TESIS DOCTORAL

Programa de Doctorado en Biología Fundamental y de Sistemas

Variación genética asociada a la metilación diferencial del ADN

Cristina A. Gómez Martín

Directores

Michael Hackenberg

José L. Oliver Jiménez

Departamento de Genética

Facultad de Ciencias

—

Granada, Noviembre de 2020

Editor: Universidad de Granada. Tesis Doctorales
Autor: Cristina Gómez Martín
ISBN: 978-84-1306-748-3
URI: http://hdl.handle.net/10481/66384

Esta tesis se ha desarrollado en el grupo *Genética Computacional y Bioinformática* del Departamento de Genética de la Universidad de Granada. La investigación realizada ha sido financiada por dos proyectos de investigación del Ministerio de Economía, Industria y Competitividad, a través del Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad (AGL2013-49090-C2-2-R y AGL2017-88702-C2-2-R) así como el Programa de Empleo Juvenil 2018/19 de la Junta de Andalucía.

A mis padres, a mi hermano
y a Nacho

If you can't do Bioinformatics,
you can't do Biology

J. D. Tisdall

Agradecimientos

A lo largo de estos años han sido muchas las personas que se han cruzado en mi camino, pero si a alguien tengo que agradecer en especial esta tesis y el haber llegado hasta aquí es al Dr. José L. Oliver y al Dr. Michael Hackenberg. Habéis sido mucho más que unos directores de tesis para mí y sin la oportunidad que me disteis y vuestra paciencia y dedicación, no sería lo que soy hoy, ni como investigadora ni como persona. Ha sido un placer compartir con vosotros esta etapa, que ojalá no acabase nunca. Y aunque la palabra gracias siempre se quedará pequeña con vosotros, gracias Pepe, gracias Mic.

A Ernesto. Hace ya más de 10 años que empezamos este camino juntos y no podría haber tenido mejor compañía. Espero tener la suerte de que la ciencia nos vuelva a juntar en el tiempo y el espacio algún día, y sobre todo nunca perder la amistad que tenemos. Gracias por ser y estar.

A todos los colaboradores con los que he tenido la suerte de trabajar y me han ayudado de una forma u otra. En especial a Guillermo, que siempre ha estado ahí dispuesto a ayudar en el tiempo que he tenido la suerte de compartir con él, a Ángel, que tantas cosas me ha enseñado, y tantos desastres nos ha arreglado, y a mis compañeros Chema, Stavros y Nicolas.

Al departamento de Genética, donde me he sentido siempre como en casa, en especial a su director Rafael Jiménez y a Francisco Barrionuevo que han hecho sencillo cada trámite de esta tesis, muchas gracias.

A todos los amigos que, sin saber bien de qué iba esto, han escuchado mi desesperación cuando las cosas no iban bien y han celebrado las buenas noticias como si fueran suyas. Gracias Priscila, contigo las penas siempre son mucho menos y las alegrías doblemente buenas.

A mis padres y a mi hermano por estar siempre ahí, apoyarme y darme un cariño incondicional. Gracias por confiar siempre en mí y enseñarme tantas cosas. Soy lo que soy gracias a vosotros, os quiero.

A Nacho. Si alguien ha disfrutado, sufrido y vivido esta tesis tanto como yo, ese eres tú. Sin cada palabra de ánimo, o tu comprensión y apoyo, hubiera sido imposible. Gracias por querer acompañarme en este camino que es la vida, te quiero.

Índice general

Abreviaturas y glosario

***back-end*:** Parte del software que procesa la entrada que proviene del usuario (*front-end*).

***CGIs*:** Islas CpG.

***DNMTs*:** Metiltransferasas de ADN, enzimas responsables de la transmisión del grupo metilo a los nucleótidos de ADN.

***eQTL*:** E*xpression quantitative trait loci*. Loci genómicos asociados a la variación en los niveles de expresión de un gen.

***FET*:**Test Exacto de Fisher.

***FDR*:** *False Discovery Rate*.

***front-end*:** Parte del software que interactúa con los usuarios.

***GWAS*:** *Genome Wide Association Studies*. Estudios de asociación a genoma completo.

***k-mero*:** Secuencia de ADN con una longitud k y una composición determinada.

***meQTLs*:** *Methylation quantitative trait loci*. Loci genómicos asociados a la variación en los niveles de metilación.

***QTL*:** *Quantitative Trait Loci*.

***PCR*:** *Polymerase Chain Reaction*. Reacción en cadena de la polimerasa.

***RRBS*:** *Reduced Representation Bisulfite Sequencing*.

***SNP*:** *Single Nucleotide Polymorphism*. Polimorfismo de un único nucleótido.

***TL-CpGs*:** *Traffic-Light CpGs*. Los semáforos CpG (CpG TL, por sus

siglas en inglés) son dinucleótidos CpG que muestran una correlación significativa entre su metilación y el perfil de expresión del gen más próximo.

***TF*:** Factor de transcripción.

TFBS: Sitio de unión a factores de transcripción.

***WGBS*:** *Whole Genome Bisulphite Sequencing.*

Resumen

En los últimos años, los estudios de asociación en genoma completo (*GWAS*, por sus siglas en inglés) han revolucionado el estudio de los caracteres cuantitativos o multigénicos, que tienen a menudo un importante componente ambiental. En nuestra especie esto ha permitido identificar un gran número de potenciales marcadores de numerosos caracteres complejos, entre ellos enfermedades como el Alzheimer, las enfermedades autoinmunes o el cáncer. La relación estadísticamente significativa entre las frecuencias alélicas en muchos loci y las enfermedades genéticas les proporciona a estos marcadores (la mayoría de ellos *Single Nucleotide Polymorphisms* o *SNPs*) un gran potencial valor diagnóstico y pronóstico. Sin embargo, la inmensa mayoría de estos *SNPs* se localizan lejos de los genes o de las regiones que los regulan, haciendo por tanto difícil identificar el vínculo molecular o funcional que pueda existir entre el marcador y la enfermedad.

Basándonos en estudios recientes de varios autores, en esta Tesis Doctoral se presenta una aproximación alternativa a este problema. La hipótesis es que la metilación del ADN podría actuar como mediador entre la variación genética y la expresión génica. En el modelo que se propone, el genotipo en los *SNPs* estaría asociado con la metilación de los dinucleótidos CpG y, a su vez, la metilación de los sitios CpG podría estar asociada con los cambios en la expresión génica. La estrategia ha sido, por tanto, determinar primero la asociación entre el genotipo de los *SNPs* y la metilación de las citosinas. Esto permitió identificar un gran número de pares *SNP-CpG* que muestran una asociación estadística altamente significativa. En una segunda fase, se estudió la co-localización de estos pares con otros elementos genómicos: promotores, potenciadores (*enhancers*) y semáforos CpG (*TL-CpGs*, sitios CpG cuyo estado de metilación está directamente relacionado

con la expresión génica), lo que ha permitido identificar aquellos pares *SNP-CpG* con un mayor significado biológico. Por último, analizando en detalle estos pares *SNP-CpG* se han podido identificar varios ejemplos que muestran nuevos vínculos funcionales previamente no descritos en los estudios de *GWAS*.

Para alcanzar estos resultados ha sido necesario el desarrollo de nuevas herramientas bioinformáticas y la modificación o mejora de otras ya disponibles en el grupo. En primer lugar, son necesarios metilomas de alta calidad de citosinas individuales, provenientes de múltiples tejidos y de diferentes individuos para optimizar la variación genética disponible. Para ello se ha desarrollado un flujo de datos automatizado, *MethylExtract2*, que, utilizando software propio y de terceros, integra todas las etapas necesarias para la obtención de los mapas de metilación. Por otra parte, los metilomas procesados mediante *MethylExtract2* se han almacenado en la base de datos *NGSmethDB*, la cual ha sido objeto de dos actualizaciones a lo largo del desarrollo de esta Tesis Doctoral: *NGSmethDB 2017* y *NGSmethDB20*, esta última aún en desarrollo y de la que se presentan los principales resultados preliminares. Además, en el transcurso de esta Tesis se ha desarrollado una importante mejora del método de predicción de islas CpG, *gCluster*, que incluye algunas mejoras respecto a su predecesor *CpGcluster* sobre todo en el modelo de distancias en el caso de *k-meros* solapantes.

Para el estudio de asociación entre variación genética y metilación del ADN se procesaron, utilizando *MethylExtract2*, los metilomas de 58 muestras, cada una de ellas de un individuo (y por tanto genotipo) distinto. A continuación, y tras explorar varios métodos, se desarrolló un protocolo basado en el Test Exacto de Fisher para determinar si existe asociación estadística entre los valores de metilación y el genotipo en el conjunto de muestras estudiadas, encontrándose un total de 51.585 *SNPs* (1,3 %) asociados con al menos un CpG ($FDR \leq 0.05$).

Con objeto de almacenar todos estos resultados y hacerlos disponibles a la comunidad científica, se ha desarrollado un nuevo recurso web, $geno^5mC$, donde se puede explorar la asociación entre genotipo y metilación mediante una aplicación interactiva que incorpora distintos métodos de búsqueda jerarquizada. Esto permite buscar nuevas conexiones funcionales entre los

SNPs asociados, por un lado, y las enfermedades u otros rasgos, por otro, lo que lo convierte en un recurso clave especialmente en aquellos casos en que no se conoce aún un vínculo molecular o funcional directo.

Palabras clave: Variación genética, metilación del ADN, expresión génica, metilomas, pares *SNPs-CpGs* asociados

Capítulo 1

Introducción

Aunque muchas enfermedades humanas muestran herencia mendeliana simple, prácticamente todos los rasgos cuantitativos son complejos, es decir, multigénicos e influenciados por el medio ambiente. Esto incluye enfermedades complejas como el Alzheimer [Bird, 2008], las enfermedades autoinmunes [Goddard et al., 2016] y la mayoría de los tipos de cáncer [Wu et al., 2016a].

Con el objetivo de identificar el componente genético de estos rasgos complejos, en los últimos años se han llevado a cabo estudios de asociación en genoma completo o *Genome Wide Association Studies* (*GWAS*), que se basan en relacionar un determinado fenotipo con las frecuencias alélicas de una serie de *loci*. Estos estudios contribuyen a ampliar el conocimiento acerca de la predisposición genética a las enfermedades complejas a través del descubrimiento de nuevas variantes genéticas, a menudo variantes de un único nucleótido o *SNPs* (*Single Nucleotide Polymorphisms*), y que tienen un potencial valor diagnóstico y pronóstico. Sin embargo, este tipo de estudios también tiene una serie de limitaciones [Tam et al., 2019]. Con gran frecuencia los *SNPs* encontrados se localizan fuera de las regiones codificantes o reguladoras ya conocidas [Tak and Farnham, 2015]. Es por ello que es muy habitual que no se pueda establecer fácilmente un mecanismo por el cual estas variantes estén influyendo en el fenotipo a estudiar. Otro enfoque utilizado son los *QTLs* de expresión o *Expression Quantitative Trait Loci* (*eQTLs*), que relacionan estadísticamente la variación genética con los niveles de expresión génica, y que, *a priori*, podrían ayudarnos a dilucidar el

mecanismo subyacente tras los resultados obtenidos por *GWAS*.

La expresión génica en eucariotas es un proceso muy bien regulado, tanto a nivel pre- como post-transcripcional. Usualmente, para que un gen pueda transcribirse se requiere que suceda un cambio en el estado de la cromatina, lo que se regula principalmente mediante mecanismos epigenéticos. Este es el caso de la metilación de ADN, que aunque tradicionalmente se ha asociado con la represión génica, en estudios recientes [Fisher et al., 2018] se ha observado que dependiendo del contexto genómico también puede provocar un aumento en los niveles de expresión, como se detalla en el apartado 1.1. Existen también evidencias que sugieren que la variación de secuencia podría provocar cambios en la metilación del ADN [McRae et al., 2018].

Esta Tesis Doctoral se ha centrado en ofrecer un enfoque alternativo a esta cuestión, con el objetivo de dilucidar la posible relación entre estos tres aspectos: la variación genética, la metilación del ADN y la expresión génica (Figura 1).

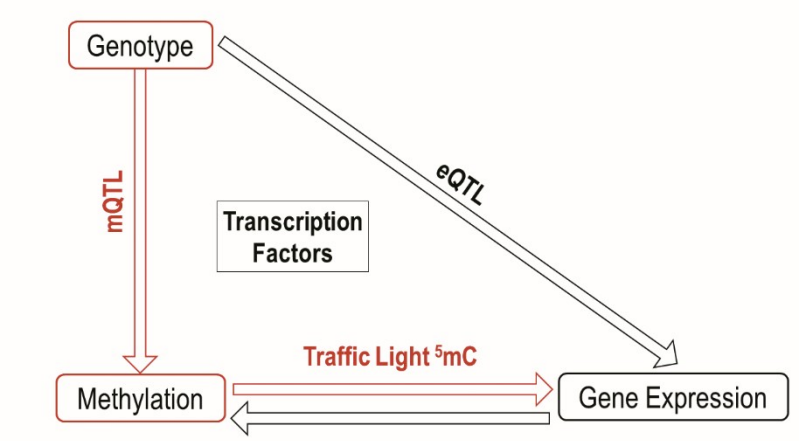


Figura 1: Esquema simplificado de la relación entre genotipo, metilación del ADN y expresión génica.

1.1. Metilación del ADN

La metilación del ADN es una de las marcas epigenéticas por excelencia y por tanto de las más estudiadas. Esta consiste en la adición de un grupo metilo a los nucleótidos de ADN, proceso que tradicionalmente se había

considerado restringido a la citosina, pero que actualmente se sabe puede producirse también en el nucleótido adenina [Wu et al., 2016b]. Se trata de un proceso enzimático que es llevado a cabo por metiltransferasas de ADN (*DNMTs*). En el caso de la citosina la metilación se produce en el carbono 5' dando lugar a la metilcitosina, a la que comúnmente se ha llamado la "quinta base del ADN".

Los patrones de metilación, distribución de las citosinas metiladas a lo largo de la secuencia, no son iguales en todos los eucariotas [Suzuki and Bird, 2008]. En el caso de los mamíferos nos encontramos ante un patrón de metilación global, en torno a un 70-80 % [Bird, 2002], que ocurre principalmente en aquellas citosinas que se encuentran en el contexto genómico CpG.

Sin embargo, como marca epigenética, la metilación es dinámica y este patrón varía con el tipo celular y/o el momento del desarrollo y en algunos tejidos nos encontramos metilación en los otros contextos: CHG y CHH (donde H puede ser cualquier nucleótido distinto de la citosina). Este es el caso de las células germinales o las células madre embrionarias [Lister et al., 2009, Laurent et al., 2010, Ziller et al., 2013].

En contraste con este patrón de metilación global, existen ciertas regiones llamadas islas CpG, que se caracterizan por su alto contenido en G+C y por una elevada frecuencia del dinucleótido CpG en comparación con el resto del genoma, y que se encuentran principalmente no metiladas. Esta menor frecuencia del dinucleótido CpG es debida a que la citosina cuando se encuentra metilada sufre fácilmente una desaminación espontanea, convirtiéndose en timina [Duncan and Miller, 1980], lo que provoca que principalmente se conserven aquellos CpGs que se encuentran no metilados de forma más habitual. Aproximadamente un 70 % de los promotores de genes humanos presentan una isla CpG y se ha observado que, de forma general, su desmetilación correlaciona con la expresión génica [Deaton and Bird, 2011]. Sin embargo, gracias a las nuevas técnicas de secuenciación masiva, que han permitido estudiar el papel de la metilación en diferentes contextos genómicos, se ha observado que la función de la metilación varía dependiendo del contexto en el que se sitúe [Jones, 2012]. Así pues la metilación en la región 3' de algunos genes podría incluso favorecer la transcripción [Yu et al., 2013].

Esta relación de la metilación y la transcripción también se puede observar en regiones potenciadoras [Hon et al., 2013] y aisladoras [Wang et al., 2012], donde regula la unión de factores de transcripción.

En el caso del cuerpo génico, que suele tener una baja densidad de CpGs, se ha observado un alto grado de metilación al que aún no se le ha encontrado una explicación funcional [Jones, 2012]. Una de las posibles explicaciones que se ha dado a este hecho es que corresponda a "promotores huérfanos" que se usen solamente en etapas anteriores del desarrollo y por tanto mantengan su densidad de CpGs, aunque no su función, en etapas posteriores [Illingworth et al., 2010].

Recientemente se ha observado también la implicación de la metilación del ADN en la regulación del *splicing* alternativo [Singer et al., 2015, Lev Maor et al., 2015]. Aproximadamente un 22 % de los exones alternativos estarían regulados por la metilación del ADN debido a sus altos niveles de metilación en comparación con los intrones flanqueantes. Se han observado algunos mecanismos por los que esto podría ocurrir como los mediados por el factor *CTCF* [Shukla et al., 2011] o la proteína *MeCP2* [Maunakea et al., 2013], así como la formación de puentes proteicos por la proteína *HP1* [Yearim et al., 2015], pero aún queda mucho por estudiar en este campo.

