Determinación de marcas epigenéticas específicas de tejido para la identificación del origen de las vesículas circulantes

Neus Martínez Micaelo
Máster en Bioinformática y Bioestadística
Área: Estudios genéticos de enfermedades humanas

Dr. Helena Brunel Montaner
Dr. David Merino Arranz

05/06/2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	Determinación de marcas epigenéticas específicas de tejido para la identificación del origen de las vesículas circulantes
Nombre del autor:	*Neus Martínez Micaelo*
Nombre del consultor/a:	**Dr. Helena Brunel Montaner**
Nombre del PRA:	**Dr. David Merino Arranz**
Fecha de entrega (mm/aaaa):	06/2018
Titulación::	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	Área: Estudios genéticos de enfermedades humanas
Idioma del trabajo:	*Español*
Palabras clave	*Metilación del ADN, vesículas extracelulares*

Resumen del Trabajo (máximo 250 palabras):

El presente Trabajo Final de Máster (TFM) pretende dar respuesta a la necesidad de identificar el origen de las vesículas extracelulares circulantes. Así teniendo en cuenta que las vesículas extracelulares contienen material genético de la célula progenitora, incluyendo ADN genómico, y que cada tipo celular tiene un patrón epigenético específico, el objetivo principal de este TFM es la aplicación de métodos y herramientas bioinformáticas en datos epigenéticos depositados en bases de datos públicos, para el diseño y el desarrollo de un algoritmo clasificador de marcadores epigenéticos capaz de diferenciar entre tejidos. Para alcanzar el objetivo principal se han analizado los datos referentes a la metilación de DNA de un total de 252 experimentos, agrupados en 24 clústers específicos de tejido o tipo celular, y mediante la creación de una función (*BiosourceMapper*) se han identificado posiciones o *loci* diferencialmente metilados para el tejido o tipo celular seleccionado, así como las regiones enriquecidas en posiciones diferencialmente metilades. Además para mejorar la visualización e interpretación de los datos se ha creado la aplicación shiny *Biosource Mapper App*, que permite poden acceder y analizar los resultados generados durante este TFM a cualquier usuario. La principal conclusión obtenida es que mediante el análisis bioinformático del patrón de metilación con la función *BiosourceMapper* implementada en la aplicación *Biosource Mapper App* es posible obtener marcas específicas de tejido.

El acceso a *Biosource Mapper App* puede realizarse a través del siguiente link: https://biosourcemapper.shinyapps.io/biosourcemapper_app/

Abstract (in English, 250 words or less):

The aim of this Final Master's Project (FMP) is to identify the origin of circulating extracellular vesicles. Taking into account that the extracellular vesicles contain genetic material from the progenitor cell, including genomic DNA, and that each cell type has a specific epigenetic pattern, the main objective of this FMP is the application of bioinformatic methods and tools on epigenetic data deposited in public repositories, for the design and development of an algorithm based on epigenetic marks to differentiate between tissues. To achieve this objective, the DNA methylation data from a total of 252 experiments, grouped in 24 tissue-specific or cell-type clusters, were analyzed and positions or loci were identified by creating a function (BiosourceMapper). Differentially methylated positions specific for the tissue or cell type selected, as well as regions enriched in differentially methylated positions, were found. Furthermore, to improve the visualization and interpretation of the data, the shiny *Biosource Mapper App* has been created, allowing users to access and analyze the results generated within this TFM. The main conclusion obtained is that bioinformatic analysis of the methylation pattern with the *BiosourceMapper* function implemented in the *Biosource Mapper App* identify specific tissue marks.

Biosource Mapper App can be accessed via the following link:
https://biosourcemapper.shinyapps.io/biosourcemapper_app/

Índice

Lista de figuras

1. Introducción

1.1. Contexto y justificación del Trabajo

Introducción a las vesículas extracelulares

Las vesículas extracelulares son un grupo heterogéneo de partículas de diámetros entre 30 y 5000 nm recubiertas por una bicapa lipídica, secretadas tanto por células sanas como patológicas, presentes en todos los fluidos corporales (Figura 1), incluyendo la sangre periférica y, cuya función principal es el transporte de moléculas funcionalmente activas, incluyendo; ADN, ARN mensajero, microARNs, proteínas y lípidos, de la célula de origen a células receptoras (1).

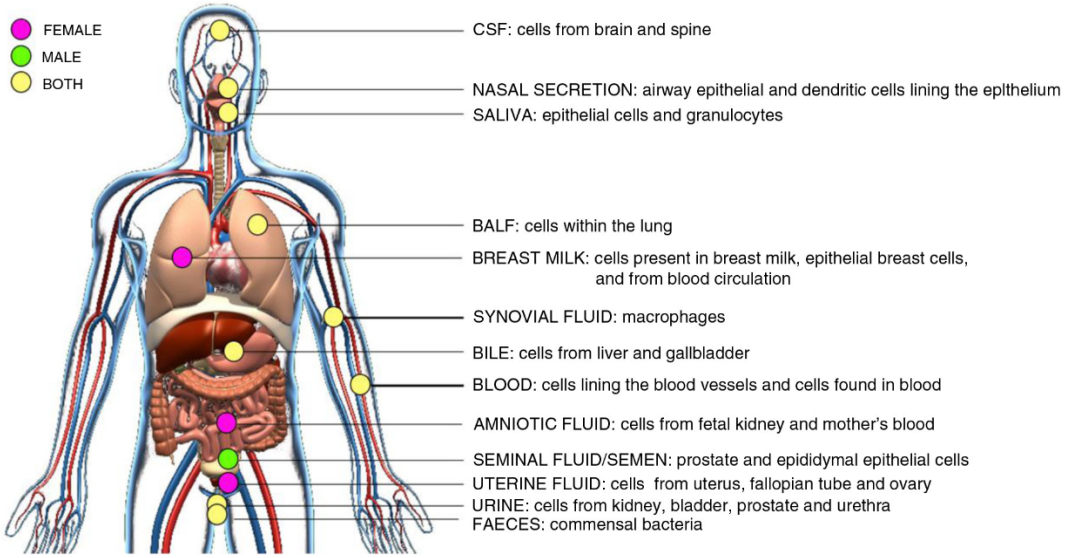

Figura 1. Uno delos mecanismos por los cuales las células que conforman los diferentes tejidos del organismo pueden comunicarse es a través de la secreción de vesículas extracelulares. Actualmente se han identificados vesículas extracelulares en prácticamente todos los fluidos del cuerpo humano. En la figura se destacan dichos fluidos y su posible origen celular. CSF = fluido cerebroespinal; BALF = fluido del lavado broncoalveolar. Fuente: (2)**.**

Las vesículas extracelulares se originan en diferentes compartimentos sub-celulares y su proporción varía en función del estado fisiológico así como de las células de origen. A parte de por su biogénesis, la vesículas extracelulares también pueden ser diferenciadas por su tamaño, diferenciando entre cuerpos

apoptóticos (1-5μm), microvesículas (100nm-1μm) y exosomas (30-150nm) (Figura 2) (3).

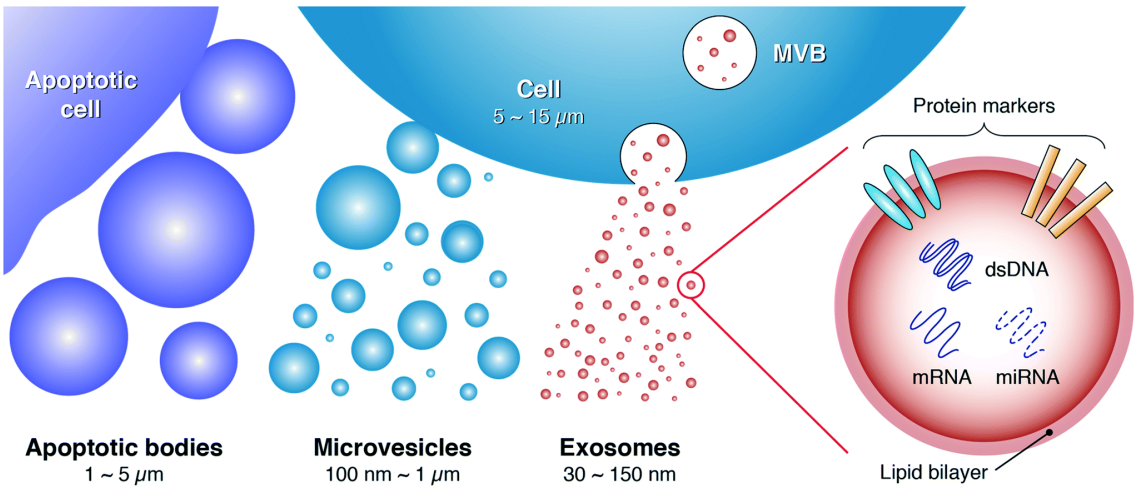

Figura 2. Dentro de las vesículas extracelulares se encuentran los cuerpos apoptóticos, las microvesículas y los exosomas la cuales desempeñan un papel clave en la mediación de la comunicación intercelular. Cabe destacar que tanto sus marcadores de superficie como su carga se asemejan al contenido de la célula de la cual han sido liberadas. Fuente (4).

Inicialmente la secreción de las células extracelulares se atribuyó a un mecanismo de eliminación de compuestos innecesarios por parte de las células (5). Sin embargo, ahora sabemos que las vesículas extracelulares son más que simples portadores de desechos, y el interés principal en el campo se centra en la capacidad que dichas vesículas tienen de intercambiar componentes entre células, llegando a convertirse en un nuevo mecanismo de comunicación inter-celular, actuando como vehículos de señalización tanto en procesos homeostáticos como en desarrollos patológicos (6).

Pese a que actualmente existe un gran interés científico en el uso de estas vesículas como una valiosa fuente no invasiva de información tisular a nivel molecular a través de una simple muestra de sangre o cualquier otro fluido (7), la metodología para estudiarlas no está lo suficientemente desarrollada y hasta el momento la mayoría de la bibliografía publicada se ha centrado en estudiar sus funciones potenciales más que su origen. De hecho los protocolos actuales disponibles para su aislamiento dan como resultado una población heterogénea de vesículas de origen desconocido.

De esta manera, aunque se ha avanzado mucho en los últimos años en la caracterización de las vesículas extracelulares, una de las mayores limitaciones

es la identificación del origen de las vesículas extracelulares aisladas, puesto que, podemos conocer y descifrar el mensaje pero no el emisor, información indispensable para relacionar el cargo molecular que contienen con el estado funcional del tejido originario. Situando así la búsqueda de posibles biomarcadores que ayuden a identificar el origen de dichas partículas incrementaría su potencial para su posible aplicación clínica.

Epigenética y metilación del ADN

Todas las células del organismo contienen esencialmente el mismo genoma, proveniente de una única célula, y es la manera de interpretar esta información genética en cada momento por cada una de ellas lo que les proporciona su identidad, estructura y funcionalidad específica. Así, de los cerca de 25.000 genes que codifican por proteínas que han sido identificados en el genoma de los mamíferos, aproximadamente la mitad se expresan en cualquier tipo celular y el resto se expresan únicamente en determinados tipos celulares, o presentan diferentes patrones de expresión en función de la célula en la que expresen (8).

