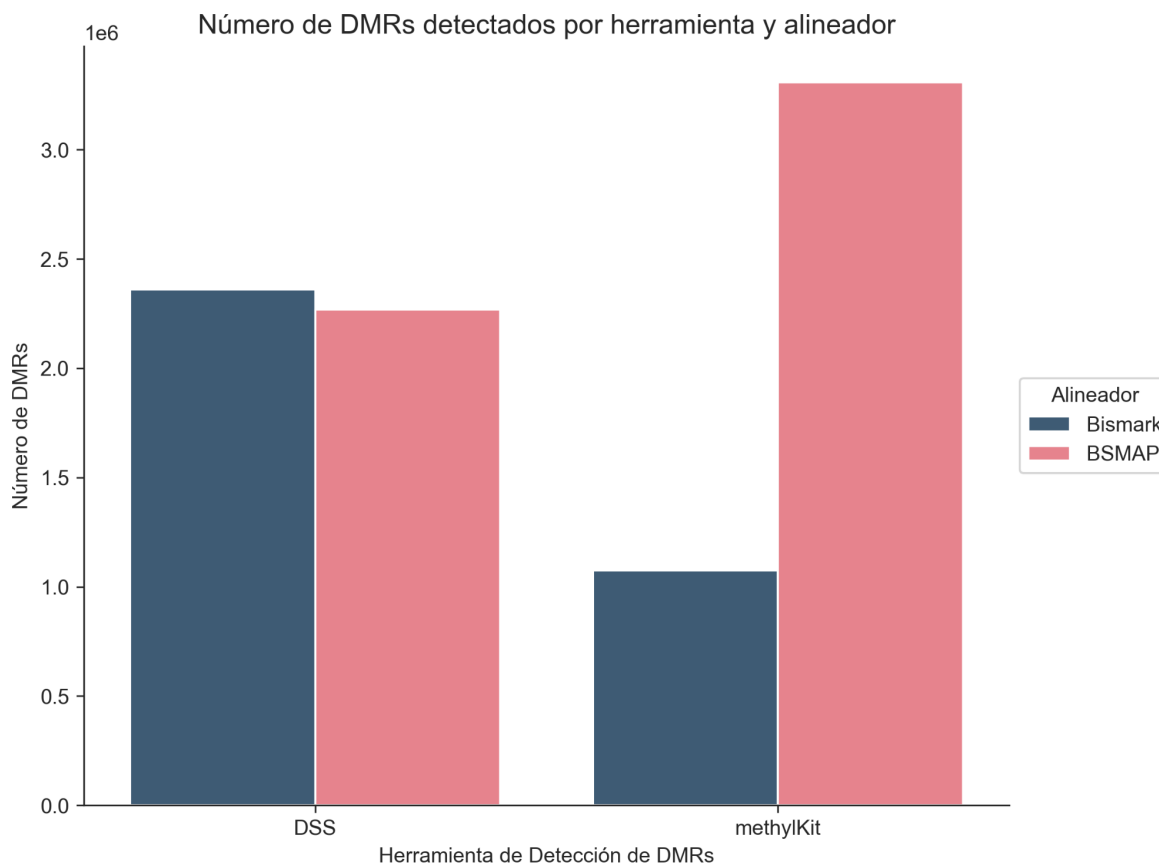


Figura 3.7. Número de sitios DMRs detectados en los diferentes flujos de análisis.

Posteriormente se evaluó el solapamiento de las DMRs identificadas por cada alternativa, resumido en la Figura 3.8. El conjunto de DMRs compartido por todos los flujos de análisis fue de 512.779 sitios, lo que representa aproximadamente un 10% del total de DMRs únicos detectados. Además, independientemente de la herramienta de detección de DMRs, 1.065.433 sitios (22%) fueron identificados a partir de los resultados de BSMAP, mientras que solo 531.521 sitios (11%) fueron exclusivamente encontrados a partir de los resultados de Bismark. Por otro lado, 1.030.704 sitios (21%) son comunes entre BSMAP+methylKit y Bismark+DSS. También, 1.049.175 sitios (22%) fueron compartidos entre BSMAP+methylKit y Bismark+methylKit y 2.021.304 sitios (42%) fueron comunes entre BSMAP+DSS y Bismark-DSS. Finalmente, la combinación BSMAP+methylKit fue la que evidenció el mayor número de DMRs exclusivos.

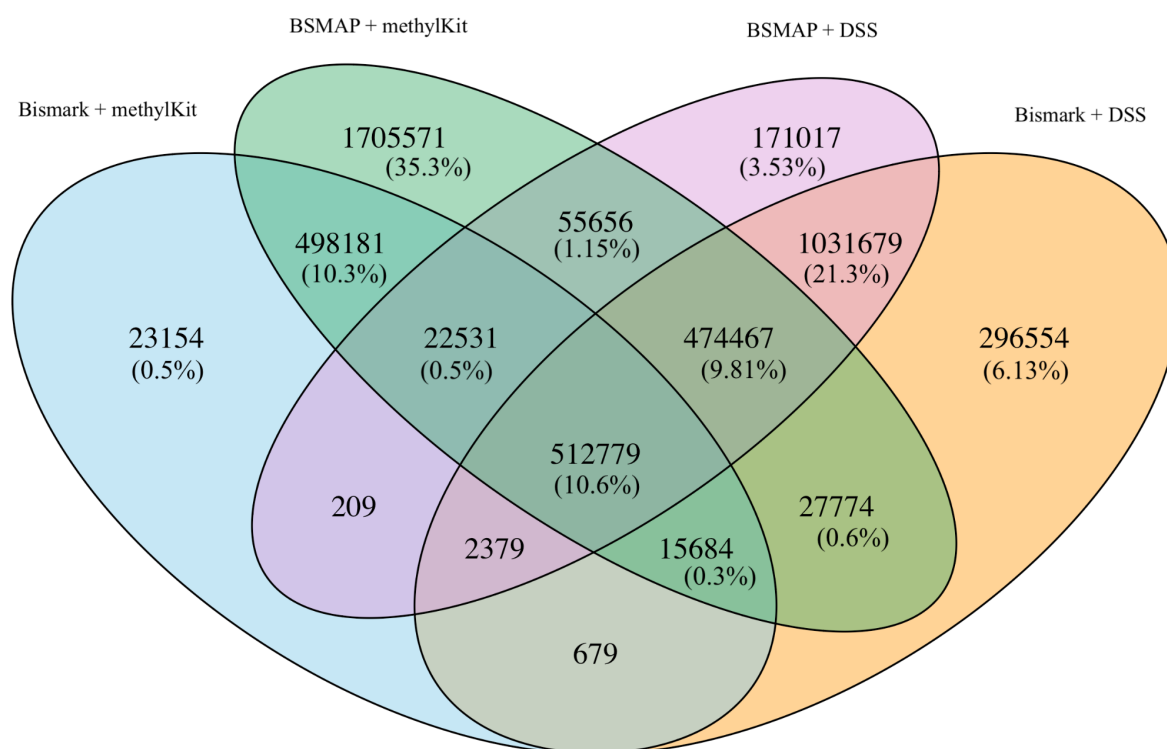


Figura 3.8. Diagrama de Venn que muestra el solapamiento de las DMRs detectadas por los cuatro flujos de análisis. En azul se muestran las DMRs detectadas por el flujo Bismark+methylKit, en verde se muestran las regiones detectadas por BSMAP+methylKit, en rosa las obtenidas utilizando BSMAP+DSS y en naranja las extraídas por Bismark+DSS.