A la vista de estos estudios queda en evidencia el papel claro de la metilación como reguladora de la transcripción génica en diferentes sentidos, dependiendo del contexto genómico en el que se produzca. Esta relación parece estar mediada por la unión de distintos factores de transcripción, como se abordará en más profundidad en el apartado 1.2.

Además de su papel regulador en la transcripción, la metilación es también un elemento clave en la estabilidad genómica. Tiene un papel muy importante silenciando elementos repetidos, como los retrotransposones [Yoder et al., 1997, De Mendoza et al., 2018] o los que se encuentran en los centrómeros, lo cual es necesario para la correcta segregación cromosómica. Además está implicada en la compensación de dosis de los cromosomas sexuales [Sharp et al., 2011] y la impronta de genes autosómicos [Li et al., 1993].

1.2. Metilación diferencial del ADN, variación de secuencia y expresión génica

Podemos definir la metilación diferencial del ADN como los cambios en el estado de metilación que encontramos entre distintos tejidos dependiendo de su origen y/o estado de desarrollo (Figura 2). El mecanismo que subyace, como ya se apuntó en el apartado 1.1, parece estar mediado por la interacción de distintos factores de transcripción (TFs) con el ADN.

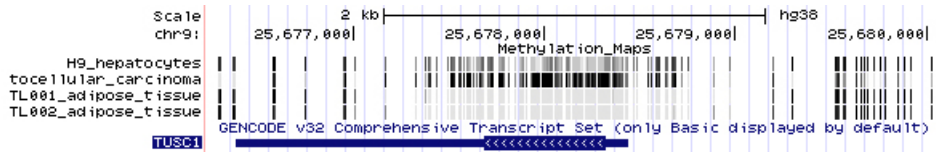


Figura 2: Región de metilación diferencial.

Recientemente, los estudios de caracteres cuantitativos en genoma completo han asociado variantes de secuencia (*SNPs*) a cambios en los niveles de metilación, los llamados *meQTLs* (*Methylation Quantitative Trait Loci*) [McRae et al., 2018]. Cuando se estudia la localización genómica de estas regiones se ha observado que muchas de ellas co-localizan con sitios de unión de factores de transcripción (*TFBSs*) u otros elementos reguladores [Corradin and Scacheri, 2014].

Por otra parte, los *loci* cuantitativos de expresión o *Expression Quantitative Trait Loci* (*eQTLs*) [Nica and Dermitzakis, 2013], a priori podrían ayudarnos a dilucidar el mecanismo subyacente a la relación entre expresión génica y variación genética. Sin embargo, un estudio reciente [Strunz et al., 2018] encontró 202.489 variantes asociadas con los niveles de expresión de 1.959 genes en hígado, lo que sugiere que una gran proporción de los genes humanos tendría al menos un *eQTL* asociado.

En estudios recientes se ha observado que a menudo estos *eQTLs* co-localizan con los *meQTLs* [Pierce et al., 2018], habiéndose encontrado más de 400 parejas *eQTL-meQTL* que comparten una variante causal común, lo que podría deberse a la existencia de un vínculo funcional. Ante las abundantes evidencias de la interacción metilación-expresión podríamos plantearnos dos posibilidades: i) los cambios en la metilación del ADN son los que conducen

a cambios en la expresión génica, o, ii) por el contrario, la metilación del ADN no sería otra cosa que el resultado de la regulación génica. Por medio de análisis de correlaciones parciales y mediación estos autores encontraron que ambas posibilidades coexisten en el genoma, y que muchos *SNPs* afectan a múltiples CpGs en direcciones opuestas.

Aunque *eQTLs* y *meQTLs* pueden extender notablemente nuestro conocimiento de los mecanismos moleculares de las enfermedades complejas, estos estudios generalmente conllevan el estudio de un gran número de muestras lo que los hace costosos y de mucho trabajo. Además no es frecuente poder estudiar ambos en el tejido o tipo celular relevante.

En cuanto al papel del genotipo, se sabe que los *SNPs* pueden afectar a la unión de factores de transcripción localmente y con ello provocar cambios en la metilación del ADN [Nica and Dermitzakis, 2013]. Además, recientemente se ha observado que la metilación del ADN puede correlacionar con la expresión génica [Yang et al., 2020], como es el caso de los semáforos CpG o *TL-CpGs* [Medvedeva et al., 2014, Lioznova et al., 2019], CpGs individuales que correlacionan positiva o negativamente con la expresión de uno o más genes.

1.3. Métodos y algoritmos para la estimación de los niveles de metilación

Cuando nos enfrentamos a la detección de los niveles de metilación nos encontramos ante dos problemas principalmente:

a. La hibridación es insensible frente a la metilación, lo que hace imposible el uso de chips de ADN para su detección.

b. La *PCR* elimina la información acerca del estado de metilación de las citosinas con las sucesivas rondas de replicación.

Es por ello que aunque a lo largo de los años se han ido desarrollando distintas técnicas, todas ellas están basadas en la realización de un pretratamiento del ADN que después permita la detección de la metilcitosina al secuenciar [Laird, 2010]. Inicialmente estos pre-tratamientos comprendían la digestión por endonucleasas y el enriquecimiento por afinidad, las cuales

tienen una serie de inconvenientes que se detallarán a continuación. Estos se solventaron mediante el uso del pre-tratamiento con bisulfito, que aunque se conocía desde hace bastantes años [Frommer et al., 1992], no ha sido hasta el desarrollo de las técnicas de secuenciación masiva cuando ha tenido un gran impacto.

1.3.1. Técnicas previas al bisulfito

Digestión por endonucleasas sensibles a la metilación del ADN

Esta técnica tiene el principal inconveniente de solo poder analizar aquellos fragmentos de ADN donde se encuentran los sitios de restricción de la enzima utilizada, lo que cubre una pequeña fracción del genoma. Además no permite distinguir entre los distintos contextos de metilación.

Técnicas basadas en enriquecimiento por afinidad

Las técnicas basadas en enriquecimiento por afinidad se basan en la utilización de proteínas con dominios de unión en sitios CpG metilados. Estas a pesar de ser rápidas y eficientes tampoco nos ofrecen datos de citosinas individuales y tienen poca sensibilidad en las regiones con baja densidad de CpGs.

1.3.2. Tratamiento con bisulfito: *WGBS* y *RRBS*

El tratamiento del ADN con bisulfito provoca la desaminación de aquellas citosinas no metiladas, que pasan a ser uracilos (y timinas tras la amplificación por *PCR*), mientras que no altera las citosinas metiladas [Frommer et al., 1992]. Tras el tratamiento se procede a la secuenciación del ADN y el estado de metilación se infiere a partir de las lecturas alineadas frente al genoma de referencia [Clark et al., 2006].

La secuenciación masiva tras el tratamiento con bisulfito (*WGBS*, *Whole-Genome Bisulphite Sequencing*) [Urich et al., 2015] es capaz de detectar, teóricamente, el estado de metilación de cada una de las citosinas del genoma, independientemente del contexto genómico en el que se encuentren o

del contexto de metilación de que se trate (CG, CHG o CHH). Esta técnica ha revolucionado el campo de la metilación del ADN, permitiendo la obtención de datos a gran escala en numerosos tejidos y condiciones celulares, y es la base de grandes proyectos como es el caso de *Roadmap Epigenomics* [Roadmap Epigenomics Consortium et al., 2015], *ENCODE* [Feingold et al., 2004, Abascal et al., 2020], o *BLUEPRINT* [Adams et al., 2012], entre otros. Sin embargo tiene algunos inconvenientes, entre ellos su gran coste. Este último inconveniente motivó el desarrollo del método *RRBS* (*Reduced-Representation Bisulphite Sequencing*) [Meissner et al., 2005]. Este método incorpora un paso previo a la conversión con bisulfito, la digestión mediante la enzima *MspI*, con el objetivo de secuenciar únicamente aquellos fragmentos con una alta densidad de CpGs. Estos fragmentos que conforman el "genoma reducido" incluyen la mayoría de los promotores y regiones con secuencias repetidas que con la técnica de *WGBS* son difíciles de analizar. Sin embargo tiene el inconveniente de no cubrir aquellas regiones con baja densidad de CpGs, y que pueden tener funciones relevantes.

Además de estos dos métodos existen otros basados en micromatrices o arrays que combinan también el pre-tratamiento con bisulfito. Estos se basan en la hibridación (tras el tratamiento por bisulfito y la amplificación por *PCR*) de los fragmentos de ADN en arrays que contienen sondas tanto para los sitios CpGs metilados como los no metilados. Los arrays de metilación son métodos económicos y que permiten la detección de un buen número de citosinas, en concreto con el último chip de *Illumina*, *Infinium MethylationEPIC BeadChip* (*EPIC*) se puede inferir el nivel de metilación de aproximadamente 850.000 citosinas situadas en regiones de interés.

1.3.3. Algoritmos de detección de los niveles de metilación

La aparición de las técnicas descritas en el apartado anterior, basadas en el tratamiento por bisulfito y la secuenciación masiva, han permitido obtener metilomas de alta resolución, con datos de citosinas individuales. Estos experimentos generan una gran cantidad de datos en forma de lecturas cortas, que deben ser procesadas adecuadamente para obtener los mejores resultados posibles, lo cual supone un reto bioinformático importante.

Habitualmente el procesado de los datos de metilación consta de tres pasos: pre-procesado de las lecturas, alineamiento frente al genoma de referencia e inferencia de los niveles de metilación. Además deben incluirse controles de calidad para asegurar que los niveles de metilación inferidos son correctos. Existen numerosos programas que pueden llevar a cabo cada una de estas etapas, especialmente en el caso del alineamiento e inferencia de los niveles de metilación. En el caso de los alineadores podemos encontrar *Bismark* [Krueger and Andrews, 2011] o *BSMAP* [Xi and Li, 2009], y en el caso de los programas para inferir los niveles de metilación tenemos también distintas alternativas como *BisSNP* [Liu et al., 2012] o *MethylExtract* [Barturen et al., 2013], e incluso tuberías o *pipelines* que conectan ambos procesos, como *GemBS* [Merkel et al., 2019].

En esta Tesis Doctoral se ha desarrollado un flujo de datos automatizado que integra todas las etapas de este proceso, así como todo el software necesario (propio o de terceros) basado en la utilización de *Bismark* [Krueger and Andrews, 2011] como alineador y *MethylExtract* [Barturen et al., 2013] como software para inferir la metilación, ya que se trata de los mejores métodos en términos de velocidad, funcionalidad y eficacia. Además en el caso de *MethylExtract* nos permite la detección simultánea (en la misma muestra) de los niveles de metilación y la variación de secuencia, lo que es una gran ventaja para el estudio de asociación entre variación genética y metilación que se propone en esta Tesis Doctoral.

1.4. Islas CpG y sus métodos de detección

Como se introdujo en el apartado 1.1, los genomas de los mamíferos están caracterizados por su patrón de metilación global excepto en las denominadas islas CpG. Estas se caracterizan por desviarse significativamente de la media genómica por su alto contenido en G+C, siendo ricas en dinucleótidos CpG y estando predominantemente hipometiladas [Deaton and Bird, 2011].

A lo largo de los años se han desarrollado diversos métodos para predecirlas. Los primeros fueron los llamados "métodos de ventana" [Takai and Jones, 2002], altamente paramétricos y basados todos ellos en los criterios umbrales de Gardiner-Frommer [Gardiner-Garden and Frommer, 1987]: con-

tenido G+C≥50 %, proporción de CpGs observados/esperados (O/E)≥0.6 y longitud≥200pb. Estos métodos tienen un principal inconveniente, su gran parametrización, además de otros: i) los umbrales deben ajustarse para cada especie individualmente, lo que impide la comparación de especies; ii) la predicción cambia notablemente cuando se cambia cualquiera de los umbrales.

Debido a estos inconvenientes se comenzó el desarrollo de nuevos métodos, en este caso basados en la clusterización de CpGs. Estos, en contraste con los anteriores, no se basan en umbrales pre-establecidos y definen y predicen las islas CpG como clústers de CpGs que ocurren en las secuencias de ADN de forma natural. *CpGcluster* [Hackenberg et al., 2006] fue el primer método en detectar islas CpG como clústers de dinucleótidos CpG estadísticamente significativos. Este método se basa en un único parámetro, la significación estadística, dado que la distancia umbral se obtiene directamente de las secuencias de ADN. Esto tiene la ventaja de que las predicciones más estrictas son simplemente un subconjunto de las predicciones más laxas (es decir, si se usara un valor-p más bajo como umbral), además de permitirnos la comparación entre especies.

Este marco teórico puede ser extendido a otras palabras de ADN (*k-meros*), como se utiliza en *WordCluster* [Hackenberg et al., 2011b], e incluso a la clusterización de cualquier elemento genómico, como en *GenomeCluster* [Dios et al., 2014]. A lo largo de esta Tesis Doctoral y basado en *CpGcluster* y *WordCluster* y se ha desarrollado *gCluster* [Gómez-Martín et al., 2018], que nos permite la predicción de clústers de cualquier "palabra" biológica o combinación de ellas, basándose solamente en la secuencia de ADN y usando como único umbral la significación estadística, pero usando un modelo de distancias mejorado respecto a estos.

Capítulo 2

Objetivos

El objetivo principal de esta Tesis Doctoral consiste en explorar el impacto de la variación genética sobre la expresión génica a través de los cambios de la metilación del ADN. Para ello se abordarán los siguientes objetivos específicos:

1. Desarrollar un flujo de datos para automatizar el proceso de obtención de mapas de metilación y variación genética a partir de la misma muestra. Este debe integrar tanto la descarga automática de los repositorios públicos a partir de un identificador, su procesado y diferentes controles de calidad.

2. Almacenar los mapas de metilación y la variación genética de todas las muestras en una base de datos local que cumpla dos requisitos: i) acceso rápido a los datos y ii) alta flexibilidad para poder añadir nuevos datos si es preciso.

3. Actualizar el contenido de la base de datos *NGSmethDB* y mejorar su interfaz web. Esta nueva versión debe incluir diferentes modos de interrogar la base de datos a partir de un conjunto de tejidos o regiones genómicas definidas por el usuario y diferentes formas de presentar los resultados, como tablas y gráficos interactivos y descargables.

4. Generalizar el método de detección de islas CpG *CpGcluster* a cualquier *k-mero* o combinación de ellos. Este nuevo algoritmo permitirá,

entre otras utilidades, obtener no solamente islas CpG si no regiones genómicas densas en todos los contextos de metilación (CHG y CHH). Para ello es preciso desarrollar un nuevo modelo de distancias que sea capaz de tomar en cuenta el solapamiento de *k-meros* ya sea en la misma hebra o entre las dos hebras.

5. Determinar el test estadístico adecuado para calcular la asociación entre variación genética y metilación del ADN. Se propone analizar el comportamiento del Test Exacto de Fisher (FET) y un test altamente usado basado en regresión lineal.

6. Calcular la asociación estadística entre la variación genética y la metilación empleando el test elegido en el objetivo anterior. Para este cálculo se usan los datos generados en los primeros tres objetivos. El resultado debe ser un conjunto de *SNPs* asociados significativamente (tras aplicar la corrección por ensayo múltiple) a al menos un CpG.

7. Explorar y caracterizar las asociaciones *SNP-CpG*. A parte de una estadística básica, se propone analizar la distribución de distancias entre los pares *SNP-CpG* asociados y la distribución a lo largo del cromosoma. La comparación con una distribución nula generada aleatorizando la composición de las parejas *SNP-CpG* demostrará si los resultados son meramente espurios (la distribución observada es igual a la esperada) o no (existen distancias enriquecidas).

8. Desarrollar una base de datos para almacenar las asociaciones significativas entre variación genética y metilación del ADN. Debe contar con una interfaz web que implemente diferentes formas de búsqueda; centrada en el *SNP* o en un gen de interés. Los resultados reflejarán la interacción entre variación genética, metilación del ADN y expresión génica, permitiendo al usuario obtener conocimiento biológico a partir de los datos de asociación.

Capítulo 3

Material y métodos

3.1. Muestras utilizadas

Las muestras utilizadas para esta tesis proceden del repositorio público *SRA* (NCBI Sequence Read Archive) [Leinonen et al., 2011]. Como se comentó en el apartado 1.3 el método que más citosinas nos permite analizar es la secuenciación a genoma completo (*WGBS*) por lo que para el estudio de asociación entre variación genética y metilación solo se seleccionaron datos obtenidos con esta técnica. Sin embargo en el caso de la base de datos *NGSmethDB* se han incluido muestras tanto de *WGBS* como de *RRBS*.

Para el estudio de asociación se seleccionaron 58 muestras de *WGBS* de diferentes tejidos y proyectos, cuya información se recoge de forma detallada en la Tabla Suplementaria SI. Cada una de las muestras proviene de un individuo o línea celular diferente, con la intención de maximizar el número de haplotipos analizados.