Muchas de las diferencias observadas en cuanto a la expresión génica surgen durante el desarrollo y posteriormente se mantienen a través de la mitosis. Alteraciones estables de este tipo se conocen como marcas epigenéticas, ya que son hereditarias a corto plazo, pero no implican cambios en la secuencia del ADN (9). Estos cambios epigenéticos pueden modular la expresión génica mediante la regulación del estado de metilación del ADN, cambiando la organización y le grado de empaquetamiento de la cromatina o la regulación post-transcripcional del ARN mensajero a través de ARN no codificante. De esta manera, se puede definir la epigenética como una regulación compleja de la expresión génica mediante un código de encriptación físico y químico escrito sobre la secuencia de ADN, provocando que, antes de transcribir la información biológica contenida en el genoma de cualquier célula sea indispensable desencriptar y leer dicho código genético.

La metilación del ADN, en la que se transfiere de manera enzimática un grupo metilo al carbono 5 de la citosina presente en el dinucleótido de citosina-fosfato-

guanina (CpGs) de las cadenas de DNA, es una forma única de regulación génica ya que, a diferencia de otros mecanismos de control, implica cambios covalentes en el genoma que proporcionan estabilidad a largo plazo (Figura 3). Aunque la mayoría de 5-metilcitosinas están presentes en los los dinucelótidos CpGs, también se encuentran con menos frecuencia en contextos que no son CpG, por ejemplo, CHG y CHH (donde H = A, T o C). La metilación del DNA no regula únicamente la transcripción de los genes cercanos a la marca de metilación (o regulación en *cis*), sino que también puede actuar en *trans*, participando de la organización nuclear y en el establecimiento de territorios cromosómicos específicos. La consecuencia funcional de la metilación es la represión de la transcripción génica. Alteraciones o colocaciones aberrantes en las marcas de metilación han sido implicadas con el desarrollo de diversas enfermedades.

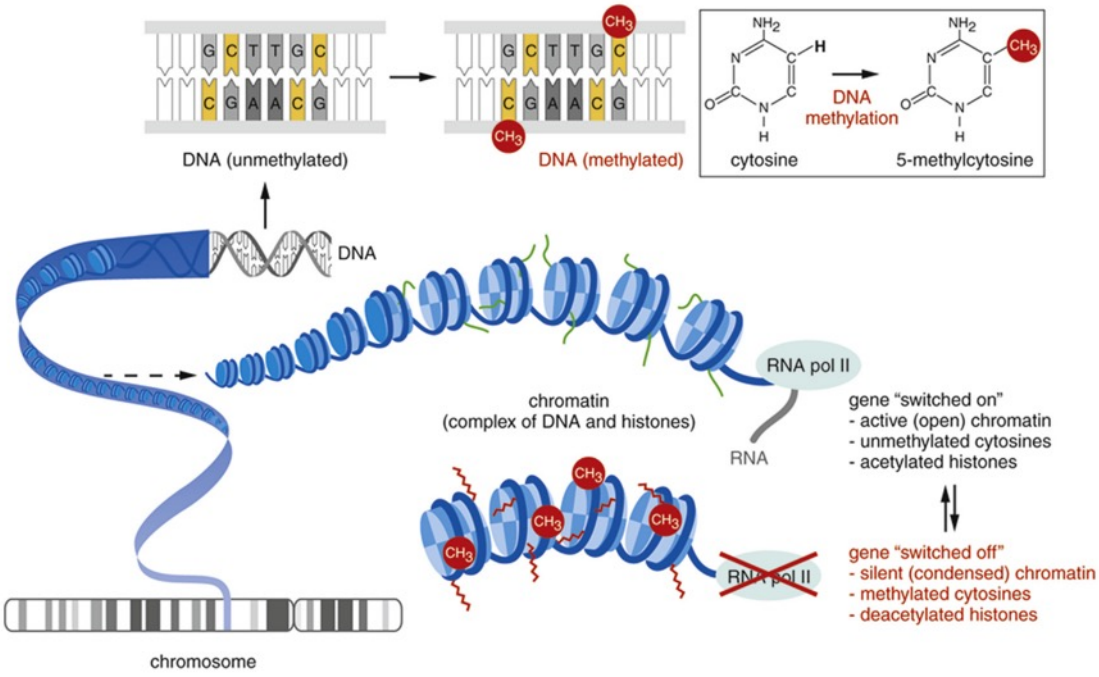

Figura 3. La metilación del ADN es uno de los mecanismos por los que la expresión génica puede ser regulada antes del inicio de la transcripción. La adición de un grupo metilo (CH_3) a residuos de citosinas forma la 5-metilcitosina, modificando la forma en la que las proteínas que controlan la expresión génica (como la ARN polimerasa II o RNA pol II) interactúan con la secuencia de ADN. Adaptado de: (10).

Durante las últimas décadas, las ciencias biomédicas han estado centradas en la identificación de genes, proteínas o metabolitos que, mediante su regulación diferencial, pudieran contribuir a un rasgo o fenotipo determinado. De esta manera enfoques como el de las ómicas, definidas como el conjunto de técnicas

basadas en el análisis de todo el conjunto de transcritos (Transcriptómica), proteínas (Proteómica) o metabolitos (Metabolómica), se han basado en analizar rasgos diferenciales resultantes de la regulación de la parte codificante del genoma, sin tener en cuenta la parte no codificante, que en humanos puede representar aproximadamente el 98% del genoma.

Dado que las tecnologías de secuenciación de última generación o *next-generation sequencing technologies* (NGS) se han convertido en poderosas herramientas que permiten el mapeo de marcas de metilación del ADN de todo el genoma, incluyendo las regiones de codificación y no codificación, junto con el hecho de que existen patrones de metilación únicos para cada tipo celular (11), que se conservan entre las células del mismo tipo, ya sea en el mismo individuo o entre individuos, y que son muy estables en condiciones fisiológicas o patológicas (12), hacen de la identificación de estos patrones de metilación de ADN diferenciales como una herramienta muy potente para proporcionar una gran cantidad de nuevas hipótesis y marcadores tanto para determinar el tipo celular del que proviene el material genético como para testarlo en relación con la salud y la enfermedad.

Justificación del Trabajo Final de Máster

Teniendo en cuenta todo lo descrito con anterioridad, el presente Trabajo Final de Máster (TFM) pretende dar respuesta a la necesidad de identificar el origen de las vesículas extracelulares circulantes, por lo que mediante la aplicación de métodos y herramientas bioinformáticas en datos epigenéticos depositados en bases de datos públicos, se ha diseñado y desarrollado un algoritmo clasificador de marcadores epigenéticos capaz de diferenciar entre tejidos.

La elección del área y el tema del TFM está directamente asociada con la experiencia profesional y las inquietudes de la estudiante, puesto que, tras identificar en primera persona la limitación actual existente, se ha considerado la elaboración de este TFM tanto un reto como una oportunidad única para aplicar los conocimientos adquiridos durante el transcurso de éste Máster, y dar así, una solución bioinformática a una necesidad experimental.

1.2. Objetivos del Trabajo

Objetivos generales

Si se tiene en cuenta que las vesículas extracelulares contienen material genético de la célula progenitora, incluyendo ADN genómico, y que cada tipo celular tiene un patrón epigenético específico, el objetivo principal de este TFM es la aplicación de métodos y herramientas bioinformáticas en datos epigenéticos depositados en bases de datos públicos, para el diseño y el desarrollo de un algoritmo clasificador de marcadores epigenéticos capaz de diferenciar entre tejidos, con la finalidad de identificar el tejido de procedencia de las vesículas extracelulares circulantes. Para ello se han planteado los siguientes objetivos generales:

1. Obtener y preparar datos epigenéticos representativos de muestras de tejido o tipos celulares depositados en repositorios públicos.
2. Diseñar, desarrollar e implementar un algoritmo clasificador de biomarcadores epigenéticos específicos.

Objetivos específicos

Para alcanzar los objetivos generales se deberán alcanzar los siguientes objetivos específicos:

1. Obtener datos epigenéticos representativos depositados en repositorios públicos
2. Pre-procesar y normalizar de los datos
3. Diseñar y desarrollar el algoritmo
4. Implementar el algoritmo y generar un catálogo de marcadores epigenéticos específicos
5. Crear una interfaz gráfica en la que poder consultar el catálogo de marcadores

1.3. Enfoque y método seguido

La ejecución de los objetivos planteados para este TFM pueden realizarse mediante diferentes estrategias y metodologías, es por ello que en primer lugar se realizará un estudio del estado del arte mediante el cual se evaluaran las diferentes estrategias. El enfoque y los métodos escogidos irán en consonancia con la metodología y la naturaleza de los datos obtenidos así como de la finalidad de este TFM. En este sentido, *a priori* entre las estrategias más importantes para la consecución de los objetivos planteados son los relacionados con el análisis de los datos de metilación incluyendo, la elección del tipo de datos de partida, el método utilizado para corregir los datos y eliminar el posible *batch effect*, así como el modelo estadístico a utilizar en el algoritmo que se acabe implementado.

1.4. Planificación del Trabajo

Una vez definidos los objetivos y el enfoque del TFM, a continuación se desglosan las diferentes tareas y la planificación realizada.

Tarea 1. Estudio del estado del arte relacionado con la temática y la metodología propuesta
Tarea 2. Identificación y obtención de los datos
Tarea 3. Pre-procesado y normalización de los datos
Tarea 4. Mapeado y *clustering* del conjunto de datos.
Tarea 5. Diseño del algoritmo.
Tarea 6. Identificación de citosinas diferencialmente metiladas en los tejidos de interés (*Differential methylated positions* o DMPs)
Tarea 7. Selección de las regiones que contengan la DMPs o *slicing*
Tarea 8. Identificación de las regiones diferencialmente metiladas (*Differential methylated regions* o DMRs) con las que se creará un catálogo de regiones específicas (marcadores) para los diferentes tejidos incluidos en el análisis.
Tarea 9. Crear una interfaz gráfica que facilite la consulta de los marcadores específicos de tejido.

Los hitos planificados en base al calendario del plan docente, se describen a continuación:

- Entrega de la propuesta de TFM (PAC 0)
- Entrega del plan de trabajo (PAC1)
- Fase 1 (PAC 2)
- Fase 2 (PAC3)
- Redacción de la memoria (PAC4)
- Preparación de la presentación (PAC5)

A continuación (ver Tabla 1 y Figura 4) se incluye la temporalización de las diferentes hitos y tareas planteados así como el diagrama de Gantt correspondiente.

Tabla 1. Temporalización de los diferentes hitos y tareas incluidos en el TFM.

	Fecha de inicio	**Fecha final**	**Horas estimadas**
PEC 0 - Propuesta de TFM	21/2/18	5/3/18	10
PEC 1 - Plan de trabajo	6/3/18	19/3/18	10
PEC 2 - Desarrollo del trabajo - Fase 1			
Tarea 1. Estudio del estado del arte	20/3/18	24/3/18	22
Tarea 2. Identificación y obtención de datos	25/3/18	30/3/18	29
Tarea 3. Pre-procesado y normalización de datos	1/4/18	6/4/18	29
Tarea 4. Mapeado y clustering de los datos	7/4/18	14/4/18	35
Tarea 5. Diseño del algoritmo	15/4/18	23/4/18	45
PEC 3 - Desarrollo del trabajo - Fase 2			
Tarea 6. Identificación de DMPs	24/4/18	28/4/18	20
Tarea 7. Selección de regiones o slicing	29/4/18	3/5/18	20
Tarea 8. Identificación de DMRs	4/5/18	12/5/18	40
Tarea 9. Creación de una interfaz gráfica	13/5/18	21/5/18	40
PEC 4 - Redacción de la memoria			
Elaboración de la memoria	22/5/18	5/6/18	50
Memoria del TFM	5/6/18	5/6/18	
PEC 5 - Elaboración de la presentación			
Preparación de la presentación	6/6/18	13/6/18	25
Defensa pública	14/6/18	25/6/18	---
		TOTAL	**375**

En la siguiente figura (Figura 4), se muestra el diagrama de Gantt con la temporalización de las tareas y los hitos propuestos.