Para evaluar la concordancia en la magnitud y dirección del cambio de metilación detectado por DSS y methylKit, se compararon los valores de cambio de metilación para los DMRs comunes detectados por ambos métodos. Esta comparación se realizó por separado para BSMAP y Bismark, generando gráficos de dispersión y calculando el coeficiente de correlación de Pearson (r).

En el caso de Bismark se determinó un valor de $r = 0,97$. Como puede apreciarse en el diagrama de dispersión de los porcentajes de cambio de metilación detectados por ambas herramientas de la Figura 3.9, la mayoría de los puntos se encuentran en torno a la recta de identidad, lo que revela una elevada concordancia en la dirección y magnitud del cambio de metilación detectado por DSS y methylKit.

En el caso de BSMAP, también se observó un coeficiente de correlación elevado ($r = 0,96$) aunque ligeramente menor en comparación con Bismark. Adicionalmente, a partir del diagrama de dispersión de la Figura 3.9, se identificaron una mayor cantidad de puntos en rojo, correspondientes a sitios en los que DSS y methylKit estimaron cambios de metilación con signo opuesto, lo que podría indicar una mayor variabilidad entre las estimaciones de

DSS y methylKit en algunos sitios, sin embargo, se observa una buena correlación en general.

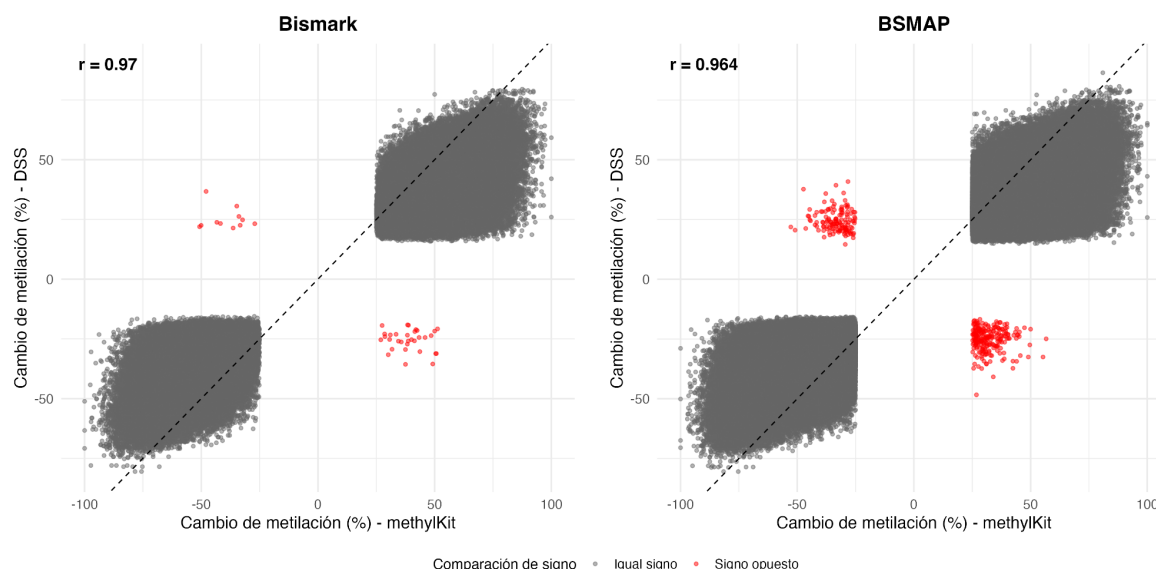


Figura 3.9. Comparación del cambio de metilación detectado por DSS y methylKit a partir de la cuantificación generada por Bismark y BSMAP. Los puntos en gris representan DMRs en los que ambos métodos coinciden en el signo del cambio de metilación, mientras que en rojo indican DMRs con signos opuestos entre ambos métodos.

Estos resultados indican que, independientemente del alineador utilizado, DSS y methylKit tienden a detectar cambios de metilación similares en los sitios que comparten. En el caso del alineador Bismark, se identificaron 531.521 sitios compartidos entre ambas herramientas, de los cuales solo 43 presentaron signos opuestos en el cambio de metilación. En el análisis con BSMAP, se encontraron 1.065.433 sitios en común, con 367 discrepancias en el signo del cambio de metilación, lo que representa un 0,034% del total.

3.2.7 Comparación con datos de metilación obtenidos de ENCODE.

Con el objetivo de evaluar la concordancia entre los niveles de metilación obtenidos mediante el flujo de análisis desarrollado y los datos originales del experimento, se realizó una comparación directa entre los valores de metilación en sitios CpG reportados por ENCODE y los generados por el pipeline aquí presentado. Como se muestra en la Figura 3.10, las distribuciones globales de metilación estimadas por BSMAP (A) y Bismark (B) son consistentes con los valores de referencia de ENCODE. En ambos métodos, las medianas de metilación se ubicaron entre 75% y 80%, con variaciones mínimas entre tejidos (corazón y riñón) y entre réplicas biológicas. La presencia de valores con baja metilación (<20%) se mantuvo en todos los casos, lo que refleja la coexistencia de regiones hipometiladas. Estos resultados sugieren que, pese a posibles diferencias puntuales en la cuantificación de sitios individuales, ambas herramientas reproducen adecuadamente los patrones globales de metilación reportados por ENCODE.