Para la actualización de la base de datos *NGSmethDB* además de las muestras incluidas en el estudio de asociación (que también han sido incluidas) se han procesado muestras de dos proyectos. En primer lugar el proyecto *Cancer Cell Line Encyclopedia* (*CCLE*) [Li et al., 2019, Ghandi et al., 2019]. Este está compuesto por datos de metilación obtenidos por la técnica *RRBS* (entre otros) de 919 líneas celulares de 24 órganos distintos y 109 tipos de cáncer que se resumen en la Tabla Suplementaria SII. Y en segundo lugar el proyecto *PRJNA421218* que está compuesto de 55 mues-

tras de *WGBS* provenientes de neuronas y oligodendrocitos de pacientes con esquizofrenia y pacientes control [Mendizabal et al., 2019].

3.2. Anotaciones genómicas

A lo largo de la Tesis Doctoral se han utilizado distintas anotaciones genómicas en distintos pasos del estudio de asociación entre variación genética y metilación del ADN.

En primer lugar, para filtrar las variantes de secuencia detectadas en las 58 muestras de *WGBS* se utilizó *dbSNP* [Sherry et al., 2001] en su versión 151, de forma que solo se consideraron variantes de secuencia conocidas.

Para clasificar los CpGs asociados de acuerdo con su localización genómica se usaron las siguientes anotaciones:

- Promotores procedentes de la base de datos *EPD* (versión 006) [Dreos et al., 2017].
- *Enhancers* de *GeneHancer* (versión 4.4) [Fishilevich et al., 2017].
- Semáforos CpG (*TL-CpGs*) o CpGs que correlacionan con expresión génica. [Medvedeva et al., 2014, Lioznova et al., 2019]
- Rasgos (*traits*) de *PheGenI* [Ramos et al., 2014] con el objetivo de poder ofrecer una búsqueda por estos *traits*. En esta base datos se pueden encontrar rasgos o *traits* que anteriormente han sido asociados (por estudios de *GWAS* en su mayoría) con *SNPs*.

En el caso de la base de datos *NGSmethDB20* se han utilizado las islas CpG predichas por *gCluster* [Gómez-Martín et al., 2018] por el momento, aunque se planea incorporar más anotaciones.

3.3. Obtención de mapas de metilación

En esta tesis se ha desarrollado un flujo de datos automatizado que integra todas las etapas del proceso de obtención de mapas de metilación a partir de

los datos crudos de *WGBS* obtenidos de *SRA*, y que incluye tanto software propio como de terceros.

Uno de los principales problemas que nos encontramos al tratar de integrar todo el software era la diversidad de versiones de los distintos programas, que necesitan de un entorno estable y con todas las dependencias necesarias instaladas. Para solventarlo se implementó un contenedor *Docker* [Merkel, 2014], que es estable e independiente del sistema operativo en el que se ejecute, permitiendo que se pueda reproducir el protocolo de obtención de mapas de metilación en cualquier equipo, sea cual sea su sistema operativo.

Además se desarrolló un protocolo (Figura 3) en Python, *methylExtract2* que se encuentra disponible en el repositorio *Dockerhub* como ugrbioinfo/-methylextract y que aúna todas las etapas del proceso, i) pre-procesado de los datos, ii) alineamiento frente al genoma de referencia, iii) detección de la metilación y variantes genéticas y iv) subida de los datos a la base de datos *NGSmethDB*.

En el caso de los metilomas utilizados para el cálculo de la asociación entre variación genética y metilación además se añadieron dos pasos intermedios: i) eliminación de duplicados y realineamiento de indels mediante herramientas de los paquetes *Picard tools* [Broad Institute, 2019] y *GATK* [McKenna et al., 2010, Depristo et al., 2011] y ii) eliminación de sesgos de metilación mediante el software *BSeQC* [Lin et al., 2013]. El protocolo seguido en este caso se puede encontrar en la Figura 4. Estos dos pasos no se han realizado sobre el resto de datos incluidos en *NGSmethDB* para mantener la coherencia en el modo de análisis de los resultados incluidos que la base de datos llevaba a lo largo del tiempo, de forma que todos los metilomas incluidos estén analizados con el mismo protocolo. Sin embargo, se está considerando el añadir estos pasos y en el futuro reprocesar en la medida de lo posible los datos ya incluidos para adaptarlos al nuevo protocolo.

En los siguientes apartados se describen en profundidad cada una de estas etapas.

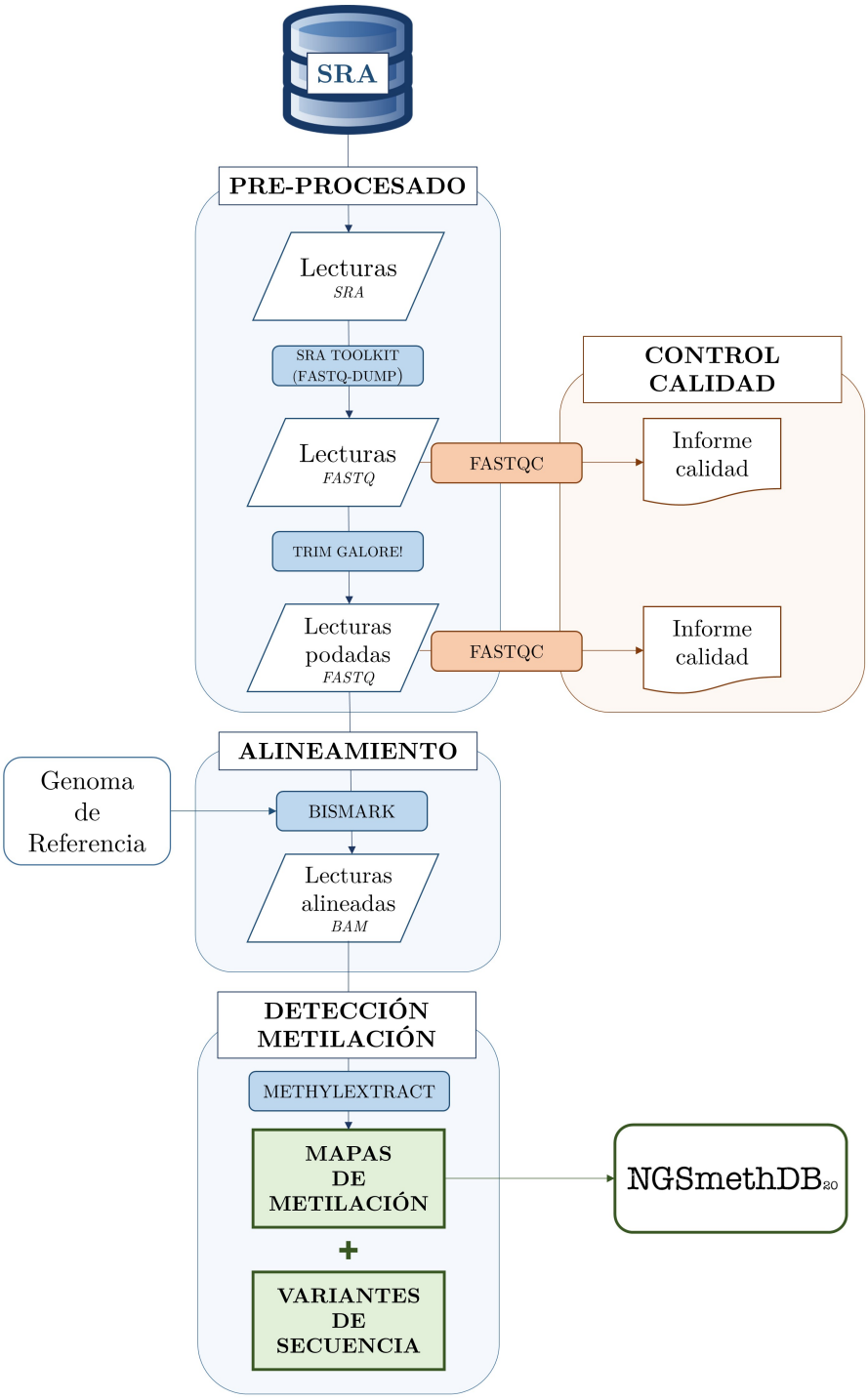


Figura 3: Protocolo de obtención de mapas de metilación. En azul las tres etapas de las que se compone el proceso (Pre-procesado, alineamiento y detección de la metilación), en naranja los controles de calidad efectuados y en verde el resultado del proceso, la obtención de los mapas de metilación, que son incorporados a la base de datos *NGSmethDB* [Hackenberg et al., 2011a, Geisen et al., 2014, Lebrón et al., 2017] y las variantes de secuencia en la misma muestra..

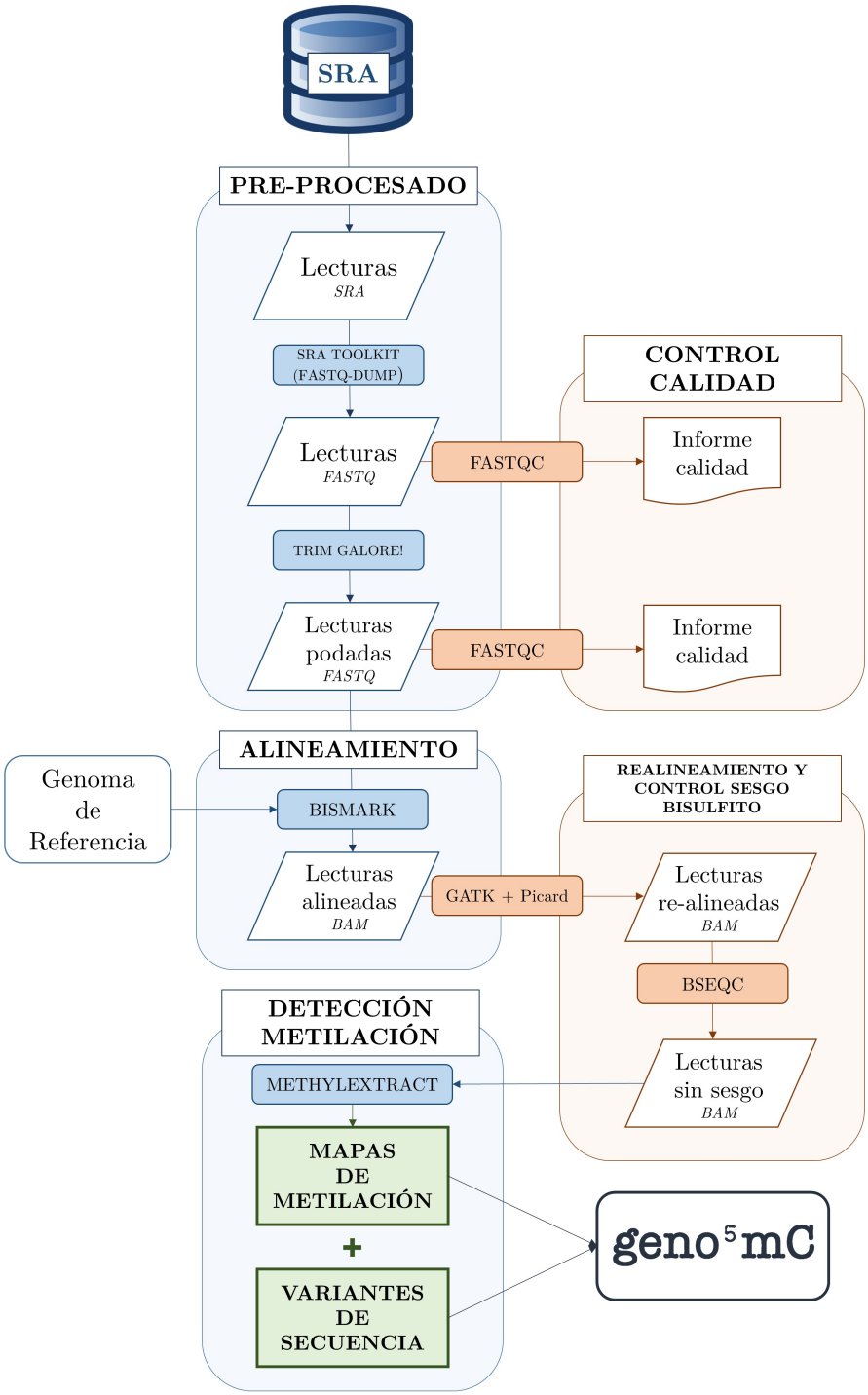


Figura 4: Protocolo de obtención de mapas de metilación para el estudio de asociación entre metilación del ADN y genotipo. En azul las tres etapas de las que se compone el proceso (Pre-procesado, alineamiento y detección de la metilación), en naranja los controles de calidad efectuados y el realineeamiento de indels y control del sesgo por bisulfito y en verde el resultado del proceso, la obtención de los mapas de metilación y variantes de secuencia a partir de los cuales se obtienen los resultados recogidos en $geno^5mC$ [Gómez-Martín et al., 2020].

3.3.1. Pre-procesado de los datos

En primer lugar y tras la descarga de los datos de secuenciación indicados en el apartado 3.1 de la base de datos *SRA* [Leinonen et al., 2011] se procede a su pre-procesado.

La base de datos nos provee los datos en formato *sra* por lo que el primer paso consiste en convertirlos al formato *fastq* utilizando la utilidad *Fastq-dump* perteneciente a *SRA Toolkit* [NCBI, 2017, Leinonen et al., 2011]. A continuación se realiza el primer control de calidad, utilizando el software *FastQC* [Andrews, 2018]. Este nos proporciona un extenso informe donde se recogen datos como la calidad de las lecturas, el contenido en G+C o la presencia de lecturas duplicadas entre otros.

El siguiente paso es el podado de las lecturas mediante *Trim Galore!* [Krueger, 2019]. Este paso es necesario para eliminar las posiciones con baja calidad en el extremo 3' de las lecturas y para eliminar los adaptadores en el extremo 5' y otras secuencias contaminantes producto del proceso de secuenciación. *Trim Galore!* está integrado además con *FastQC*, con lo que podemos realizar de forma automática un segundo control de calidad para asegurar que el proceso se ha ejecutado correctamente. Los parámetros empleados en este proceso se pueden ver en la Tabla 1.

Tras estos pasos de pre-procesado se obtienen las lecturas en formato *fastq* y preparadas para el alineamiento.

Tabla 1: Parámetros utilizados en *Trim Galore!* para el podado de las lecturas, su definición y el valor utilizado.

Parámetro	Definición	Valor
illumina	Indica al programa que la secuencia adaptadora que debe ser podada son las 13 primeras pb del adaptador universal de Ilumina.	True

length	Descarta las lecturas con una longitud menor del valor que se le haya dado tras podarlas ya sea por la baja calidad o por la presencia del adaptador.	35
paired/single	Se indica el método con el que se ha producido la secuenciación, paired-end o single-end.	Dependiendo de la muestra

3.3.2. Alineamiento frente al genoma de referencia

En esta etapa se procede a alinear las lecturas frente al genoma de referencia, en nuestro caso al tratarse de muestras de humano, *hg38* o *GRCh38*. Este paso se lleva a cabo usando el software *Bismark* [Krueger and Andrews, 2011] que es específico para muestras de metilación tratadas con bisulfito y que a su vez utiliza *Bowtie2* [Langmead and Salzberg, 2012] como alineador.

En primer lugar y previo al alineamiento es necesario preparar el genoma de referencia para poder usarlo con *Bismark*. Durante el tratamiento con bisulfito, se produce la desaminación de la citosina a timina y esto reduce la secuencia del ensamblado a un alfabeto de 3 letras (A, T y G), con la consiguiente pérdida de complejidad de secuencia y problemas en el alineamiento. Para poder enfrentarse a este problema *Bismark* usa dos índices de cada ensamblado, también transformados a un alfabeto de tres letras, uno sustituyendo la C por T y otro sustituyendo la G por A. De esta forma consigue alinear las lecturas provenientes del tratamiento por bisulfito de forma correcta. Para obtener dichos índices se utilizó la utilidad *bismark_genome_preparation* que viene incluida con *Bismark*.

A continuación se procede propiamente al alineamiento frente al genoma de referencia con el programa principal de *Bismark* y especificando como opción que se utilice *Bowtie2* como alineador. Al finalizar este paso obtenemos las lecturas alineadas en formato *bam*. Los parámetros utilizados en este paso se resumen en la Tabla 2.

Tabla 2: Parámetros utilizados en *Bismark* para el alineamiento de las lecturas, su definición y el valor utilizado.