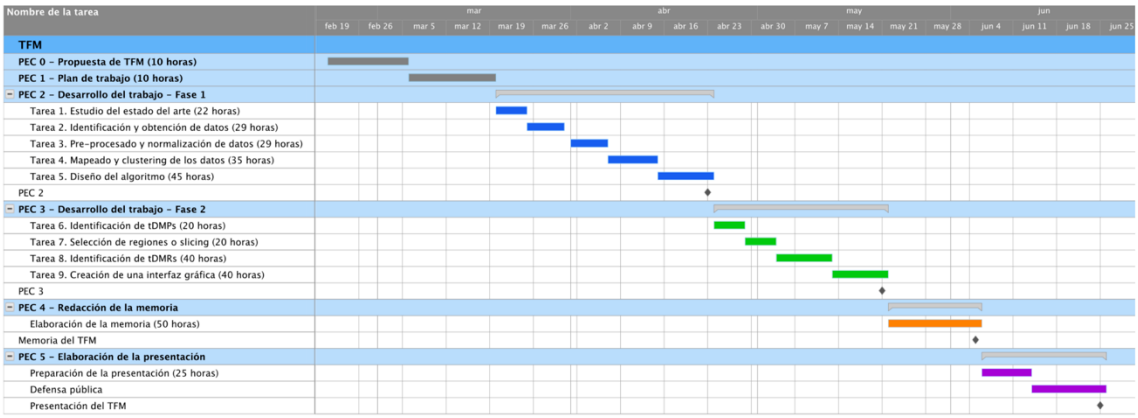

Figura 4. Diagrama de Gantt con la temporalización de los diferentes hitos y tareas incluidos en el TFM.

1.5. Breve sumario de productos obtenidos

Tal y como se ha indicado en la descripción tanto de los objetivos como en la planificación del trabajo, mediante la ejecución de las diferentes tareas ha permitido obtener un catálogo de posiciones y regiones diferencialmente metiladas específicas de tejido.

Para ello se ha diseñado una función de R, *BiosourceMapper* que identifica posiciones diferencialmente metiladas, las regiones en las que se agrupan dichas posiciones y permite la visualización y contextualización a nivel genómico.

Además dichos resultados han sido plasmados en una aplicación *shiny*, llamada *BioSource Mapper App*, a la que se le ha implementado la función *BiosourceMapper* y mediante la cual permite determinar y visualizar los resultados obtenidos acercando su uso a cualquier usuario. *BioSource Mapper App* permite que seleccionando un tejido o tipo celular de los disponibles, identificar las posiciones y regiones diferencialmente metiladas para ese tejido permitiendo además su visualización y contextualización a nivel de localización cromosómica.

1.6. Breve descripción de los otros capítulos de la memoria

En el resto de capítulos de la memoria se describirá la realización y consecución de las diferentes tareas, integrando el estado del arte para cada una de ellas así como las alternativas y los criterios que se han utilizado para tomar las diferentes decisiones.

2. Identificación y obtención de los datos

Una de las estrategias más importantes para la consecución de los objetivos planteados en este trabajo es la elección del tipo de datos y los recursos públicos que serán utilizados para posteriormente trabajar con ellos.

Existen diferentes enfoques a partir de los cuales puede determinarse el perfil de metilación del ADN, incluyendo la digestión con endonucleasas de restricción sensibles a la metilación, mediante enfoques basados en el enriquecimiento y conversión con bisulfito seguida de secuenciación de ADN o mediante ensayos basados en *arrays*. En base a estos enfoques existen diferentes técnicas experimentales para la determinación del perfil de metilación de todo el genoma en general o de alto rendimiento (*high-throughput),* entre las más utilizadas destacan las técnicas basadas en el enriquecimiento mediante proteínas o anticuerpos al dominio de unión metil-CpG, MBD y MeDIP correspondientemente, las cuales pueden ser determinadas acopladas a *arrays* o secuenciación (MBD-chip/MBDcap-seq o MeDIP/MeDIP-seq consecuentemente), las técnicas basadas en secuenciación en base a enzimas de restricción sensibles a metilación (MRE-seq), la secuenciación de bisulfito oxidativo (oxBS-seq) y la oxidación mediada por la proteína TET o TAB-seq, las técnicas basadas en secuenciación de bisulfito de representación reducida (RRBS), la secuenciación de bisulfito de genoma completo (WGBS) o mediante los arrays de Illumina, el Infinium HumanMethylation450 (450K) BeadChip y el Infinium MethylationEPIC (850K) BeadChip (Figura 5). En base a la técnica experimental seleccionada variará la cantidad de partida y el procesado experimental de la muestra de ADN, la resolución, la cobertura de la región genómica y los análisis bioinformáticos (13).

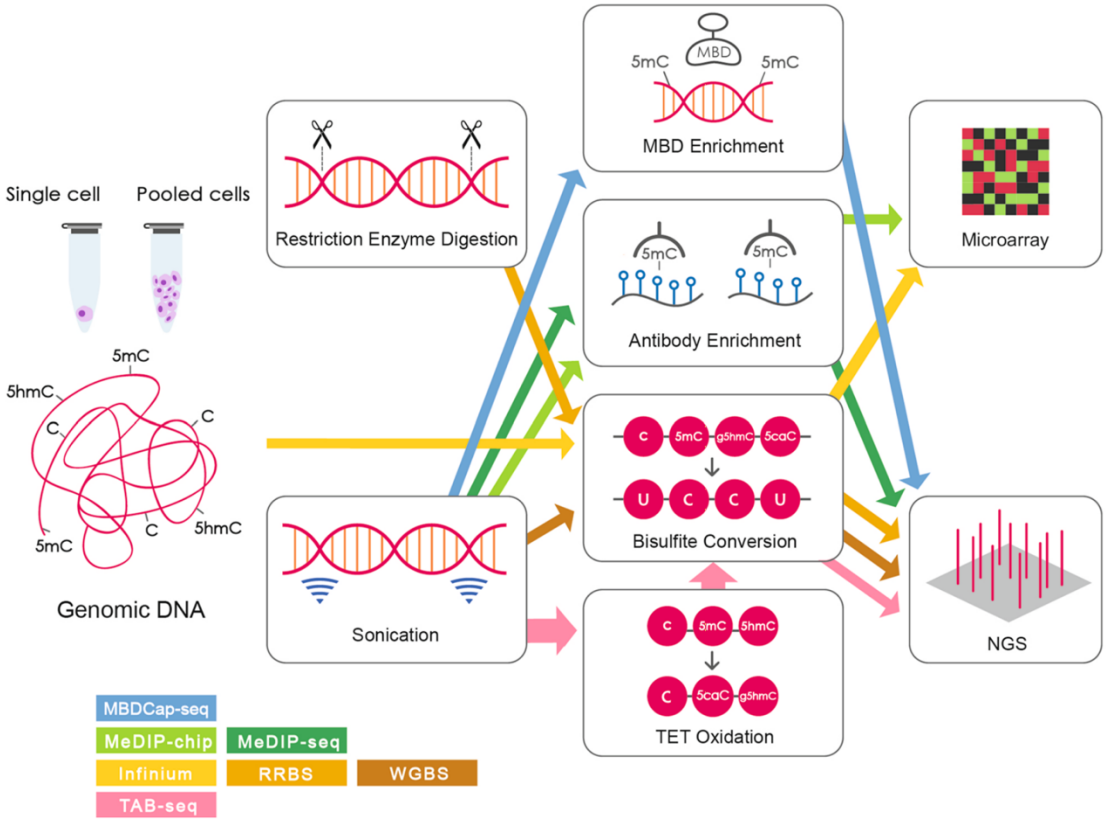

Figura 5. Métodos más utilizados para el análisis de metilación del ADN de todo el genoma. a) Los procedimientos pueden implicar la fragmentación del ADN genómico mediante digestión con enzimas de restricción o sonicación. El ADN genómico puede ser sometido a enriquecimiento de MBD, inmunoprecipitación, conversión de bisulfito o oxidación de TET antes del análisis por *arrays* o la plataforma de secuenciación de próxima generación. Adaptado de: (13).

Recursos públicos asociados a datos de metilación

La determinación del perfil de metilación del ADN utilizando estrategias de alto rendimiento como los acoplados a la secuenciación de última generación (*NGS*) resulta en la producción de enormes cantidades de información que pueden utilizarse para resolver muchas cuestiones pero que a su vez requiere de herramientas bioinformáticas que faciliten el análisis de este tipo de datos masivos. Es por ello que para resolver esta creciente demanda se han desarrollado diversos recursos públicos que proporcionan acceso al almacenamiento, la exploración y recuperación de datos de metilación. Alguno de los recursos públicos existentes son por ejemplo la *Database of CpG islands and Analytical Tools* (DBCAT), *MethDB*, *MethBank*, *MethBase*, *MethylomeDB*, *NCBI Epigenomics*, *EpiFactors*, *NGSmethDB*, *DeepBlue* o la *Gametogenesis*

Epigenetic Modification Database (14–20). Así mediante el acceso a los recursos mencionados se puede acceder a datos depositados procedentes de los consorcios *Epigenome datasets* del *International Human Epigenome Consortium* (*IHEC*), del *Canadian Epigenetics, Environment and Health Research Consortium* (*CEEHRC*), del proyecto europeo *BLUEPRINT*, del proyecto alemán *DEEP*, de la base de datos *NCBI Gene Expression Omnibus* (*GEO*), de la base de datos del *European BioinIormatics Institute* (*EBI*), del *European Genome-Phenome Archive* (*EGA*), o del proyecto de la Enciclopedia de los Elementos de ADN (ENCODE) (21–23). Estos proyectos epigenómicos a gran escala generan cantidades masivas de datos de una amplia gama de tipos de células y tejidos, como células sanguíneas, células madre, líneas celulares cultivadas, así como tejidos primarios, para mapear los cambios epigenéticos asociados con enfermedades específicas.

Secuenciación basada en el tratamiento con bisulfito

El tratamiento con bisulfito junto con la secuenciación es la técnica de referencia para la determinación del metiloma, y, en los recursos mencionados anteriormente destaca el uso de las técnicas RRBS y WGBS, a partir de las cuales se han generado cantidades masivas de datos de una amplia gama de tipos de células y tejidos, como células sanguíneas, células madre, líneas celulares cultivadas, así como tejidos primarios, para mapear los cambios epigenéticos asociados con enfermedades específicas. La particularidad del bisulfito de sódico es que convierte las citosinas (Cs) no metiladas presentes en los fragmentos de ADN en uracilos (Us), que posteriormente se reconocerán como timinas (Ts) en las lecturas de secuenciación, mientras que no afecta a las citosinas metiladas (5mC) (24). Así mediante el mapeo y la alineación de las lecturas lecturas de bisulfito contra la secuencia del genoma de referencia permite calcular el porcentaje de Cs y Ts, y por tanto el porcentaje de metilación de aquel fragmento de ADN.

Estrategias seleccionadas para la elaboración del presente TFM

Como se ha especificado anteriormente, cada técnica utilizada en la obtención de los datos de metilación tiene unas características específicas, y la elección dependerá de la pregunta biológica que se quiera responder, así como de los recursos tanto informáticos como económicos. Así en función de la técnica seleccionada se obtendrán datos con diferentes características lo que implicará enfocar el TFM de una manera u otra.