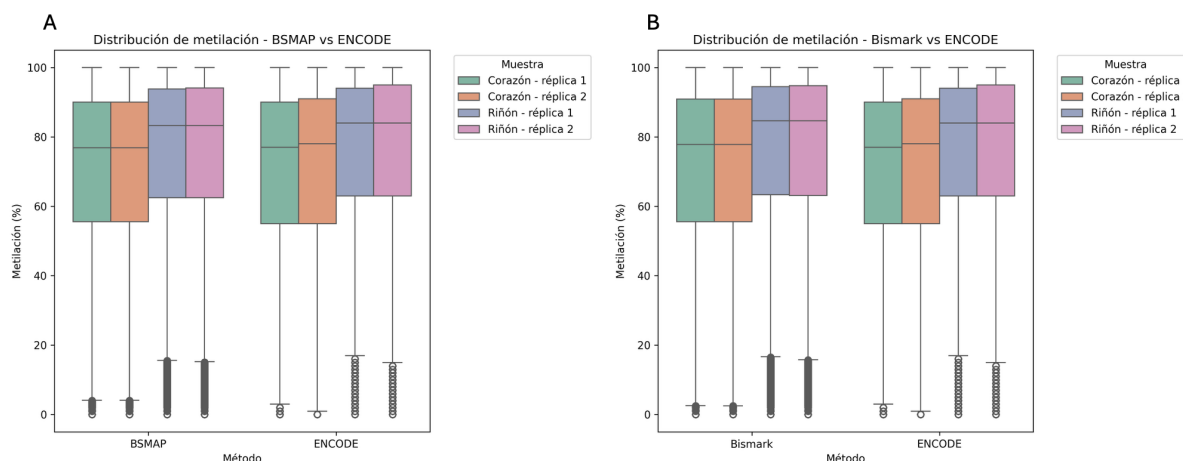


Figura 3.10. Distribución de niveles de metilación estimados por BSMAP (A) y Bismarck (B) en comparación con los datos de referencia de ENCODE. Se muestran los valores de metilación (%) para las muestras de corazón y riñón.

Con el objetivo de evaluar la consistencia en los niveles de metilación, se compararon los perfiles obtenidos a partir de las herramientas estudiadas con los reportados en ENCODE. En todos los casos se observó una fuerte correlación positiva ($r > 0,91$) entre los niveles de metilación estimados por cada herramienta y los datos de ENCODE. En las Figuras 3.11 y 3.12, se observa una tendencia lineal positiva, indicando que ambas herramientas reproducen de manera consistente los niveles de metilación. Estos resultados indican que, a pesar de haber discrepancias en sitios puntuales, la tendencia general muestra que ambos métodos reproducen adecuadamente los perfiles de metilación en comparación a los definidos en ENCODE.

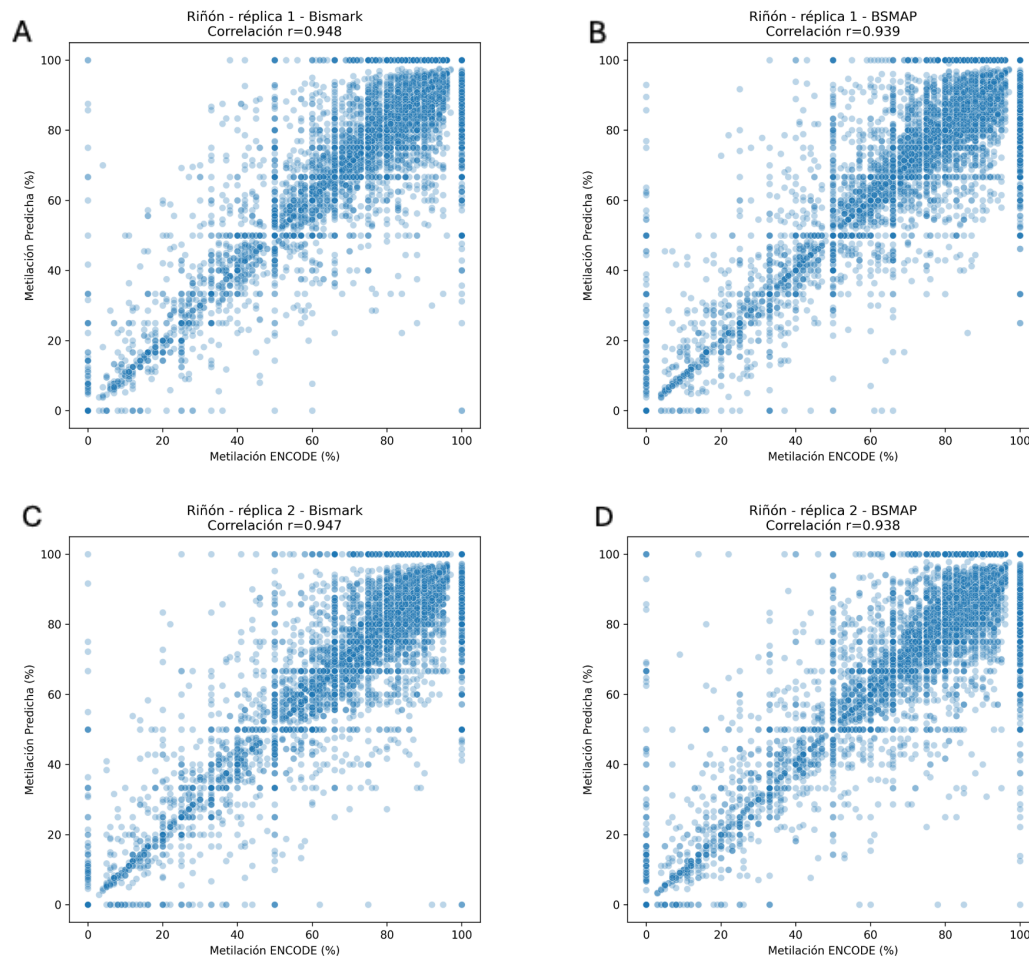


Figura 3.11. Correlación entre los niveles de metilación predichos por Bismark (A y C) y BSMAP (B y D) y los valores reportados en ENCODE en sitios CpG comunes para las muestras de riñón. Cada punto representa un sitio CpG.