Parámetro	Definición	Valor
bowtie2	Uso del alineador *bowtie2* en lugar de *bowtie1*.	True
N	Número de desemparejamientos permitidos en el alineamiento de la semilla.	1
L	Longitud de las subcadenas de la semilla durante el alineamiento. Valores más pequeños hacen el alineamiento más lento pero más sensitivo.	20

3.3.3. Realineamiento de *indels* y corrección del sesgo por bisulfito

Como se comentó anteriormente, los datos utilizados para el estudio de asociación entre metilación y genotipo, después del alineamiento, se pasan por un proceso de realineamiento de *indels* y corrección del sesgo por bisulfito. Ambos pasos mejoran la calidad del alineamiento y el primero de ellos, además, mejora el alineamiento de las variantes de secuencia, lo cual es muy importante de cara al estudio de asociación.

En primer lugar se realiza la eliminación de lecturas duplicadas durante el proceso de secuenciación mediante las herramientas *AddOrReplaceReadGroups* y *Markduplicates* pertenecientes a *Picard tools* [Broad Institute, 2019]. A continuación se realiza el realineamiento de *indels* mediante las herramientas *RealignerTargetCreator* e *IndelRealigner* de *GATK* [McKenna et al., 2010]. En ambos pasos utilizamos los parámetros por defecto del software.

Por último se realiza la corrección del sesgo de bisulfito utilizando el software *BSeQC* [Lin et al., 2013], utilizando los parámetros recogidos en la Tabla 3. La salida de este software ya se encuentra en formato *bam* y por tanto puede utilizarse de forma directa para el paso de detección de

metilación.

Tabla 3: Parámetros utilizados en *BSeQC* para la correción del sesgo por bisulfito

Parámetro	Definición	Valor
l	Longitud original de los alineamientos previa al podado del adaptador o por baja calidad.	Dependiendo de la muestra, ej.: 151
p	valor-p máximo utilizado para la corrección, expresado como el exponente. Un valor de 2 equivale a un valor-p de 0,01.	2

3.3.4. Detección de la metilación y las variantes de secuencia

El último paso para la obtención de los mapas de metilación es la detección como tal del estado de metilación a partir de las lecturas alineadas que obtuvimos en el paso anterior.

Este paso se lleva a cabo gracias al software *MethylExtract* [Barturen et al., 2013], que además de detectar los niveles de metilación de todas las citosinas es capaz de detectar las variantes de secuencia de la misma muestra, característica que lo hace especialmente útil para el estudio de asociación. Esto último lo hace mediante el uso de un Test Exacto de Fisher de forma análoga a *varScan* [Koboldt et al., 2012]. Además integra una serie de controles de calidad que podemos ajustar en función de diversos parámetros. En la Tabla 4 se pueden ver los que se han utilizado en este caso.

Tabla 4: Parámetros usados en *MethylExtract* para la obtención de los mapas de metilación, sus definiciones y el valor utilizado.

Parámetro	Definición	Valor
flagW	Tag usado en el fichero SAM para marcar los alineamientos procedentes de la hebra Watson.	0 para lecturas *single-end* y 99,147 para *paired-end*
flagC	Tag usado en el fichero SAM para marcar los alineamientos procedentes de la hebra Watson.	16 para lecturas *single-end* y 83,163 para *paired-end*
minDepthMeth	Mínimo número de lecturas alineadas a una citosina para inferir su nivel de metilación.	1
methNonCpGs	Límite de la fracción de citosinas fuera del contexto CpG que se encuentran metiladas. Un alto nivel de estas indica un fallo del bisulfito, por lo que se eliminan aquellas lecturas que sobrepasen este valor.	0,9
Context	Contexto de secuencia en el que se van a inferir los niveles de metilación.	CG
minQ	Valor de calidad PHRED score mínimo considerado. Las lecturas con un nivel menor son descartadas.	20
delDup	Activa la detección de reads duplicados.	Y
simDupPb	Número mínimo de nucleótidos iguales en la región 5' de dos lecturas que mapean en la misma posición para considerar que se trata de lecturas duplicadas.	32

La salida de *MethylExtract* produce los siguientes ficheros:

- **CG.output:** Mapa de metilación de la muestra en el contexto CpG.
- **CHG.output / CHH.output:** Mapas de metilación en los contextos CHG o CHH si se ha indicado como opción. Dado que en humanos la metilación en los otros contextos no es especialmente relevante como ya se comentó en el apartado 1.1 esta no se ha calculado.
- **SNVs.vcf**: Variantes de secuencia encontradas en la muestra en formato *vcf*.
- **RatiosCGStats.log**: *Log* del proceso que aúna los parámetros utilizados y estadísticas generales de los resultados.

3.4. Almacenamiento en *MongoDB* de los datos de metilación

Una vez obtenidos los mapas de metilación es preciso almacenarlos en una base de datos para que puedan ser servidos por la base de datos *NGSmethDB* a los usuarios y para que se puedan usar por el software que calcula la asociación para obtener los resultados.

En el caso de los datos de metilación se ha optado por usar como base de datos *MongoDB*. *MongoDB* es una base de datos del tipo *NoSQL* que, frente al esquema tradicional de las bases *SQL*, tiene una estructura en documentos con formato *JSON* y con esquemas dinámicos. Esto hace posible que la comparación de datos de diferentes muestras sea mucho más rápida, así como la inclusión de nuevos datos en la base de datos. Este punto es de vital importancia ya que la base de datos *NGSmethDB* se encuentra en continua actualización, por lo que al optar por este esquema se facilita mucho el añadir nuevas muestras.

El esquema de la base de datos desarrollada se basa en que cada ensamblado comprende una base de datos (por ejemplo *hg38*) y dentro de cada una de esas bases de datos podemos encontrar una colección para cada cromosoma, así como colecciones para las tablas de contenidos y otras anotaciones (como los *SNPs* de la base de datos *dbSNP* [Sherry et al., 2001]).

Dentro de una colección, cada documento representa una citosina y contiene todas las muestras y los datos que se tienen sobre ella: genotipo, metilación o metilación diferencial.

Además la base de datos *MongoDB*, siguiendo el esquema de *MethylExtract2* también se encuentra instalada en un contenedor *Docker*, permitiendo su portabilidad y su independencia del sistema madre en el que se encuentre.

Para poblar esta base de datos y su posterior consulta se han desarrollado una serie de scripts en Python, que se encuentran disponibles en *Github* (https://github.com/cris12gm/mongoTools).

3.5. Asociación genotipo – metilación: *SNPs* Asociados

Para estudiar la asociación entre los datos de genotipo y de metilación que se encuentran almacenados en la base de datos *MongoDB*, se desarrolló un script basado en el modelo estadístico mostrado en la Figura 5.

A nivel celular solo son biológicamente posibles tres niveles de metilación: 0 (no metilado), 0,5 (metilación alelo específica) y 1 (metilado). En base a esto, en primer lugar se clasifican los valores de metilación de cada dinucleótido CpG en cada muestra en tres grupos:

- **Metilado (M)**: Valores de $methRatio > 0{,}65$
- **Intermedio (I)**: Valores de $methRatio \leq 0{,}65$ y $\geq 0{,}35$
- **No metilado (U)**: Valores de $methRatio < 0{,}35$

Solo se analizan aquellos CpGs con una cobertura mayor o igual a 5, es decir, que al menos 5 lecturas mapeen en esa posición. Además, con la intención de mejorar el tiempo de cómputo, también se descartan aquellas CpGs que en las 58 muestras muestran el mismo estado de metilación, es decir, están metiladas/no metiladas en todas las muestras. En total se analizaron 15.663.849 citosinas una vez realizados estos filtros.

A continuación se filtra para todas las muestras aquellas variantes que aparecen anotadas en *dbSNP* (versión 151), y que tengan además una fre-

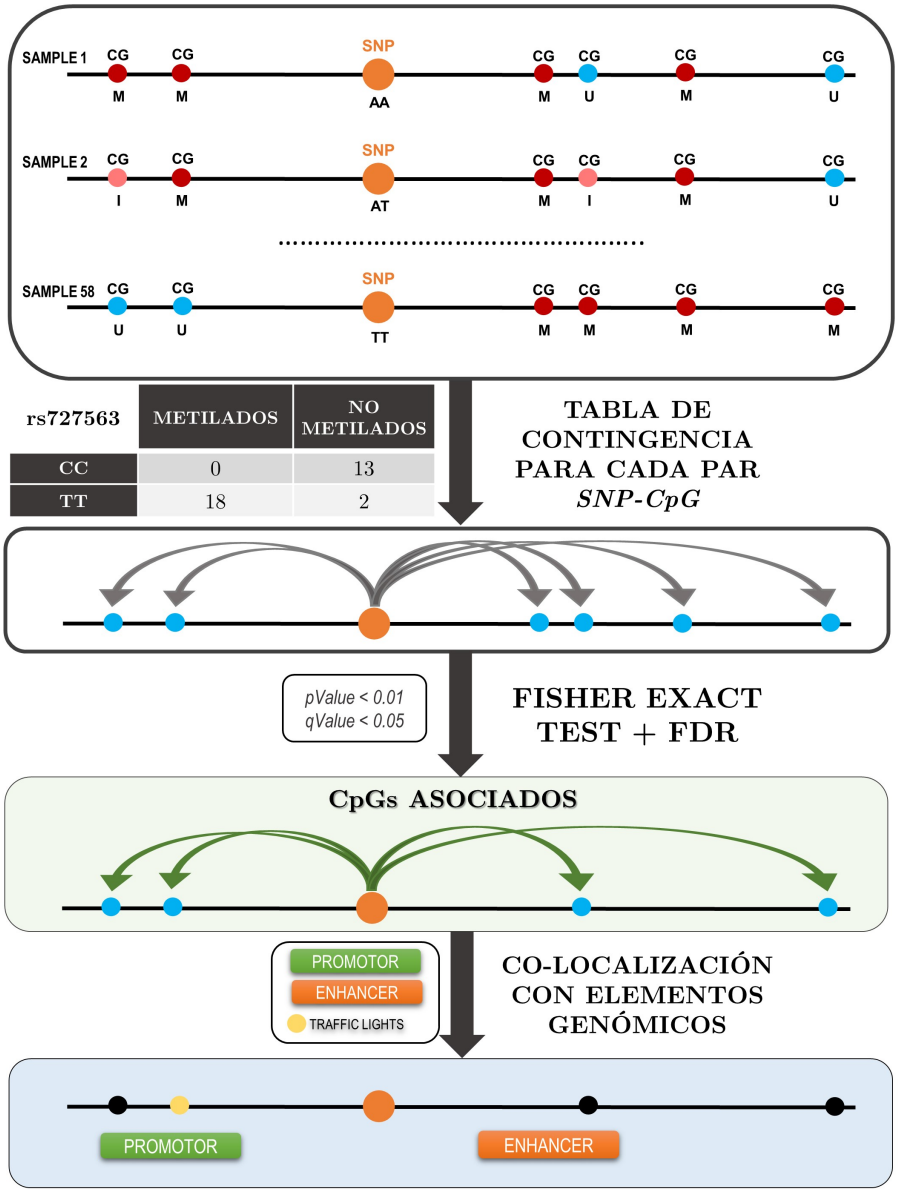


Figura 5: Representación esquemática del modelo estadístico para el test de asociación. El *SNP rs727563*, localizado en *chr22:32771861* se ha usado para este ejemplo.

cuencia mínima del alelo menor de 0,1 en nuestro set de 58 muestras. El número final de *SNPs* analizados es de 4.086.616.

El siguiente paso es obtener una tabla de contingencia (2x2) para cada par *SNP-CpG* donde se tienen en cuenta los homocigotos para los dos alelos, y los estados de metilación M y U. Se puede ver un ejemplo en la Tabla 5. Para simplificar el estudio tanto los heterocigotos como los estados intermedios de metilación no se consideran en el test, y por tanto no intervienen en la significación estadística.

Esta simplificación se realiza porque es difícil distinguir cuando los valores intermedios de metilación provienen de la heterogeneidad celular o si realmente se da el caso de una metilación alelo específica.

Tabla 5: Ejemplo de tabla de contingencia para el *SNP rs727563* y la citosina *chr22:32771861*.

rs727563	**Metilado**	**No metilado**
Homocigoto CC	0	18
Homocigoto TT	13	2

Con esta tabla de contingencia se realiza un Test Exacto de Fisher [Fisher, 1992] a partir del cual se puede calcular un valor-p para cada asociación *SNP-CpG*. Finalmente, y dado el elevado número de tests, se realiza una corrección de los valores-p usando la *FDR* (*False Discovery Rate*) [Benjamini and Hochberg, 1995], corrigiéndolos con el número de test realizados para cada *SNP*. Solo las parejas *SNP-CpG* con una FDR$\leq$0,05 se consideran para los análisis posteriores y por tanto se añaden a la base de datos.

3.6. Bloques de *SNPs* asociados

Algunos *SNPs* asociados pueden pertenecer al mismo haplotipo, y por tanto tender a ser heredados juntos. Es por ello que calculamos los que llamamos "bloques de *SNP*" en nuestros datos de asociación.

Para ello se utilizó un script propio `https://github.com/cris12gm/snpsAssociatedTools/getBlocks.py` basado en ventanas móviles. Se establece una ventana de 5Kb y se agrupan en un mismo bloque todos aquellos

SNPs que, estando a una distancia menor que el tamaño de la ventana, se encuentren asociados a un mismo CpG.

3.7. Co-localización de los datos de asociación con elementos genómicos

Una vez identificados los pares de *SNP-CpG* asociados se determina su co-localización con los elementos genómicos descritos en el Apartado 3.2: promotores, *enhancers* y *TL-CpGs*. Esto se realizó mediante el uso de las distintas herramientas del software *BEDTools* [Quinlan and Hall, 2010], y luego se incorporaron a la base de datos para dar lugar a los distintos niveles de resultados que se ofrecen en $geno^5mC$.

3.8. Almacenamiento en bases de datos de los datos de asociación: MariaDB

Una vez calculados los datos de asociación es necesario almacenarlos en una base de datos para su posterior utilización tanto para el cruce con las anotaciones genómicas como se ha descrito en el apartado 3.7 como para servirlos a través de la página web $geno^5mC$ [Gómez-Martín et al., 2020]. En este caso la base de datos elegida es del tipo SQL, en concreto *MariaDB* que se basa en la conocida base de datos *MySQL*.

La elección de una base de datos *SQL* en este caso se justifica porque en que el esquema organizado en tablas es más adecuado para este tipo de datos, que no necesitan de actualización continua.

De nuevo, y siguiendo el esquema del resto de servicios, la base de datos forma parte de un contenedor *Docker* que es independiente de la máquina madre en la que se ejecute.

3.9. Desarrollo de servidores web: *geno*^{5}*mC* y *NGS-methDB*

Tanto el servidor web *geno*^{5}*mC* como el que soporta la base de datos *NGS-methDB* se implementaron usando el entorno de trabajo *Django* [Django Software Foundation, 2020]. Además para el desarrollo de las plantillas y mejorar la interactividad de la página se usaron también *Bootstrap* y *Javascript*.

Para conectar la base de datos *MariaDB* y Python se utilizó el *ORM* (*Object Relational Mapper*) *SQLAlchemy* [Michael Bayer, 2012]. Para la visualización de los datos y también para incrementar la interactividad de la aplicación web se utilizó el paquete *Plotly*.

Capítulo 4

Resultados

4.1. Protocolo automatizado para la obtención de mapas de metilación: *methylExtract2*

En esta Tesis Doctoral se ha desarrollado el protocolo *methylExtract2* como una mejora del anterior protocolo del que disponíamos para obtener los mapas de metilación.

El nuevo protocolo consiste en un flujo de datos automatizado que integra todas las etapas del proceso de obtención de mapas de metilación, desde la descarga de los datos crudos de la base de datos *SRA* hasta la subida de los resultados a la base de datos *NGSmethDB*. Todo ello se encuentra integrado en un contenedor *Docker* [Merkel, 2014] que se encuentra disponible en el repositorio ugrbioinfo/methylextract2 de *Dockerhub*. Este paso es importante ya que facilita que se mantengan las versiones del software estables, y por tanto todos los metilomas que se añaden a la base de datos o se usan para el estudio de asociación se procesan del mismo modo, y además permite la portabilidad del software de un equipo a otro.

El protocolo *MethylExtract2* (Figura 3 y Figura 4 en el apartado 3.3) se ha desarrollado en Python, y aúna todas las etapas del proceso de obtención de los perfiles de metilación, desde el pre-procesado de los datos hasta su subida a la base de datos *NGSmethDB*. Este último paso no se ofrece en la versión disponible en *Dockerhub* dado que solamente se usa de forma interna

en nuestro grupo de investigación.

4.1.1. Obtención de mapas de metilación

El proceso de obtención de los mapas de metilación conlleva 4 pasos esenciales y uno optativo, que *MethylExtract2* realiza de forma automática. A continuación se describen cada uno de estos pasos de forma breve. En el apartado 3.3 se encuentra más información acerca de los parámetros utilizados y el software utilizado en cada caso.

1. **Descarga de los datos y cambio a formato *fastq***: *MethylExtract2* admite dos tipos de entrada de datos, una (o varias) IDs de *SRA* o datos propios. En el primer caso, el software verifica que las IDs proporcionadas pertenecen a *SRA* y se encuentran disponibles para la descarga, procediendo entonces a realizarla, seguida del procesado hasta formato *fastq*. En el caso de los datos propios, se chequea cual es el formato en el que están, y se hace el cambio a *fastq* de ser necesario.