La obtención de datos basados en técnicas que utilizan secuenciación es mucho más complejo y costoso, tanto experimentalmente como a nivel de pre-procesado de datos y coste económico y computacional en comparación de las técnicas basadas en *arrays*, pero a su vez la resolución a nivel de posiciones determinadas en la secuencia de ADN es mucho más profunda con el uso de secuenciación que no con *arrays*. Es por ello que teniendo en cuenta los aspectos mencionados, en un primer momento se optó por utilizar datos procedentes de *arrays* pero durante la ejecución de dicha tarea, se pudo observar que, a diferencia de los datos derivados de secuenciación, los datos derivados de *arrays* por lo general consisten en experimentos con pocas muestras en los que se pretende comparar el efecto de algún agente (enfermedad, tratamiento, dieta, ambiente…) en estado de metilación de algún o algunos tejidos, lo que se puede resumir como una mezcla heterogénea de pequeños experimentos realizados de manera independiente por grupos independientes, los cuales varían muy probablemente en metodología tanto de preparación de los *arrays* como de obtención de los datos. Es por ello que tras observar dicha problemática, se optó por reconsiderar la opción de utilizar datos derivados de secuenciación, y concretamente aquellos que derivan del tratamiento con bisulfito porque son los que mejor se ajustan a la respuesta biológica a la que se pretende dar respuesta con el presente TFM, así como por su amplio uso, y por tanto gran disponibilidad de datos, así como por su gran resolución a nivel de posiciones y la gran sensibilidad de la técnica para detectar pequeños cambios de metilación. Así teniendo en cuenta el enfoque seleccionado para la obtención de datos, y sus características se tuvieron en cuenta las siguientes opciones para la obtención de los datos:

1. Obtener los datos depositados en repositorios públicos, y analizarlos directamente en el ordenador. Esta opción resultó inviable, puesto que el análisis de dichos datos requiere unas necesidades computacionales que sobrepasan las características de un ordenador personal.

2. Otra opción que se planteó fue la creación de una base de datos unificada para los diferentes repositorios, conectarla a un servidor y a partir de él realizar las consultas requeridas, disminuyendo así el coste computacional y el tiempo de análisis. El principal inconveniente de dicha propuesta es el tiempo/recursos invertidos para llevarla a cabo serian excesivos teniendo en cuenta que no es uno de los objetivos planteados para este TFM.

3. La tercera opción es realizar una búsqueda de bases de datos o *web tools* que permitieran realizar este tipo de consultas sin la necesidad de descargar los datos crudos. Entre los recursos consultados se optó por el uso del servidor de datos epigenómicos DeepBlue (14) y concretamente el acceso mediante el correspondiente paquete de R (*DeepBlueR*) (25). DeepBue es una plataforma que proporciona acceso programático a datos epigenéticos sin alterar depositados en los principales consorcios epigenéticos, incluyendo: *BLUEPRINT Epigenome Project* (21), *The German Epigenome Programme* (*DEEP*) (http://www.deutsches-epigenom-programm.de), *Canadian Epigenetics, Environment and Health Research Consortium* (*CEEHRC*), *The Encyclopedia of DNA Elements* (*ENCODE*) (26) o el *NIH Roadmap Epigenomics Mapping Consortium* (*ROADMAP*) (22).

Teniendo en cuenta la planificación y la viabilidad de las alternativas planteadas se seleccionó la tercera de ellas, es decir, mediante el uso del paquete *DeepBlueR*, acceder al servidor de datos epigenómicos *DeepBlue* y sobre el cual hacer las diferentes consultas y análisis para experimentos que utilizando las técnicas RRBS y WGBS en la versión del genoma *hg19* hayan determinado el

perfil de metilación global. Mediante este enfoque se ha podido acceder a datos depositados en los consorcios *Roadmap Epigenomics*, *CEEHRC* y *DEEP*. Tal y como se irá detallando a lo largo del presente TFM, se han utilizado dos estrategias para la elaboración de este TFM, en una primera aproximación los datos de metilación se seleccionaron únicamente a nivel de dinucleótidos CpGs (también nombrada como aproximación CpG), utilizando estos datos se pusieron a punto las diferentes tareas y objetivos específicos del TFM. Posteriormente y con la finalidad de mejorar el resultado del TFM, se planteó una segunda aproximación, en la que aparte de incluir experimentos representando tejidos que no habían sido considerados en la primera aproximación, se aumentó la resolución del análisis, es decir, a parte de las citosinas contenidas en dinucleótidos CpGs, también se añadieron las citosinas no-CpGs, o CpHs (donde H = C, G, T y A), aumentando de manera considerable la cantidad de datos disponibles (también nombrada como aproximación CpH).

En la figura siguiente (Figura 6) se muestra un diagrama con los experimentos seleccionados de los diferentes proyectos, así como las 2 aproximaciones utilizadas para la realización de este TFM, y las diferentes tareas realizadas hasta llegar al producto final.

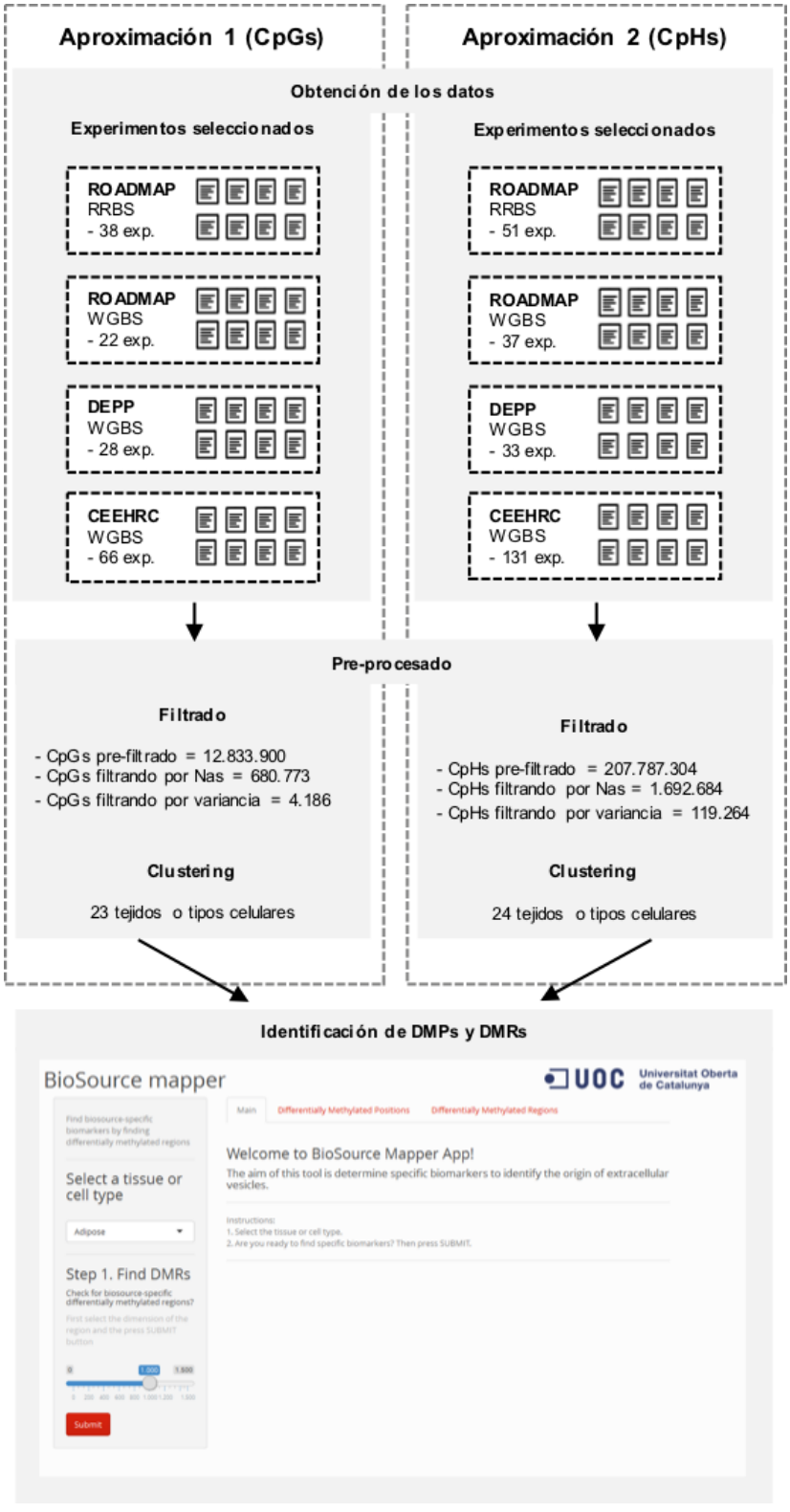

Figura 6. Diagrama donde se muestran el proceso que se ha llevado a cabo durante la ejecución de este TFM. En este TFM se han utilizado dos aproximaciones diferenciadas en el número de experimentos incluidos y en el número y tipo de posiciones analizadas. Una vez seleccionado los datos se han pre-procesado, donde se han filtrado, se ha eliminado el batch effect y se han agrupado por clústers biológicos en diferentes grupos y finalmente la determinación de DMPs y DMRs específicas de tejido puede realizarse mediante la aplicación shiny *BioSource Mapper App*.

En la primera aproximación (aproximación CpG) se parte de 12.833.900 posiciones CpGs analizadas en 154 experimentos con las técnicas RRBS y WGBS de los proyectos Roadmap Epigenomics, ENCODE, CEEHRC y DEEP (IHEC) y para el genoma hg19. Y, en la segunda aproximación (aproximación CpH) se parte de 207.787.304 posiciones CpHs analizadas en 252 experimentos con las técnicas RRBS y WGBS de los proyectos Roadmap Epigenomics, ENCODE, CEEHRC y DEEP (IHEC) y para el genoma hg19.

3. Pre-procesado de los datos: filtrado, corrección de la variabilidad no biológica y clustering

En este apartado se describen los procesos que se han llevado a cabo por tal de que los datos crudos o *raw data* obtenidos mediante *DeepBlueR* sean aptos para analizar las posiciones diferencialmente metiladas específicas de tejido. El pre-procesado consiste en los procesos de filtrado, normalización y clustering. Como se ha descrito anteriormente dichos procesos han sido aplicados tanto para la aproximación CpG como para a aproximación CpH, por lo que se describirá en paralelo los resultados obtenidos para cada una de ellas.

Filtrado de los datos

Tal y como se puede observar en la Figura 6, en ambas aproximaciones se parte de una gran cantidad de posiciones analizadas (12.833.900 posiciones para la aproximación CpG y 207.787.304 posiciones para la aproximación CpH), pero es posible que alguno de los experimentos seleccionados no presenten medida para una determinada posición (definido como *missing data* o NAs), así como no se espera que todas las posiciones medidas muestren variación biológica. Mediante el filtrado de los datos se pretende minimizar el número de datos ausentes para cada posición así como seleccionar sólo aquellas posiciones que presenten más varianza entre experimentos. Además en el contexto de este TFM, no interesa identificar regiones asociadas a los cromosomas sexuales, por

lo que se filtraran todas aquellas regiones correspondientes al cromosoma X e Y. Los criterios de filtrado de datos han sido los siguientes:

- Eliminar todas las posiciones localizadas en los cromosomas X e Y.
- Conservar únicamente aquellas posiciones con un contenido en "*missing data*" o NAs inferior al 20%.
- Eliminar aquellas posiciones que presentaran una varianza inferior al 0.05 para el conjunto de los experimentos seleccionados.