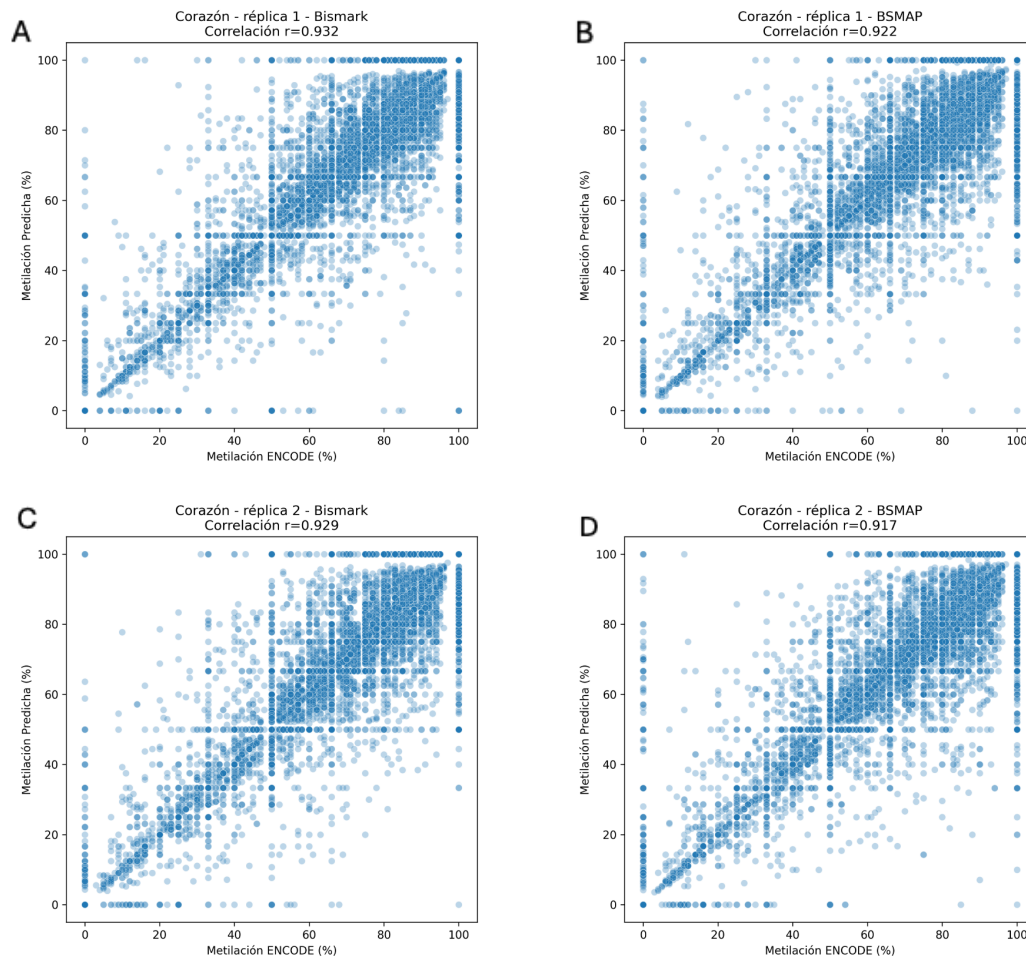


Figura 3.12. Correlación entre los niveles de metilación predichos por Bismark (A y C) y BSMAP (B y D) y los valores reportados en ENCODE en sitios CpG comunes para las muestras de corazón. Cada punto representa un sitio CpG.

3.3 Análisis de los patrones de metilación del ADN de elementos regulatorios genómicos anotados en Ensembl, utilizando el flujo de análisis desarrollado

Con el objetivo de evaluar la distribución funcional de los DMRs, se analizó el solapamiento de los resultados obtenidos por los cuatro flujos con el conjunto de elementos regulatorios anotados en el genoma de ratón en Ensembl. La Figura 3.13 muestra el porcentaje de DMRs identificados por cada estrategia que solapan cada uno de los tipos de elementos regulatorios. Este patrón de distribución refleja, en parte, la abundancia relativa de los distintos tipos de elementos en el genoma de ratón (GRCm39), donde los *enhancers* constituyen la categoría más representada (161.385), seguidos por promotores (31.408) y regiones de cromatina abierta (13.677). La mayor proporción de DMRs se encontró en regiones anotadas como *enhancers*, seguido por promotores y regiones de cromatina abierta, sin embargo, la mayor parte de los DMRs no se solapan con elementos regulatorios anotados. En particular, la

combinación BSMAP+methylKit mostró el mayor porcentaje de DMRs en *enhancers* (~50%) y un porcentaje mayor en promotores y cromatina abierta, con un menor porcentaje en regiones no anotadas. Este patrón puede interpretarse en parte como un efecto de enmascaramiento, ya que esta estrategia es la que reportó un mayor número de DMRs, aumentando la sensibilidad aparente. Sin embargo, una fracción considerable de estos DMRs se encuentra en regiones regulatorias, sugiriendo que podrían corresponder a DMRs biológicamente relevantes.

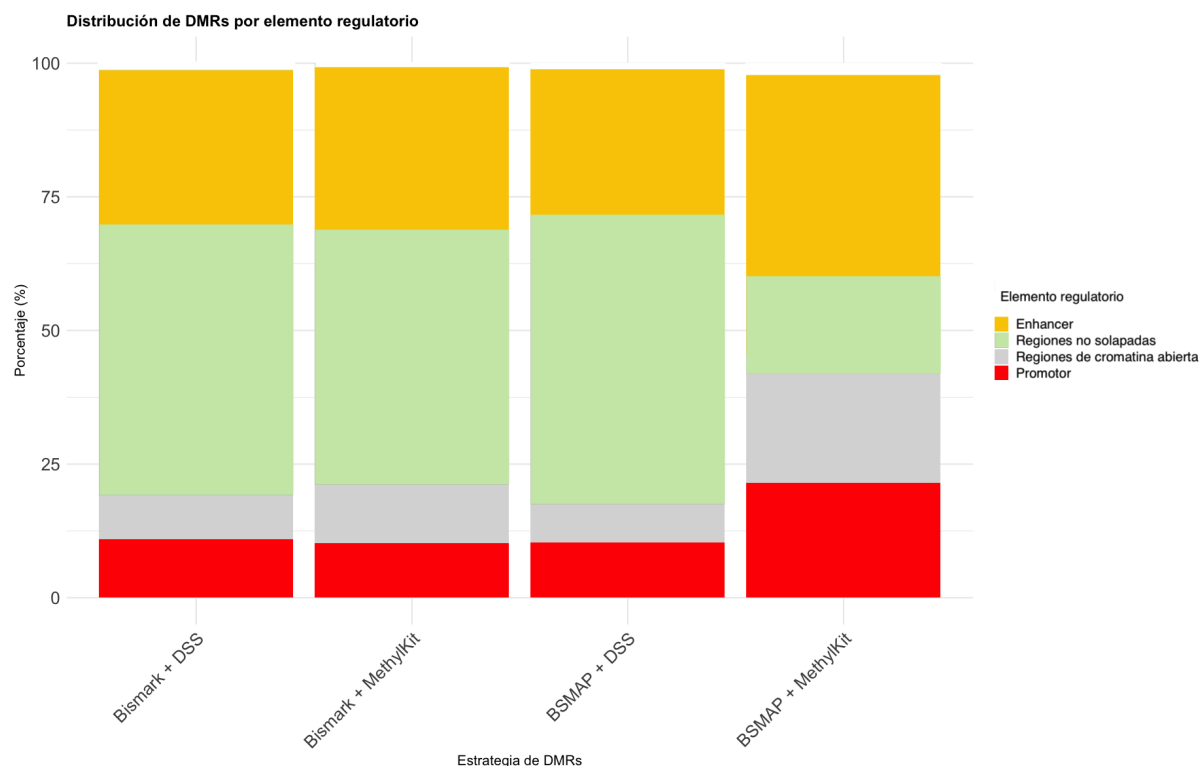


Figura 3.13. Distribución funcional de los DMRs para cada flujo de análisis estudiado. En rojo se encuentran los promotores, en amarillo los *enhancers*, en gris claro las regiones de cromatina abierta y en verde las regiones que no solapan con ningún elemento regulatorio.