2. **Pre-procesado de los datos y análisis de calidad**: Una vez que los datos están en formato *fastq* se realiza un primer control de calidad, y a continuación, el podado de las lecturas para eliminar los adaptadores en el extremo 5' y las posiciones con baja calidad del extremo 3'.

3. **Alineamiento frente al genoma de referencia.**

4. **Eliminación de duplicados, realineamiento de *indels* y eliminación del sesgo por bisulfito**. Como se indicó en el apartado 3.3 este paso solo se realiza en el caso de aquellos metilomas utilizados para el estudio de asociación.

5. **Obtención de los mapas de metilación**, utilizando *MethylExtract* [Barturen et al., 2013].

Como cada uno de los pasos conlleva la elección de una serie de parámetros (que se detallaron en el apartado 3.3), *MethylExtract2* se basa en la utilización de un fichero de configuración donde el usuario puede editarlos. En la página de *Dockerhub* además se encuentra un pequeño manual de uso

del protocolo, y dentro del *Docker* se puede encontrar un conjunto de datos de ejemplo.

4.1.2. Conexión con *NGSmethDB20*

Para subir los resultados a la base de datos *NGSmethDB20* se han desarrollado una serie de scripts disponibles en un repositorio *Github* (`https://github.com/cris12gm/mongoTools`) que permiten realizar estos pasos de forma automática, y que se recogen de forma resumida en la Tabla 6.

Tabla 6: Herramientas desarrolladas para la conexión entre la base de datos *NGSmethDB* y el motor de base de datos *MongoDB*.

Software	Función
updateDatabaseContent.py	Actualiza la tabla "*Database content*" con las muestras que se le indique
createJson.py	Crea un archivo *.json* para cada cromosoma a partir del archivo de salida de *MethylExtract*
createMongoImportBash.py	Crea un archivo *.sh* con los comandos necesarios para subir las muestras a *MongoDB*

Todos estos *scripts* se encuentran integrados en MethylExtract2 y se lanzan de manera automática para subir la muestra de forma directa a *NGSmethDB* si así se indica.

4.2. Asociación entre metilación y genotipo: *geno^{5}mC*

El principal objetivo de esta Tesis Doctoral, como se comentó en apartados anteriores, es dilucidar la posible relación entre variación genética, metilación del ADN y expresión génica (Figura 1). Para abordar esta cuestión, en primer lugar se desarrolló un protocolo (Apartado 3.5)) mediante el cual se puede determinar si en nuestro conjunto de muestras existe una asociación estadística entre los valores de metilación y el genotipo. Dichos resultados se almacenan en la base de datos *geno^{5}mC* [Gómez-Martín et al., 2020].

En este apartado se abordarán los principales resultados del estudio de asociación:

- Elección del método más adecuado para el cálculo de la asociación entre metilación del ADN y genotipo
- Visión estadística general de las asociaciones *SNP-CpG*, abordando su distribución de distancias o de valores-p entre otros resultados.
- Co-localización de los resultados de asociación con diferentes elementos genómicos.
- Principales características y algunos ejemplos de la base de datos $geno^5mC$.

4.2.1. Elección del método de asociación

Se han descrito distintos métodos de asociación estadística entre metilación y genotipo, pero todos ellos tienen en común que están basados en la implementación de modelos de regresión entre las dos variables. Uno de los más utilizados es *Matrix eQTL* [Shabalin, 2012], que aunque ha sido desarrollado para analizar *eQTLs* también puede utilizarse para *meQTLs* ya que ambos datos son cuantitativos y se pueden tratar del mismo modo.

Sin embargo, estos métodos tratan los datos de metilación como si de una variable continua se tratara, algo que no se corresponde con los tres estados biológicos que presenta la metilación: M (Metilado), I (Metilación alelo específica) y U (No metilado). Basándonos en esto implementamos un nuevo modelo basado en el Test Exacto de Fisher (*FET*) [Fisher, 1992] que solo considera un número determinado de estados de metilación.

Con el objetivo de comparar ambas aproximaciones, realizamos el cálculo de la asociación metilación-genotipo en las muestras descritas en el Apartado 3.2 para el cromosoma 22 usando el software *Matrix eQTL* por un lado y un algoritmo basado en el Test Exacto de Fisher (Apartado 3.5) por otro, y comparamos los resultados. En la Tabla 7 se recoge el número de asociaciones encontradas por cada método y la intersección de ambos.

Tabla 7: Comparación del número de asociaciones predichas por *Matrix eQTL* y el método basado en el Test Exacto de Fisher.

Método	Número de asociaciones predichas
Matrix eQTL	3.137.879
Test Exacto de Fisher	2.034.702
Ambos	167.889

El método de *Matrix eQTL* predice aproximadamente 1 millón de asociaciones más que el basado en el *FET* y además solo existe un 8 % de asociaciones comunes. Para evaluar que método es más adecuado, en la Figura 6 se muestran algunos ejemplos de las asociaciones predichas tanto por los dos métodos como las que son exclusivas de uno de ellos.

En Figura6A podemos ver una asociación que predicen ambos métodos, donde el homocigoto para el alelo T del *SNP rs932380* es el que se encuentra asociado a la metilación de la citosina *chr22:41998798* mientras que el alelo G se encuentra asociado a la no metilación. Sin embargo si nos fijamos en la Figura 6B, donde se puede ver la asociación que solo predice *Matrix eQTL*, podemos ver como la citosina *chr22:46990842* en las muestras en las que están ambos alelos del *SNP rs743057* (C y A) se encuentra siempre metilada; el que salga una asociación mediante la regresión simplemente se debe a pequeñas diferencias dentro de la misma clase. Estas diferencias, pueden provenir de la heterogeneidad celular o de procesos de metilación parcial, dado que como se comentó anteriormente solo existen tres estados biológicos en el caso de la metilación: M (Metilado), I (Metilación alelo específica) y U (No metilado).

Por último en la Figura 6C podemos ver una asociación predicha solo por el método basado en *FET*, en la que el alelo C del *SNP rs6009831* está asociado con la metilación de la citosina *chr22:34134421* y el alelo T con la no metilación. En este caso *Matrix eQTL* no predice la asociación lo cual probablemente sea debido a la dispersión de los datos, pese a la clara tendencia que se observa.

Es por ello que nos pareció más adecuado la utilización del modelo basado en el Test Exacto de Fisher para analizar la asociación entre metilación del ADN y genotipo, y a partir del algoritmo que se desarrolló basado en el

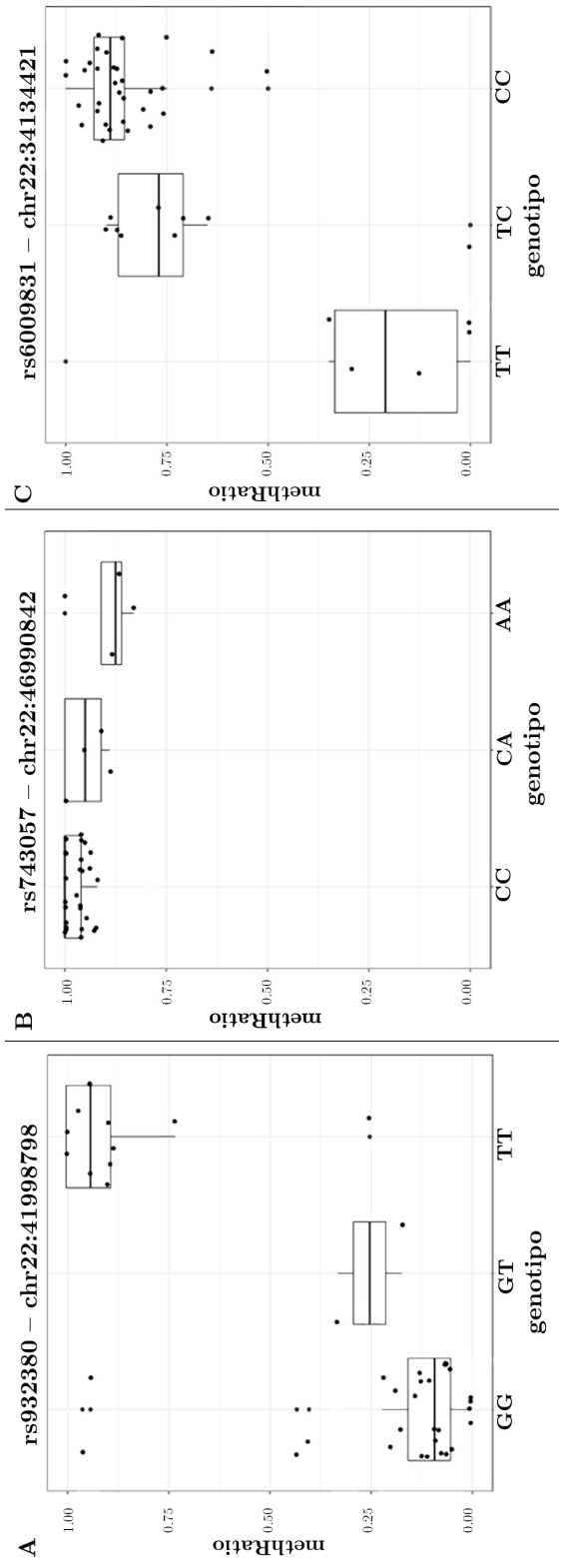


Figura 6: Comparación entre *Matrix eQTL* y el método basado en el Test Exacto de Fisher. A) Ejemplo de asociación predicha por ambos métodos. B) Ejemplo de asociación predicha solo por *Matrix eQTL*. C) Ejemplo de asociación predicha solo mediante nuestro algoritmo (basado en el Test Exacto de Fisher). .

(Apartado 3.5) se calcularon las asociaciones *SNP-CpG* que se describen en los siguientes apartados.

4.2.2. Visión general

Como se explicó en detalle en el apartado 3.5, después de filtrar los datos de variación genética de las 58 muestras por una frecuencia mínima del alelo menor de 0,1 y eliminar todos aquellos *SNPs* que no pertenecen a la base de datos *dbSNP* (versión 151), se obtuvieron 4.086.616 *SNPs* para el análisis de asociación. Tras realizar el Test Exacto de Fisher y filtrar los resultados con una $FDR \geq 0{,}05$, solo 51.585 (1,3 %) *SNPs* resultaron estar asociados con al menos un CpG.

Por otra parte, encontramos que los niveles de metilación de 5.417.468 (19,3 %) dinucleótidos CpG se asociaban con al menos un *SNP*. Cabe destacar que este número de CpGs es cinco veces mayor que el que permite examinar el método *Infinium MethylationEPIC array* de *Illumina*, por lo que nos da información de muchos más CpGs que este y otros métodos relacionados.

En total tenemos, por tanto, 506.041.598 pares *SNP-CpG* asociados distribuidos a lo largo de los 22 autosomas, como podemos ver en la Tabla 8. El cromosoma con más CpGs asociados es el cromosoma 2, mientras que el cromosoma 21 es en el que hay menos asociaciones.

Tabla 8: Número de asociaciones *SNP-CpG* por cromosoma. Se resaltan en negro los cromosomas con mayor (cromosoma 2) y menor (cromosoma 21) número de asociaciones.

Cromosoma	**Nº de Asociaciones**	**Cromosoma**	**Nº de Asociaciones**
Cromosoma 1	44.748.782	Cromosoma 12	22.742.402
Cromosoma 2	88.170.280	Cromosoma 13	16.381.969
Cromosoma 3	39.777.579	Cromosoma 14	13.673.969
Cromosoma 4	45.510.809	Cromosoma 15	9.189.023
Cromosoma 5	29.035.131	Cromosoma 16	19.100.199
Cromosoma 6	22.050.840	Cromosoma 17	8.458.175
Cromosoma 7	27.500.005	Cromosoma 18	10.123.858

Cromosoma 8	28.262.846	Cromosoma 19	3.724.623
Cromosoma 9	15.704.588	Cromosoma 20	7.143.873
Cromosoma 10	39.603.475	Cromosoma 21	1.925.738
Cromosoma 11	11.178.732	Cromosoma 22	2.034.702

En la Figura 7 podemos ver la distribución del número de CpGs asociados por *SNP* en los 22 autosomas. Los valores de la mediana oscilan entre los 979 CpGs del cromosoma 21 y los 13.684 del cromosoma 2. Solo una pequeña fracción de los *SNPs* (aproximadamente el 1,3 %) están asociados, y la mayoría de estos correlacionan con miles de CpGs diferentes.

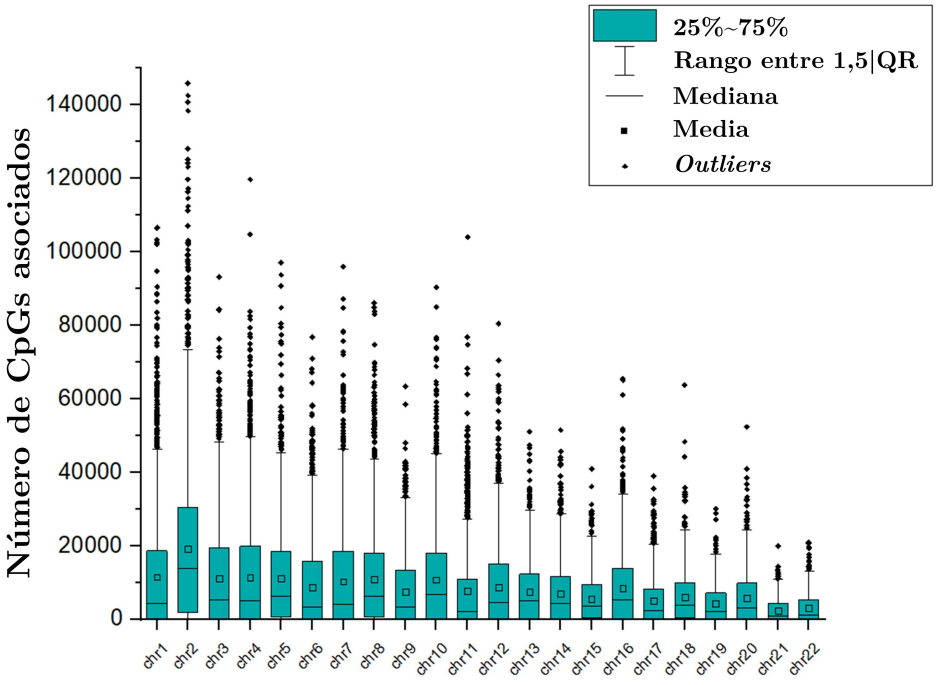


Figura 7: Distribución del número de CpGs asociados por *SNP* en función del cromosoma.

En la Figura 8 se puede ver un ejemplo de asociación entre un *SNP* (*rs854944*) y 2 CpGs asociados con él: *chr22:20430694* y *chr22:20430697*. En ambos casos el alelo G, que corresponde al alelo de referencia del *SNP* está asociado con la metilación de las citosinas y el alelo A con la no metilación.

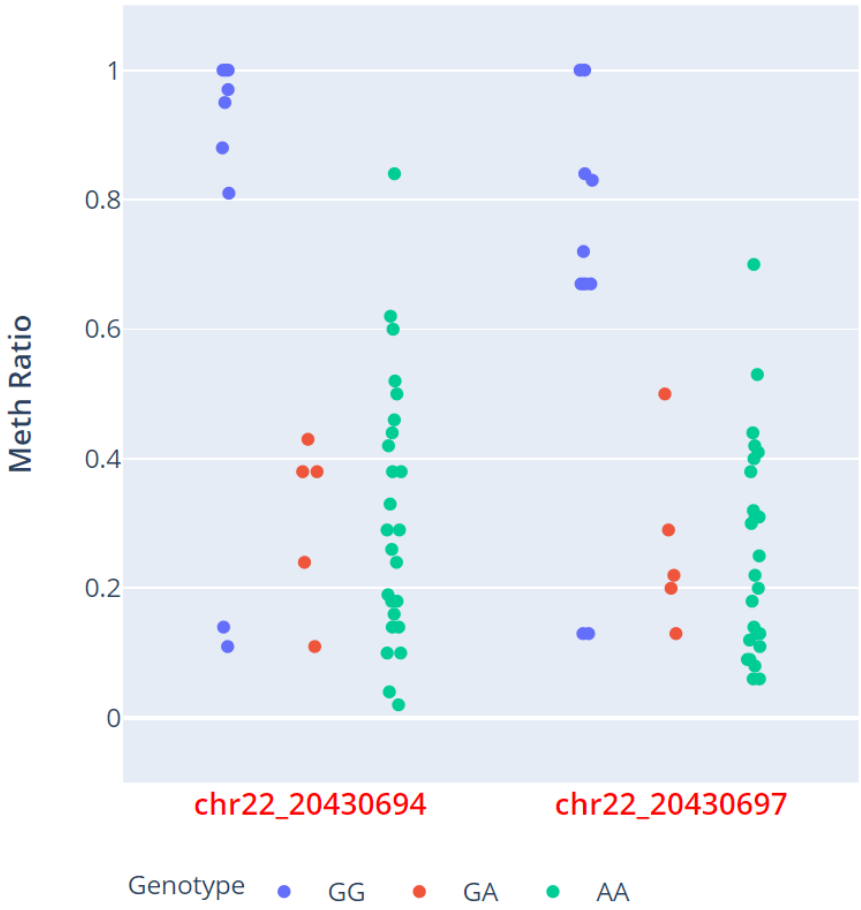


Figura 8: Ejemplo de dos pares *SNP-CpG* asociados. Se trata del *SNP rs854944* y las citosinas *chr22:20430694* y *chr22:20430697* (imagen tomada de *geno*^{5}*mC* [Gómez-Martín et al., 2020], `https://arn.ugr.es/geno5mc`)

La distribución espacial de los *SNPs* asociados a lo largo de los cromosomas se puede ver en la Figura 9. En el eje de abscisas se puede ver la posición cromosómica de los *SNPs* para cada uno de los 22 autosomas, y en el eje de ordenadas el logaritmo del valor-p corregido (o valor-q) más pequeño de todas las asociaciones de dicho *SNP* con los CpG del cromosoma (cambiado de signo). En todos los cromosomas se observan asociaciones *SNP-CpG* con valores-p muy bajos, y por tanto muy significativas.