Utilizando estos criterios, y tal como se puede observar en el apartado correspondiente de la Figura 6 así como detalladamente en el Anexo 1 (Tablas Suplementarias; Tabla S1 y Tabla S2) el número de posiciones se reduce a 4.186 posiciones para la aproximación CpG y 119.264 posiciones para la aproximación CpH.

Eliminación de la variabilidad debida al origen de los datos (batch effect)

Otro de los aspectos claves para consecución del objetivo final de este TFM y debido a que se han obtenido los datos de diversas plataformas y diferentes proyectos es la de determinar, y corregir en caso necesario, si existe variabilidad dependiente de la plataforma de origen de los datos, o lo que se conoce como *batch effect*. Existen varios métodos o herramientas que permiten corregir la variabilidad dependiente al origen de los datos, y en este TFM se quiso comparar 2 métodos el *Empirical Bayes* que es en el que se basa la función de R *ComBat(),* el cual es uno de los métodos y funciones más utilizados para la corregir el *batch effect* en datos ómicos y se puede encontrar en el paquete de R *sva* (27), entre otros, y el *Latent factor model*, incluido en el paquete de R *BEclear*, paquete diseñado para corregir el *batch effect* en datos de metilación (28).

La evaluación del posible *batch effect*, así como los dos métodos de corrección aplicados a los datos derivados de la primera aproximación (aproximación CpG) pueden observarse mediante un análisis de componentes principales o PCA, la representación de los valores Beta mediante boxplot, y a través de la representación de los valores con un *heatma*p en la Figura 7.

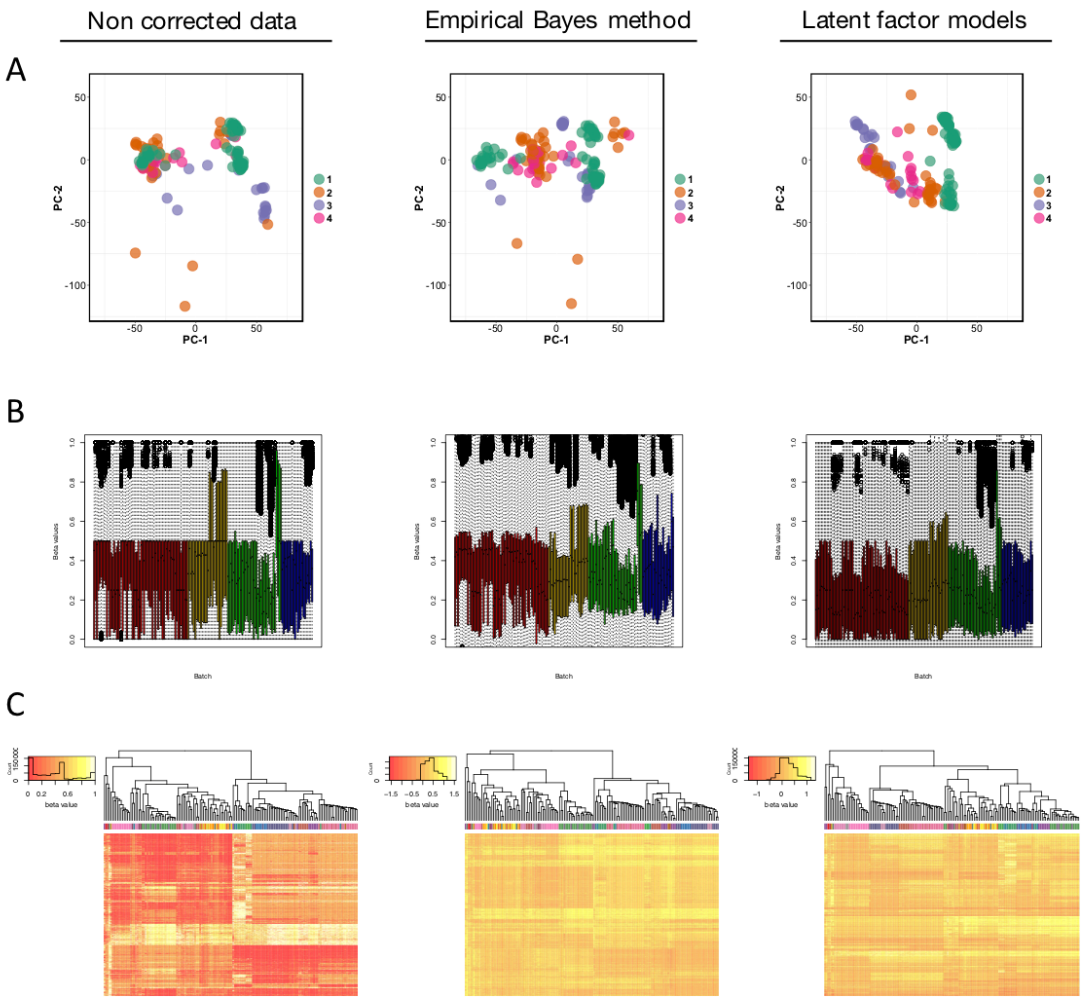

Figura 7. Comparación de los métodos para determinar e eliminar el *batch effect*. A) Evaluación del *batch effect* utilizando un análisis de componentes principales o PCA en los que los datos han sido coloreados en función del plataforma a la que corresponden. B) Boxplots para los valores Beta sin corregir o corregidos por los dos métodos utilizados, los diferentes colores de las barras corresponden a las diferentes plataforma de las que proceden los datos. C) *Heatmap* de los valores Beta para los diferentes experimentos (el dendrograma situado en la parte superior está basado en un clustering jerárquico basado en la correlación determinada por "Pearson" de los valores entre los diferentes experimentos).

Observando la anterior (Figura 7), se puede ver como los datos sin corregir presentan una pequeña variabilidad en función de la plataforma por lo que es necesario corregirla. Entre los métodos analizados se decidió que para este conjunto de datos el segundo método, o método basado en *Latent factor model* era el método que mejor eliminaba el *batch effect*. En la siguiente tabla (Tabla 2) se puede comprobar como el hecho de corregir los datos y eliminar el *batch effect* no implica que haya una pérdida de la variabilidad.

Tabla 2. Determinación de la media, desviación estándar y la varianza para los datos antes y después de corregir el batch effect por los diferentes métodos en los datos derivados de la primera aproximación (CpG).

Corrección del *batch effect*	Media	Desviación estándar	Varianza
Datos sin corregir	0.374	0.308	0.095
Empirical Bayes method	0.374	0.289	0.084
Latent factor model	0.249	0.313	0.098

Siguiendo el mismo criterio se aplicaron los dos métodos de corrección para los datos derivados de la segunda aproximación (aproximación CpH), en esta ocasión pero el método basado en *Latent factor model*, resultó inviable, tanto en el tiempo computacional requerido como en el resultado obtenido, puesto que el resultado distaba mucho de la finalidad pretendida. Es por ello que para esta segunda aproximación la corrección del *batch effect* de los datos se ha realizado mediante el método basado en *Empirical Bayes.* En la figura siguiente (Figura 8) se puede observar la distribución de los datos antes y después de aplicar la corrección del *batch effect*.

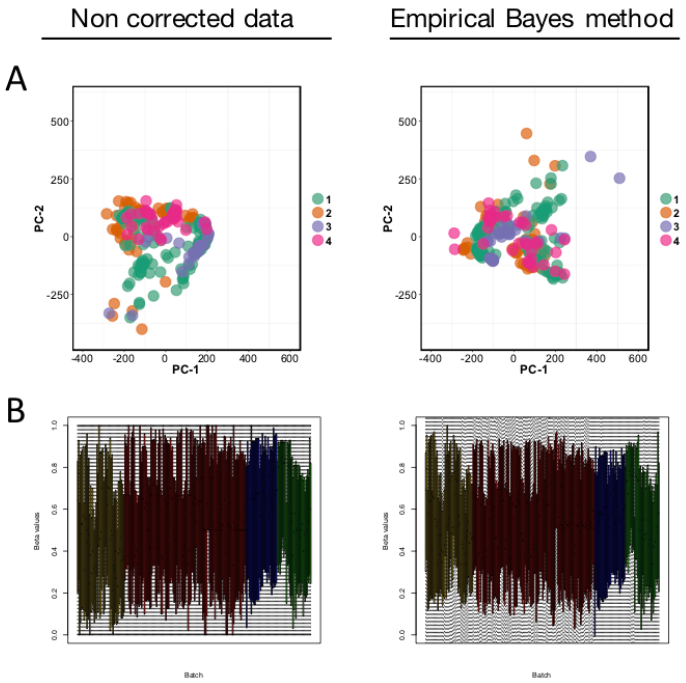

Figura 8. Evaluación del *batch effect* en los datos derivados de la segunda aproximación (CpH). A) Evaluación del *batch effect* utilizando un análisis de componentes principales o PCA en los que los datos han sido coloreados en función de la plataforma a la que corresponden. B) Boxplots para los valores Beta sin corregir o corregidos, los diferentes colores de las barras corresponden a las diferentes plataformas de las que proceden los datos.

De igual manera que para la primera aproximación, para esta segunda aproximación también se ha comprobado como el hecho de corregir los datos y eliminar el *batch effect* no implica que haya una pérdida de la variabilidad (Tabla 3).

Tabla 3. Determinación de la media, desviación estándar y la varianza para los datos antes y después de corregir el "batch effect" método de Empirical Bayes en los datos derivados de la primera aproximación (CpH).

Corrección del *batch effect*	Media	Desviación estándar	Varianza
Datos sin corregir	0.525	0.360	0.129
Empirical Bayes method	0.525	0.346	0.120

Clustering

Tal y como se esperaba, y como ha sido descrito previamente en la bibliografía, existe una gran correlación entre los valores de metilación para los tipos celulares o tejidos similares o derivados, es por ello que para mejorar el análisis y poder determinar aquellas posiciones y/o regiones específicas de tejido se han agrupado los diferentes experimentos en clústers. Para ello se ha utilizado el método de clustering jerárquico utilizando como distancia los valores de correlación determinados mediante el método de Pearson para cada par de experimentos para cada una de las dos aproximaciones consideradas. En el Anexo 2 (Figuras suplementarias; Figura S1), se muestra el clustering jerárquico para la aproximación CpG).

Además la clasificación obtenida utilizando el método jerárquico se ha optimizado manualmente por tal de obtener grupos homogéneos tanto estadísticamente como biológicamente. Los diferentes grupos representando a los tejidos o tipos celulares que serán comparados posteriormente se incluyen en la Tabla 4 para la aproximación CpG y en la Tabla 5 para la aproximación CpH.

Tabla 4. Agrupación de los diferentes experimentos en los siguientes clústers basados en el análisis de clústers jerárquicos e información biológica para los datos correspondientes a la aproximación CpG.

Clústers biológicos	Células	Tejidos
Adipose		x
B cell	x	
Brain		x
Colon		x
Duodenum		x
Eosinophil	x	
Embryonic stem cells	x	
Heart		x
Hepatocyte	x	
Hematopoietic stem cells	x	
Intestine		x
Induced Pluripotent Stem Cells	x	
Kidney		x
Lung		x
Macrophage	x	
Monocyte	x	
Muscle		x
Pancreas		x
Rectum		x
Spleen		x
Stomach		x
T cell	x	
Thymus		x

Para la primera aproximación se han obtenido datos de 23 tejidos o tipos celulares diferentes.

Tabla 5. Agrupación de los diferentes experimentos en los siguientes clústers basados en el análisis de clústers jerárquicos e información biológica para los datos correspondientes a la aproximación CpH.