Por otro lado, se estudió la relación entre los cambios en la metilación del ADN y la actividad de las regiones promotoras entre riñón y corazón. Para ello, se identificaron los DMRs encontrados en promotores, y se los clasificó según su estado de actividad en el corazón y en riñón (activo o inactivo) reportado por Ensembl. En este análisis, un valor de diferencia de metilación positivo indica que la región presenta mayor metilación en riñón que en corazón, mientras que un valor negativo indica lo contrario (mayor metilación en corazón). Así, en el caso de un promotor que cambia de Activo a Inactivo, se espera observar un aumento de metilación en riñón (diferencia de metilación > 0), consistente con la represión de la transcripción en ese tejido. A la inversa, en un promotor que cambia de Inactivo a Activo, se debería obtener una disminución de la metilación en riñón (diferencia de metilación < 0), lo que concuerda con la activación transcripcional de la región. La Figura 3.14 muestra las distribuciones de las diferencias en los valores de metilación de los DMRs que solapan promotores que mantienen su estado (Activo - Activo, Inactivo - Inactivo) y que cambian de estado de actividad entre tejidos (Activo - Inactivo e Inactivo - Activo). Como se puede observar, estos últimos presentan mayores diferencias de metilación, con medianas cercanas a

$\pm 50\%$, en comparación con los promotores que mantienen su estado (Activo - Activo o Inactivo - Inactivo), que muestran diferencias más cercanas a cero. Estos resultados son esperables teniendo en cuenta que la metilación en regiones promotoras se asocia a un mecanismo represivo de la transcripción génica. Además, la reproducibilidad del patrón en todos los métodos sugiere que las diferencias observadas no son artefactos metodológicos, sino que reflejan una tendencia dependiente del tejido.

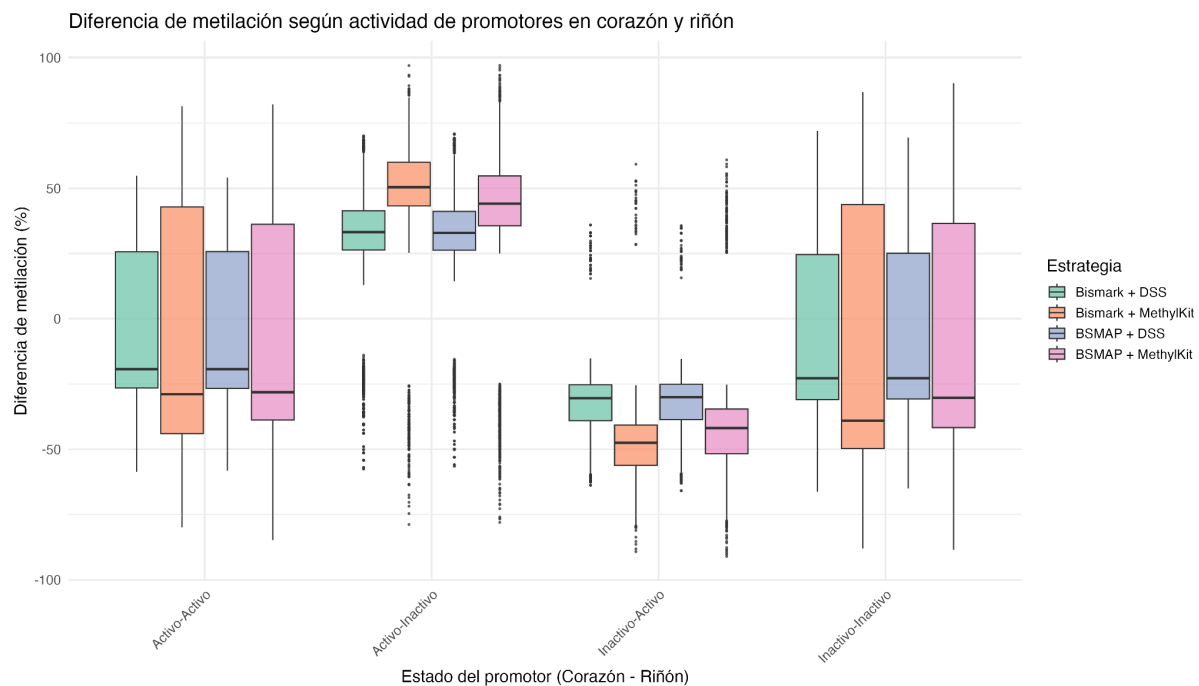


Figura 3.14. Distribución de las diferencias de metilación en promotores según el estado de actividad en corazón y riñón, caracterizado como Activo - Activo, Activo - Inactivo, Inactivo - Activo e Inactivo - Inactivo. La diferencia de metilación se calculó como Metilación en riñón - Metilación en corazón. Cada caja corresponde a las distintas estrategias estudiadas.

Al igual que con los promotores, se estudió la relación entre los cambios en la metilación del ADN y la actividad de las regiones *enhancers* entre riñón y corazón. Los resultados observados en la Figura 3.15 muestran una tendencia similar a los promotores. Los *enhancers* que tienen cambios en la actividad entre el corazón y el riñón presentan mayores diferencias de metilación, mientras que aquellos que mantienen su actividad en los dos tejidos, los valores son cercanos a cero.