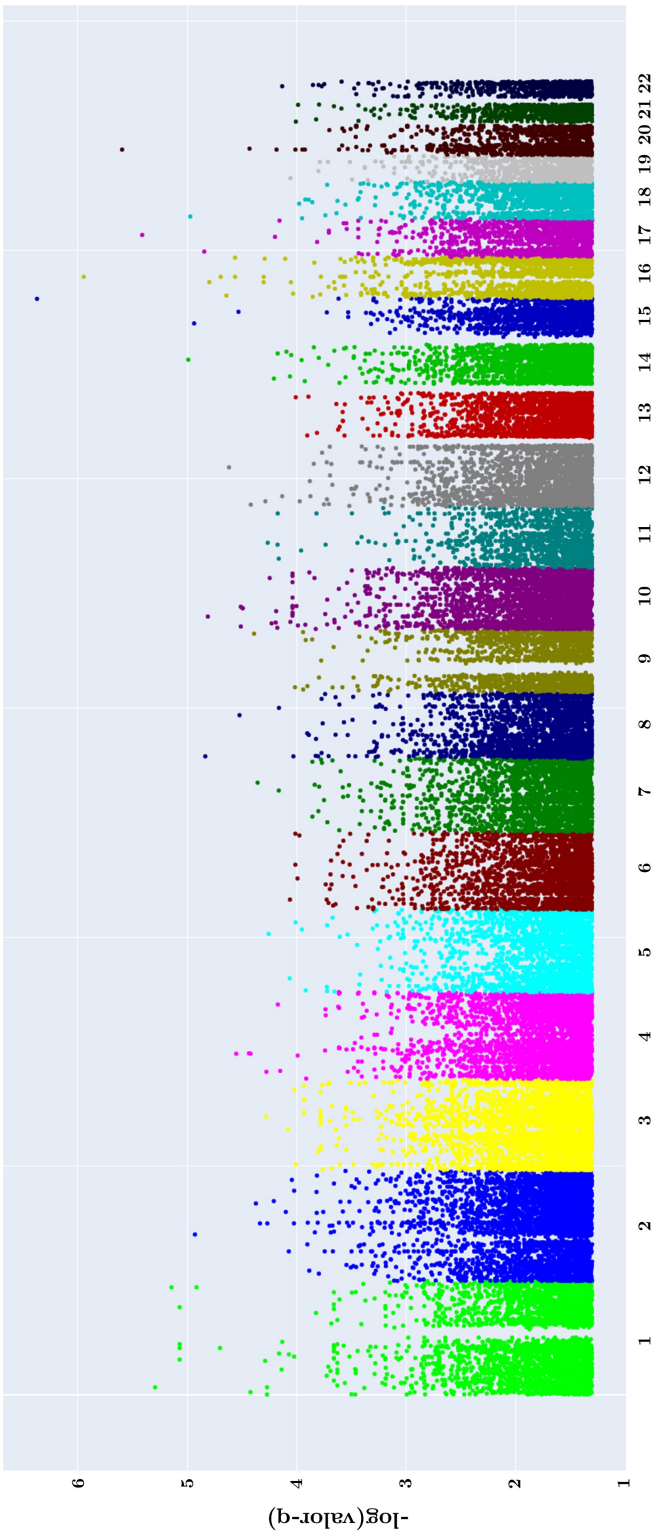


Figura 9: Distribución espacial de los *SNPs* asociados a lo largo de los cromosomas de forma análoga a un *Manhattan plot*. En el eje de abcisas se representan los 22 autosomas con sus coordenadas genómicas y en el eje de ordenadas el -log(valor-q) más significativo de todos los CpGs asociados a cada *SNP*. De esta forma cada *SNP* se encuentra solo representado por un punto.

4.2.3. Distribución de distancias *SNP-CpG*

Para caracterizar las asociaciones se analizó la distribución de distancias entre los *SNP* y sus correspondientes CpG asociados. Además, se ha comparado dicha distribución con la esperada por azar. Esto permite comprobar si existe sobre o infrarrepresentación estadísticamente significativa a ciertas distancias. El proceso seguido es el siguiente:

1. Se calcula el histograma de distancias observadas utilizando bines de 100Kb (en negro en la gráfica superior de la Figura 10).

2. Para calcular el histograma de valores esperados (en rojo en la gráfica superior de la Figura 10), se realizan 100 rondas de aleatorización para cada *SNP* asociado. Para ello, se tiene en cuenta el número de CpGs asociados que tiene cada *SNP* y se aleatoriza el mismo número de etiquetas entre todos los posibles CpGs asociados de su cromosoma. En cada ronda de aleatorización se calcula una distribución de distancias esperada, y finalmente se aúnan todas las rondas obteniendo una desviación estándar y un número esperado (que se corresponde con la media).

3. Se calcula un *z-score* para determinar las distancias a las que existen asociaciones sobre o infrarrepresentadas estadísticamente significativas (gráfica inferior de la Figura 10).

En la Figura 10 se puede ver como la distribución de distancias observada no decrece monótonamente y muestra en algunos puntos claras diferencias respecto a la distribución esperada. Todos los cromosomas muestran una clara sobrerrepresentación de las asociaciones a distancias cortas (inferiores a 2Mb en la mayoría de los casos), observándose que hasta 2Mb el número de parejas *SNP-CpG* observadas es significativamente mayor que lo esperado por azar (*z-score* 3,3).

En el caso de las distancias largas se observan tanto tramos de sobrerrepresentación como de infrarrepresentación. En el cromosoma 22 llama la atención especialmente un pico muy pronunciado entre las distancias de 30 y 33 Mb. Una posible explicación de estos picos es la existencia de bloques

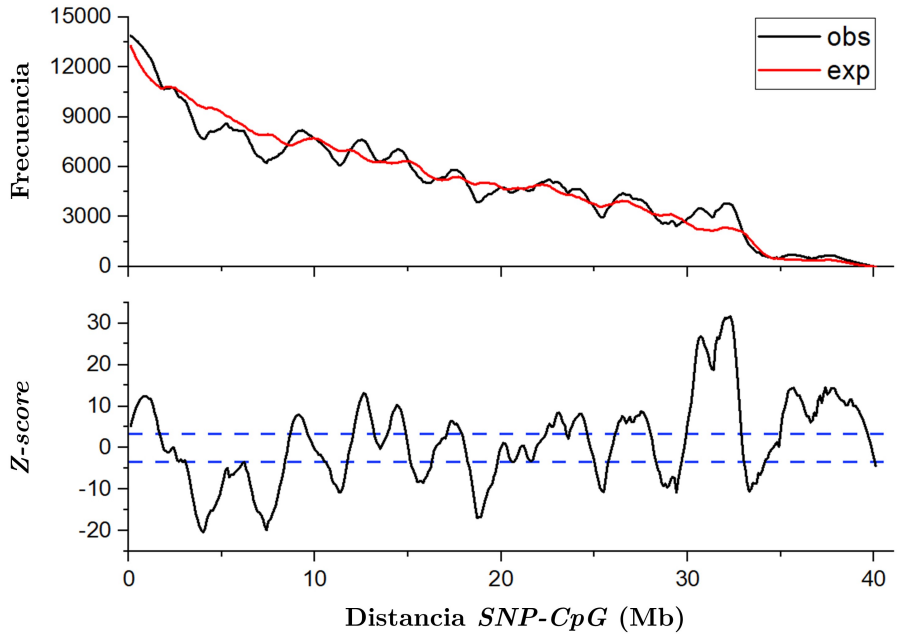


Figura 10: Distribución de distancias *SNP-CpG* para todos los pares de asociados del cromosoma 22. En la parte superior se puede observar la distribución de distancias *SNP-CpG* observada (negro) y esperada (rojo) y en la parte inferior la significación estadistica expresada en forma de *z-scores*. Se marcan con lineas puntuadas las diferencias significativas entre observados y esperados (*z-score* 3,3).

de *SNPs* que pertenezcan a un mismo haplotipo y que se encuentren todos ellos asociados a un mismo clúster de CpGs. Para evaluar si este es el caso, hemos calculado los bloques de *SNPs* para este cromosoma (Apartado 3.6 y buscado si existe alguno que se encuentre a una distancia de 30 – 33Mb. El resultado de esta búsqueda se puede ver en la Figura 11.

En esta figura se puede ver un bloque de 36 *SNPs* situados en la región *chr22:46923540-46972342* que se encuentran asociados a 26 CpGs (*chr22:16602566-16602738*) que forman parte de una isla CpG. Un ejemplo claro de asociación es el *SNP rs1569518* que se asocia con los valores de metilación de 25 de los 26 CpGs (parte superior de la Figura 11). Solo la asociación entre estas dos regiones contribuye con 560 pares *SNP-CpG*, lo que explicaría la sobrerrepresentación de asociaciones a distancias entre 30 y 33 Mb que observamos en la Figura 10.

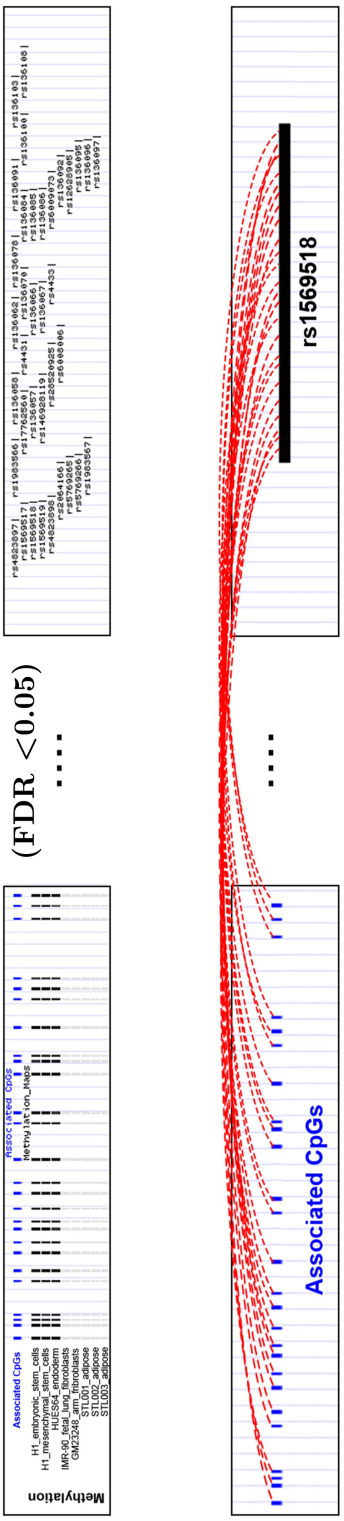


Figura 11: Representación de dos regiones genómicas, una muy densa en CpGs (en concreto 26) y otra muy densa en *SNPs* (36). Un total de 560 pares *SNP-CpGs* asociados derivan de la asociación entre estas dos regiones, explicando los altos valores de *z-score* a distancias entre 30 y 33 Mb en el cromosoma 22.

Sin embargo, aunque estas asociaciones de largo alcance pueden aparecer por la presencia de bloques de *SNPs* que pertenecen a un mismo haplotipo, como se ha visto en el caso anterior, recientemente se ha demostrado que los dominios de metilación pueden formar grandes conexiones en forma de bucle, conectando *loci* a varias decenas de megabases [Zhang et al., 2020], por lo que también podrían provenir de un vínculo funcional; este punto se discutirá más adelante mediante algunos ejemplos en el apartado 4.2.6.

Por otra parte hemos analizado si la distribución observada de distancias cambia con la significación estadística, es decir, si a valores-p más bajos se obtiene una distribución de distancias diferente. Para ello, en la Figura 12 podemos observar la distribución de distancias de las asociaciones *SNP-CpG* del cromosoma 22 para distintos valores-p umbral desde $1e^{-3}$ (en amarillo), el más laxo, hasta el más restrictivo, $1e^{-6}$ (en azul). La parte A de la Figura 12 corresponde a la distribución de distancias completa, y se puede ver como solo hay una clara diferencia entre los valores de los distintos histogramas a distancias pequeñas. Si miramos estas distancias pequeñas en detalle (Figura 12B) podemos observar como las distancias muy pequeñas están claramente asociadas a valores-p menores. Esto quiere decir, por tanto, que las asociaciones estadísticamente más significativas se encuentran a distancias cortas.

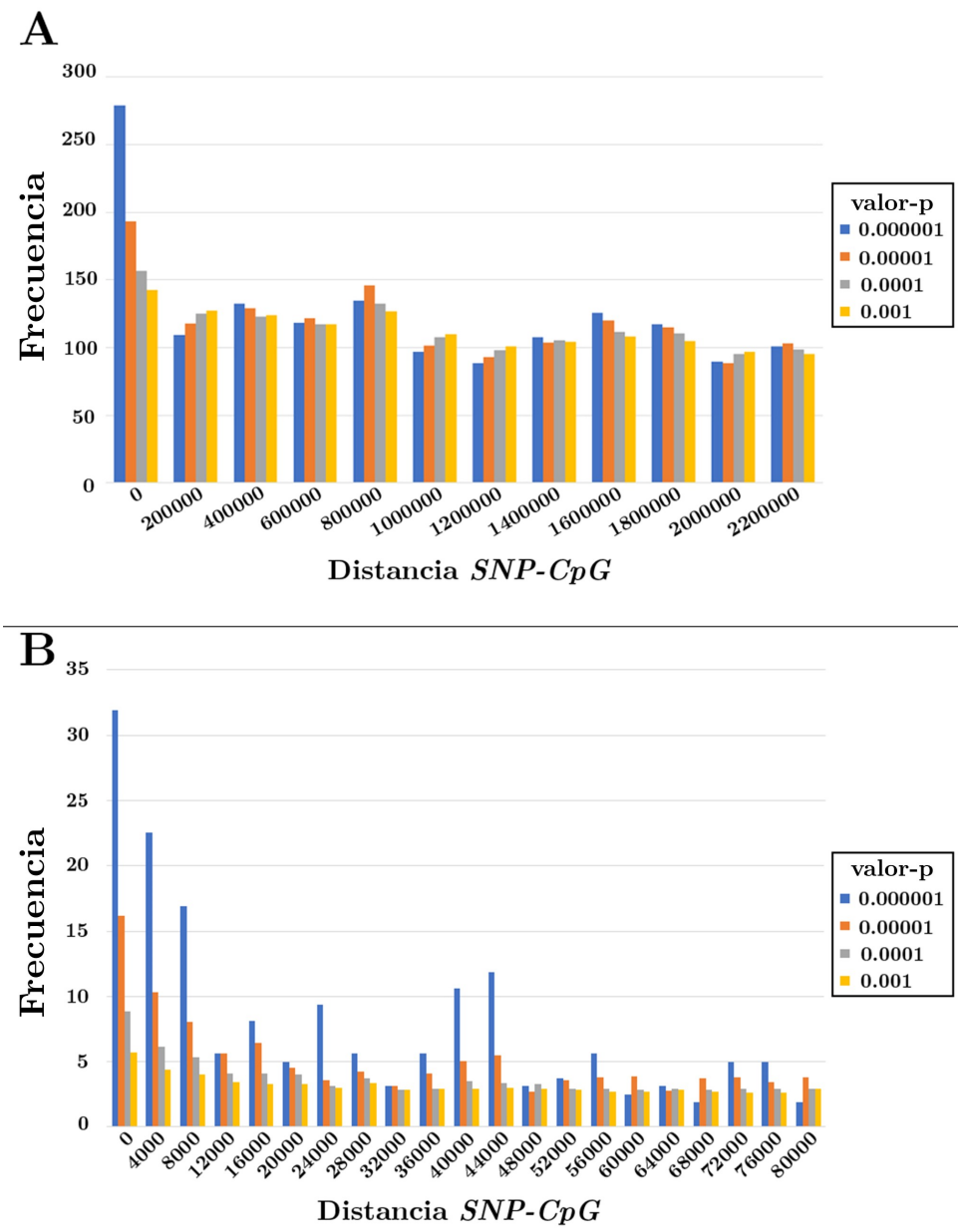


Figura 12: Distribución de distancias *SNP-CpG* a distintos umbrales de significación máxima (ver leyenda de la gráfica).