Clústers biológicos	Células	Tejidos
Adipose		x
Blood	x	
Brain		x
Breast		x
Bone		x
Colon		x
Duodenum		x
Esophagus		x
Heart		x
Immune cells	x	
Intestine		x
Kidney		x
Liver		x
Lung		x
Intestine		x
Muscle		x
Pancreas		x
Rectum		x
Skin		x
Spleen		x
Stem Cells	x	
Stomach		x
Thymus		x
Thyroid Gland		x

Para la segunda aproximación (aproximación CpH), se han clasificado los experimentos en 24 clústers específicos de tejido o tipo celular, cabe destacar que para esta aproximación se han agrupado las células del sistema inmune y los diferentes tipos de *Stem Cells*.

4. Identificación de posiciones diferencialmente metiladas (DMPs)

La identificación de posiciones diferencialmente metiladas es uno de los aspectos claves para la consecución del objetivo principal de este TFM, por lo

que una vez se pre-procesados los datos y eliminado las diversas fuentes de variación no biológica la siguiente tarea hace referencia a la identificación de DMPs, para ello en primer lugar se han analizado los diferentes métodos y aproximaciones estadísticas aplicadas al entorno R que pueden ser utilizadas para identificar posiciones o *loci* diferencialmente metilados cuando se compara un tejido seleccionado respecto al resto.

Así, para determina aquellas posiciones que se encuentran diferencialmente metiladas generalmente se compara la proporción de citosinas metiladas para la posición analizada entre dos grupos. En comparaciones dónde únicamente existe una muestra por condición se recomienda el uso de métodos como el Test de Fisher, implementado por ejemplo en los paquetes de R *methylkit* (29) y *RnBeads* (30), o los métodos basados en *Hidden Markov Models* (HMMs), como el método ComMet, incluido en el paquete MethPipe (31). Estos métodos pueden aplicarse cuando existen replicados aunque su principal desventaja es que no tienen en cuenta la variación biológica entre replicados.

Otro grupo de métodos utilizados son los métodos basados en análisis de regresión, los cuales son adecuados para modelar los niveles de metilación entre grupos teniendo en cuenta la variación existente entre repeticiones dentro del mismo grupo. Las principales diferencias entre métodos basados en análisis de regresión se derivan de la distribución utilizada para modelar los datos y la variación asociada a ellos. Así, los métodos basados en regresión lineal se utilizan para modelar la metilación por posición a través de la comparación entre diferentes grupos. El modelo ajusta los coeficientes de regresión para modelar los valores de metilación esperados para cada posición en todos los grupos considerados y la hipótesis nula referente a que los coeficientes son 0 puede probar-se utilizando un T-test. Uno de los paquetes R más populares que incluyen este tipo de regresión es el paquete *Limma* (32). Aunque *Limma* fue desarrollado inicialmente para la detección de la expresión de genes diferenciales en datos de microarrays, también se utiliza para datos de metilación, siendo el método predeterminado en el paquete *RnBeads*, o el método en el que se basan otros paquetes como *BSmooth* (33). A parte de los métodos basados en regresión lineal también se han descrito métodos que se

basan en regresión logística. Y otros métodos más complejos de regresión se basan en la distribución beta binomial y son útiles para modelar la variancia y este tipo de métodos se incluyen en los paquetes *MOABS* (34), *DSS* (35) o *BiSeq* (36).

En una revisión realizada por Wreczycka et al. 2017 (37), comparan varios métodos basados en el T-test/regresión linear, regresión logística y regresión beta binomial para dos conjuntos de datos (uno simulado y uno real), y concluyen que los resultados obtenidos utilizando los diferentes métodos son comparables, obteniendo valores de sensibilidad y especificidad parecidos.

Teniendo en cuenta la naturaleza de los datos de los que se dispone, así como la bibliografía consultada, para identificar las posiciones diferencialmente metiladas en este TFM se utilizará el método de regresión lineal con el T-test similar al aplicado en el paquete *Limma*. Para la identificación de posiciones diferencialmente metiladas se ha utilizado el método de ajustes por mínimos cuadrados para ajustar el modelo lineal y el método de *Benjamini-Hochberg* (*BH*) para corregir por múltiples comparaciones. Para ello se han adaptado las funciones contenidas en el paquete *Limma*, y se han incorporado en la función creada en este TFM (*BiosourceMapper*), función que permite automatizar la comparación de un clúster o tejido/tipo celular de interés respecto al resto.

En la tabla siguiente (Tabla 6) se muestra los el top15 de DMPs obtenidas para el clúster correspondiente a monocitos utilizando los datos derivados de la primera aproximación (CpG).

Tabla 6. Ejemplo de las top 15 DMPs para monocitos comparado con el resto de grupos.

Chromosome	Start	End	Average Beta-vaue	Adjusted p-value
chr14	95969017	95969019	0.379	3.69E-16
chr8	127569281	127569283	0.169	1.47E-13
chr12	69181525	69181527	0.318	4.84E-13
chr20	48900666	48900668	0.295	7.43E-12
chr22	35697620	35697622	0.316	9.37E-12
chr10	98405284	98405286	0.356	1.14E-11
chr6	167011311	167011313	0.675	1.97E-11
chr16	30485159	30485161	0.063	1.90E-10
chr21	44072918	44072920	0.157	1.99E-10
chr11	47350196	47350198	0.674	2.71E-10
chr16	85608122	85608124	0.287	5.50E-10
chr17	18228998	18229000	0.584	5.50E-10
chr17	76886742	76886744	0.308	1.22E-09
chr1	1695803	1695805	0.326	1.22E-09
chr11	69260084	69260086	0.331	1.22E-09

5. Identificación de regiones diferencialmente metiladas (DMRs)

La mayoría de los métodos descritos para la determinación de regiones diferencialmente metiladas o DMRs, se basan en la agregación de un número determinado de DMPs en regiones de tamaño predefinido.

Concretamente para este TFM, se ha adaptado la función DMRfinder (38), incorporándola en la función creada para este TFM (*BiosourceMapper*) en la que se considera una región como diferencialmente metilada cuando para una longitud determinada se agregan 3 o más DMPs. El diseño tanto de la función *BiosourceMapper* como de la aplicación *Biosource Mapper App* permite seleccionar el tamaño de la región que se desee (rango 100-1500 bases), ya que en función del tejido seleccionado y de la cantidad de DMPs puede interesar aumentar o disminuir la longitud de la región seleccionada. Además la función creada permite cuantificar el grado de enriquecimiento en DMPs mediante el ratio de DMPs, ya que representa el cociente entre DMP determinadas para aquella región entre el número mínimo de DMPs para que sea considerada como DMR

(que es 3), por lo tanto cuanto menor sea el ratio, mayor número de DMPs estarán presentes en dicha región.

En la tabla siguiente (Tabla 7) se muestran las 15 de las DMR obtenidas para el análisis de marcadores específicos de monocitos utilizando la aproximación CpG.

Tabla 7. tDMRS obtenidas a partir de las tDMPs para el clúster de monocitos.

Chromosome	Start	End	Ratio of DMPs
chr1	164545703	164546071	0.60
chr2	233251655	233253326	0.57
chr2	54087072	54087135	0.63
chr7	27182998	27184540	0.50
chr7	27146262	27146324	0.57
chr9	131154435	131155185	0.60
chr9	116225863	116225990	0.63
chr10	101281838	101281967	0.50
chr14	77492665	77492784	0.63
chr15	90543224	90543524	0.60
chr17	46631952	46632197	0.50
chr19	10444923	10445374	0.60
chr19	39086998	39087054	0.60
chr20	62198931	62199782	0.60
chr22	28194381	28194677	0.60

6. Creación de una interfaz gráfica para la visualización de los resultados: *BioSource Mapper App*

Una vez realizadas las diferentes tareas que han permitido la creación de una función (*BiosourceMapper*) que mediante la selección de un tejido o tipo celular determinado calcula las posiciones y regiones específicas para ese tejido, la siguiente tarea ha consistido en plasmar dichos resultados y dicha función en una aplicación *shiny* (bautizada como *Biosource Mapper App*), que mejore la

visualización de dichos datos y sobretodo que permita a cualquier usuario acceder a dicha función.

El acceso a Biosource Mapper App puede realizarse mediante el siguiente link:

https://biosourcemapper.shinyapps.io/biosourcemapper_app/

En la siguiente figura (Figura 9) se muestra la pantalla inicial una vez se accede a la *Biosource Mapper App.*

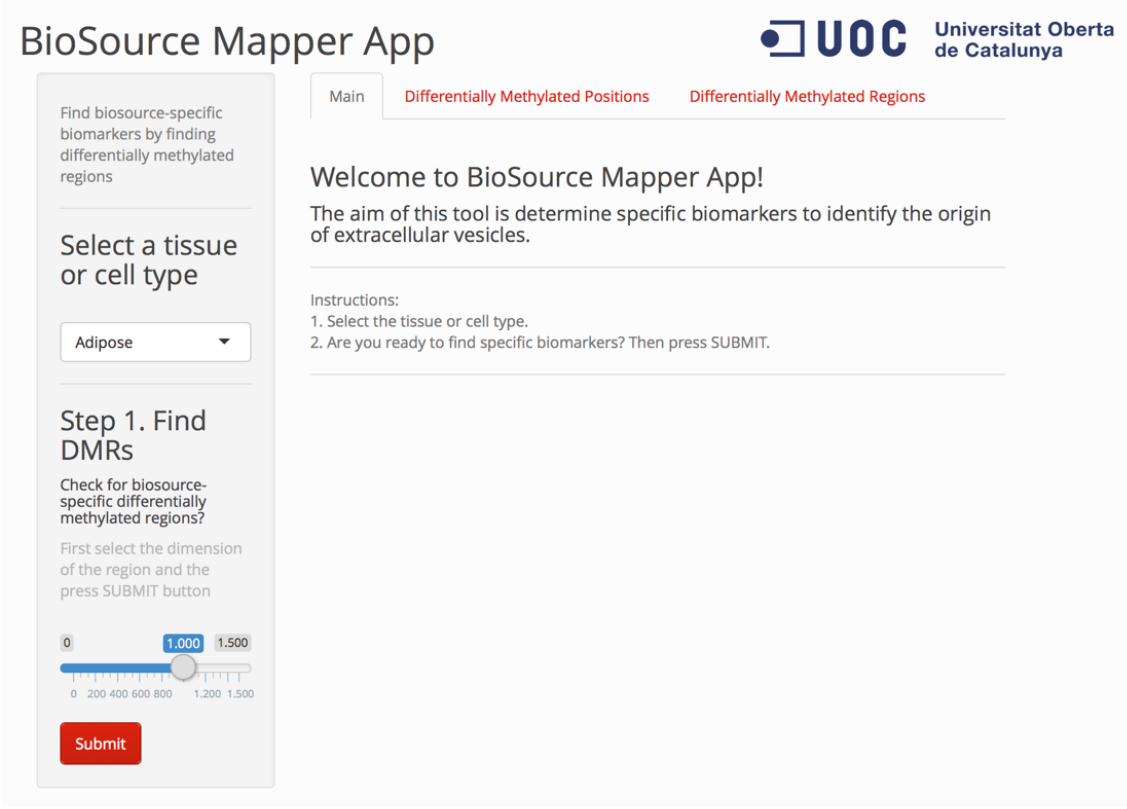

Figura 9. Pantalla inicial de acceso a la aplicación *Biosource Mapper App* desarrollada en este TFM.