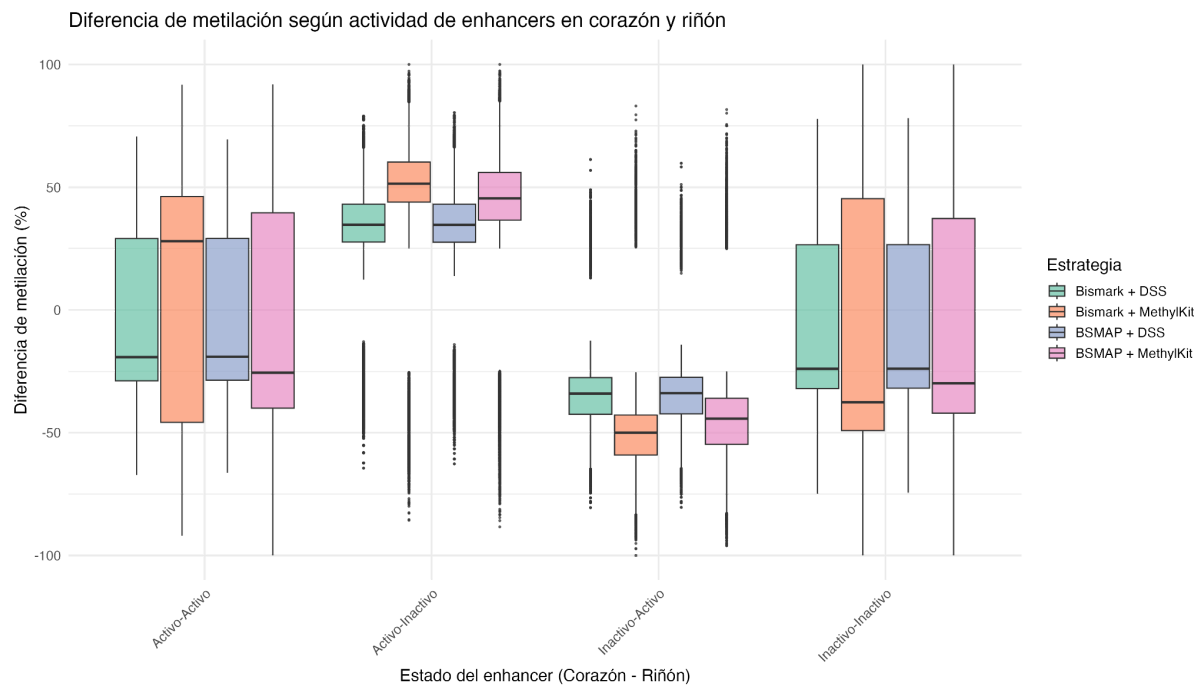


Figura 3.15. Distribución de las diferencias de metilación en *enhancers* según el estado de actividad en corazón y riñón, caracterizado como Activo - Activo, Activo - Inactivo, Inactivo - Activo e Inactivo - Inactivo. La diferencia de metilación se calculó como Metilación en riñón - Metilación en corazón. Cada caja corresponde a las distintas estrategias estudiadas.

Finalmente, se evaluó la relación entre la diferencia de metilación de los DMRs que solapan promotores y el cambio en la expresión génica (conocido como *fold change*, en escala log2) para los genes asociados a dichos promotores. La Figura 3.16 muestra cada gen coloreado según el estado de actividad del promotor. En verde se observan los genes con promotores activos para corazón e inactivos para riñón. Estos tienden a presentar cambios negativos en la expresión ($\text{Log2FC} < 0$) y mayor metilación a nivel renal, lo cual se asocia con una disminución en la expresión. En cambio, en celeste, se encuentran los genes con promotores inactivos para el tejido cardíaco y activos para el tejido renal. En este caso, la expresión de estos genes presenta cambios positivos, junto con una diferencia de metilación negativa, reflejando una hipometilación a nivel renal relacionada con un aumento de la expresión. Estos resultados evidencian la efectividad de las cuatro alternativas implementadas a la hora de detectar regiones DMRs funcionales. Es importante destacar que si bien hay una tendencia general, la relación entre metilación y expresión no es estrictamente lineal, ya que hay otros factores y procesos regulatorios que podrían también influir en la expresión génica.

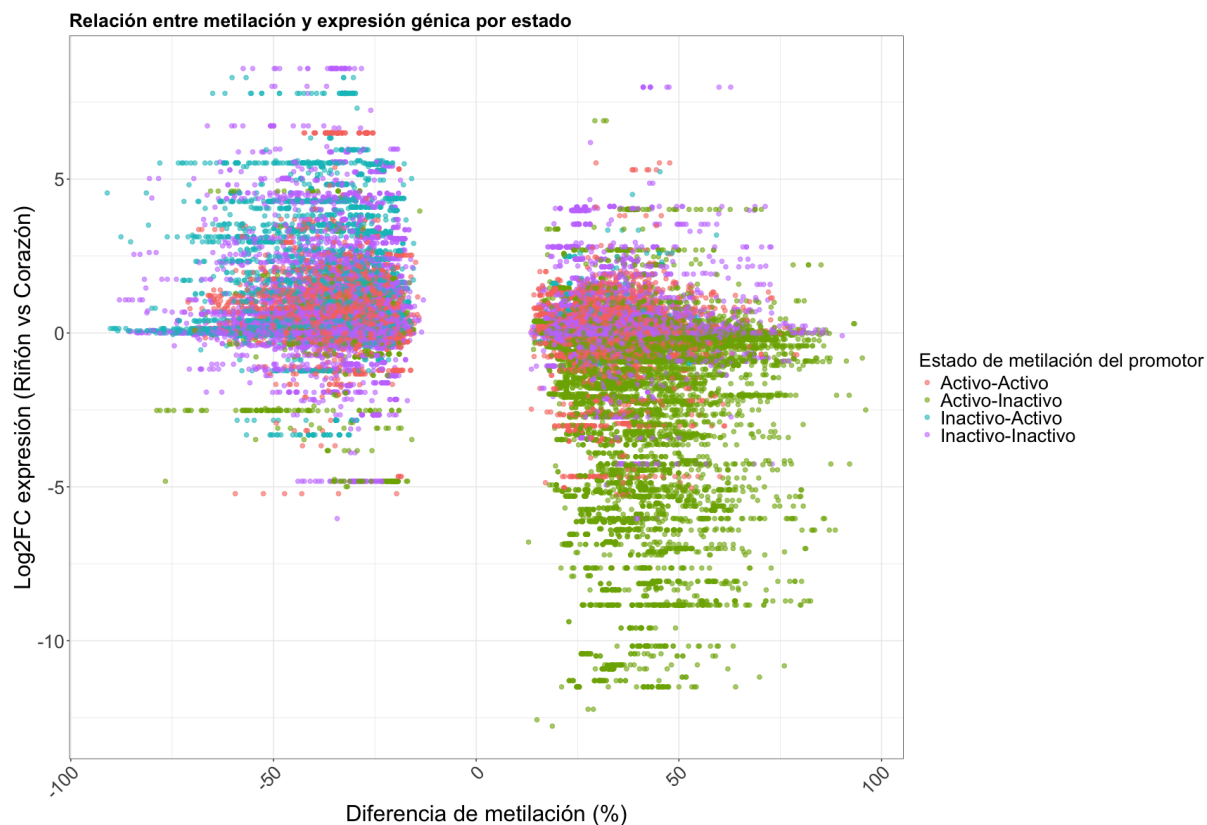


Figura 3.16. Relación entre metilación del promotor y la expresión génica. En verde se encuentran los genes que están reportados Activo en el corazón e Inactivo en riñón, en celeste aquellos que están Inactivos en corazón y Activos en riñón, en rojo aquellos que se encuentran Activos tanto en corazón como en riñón y en violeta los que están Inactivos en ambos tejidos. La diferencia de metilación se calculó como Metilación en riñón - Metilación en corazón.

Capítulo 4

Discusión y Conclusiones

La metilación del ADN es uno de los mecanismos epigenéticos que cumple un rol importante en la regulación de la expresión génica, el silenciamiento transcripcional y la estabilidad del genoma. Principalmente, los patrones de metilación en sitios CpG se encuentran relacionados con regiones regulatorias del genoma, como promotores y *enhancers*. En los últimos años, el análisis de estos patrones se potenció por el desarrollo de tecnologías como la secuenciación de bisulfito de genoma completo (WGBS), que permite cuantificar con alta resolución los niveles de metilación en millones de sitios del ADN en forma simultánea. El análisis de datos de WGBS implica una serie de etapas críticas que condicionan tanto la calidad de los resultados como la fidelidad en la interpretación biológica.