4.2.4. Co-localización con elementos genómicos

Para caracterizar las asociaciones más relevantes desde el punto de vista biológico y mostrarlas en *geno*^{5}*mC* hemos analizado su co-localización con distintos elementos genómicos, como se explica en el apartado 3.7. Una estadística general de estos resultados se recoge en la Tabla 9.

Tabla 9: Tabla resumen de la co-localización de los pares *SNP-CpG* asociados con promotores, *enhancers* y *TL-CpGs*

Elemento genómico	**Nº de pares *SNP-CpG* que solapan**	**Nº de *SNPs* que tienen CpGs asociados que solapan (y % respecto al total de *SNPs* asociados)**	**Nº de CpGs asociados que solapan**
Promotores	291.038	32.492 (64,92 %)	5.328
Enhancers	65.050.714	45.081 (87,39 %)	1.108.146
TL-CpGs	2.033.050	37.192 (72,10 %)	18.276

Podemos observar que gran parte de los *SNPs* asociados tienen al menos algún CpG asociado en los distintos elementos genómicos, con tantos por ciento que van desde el 64 % en promotores al 87 % en *enhancers*. Esto indica que las asociaciones estadísticas encontradas entre metilación y genotipo pueden tener significado biológico.

4.2.5. *geno*^{5}*mC*

La base de datos *geno*^{5}*mC* [Gómez-Martín et al., 2020] (Figura 13) (`https://arn.ugr.es/geno5mc`) recoge los datos de los pares *SNP-CpG* asociados. Además dispone de distintos métodos de búsqueda jerarquizada de los resultados que ayuda a explorar estas asociaciones y extraer de ellas conclusiones biológicas.

Figura 13: Captura de la página principal de la base de datos *geno^{5}mC* (`https://arn.ugr.es/geno5mc`).

Está compuesta de las siguientes secciones:

- *Query DB*: Donde podemos encontrar los cuatro modos de interrogar la base de datos.
- *Statistics*: Recoge algunas de las estadísticas más importantes comentadas en los apartados anteriores.
- *Primary Data*: Tabla de con las muestras usadas en el estudio.
- *Downloads*: Descarga de todos los datos de asociación por cromosoma
- *Tour*: Tour del modo de consulta *Query SNP* utilizando un ejemplo para mostrar el uso de la base de datos.

La base de datos puede interrogarse de cuatro formas distintas que se encuentran en la sección *Query DB*: i) usando un *SNP ID* de *dbSNP*; ii) a partir de un rasgo (*trait*); iii) *gene ID*; y iv) región genómica. En la Figura 14 se puede ver un ejemplo de la salida de cada modo de consulta y a continuación se describirá cada uno de ellos.

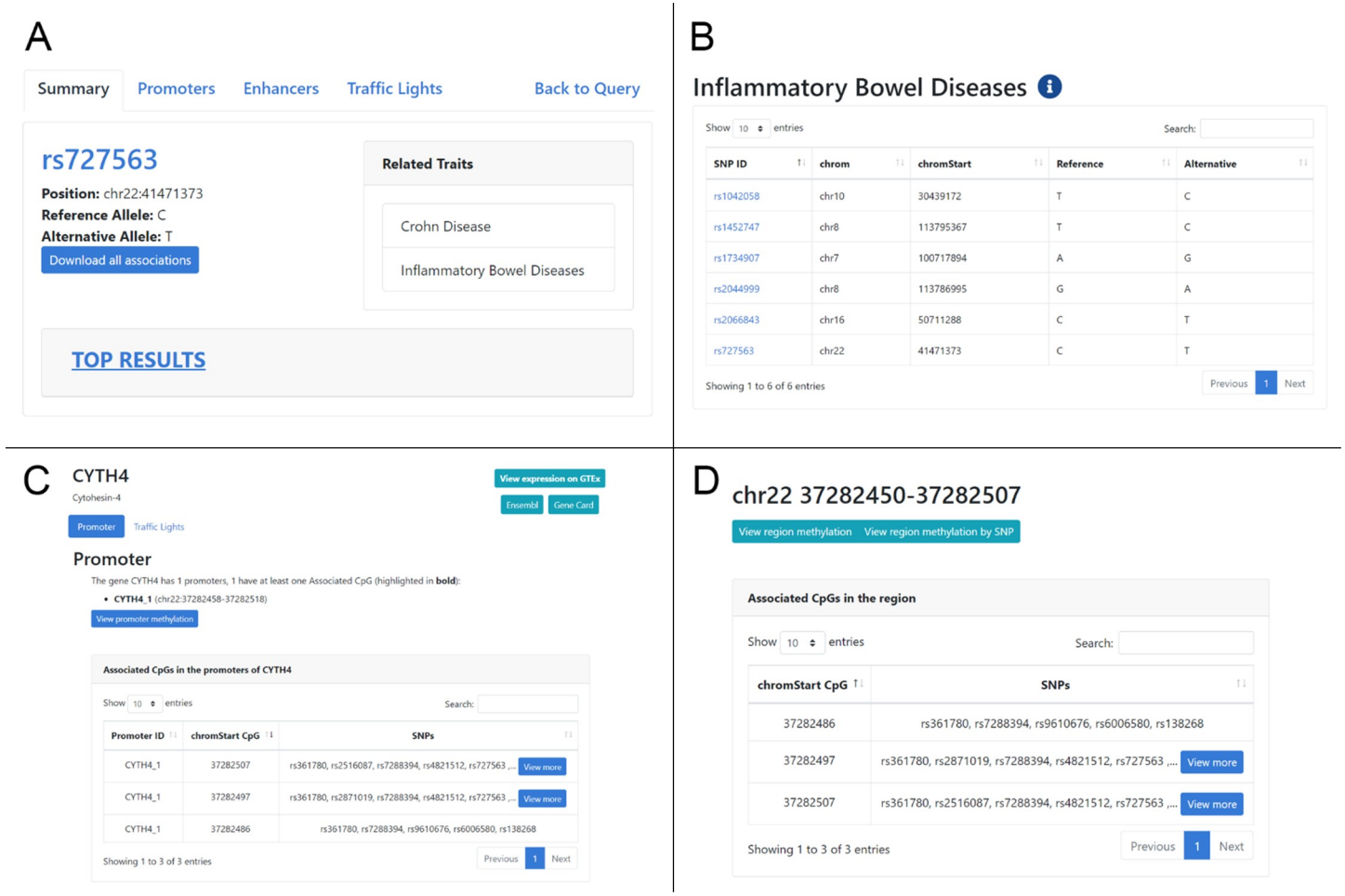


Figura 14: Ejemplo de la salida de los cuatro modos de consulta de *geno*^{5}*mC*: A) *Query SNP*, B) *Query trait*, C) *Query gene* y D) *Query region*.

Búsqueda por *SNP* (*Query SNP*)

A través de este modo de consulta (Figura 14A) se puede obtener toda la información disponible en la base de datos sobre un *SNP* concreto. Los resultados están organizados de forma jerárquica, siguiendo el esquema de la Figura 15.

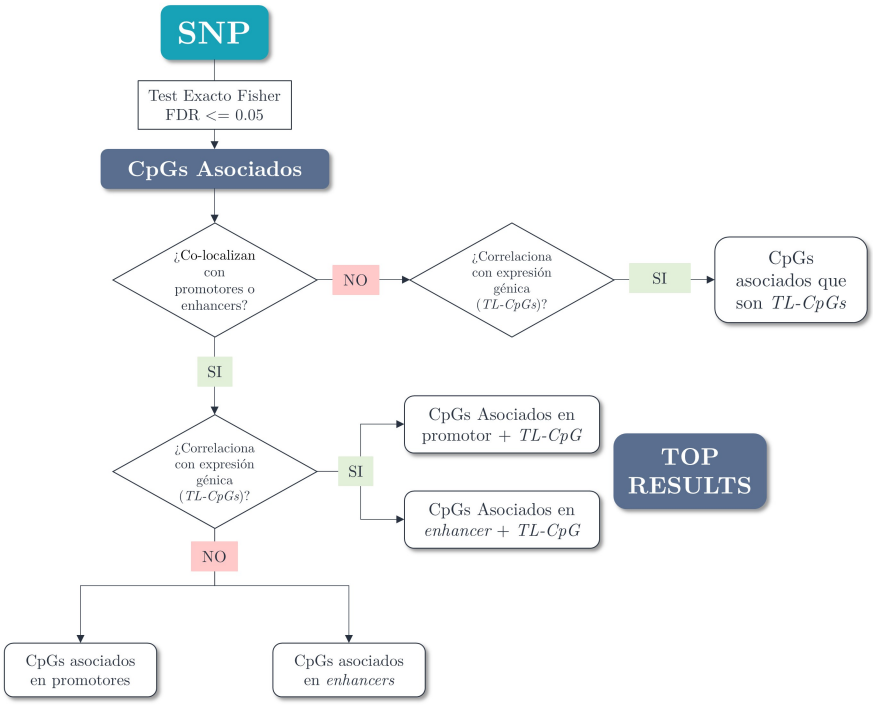


Figura 15: Jerarquía de los resultados que se muestran tras hacer una consulta en el modo de consulta por *SNP* (*Query SNP*).

De este modo se agrupan los CpGs asociados al *SNP* objeto de la búsqueda por relevancia biológica, ya sea porque se localizan en regiones reguladoras conocidas (promotores o *enhancers*) o porque están anotados como semáforos CpG (*TL-CpGs*).

La salida se reparte entre distintas pestañas, donde podremos ir encontrando los distintos niveles de información según su relevancia biológica.

En primer lugar tenemos la pestaña "*Summary*" donde además de datos generales de la asociación y la posibilidad de descargar todos los CpG asociados con ese *SNP* nos encontramos los "*Top results*". Resaltamos como

"*Top results*" aquellos CpG asociados que correlacionan con los niveles de transcripción de al menos un gen y que además están en la región promotora del mismo, o bien en un *enhancer*.

A continuación tenemos las pestañas "*Promoters*", "*Enhancers*" y "*Traffic Lights*", donde encontramos aquellos CpGs que co-localizan con alguno de estos elementos genómicos, respectivamente.

En cualquiera de los casos, los resultados se muestran también en forma de figuras interactivas, así como de tablas donde se muestran tanto la información del elemento genómico con el que co-localiza como la metilación de todos los CpGs asociados que se encuentran en ese elemento genómico (Figura 16).

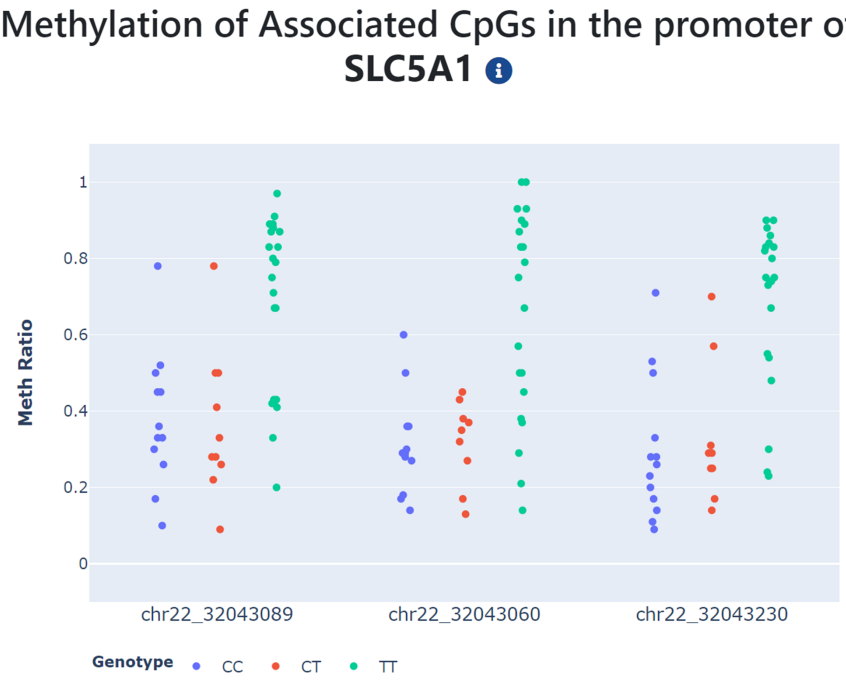


Figura 16: Ejemplo de las figuras de metilación interactivas que se encuentran en todos los modos de consulta de *geno*^{5}*mC* y esta en concreto en el modo de consulta *Query SNP*. En este caso se trata de 3 CpG asociados al *SNP rs727563* que se encuentran en el promotor del gen *SLC5A1*. Imagen tomada de: `https://arn.ugr.es/geno5mc/plotElement/?element=promoter;snp=rs727563;name=SLC5A1;start=32043031;end=32043091`.

Búsqueda por rasgo (*Query Trait*)

En este tipo de búsqueda (Figura 14B) se muestran aquellos *SNPs* asociados que anteriormente han sido anotados como asociados con algún rasgo o *trait* por la base de datos *PheGenI* [Ramos et al., 2014]. Los *SNPs* se presentan en forma de tabla y se puede pinchar en cada uno de ellos para acceder a toda la información del *SNP* seleccionado. En la Figura 17 se puede ver el esquema que se ha seguido para seleccionar estos *SNPs*.

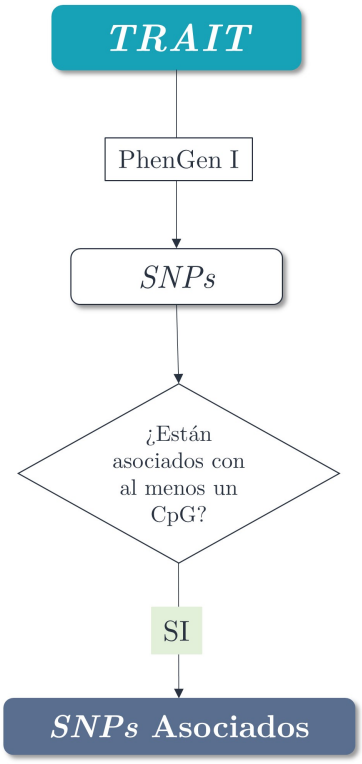


Figura 17: Esquema del modo de consulta por rasgo (*Query Trait*) en *geno^{5}mC*.

Búsqueda por gen (*Query Gene*)

En este modo de consulta (Figura 14C) se puede buscar cualquier *gene ID*, y los resultados se muestran de una forma ligeramente diferente, ya que la salida está organizada en dos pestañas. En la primera de ellas encontramos

aquellos CpGs asociados localizados en la región promotora del gen buscado, junto con una lista de todos los *SNPs* asociados a estos CpGs. En la segunda pestaña encontramos aquellos CpGs asociados que correlacionan con la expresión del gen que se ha buscado (*TL-CpGs*). Además, en cada una de las dos pestañas disponemos de un botón que nos permite visualizar los valores de metilación y el genotipo (para cada uno de los *SNPs* asociados) para todas las citosinas del promotor o promotores, en el caso de que exista más de uno, o todos los *TL-CpGs* respectivamente.

En la Figura 18 se puede ver el esquema que se ha seguido para seleccionar los resultados que se muestran en esta utilidad.

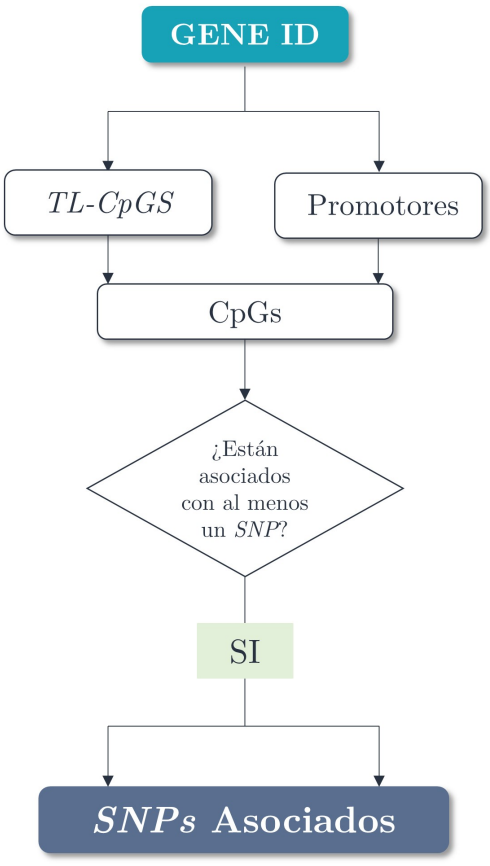


Figura 18: Esquema del modo de consulta por gen (*Query Gene*) en *geno*^{5}*mC*.