Así, tal y como se puede observar en la Figura 9, *Biosource Mapper App*, permite la obtención de los datos presentados en el presente proyecto sin necesidad de tener conocimiento de R. *Biosource Mapper App* está diseñada para que el usuario únicamente tenga que escoger de la lista de de tejidos o tipos celulares (situada en la parte superior izquierda) aquel para el que desee obtener marcas de metilación específicas y modifique, si lo desea, el tamaño de la región que

será considerado para cuantificar las DMRs (por defecto este valor es 1kb). Una vez seleccionados los 2 parámetros y apretando el botón de *Submit*, la aplicación utiliza la función *BiosourceMapper* para identificar tanto DMPs como DMRs. Si se quiere consultar a nivel individual aquellas posiciones o regiones diferencialmente metiladas, basta con acceder a la pestaña correspondiente y aparecerá la tabla con los valores que han resultado significativos.

Además y para mejorar la visualización de los resultados se han contextualizado a nivel genómico, y mediante el uso del paquete de R *BioCircos* (39), se puede observar la localización de cada DMP y DMR que ha resultado significativa.

En la figura siguiente (Figura 10) se muestra el proceso para la obtención de DMPs y DMRs específicas de tejido adiposo.

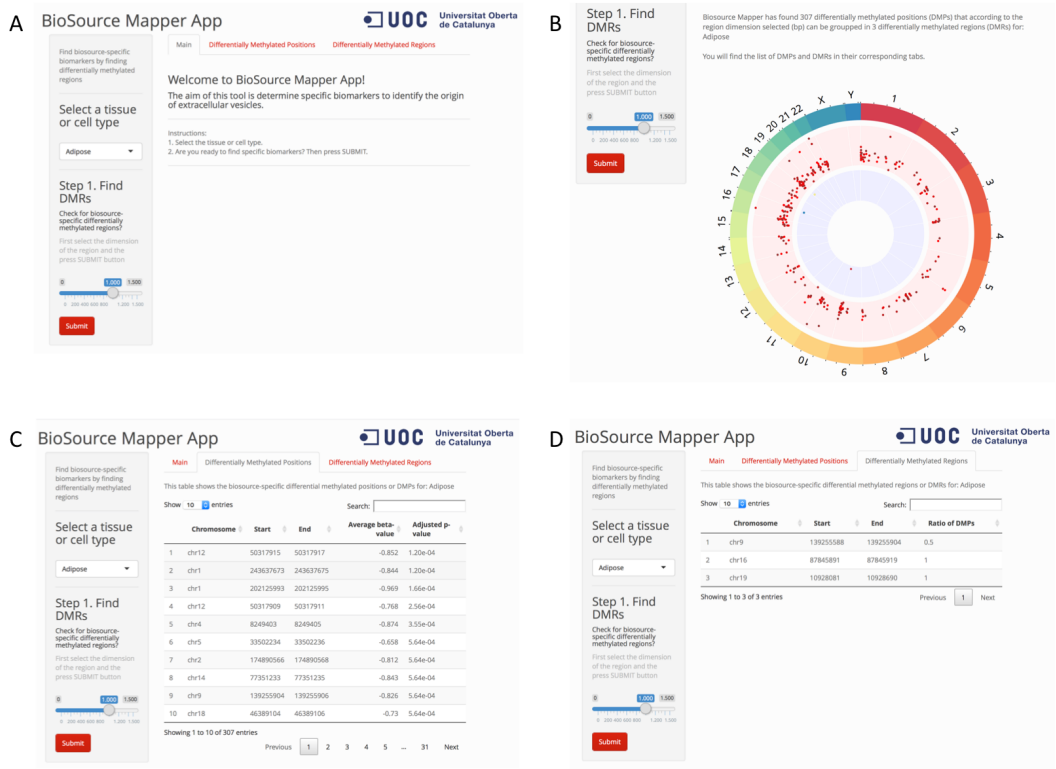

Figura 10. Proceso de obtención de DMPs y DMRs para el tejido asiposo con la aplicación *Biosource Mapper App* desarrollada en este TFM. A) Pantalla de bienvenida, en la que después de seleccionar el tejido y la región de interés, apretando *Submit* se procede a la realización del cálculo. B) El resultado que se obtiene es la cuantificación tanto de DMPs como DMRs y su localización genómica mediante el gráfico circular. El *track* exterior representan los diferentes cromosoma, en el *track* rojo se localizan las DMPs cuya posición depende de su localización y del valor beta. Y el *track* azul muestra las DMRs obtenidas en función de su localización y ratio de DMPs. C y D) En las pestaña correspondientes se muestran las DMPs y DMRs que se han identificado utilizando los parámetros seleccionados.

En la próxima figura (Figura 11), se muestra el mismo proceso pero en este caso para el hígado. Tal y como se puede observar el número de DMPs y DMRs aumenta considerablemente lo que a su vez aumenta el tiempo computacional para la obtención de los datos.

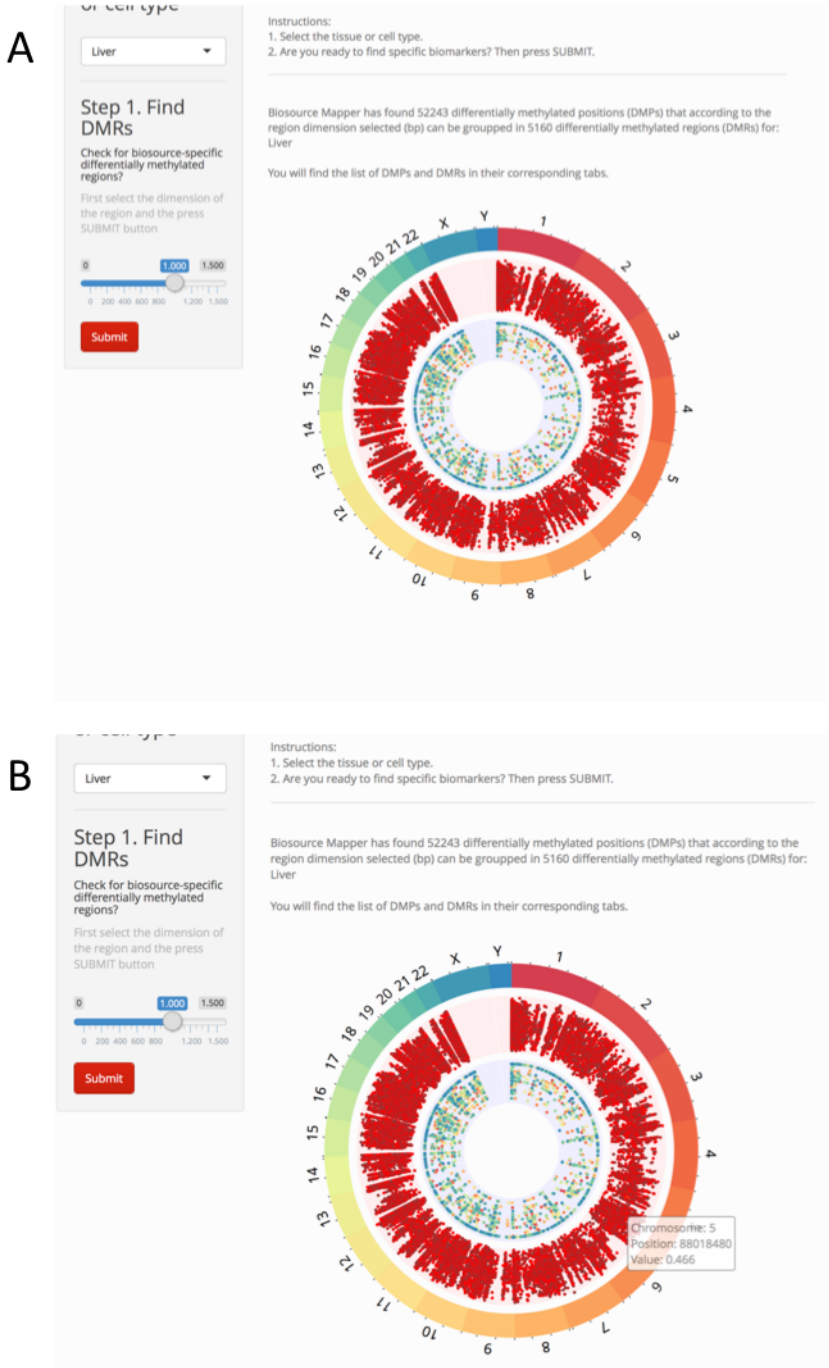

Figura 11. Proceso de obtención de DMPs y DMRs para el hígado con la aplicación *Biosource Mapper App* desarrollada en este TFM. A) Gráfico resumen mostrando las DMPs y DMRs identificada. B) Detalle mostrando que al situar el cursos encima de uno de los puntos correspondientes a las DMPs o DMRs muestras sus características.

7. Conclusiones

La principal conclusión principal que se deriva del presente TFM es que mediante el análisis bioinformático del patrón de metilación con la función *BiosourceMapper* implementada en la aplicación *Biosource Mapper App* es posible obtener marcas específicas de tejido.

A parte de esta conclusión principal durante la elaboración del TFM se ha podido ir concluyendo que es fundamental conocer la naturaleza y el origen de los datos para seleccionar los métodos más adecuados para su análisis.

Tal y como se ha comentado en la justificación de este TFM, su ejecución ha sido realmente tanto un reto como una oportunidad de poder poner en practica todo lo aprendido durante el máster y poder aplicarlo a un problema real. En términos generales se puede concluir que se han logrado los objetivos que se plantearon y que su planificación y metodología se ha seguido de manera sin grandes variaciones. Aunque se han introducidos cambios como el de ampliar las posiciones analizadas considerando a parte de los dinuccleótidos CpG, los no CpG, se ha podido lograr el objetivo principal del TFM que era la obtención de un catálogo de marcas específicas de tejido basado en datos de metilación.

En cuanto a las líneas de trabajo futuras, durante la ejecución del TFM se ha podido comprobar que existen diferentes alternativas que aplicadas podrían optimizar la identificación de marcadores, incluyendo la inclusión de nuevos datos, la modificación de los criterios de clustering, la inclusión de criterios que restrinjan la identificación de DMPs y DMRs, así como la integración de las regiones resultantes con otros datos epigenómicos o trasncriptómicos que permitan, a la vez de hacer una selección de los marcadores en base a criterios estadísticos, añadir criterios biológicos y funcionales. Además también se debe optimizar la aplicación creada puesto que para ser completamente funcional la velocidad de análisis debe ser mayor.

8. Glosario

CEEHR - *Canadian Epigenetics, Environment and Health Research Consortium*

CpG - dinucleótido de citosina-fosfato-guanina

CpH - dinucleótido de citosina-fosfato-H (donde H = A, T G o C)

DEEP - *The German Epigenome Programme*

DMPs - *Differential methylated positions*

DMRs - *Differential methylated regions*

NGS - *Next-generation Sequencing Technologies*

ROADMAP - *NIH Roadmap Epigenomics Mapping Consortium*

TFM - Trabajo Final de Máster

9. Bibliografía

1. EL Andaloussi S, Mäger I, Breakefield XO, Wood MJA. Extracellular vesicles: biology and emerging therapeutic opportunities. Nat Rev Drug Discov [Internet]. Nature Publishing Group; 2013 May 15;12(5):347–57.

2. Yáñez-Mó M, Siljander PR-M, Andreu Z, Zavec AB, Borràs FE, Buzas EI, et al. Biological properties of extracellular vesicles and their physiological functions. J Extracell vesicles. 2015;4:27066.

3. van Niel G, D'Angelo G, Raposo G. Shedding light on the cell biology of extracellular vesicles. Nat Rev Mol Cell Biol. Nature Publishing Group; 2018 Jan 17;19(4):213–28.