En este trabajo se diseñaron y evaluaron cuatro flujos de análisis de datos de WGBS para la estimación de los niveles de metilación del ADN. Adicionalmente, mediante el análisis de datos de libre acceso, se establecieron relaciones entre los patrones de metilación obtenidos y elementos regulatorios genómicos obtenidos de bases de datos públicas.

Las cuatro alternativas de análisis contempladas en esta tesis difieren principalmente en la herramienta utilizada para la alineación y cuantificación, Bismark y BSMAP, y en la herramienta empleada para la detección de regiones diferencialmente metiladas (DMRs), DSS y methylKit.

En cuanto al proceso de alineamiento, la comparación entre los alineadores Bismark y BSMAP mostró que, si bien ambos generan porcentajes de alineamientos únicos similares, cada uno presenta ventajas específicas: BSMAP es más eficiente en términos de tiempo y recursos computacionales, mientras que Bismark facilita la integración con herramientas posteriores, ofreciendo mayor flexibilidad y control en el análisis de metilación. Esta observación es importante para la planificación de proyectos de WGBS, especialmente cuando se manejan grandes volúmenes de datos o se requiere una integración completa con análisis estadísticos avanzados.

El análisis de DMRs mostró que DSS proporciona resultados consistentes y robustos, independientemente del alineador utilizado, mientras que methylKit puede ser más sensible a la estructura de los datos generados por cada alineador. Esta sensibilidad puede presentar una ventaja para la detección de DMRs en regiones de cobertura variable, pero requiere una interpretación cuidadosa para evitar sobreestimaciones. La fuerte correlación observada en los DMRs identificados con los diferentes flujos ($r > 0,96$) refuerza la reproducibilidad y la confiabilidad de los resultados, destacando que las discrepancias metodológicas afectan principalmente a regiones marginales de baja cobertura o con cambios sutiles de metilación.

Desde un punto de vista biológico, la distribución de los DMRs en elementos reguladores críticos, como promotores y *enhancers*, resalta la importancia funcional de la metilación en la regulación génica dependiente del tejido. La observación de hipometilación en promotores activos del corazón frente a hipermetilación en tejidos donde los genes permanecen inactivos refleja la coherencia con los modelos actuales de control epigenético. Además, la menor superposición de DMRs con regiones de cromatina abierta, sin función regulatoria aún definida, sugiere que los cambios más relevantes se concentran en elementos reguladores específicos, consolidando la capacidad del flujo para identificar patrones epigenéticos funcionalmente significativos.

La validación frente a los datos de referencia de ENCODE y la concordancia observada en los perfiles de metilación refuerzan la aplicabilidad del flujo desarrollado a distintos contextos biológicos y especies. Esto demuestra que las diferencias en algoritmos y estrategias de alineamiento, aunque relevantes desde el punto de vista computacional, no comprometen la interpretación biológica de los resultados a nivel global. La implementación de este flujo establece un marco confiable para futuros estudios de epigenómica, permitiendo no solo la comparación de condiciones experimentales, sino también la integración con otros niveles de información genómica, como la expresión génica o los datos de histonas y remodeladores de cromatina.

En conclusión, el presente trabajo valida un flujo de análisis de WGBS que combina precisión técnica, reproducibilidad estadística y relevancia biológica. La flexibilidad para alternar entre diferentes alineadores y herramientas de detección de DMRs permite adaptar el flujo a distintos escenarios experimentales y recursos computacionales. Además, la integración con anotaciones funcionales y datos de referencia proporciona un marco consistente para interpretar los cambios de metilación en un contexto biológico, consolidando la utilidad del pipeline como una herramienta confiable para estudios epigenómicos de alta resolución.

Bibliografía

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., & Mason, C. E. (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>

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Babraham Bioinformatics. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Assenov, Y., Müller, F., Lutsik, P., Walter, J., Lengauer, T., & Bock, C. (2014). Comprehensive analysis of DNA methylation data with RnBeads. *Nature Methods*, 11(11), 1138–1140. <https://doi.org/10.1038/nmeth.3115>

Beck, D., Ben Maamar, M., & Skinner, M. K. (2022). Genome-wide CpG density and DNA methylation analysis method (MeDIP, RRBS, and WGBS) comparisons. *Epigenetics*, 17(5), 518–530. <https://doi.org/10.1080/15592294.2021.2006757>

Benincasa, G., Marfella, R., & Paolisso, G. (2021). Epigenetics and pulmonary diseases in the horizon of precision medicine: A review. *European Respiratory Journal*, 57(6), 2003406. <https://doi.org/10.1183/13993003.03406-2020>

Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, 13(10), 705–719. <https://doi.org/10.1038/nrg3273>

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>

Chen, H., Smith, A. D., & Sun, F. (2016). WALT: Fast and accurate read mapping for bisulfite sequencing. *Bioinformatics*, 32(22), 3507–3509. <https://doi.org/10.1093/bioinformatics/btw490>

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

Dolzhenko, E., & Smith, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics*, 15, 215. <https://doi.org/10.1186/1471-2105-15-215>

Dor, Y., & Cedar, H. (2018). Principles of DNA methylation and their implications for biology and medicine. *The Lancet*, 392(10149), 777–786.

[https://doi.org/10.1016/S0140-6736\(18\)31268-6](https://doi.org/10.1016/S0140-6736(18)31268-6)

Dupont, C., Armant, D. R., & Brenner, C. A. (2009). Epigenetics: Definition, mechanisms and clinical perspective. *Seminars in Reproductive Medicine*, 27(5), 351–357. <https://doi.org/10.1055/s-0029-1237423>

Farrell, C., Thompson, M. J., & Lowe, R. (2021). BiSulfite Bolt: A bisulfite sequencing analysis platform. *GigaScience*, 10(5), giab033. <https://doi.org/10.1093/gigascience/giab033>

Feng, H., Conneely, K. N., & Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Research*, 42(8), e69. <https://doi.org/10.1093/nar/gku154>

Gong, W., Pan, X., Xu, D., Ji, G., Wang, Y., Tian, Y., Cai, J., Li, J., Zhang, Z., & Yuan, X. (2022). Benchmarking DNA methylation analysis of 14 alignment algorithms for whole genome bisulfite sequencing in mammals. *Computational and Structural Biotechnology Journal*, 20, 4704–4716. <https://doi.org/10.1016/j.csbj.2022.08.051>

Grehl, C., Wagner, F., Lemnian, I., Vornam, B., & Krezdorn, N. (2020). Performance of mapping approaches for whole-genome bisulfite sequencing data in crop plants. *Frontiers in Plant Science*, 11, 176. <https://doi.org/10.3389/fpls.2020.00176>

Hansen, K. D., Langmead, B., & Irizarry, R. A. (2012). BSmooth: From whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10), R83. <https://doi.org/10.1186/gb-2012-13-10-r83>

Hebestreit, K., Dugas, M., & Klein, H. U. (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13), 1647–1653. <https://doi.org/10.1093/bioinformatics/btt263>

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>

Jühling, F., Kretzmer, H., Bernhart, S. H., Otto, C., Stadler, P. F., & Hoffmann, S. (2016). Metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Research*, 26(2), 256–262. <https://doi.org/10.1101/gr.196394.115>

Krueger, F. (2012). 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 Bisulfite-Seq) libraries. *Babraham Bioinformatics*. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

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*, 9(2), 145–151. <https://doi.org/10.1038/nmeth.1828>

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

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), R25. <https://doi.org/10.1186/gb-2009-10-3-r25>

Li, N., Ye, M., Li, Y., Yan, Z., Butcher, L. M., Sun, J., Han, X., Chen, Q., Zhang, X., & Wang, J. (2010). Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods*, 52(3), 203–212. <https://doi.org/10.1016/j.ymeth.2010.04.003>

Li, Y. (2021). Modern epigenetics methods in biological research. *Methods*, 187, 104–113. <https://doi.org/10.1016/j.ymeth.2020.06.022>

Lin, Q. T., Yang, W., Zhang, X., Li, Q. G., Liu, Y. F., Yan, Q., & Sun, L. (2023). Systematic and benchmarking studies of pipelines for mammal WGBS data in the novel NGS platform. *BMC Bioinformatics*, 24(1), 33. <https://doi.org/10.1186/s12859-023-05163-w>

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q. M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., & 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. (2020). A comprehensive evaluation of computational tools to identify differential methylation regions using RRBS data. *Genomics*, 112(6), 4567–4576. <https://doi.org/10.1016/j.ygeno.2020.07.032>

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, 17(1), 10–12. <https://doi.org/10.14806/ej.17.1.200>

Park, Y., Figueroa, M. E., Rozek, L. S., & Sartor, M. A. (2014). MethylSig: A whole genome DNA methylation analysis pipeline. *Bioinformatics*, 30(17), 2414–2422. <https://doi.org/10.1093/bioinformatics/btu339>

Piao, Y., Liu, H., He, Y., Li, H., Zhou, Y., & Wu, H. (2021). Comprehensive evaluation of differential methylation analysis methods for bisulfite sequencing data. *International Journal*

of Environmental Research and Public Health, 18(15), 7975. <https://doi.org/10.3390/ijerph18157975>

Rauluseviciute, I., Drabløs, F., & Rye, M. B. (2019). DNA methylation data by sequencing: Experimental approaches and recommendations for tools and pipelines for data analysis. *Clinical Epigenetics*, 11(1), 193. <https://doi.org/10.1186/s13148-019-0755-8>

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47. <https://doi.org/10.1093/nar/gkv007>

Robinson, M. D., Kahraman, A., Law, C. W., Lindsay, H., Nowicka, M., Weber, L. M., & Zhou, X. (2014). Statistical methods for detecting differentially methylated loci and regions. *Frontiers in Genetics*, 5, 324. <https://doi.org/10.3389/fgene.2014.00324>

Singer, B. D. (2019). A practical guide to the measurement and analysis of DNA methylation. *American Journal of Respiratory Cell and Molecular Biology*, 61(4), 417–428. <https://doi.org/10.1165/rcmb.2019-0150TR>

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., Garvin, T., Kessler, M., Zhou, J., & Smith, A. D. (2013). A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLOS ONE*, 8(12), e81148. <https://doi.org/10.1371/journal.pone.0081148>

Sun, D., Xi, Y., Rodriguez, B., Park, H. J., Tong, P., Meong, M., Goodell, M. A., & Li, W. (2014). MOABS: Model based analysis of bisulfite sequencing data. *Genome Biology*, 15(2), R38. <https://doi.org/10.1186/gb-2014-15-2-r38>

Tolu, S. S., Viny, A. D., Amengual, J. E., & Figueroa, M. E. (2025). Getting the right combination to break the epigenetic code. *Nature Reviews Clinical Oncology*, 22, 117–133. <https://doi.org/10.1038/s41571-024-00972-1>

Waddington, C. H. (2012). The epigenotype (1942). *International Journal of Epidemiology*, 41(1), 10–13. <https://doi.org/10.1093/ije/dyr184>

Wu, Y., Qie, R., Cheng, M., Zeng, Y., Huang, S., Guo, C., Zhou, Q., Li, Q., Tian, G., Han, M., Zhang, Y., Wu, X., Li, Y., Zhao, Y., Yang, X., Feng, Y., Liu, D., Qin, P., Hu, D., Hu, F., ... Zhang, M. (2021). Air pollution and DNA methylation in adults: A systematic review and meta-analysis of observational studies. *Environmental Pollution*, 284, 117152. <https://doi.org/10.1016/j.envpol.2021.117152>

Xi, Y., & Li, W. (2009). BSMAP: Whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, 10, 232. <https://doi.org/10.1186/1471-2105-10-232>

Yang, X., Shao, X., Gao, L., & Zhang, S. (2016). Comparative DNA methylation analysis to decipher common and cell type-specific patterns among multiple cell types. *Briefings in Functional Genomics*, 15(6), 399–407. <https://doi.org/10.1093/bfgp/elw018>

Yang, Y., Liu, H., Liu, Y., Zhou, L., Zheng, X., Yue, R., Mattson, D. L., Kidambi, S., Liang, M., Liu, P., & Pan, X. (2023). E-value: A superior alternative to P-value and its adjustments in DNA methylation studies. *Briefings in Bioinformatics*, 24(4), bbad241. <https://doi.org/10.1093/bib/bbad241>

Zhang, F., Cui, Y., Zhang, T., & Yin, W. (2024). Epigenetic regulation of macrophage activation in chronic obstructive pulmonary disease. *Frontiers in Immunology*, 15, 1445372. <https://doi.org/10.3389/fimmu.2024.1445372>

Zhang, K., Chen, G., He, J., Chen, Z., Pan, M., Tong, J., Liu, F., & Xiang, H. (2024). DNA methylation mediates the effects of PM2.5 and O3 on ceramide metabolism: A novel mechanistic link between air pollution and insulin resistance. *Journal of Hazardous Materials*, 469, 133864. <https://doi.org/10.1016/j.jhazmat.2024.133864>

Zhang, Y., Baheti, S., & Sun, Z. (2018). Statistical method evaluation for differentially methylated CpGs in base resolution next-generation DNA sequencing data. *Briefings in Bioinformatics*, 19(3), 374–386. <https://doi.org/10.1093/bib/bbw133>

Zhou, L., Ng, H. K., Drautz-Moses, D. I., Schuster, S. C., Beck, S., Kim, C., Chambers, J. C., & Loh, M. (2019). Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. *Scientific Reports*, 9(1), 10383. <https://doi.org/10.1038/s41598-019-46897-6>