4. Contreras-Naranjo JC, Wu H-J, Ugaz VM. Microfluidics for exosome isolation and analysis: enabling liquid biopsy for personalized medicine. Lab Chip. Royal Society of Chemistry; 2017 Oct 25; 17(21):3558–77.

5. Johnstone RM, Adam M, Hammond JR, Orr L, Turbide C. Vesicle formation during reticulocyte maturation. Association of plasma membrane activities with released vesicles (exosomes). J Biol Chem. 1987 Jul 5;262(19):9412–20.

6. Lo Cicero A, Stahl PD, Raposo G. Extracellular vesicles shuffling intercellular messages: for good or for bad. Curr Opin Cell Biol. 2015 Aug;35:69–77.

7. Marca V La, Fierabracci A. Insights into the Diagnostic Potential of Extracellular Vesicles and Their miRNA Signature from Liquid Biopsy as Early Biomarkers of Diabetic Micro/Macrovascular Complications. Int J Mol Sci. Multidisciplinary Digital Publishing Institute; 2017 Sep 14;18(12):1974.

8. Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. Epigenomics: Roadmap for regulation. Nature. Nature Publishing Group; 2015 Feb 18;518(7539):314–6.

9. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. Nature Publishing Group; 2003 Mar 1;33(3s):245–54.

10. Mukherjee K, Twyman RM, Vilcinskas A. Insects as models to study the epigenetic basis of disease. Prog Biophys Mol Biol. Pergamon; 2015 Jul 1;118(1–2):69–78.

11. Schübeler D. Function and information content of DNA methylation. Nature. 2015 Jan 15;517(7534):321–6.

12. Bergman Y, Cedar H. DNA methylation dynamics in health and disease. Nat Struct Mol Biol. Nature Publishing Group; 2013 Mar 1;20(3):274–81.

13. Yong W-S, Hsu F-M, Chen P-Y. Profiling genome-wide DNA methylation. Epigenetics Chromatin. BioMed Central; 2016;9:26.

14. Albrecht F, List M, Bock C, Lengauer T. DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. Nucleic Acids Res. Oxford University Press; 2016 Jul 8;44(W1):W581-6.

15. Kuo H-C, Lin P-Y, Chung T-C, Chao C-M, Lai L-C, Tsai M-H, et al. DBCAT: Database of CpG Islands and Analytical Tools for Identifying Comprehensive Methylation Profiles in Cancer Cells. J Comput Biol. 2011 Aug;18(8):1013–7.

16. Grunau C, Renault E, Rosenthal A, Roizes G. MethDB--a public database for DNA methylation data. Nucleic Acids Res. Oxford University Press; 2001 Jan 1;29(1):270–4.

17. Zou D, Sun S, Li R, Liu J, Zhang J, Zhang Z. MethBank: a database integrating next-generation sequencing single-base-resolution DNA methylation programming data. Nucleic Acids Res. 2015 Jan 28;43(D1):D54–8.

18. Xin Y, Chanrion B, O'Donnell AH, Milekic M, Costa R, Ge Y, et al. MethylomeDB: a database of DNA methylation profiles of the brain. Nucleic Acids Res. 2012 Jan;40(D1):D1245–9.

19. Fingerman IM, McDaniel L, Zhang X, Ratzat W, Hassan T, Jiang Z, et al. NCBI Epigenomics: a new public resource for exploring epigenomic data sets. Nucleic Acids Res. Oxford University Press; 2011 Jan;39(Database issue):D908-12.

20. Bai W, Yang W, Wang W, Wang Y, Liu C, Jiang Q, et al. GED: a manually curated comprehensive resource for epigenetic modification of gametogenesis. Brief Bioinform. 2017 Jan ;18(1):98–104.

21. Adams D, Altucci L, Antonarakis SE, Ballesteros J, Beck S, Bird A, et al. BLUEPRINT to decode the epigenetic signature written in blood. Nat Biotechnol. 2012 Mar 1;30(3):224–6.

22. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb 18;518(7539):317–30.

23. Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. Nature. Nature Publishing Group; 2012 Sep 6;489(7414):57–74.

24. Li Y, Tollefsbol TO. DNA Methylation Detection: Bisulfite Genomic Sequencing Analysis. In: Methods in molecular biology (Clifton, NJ). 2011. p. 11–21.

25. Albrecht F, List M, Bock C, Lengauer T. DeepBlueR: large-scale epigenomic analysis in R. Bioinformatics. 2017 Jul 1;33(13):2063–4.

26. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science (80-). 2004 Oct 22;306(5696):636–40.

27. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatic. Oxford University Press; 2012 Mar 15;28(6):882–3.

28. Akulenko R, Merl M, Helms V. BEclear: Batch Effect Detection and Adjustment in DNA Methylation Data. Deng D, editor. PLoS One. Public Library of Science; 2016 Aug 25;11(8):e0159921.

29. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 2012 Oct 3;13(10):R87.

30. Assenov Y, Müller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. Nat Methods. Nature Publishing Group; 2014 Nov 28;11(11):1138–40.

31. Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, et al. A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics. El-Maarri O, editor. PLoS One. Public Library of Science; 2013 Dec 6;8(12):e81148.

32. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Jan 20;43(7):e47.

33. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 2012 Oct 3;13(10):R83.

34. Sun D, Xi Y, Rodriguez B, Park H, Tong P, Meong M, et al. MOABS: model based analysis of bisulfite sequencing data. Genome Biol . 2014 Feb 24;15(2):R38.

35. Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. 2014 Apr;42(8):e69–e69.

36. Hebestreit K, Dugas M, Klein H-U. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics. 2013 Jul 1;29(13):1647–53.

37. Wreczycka K, Gosdschan A, Yusuf D, Grüning B, Assenov Y, Akalin A. Strategies for analyzing bisulfite sequencing data. J Biotechnol. 2017 Nov 10;261:105–15.

38. Slieker RC, Bos SD, Goeman JJ, Bovée JV, Talens RP, van der Breggen R, et al. Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. Epigenetics Chromatin. BioMed Central; 2013 Aug 6;6(1):26.

39. Yu Y, Ouyang Y, Yao W. shinyCircos: an R/Shiny application for interactive creation of Circos plot. Bioinformatics. 2018 Apr 1;34(7):1229–31.

10. Anexos

Anexo 1: Tablas suplementarias

Tabla S1. Relación de posiciones conservadas antes y después de aplicar los filtros asociados al número de "missing data" o NAs, así como la varianza de cada posición para el conjunto de experimentos incluidos en la primera aproximación o aproximación CpG.

	Posiciones antes de filtrar	**Posiciones después de filtrar (<20% NAs)**	**(Var>0.05)**
chr1	1081731	61866	365
chr2	1052967	47501	253
chr3	723541	29742	105
chr4	642580	25765	133
chr5	697600	30874	149
chr6	690627	26821	110
chr7	746012	37569	276
chr8	631682	28116	135
chr9	590343	32006	232
chr10	682206	34224	191
chr11	637754	38293	211
chr12	569683	29310	211
chr13	383988	14766	48
chr14	410353	22272	142
chr15	425901	20591	83
chr16	563983	38244	253
chr17	582827	42878	317
chr18	341865	13790	85
chr19	506202	50233	504
chr20	364314	23560	113
chr21	198168	11238	72
chr22	309663	21114	198
TOTAL	**12833990**	**680773**	**4186**

Tabla S2. Relación de posiciones conservadas antes y después de aplicar los filtros asociados al número de "missing data" o NAs, así como la varianza de cada posición para el conjunto de experimentos incluidos en la segunda aproximación o aproximación CpH.

	Posiciones antes de filtrar	**Posiciones después de filtrar (<20% NAs)**	**(Var>0.05)**
chr1	22991912	152647	11042
chr2	13961119	84004	5980
chr3	13413635	96572	7230
chr4	12517885	72079	4610
chr5	9235960	36113	2154
chr6	8845676	55162	4004
chr7	8774146	49781	3054
chr8	8801377	97619	7751
chr9	9128984	109166	8162
chr10	7707283	34059	2114
chr11	5939329	129327	10228
chr12	24308858	115496	7580
chr13	6546675	58886	4704
chr14	3667463	28553	2146
chr15	4403859	55249	5163
chr16	18724507	71579	3588
chr17	16748012	61965	3700
chr18	17033388	74435	5034
chr19	16349446	64351	3313
chr20	15485342	94558	6744
chr21	13992317	69598	4596
chr22	12210131	81485	6367
TOTAL	**270787304**	**1692684**	**119264**

Anexo 2: Figuras suplementarias

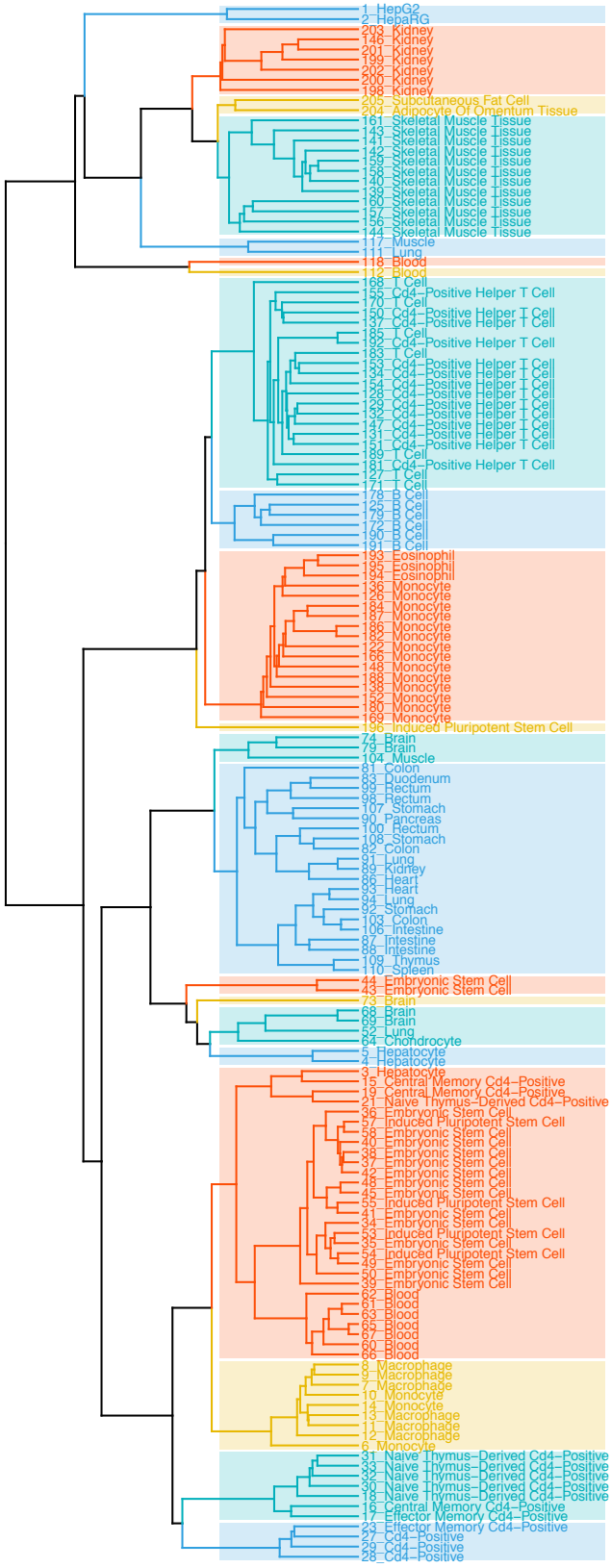

Figura S1. Clustering jerárquico de los experimentos incluidos en la primera aproximación o aproximación CpG en base a la correlación medida mediante Pearson.