Búsqueda por región genómica (*Query Region*)

Finalmente se encuentra el modo de consulta *Query Region* (Figura 14D) que permite al usuario introducir una región genómica utilizando sus coordenadas e interrogar con ella la base de datos.

Los resultados se organizan igual que en el caso de *Query Gene* pero con la excepción de que no se reportan los CpGs semáforo (*TL-CpGs*) dado que estos se limitan a regiones centradas alrededor de los genes. El esquema jerárquico usado para definir los resultados en este tipo de búsqueda se puede ver en la Figura 19.

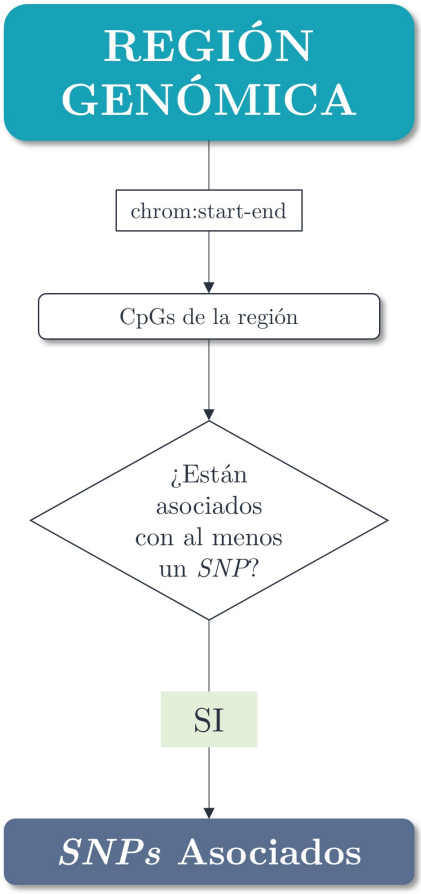


Figura 19: Esquema del modo de consulta por región (*Query Region*) en *geno^{5}mC*.

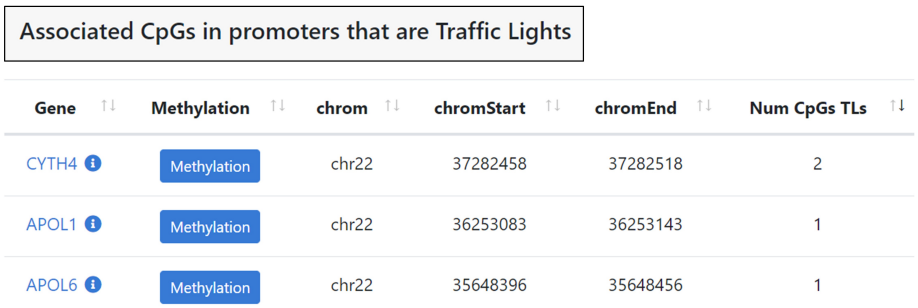
4.2.6. Ejemplos

Para ilustrar el funcionamiento de *geno*^{5}*mC* y cómo su uso puede ayudar a extender el conocimiento sobre las implicaciones funcionales de algunos *SNPs*, en este apartado se recogen algunos ejemplos.

El *SNP rs727563* se conoce que está asociado estadísticamente (mediante *GWAS*) con la enfermedad inflamatoria intestinal [Liu et al., 2015]. Este *SNP*, situado en el cromosoma 22 (*chr22:41471373*, *GRCh38.p12*) tiene dos alelos posibles, C y T, siendo C el alelo de riesgo. Y aunque el análisis por *GWAS* muestra una asociación estadística muy significativa no se conoce nada acerca de la relación funcional entre el *SNP* y la enfermedad.

Este *SNP* es una variante intrónica del gen *ACO2*, que codifica para la proteína Aconitasa 2. Esta proteína pertenece a la familia de las aconitasas/isomerasas IPM y cataliza la interconversión de citrato a isocitrato a través de cis-aconitato en el segundo paso del ciclo de Krebs. Algunas enfermedades asociadas a esta proteína son la degeneración cerebro-retinal infantil o la atrofia óptica 9 [Spiegel et al., 2012, Metodiev et al., 2014], pero no se ha encontrado ninguna relación aparente con la enfermedad inflamatoria intestinal.

Si se hace una búsqueda de este *SNP* en *geno*^{5}*mC* encontramos que está asociado a un total de 6280 CpGs. Estos, como se explicó anteriormente, se jerarquizan siguiendo su relevancia biológica (Figura 15). De acuerdo con esto, encontramos 2299 CpGs asociados localizados en *enhancers*, 53 que son CpGs semáforo (*TL-CpGs*) y 16 en promotores. Si nos centramos en los "*Top results*", en la Figura 20 se muestran los 3 genes que tienen CpGs asociados a este *SNP* en su promotor y que además son CpGs semáforo.



Associated CpGs in promoters that are Traffic Lights

Gene	Methylation	chrom	chromStart	chromEnd	Num CpGs TLs
CYTH4	Methylation	chr22	37282458	37282518	2
APOL1	Methylation	chr22	36253083	36253143	1
APOL6	Methylation	chr22	35648396	35648456	1

Figura 20: Genes con CpGs asociados al *SNP rs727563* en sus regiones promotoras. Estos CpGs además son semáforos o *TL-CpGs*. Imagen tomada de: `https://arn.ugr.es/geno5mc/querySNP/snp/rs727563`.

El gen *CYTH4* (Figura 21) con 2 CpGs semáforos *TL-CpGs* asociados codifica para la proteína citohesina 4, que se ha relacionado anteriormente con la enfermedad inflamatoria intestinal [Peters et al., 2017].

Por otra parte, los análisis de co-localización de los CpGs asociados con promotores revelaron que la región promotora del gen *SLC5A1* presenta 3 CpGs asociados (no reportados como *TL-CpGs*), como se puede ver en la Figura 22. Este gen codifica una proteína miembro de la familia de los transportadores de glucosa dependientes de sodio (*SGLT*). Esta es una proteína integral de membrana que se encarga del transporte de la glucosa ingerida en la dieta desde el lumen del intestino, tejido donde se expresa principalmente. Esto es importante ya que esta función también ha sido relacionada con la enfermedad inflamatoria intestinal [Lee, 2013, Brzozowski et al., 2016]. En la sección de *enhancers* podemos observar que el *SNP* está asociado a CpGs que se encuentran en 6 *enhancers* de este gen, reforzando aún más la posible relación funcional entre el *SNP* y *SLC5A1* a través de cambios de metilación en el ADN de los CpGs asociados.

Para mostrar que el caso anterior no es una excepción se analizaron algunos *SNPs* asociados a otros fenotipos, que tampoco están localizados en regiones reguladoras conocidas. El *SNP rs4780401* se ha asociado estadísticamente mediante *GWAS* con la Artritis reumatoide [Okada et al., 2014] pero no se localiza en ningún gen (o cerca de él) o región reguladora que estén relacionados. En nuestra base de datos encontramos que este *SNP* se

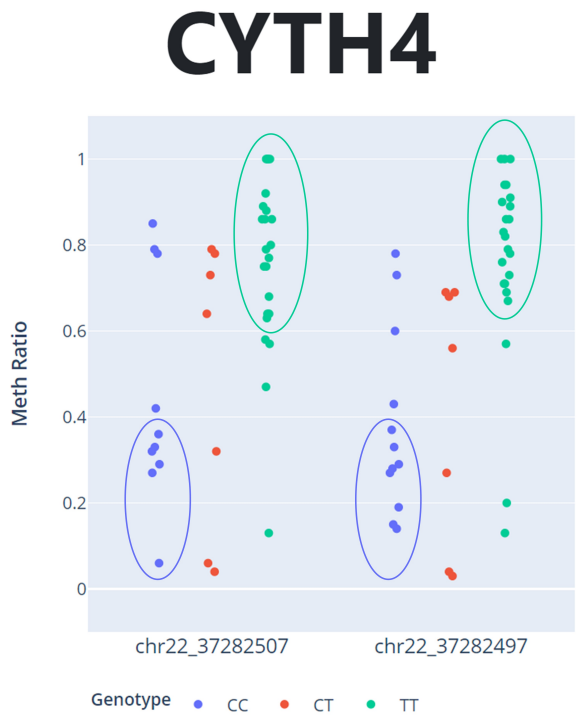


Figura 21: Distribución de los valores de metilación de dos de los *TL-CpGs* asociados que se encuentran en la región promotora del gen *CYTH4*. Puede verse claramente como el genotipo CC está asociado a la no metilación en ambos casos (*chr22:3728507* valor-p= $6{\cdot}10^{-4}$, *chr22:37282497* valor-p = $1.3{\cdot}10^{-4}$). Imagen tomada de: `https://arn.ugr.es/geno5mc/plotElement/?element=promoter;snp=rs727563;name=CYTH4;start=37282458;end=37282518`

asocia con un *CpG-TL* en el promotor del gen *MLKL* que se ha asociado a la misma enfermedad recientemente [Wang et al., 2020].

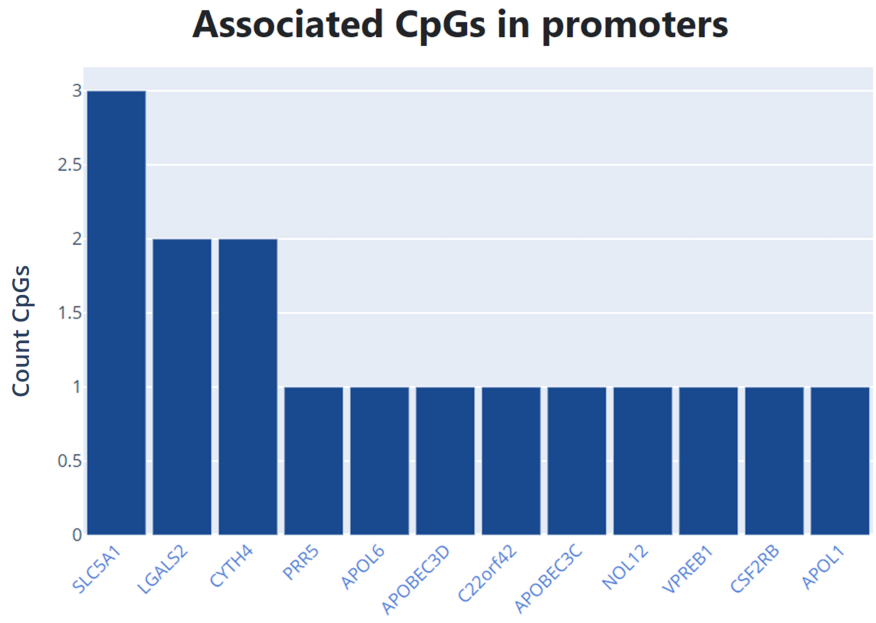


Figura 22: CpGs asociados al *SNP rs727563* que se encuentran en regiones promotoras. Imagen tomada de:
https://arn.ugr.es/geno5mc/querySNP/snp/rs727563#promoters

Un último ejemplo es el *SNP rs10746333* que se ha asociado por *GWAS* a la Artritis Reumatoide, también sin una relación funcional aparente. $geno^5mC$ muestra que se asocia con varios *CpGs-TL* situados en los promotores de los genes *LRMP* y *MGP* que anteriormente han sido descritos como asociados con la misma enfermedad [Grimm et al., 2003, Sardana et al., 2017].

4.3. *NGSmethDB20*

NGSmethDB es una base de datos de metilación de citosinas individuales provenientes del tratamiento del ADN con bisulfito, publicada por primera vez en el año 2011 [Hackenberg et al., 2011a]. A lo largo de los años se han realizado distintas actualizaciones [Geisen et al., 2014, Lebrón et al., 2017], las dos últimas de ellas durante la realización de la presente Tesis Doctoral, *NGSmethDB 2017* [Lebrón et al., 2017] y *NGSmethDB20*; esta última en periodo de desarrollo.

En esta memoria se presentan los resultados preliminares de la nueva

actualización *NGSmethDB20* que incluye cambios significativos a distintos niveles.

En primer lugar se implementado una nueva interfaz (Figura 23) con el objetivo de mejorar la usabilidad de la base de datos, mediante el uso del entorno de trabajo *Django*. Se intenta con esto que sea una base de datos interactiva, con distintos métodos de búsqueda y que incluya gráficos interactivos, haciéndola más útil para los potenciales usuarios. Todo ello se describe en más profundidad en los próximos apartados.

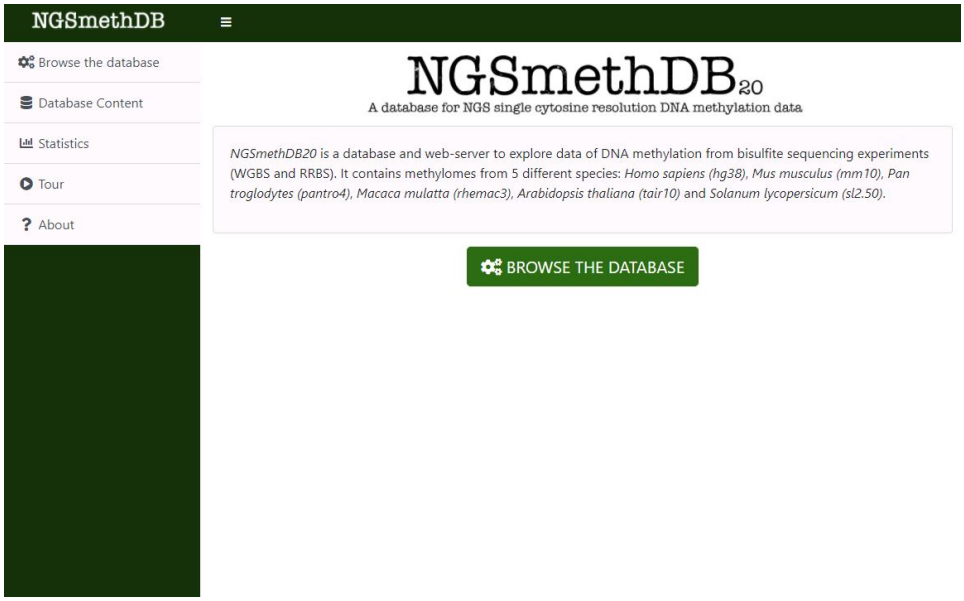


Figura 23: Captura de la página principal de la base de datos *NGSmethDB20* (`https://arn.ugr.es/NGSmethDB`).

Por otra parte se han añadido nuevos datos (incluidos los metilomas que se han usado en el estudio de asociación entre metilación y genotipo), ya descritos en el apartado 3.1. Para actualizar los datos de manera continua, se han desarrollado una serie de *scripts* (`https://github.com/cris12gm/mongoTools`) que permiten poblar la base de datos, y actualizar las tablas de contenidos (Apartado 4.1.2).

4.3.1. Nueva interfaz y base de datos *MongoDB*

La nueva interfaz (Figura 23) se está desarrollando mediante el uso del entorno de trabajo de desarrollo web *Django* [Django Software Foundation, 2020]. Basado en código Python, *Django* permite usar todos los paquetes de los que este dispone ofreciendo una gran versatilidad. Se basa en el patrón de diseño *MVC* (Modelo-Vista-Controlador), de forma que actúa al mismo tiempo como *back-end* y *front-end* y nos permite la conexión con la base de datos *MongoDB*.

En el caso del *front-end* se ha desarrollado conjuntamente con el uso de *Javascript*, *CSS* y *Bootstrap* para mejorar las posibilidades de diseño.

En cuanto a la base de datos utilizada, y como ya se comentó en el apartado 3.4 se sigue usando de *MongoDB*, como en la versión de *NGSmethDB* de 2017. La base de datos *MongoDB* se organiza de forma jerárquica siguiendo el esquema mostrado en la Figura 24. De esta forma, cada base de datos corresponde a un ensamblado distinto (por ejemplo, *hg38*). Dentro de las bases de datos encontramos colecciones para cada una de las anotaciones disponibles para ese ensamblado (genes, promotores, *enhancers*,...), y una colección para cada cromosoma.

En el caso de las colecciones correspondientes a los cromosomas, dentro de cada colección tenemos una serie de documentos, que corresponden con las posiciones del genoma. Cada uno de ellos contiene toda la información disponible sobre esa posición, como su metilación o el genotipo.

En el caso de las colecciones de elementos genómicos, de forma análoga, cada documento se corresponde con un elemento de esa anotación. Por ejemplo, en el caso de la anotación de genes, cada documento se correspondería con un gen, y dentro de él tendríamos toda la información acerca de este: coordenadas de inicio y de fin, descripción, etc.

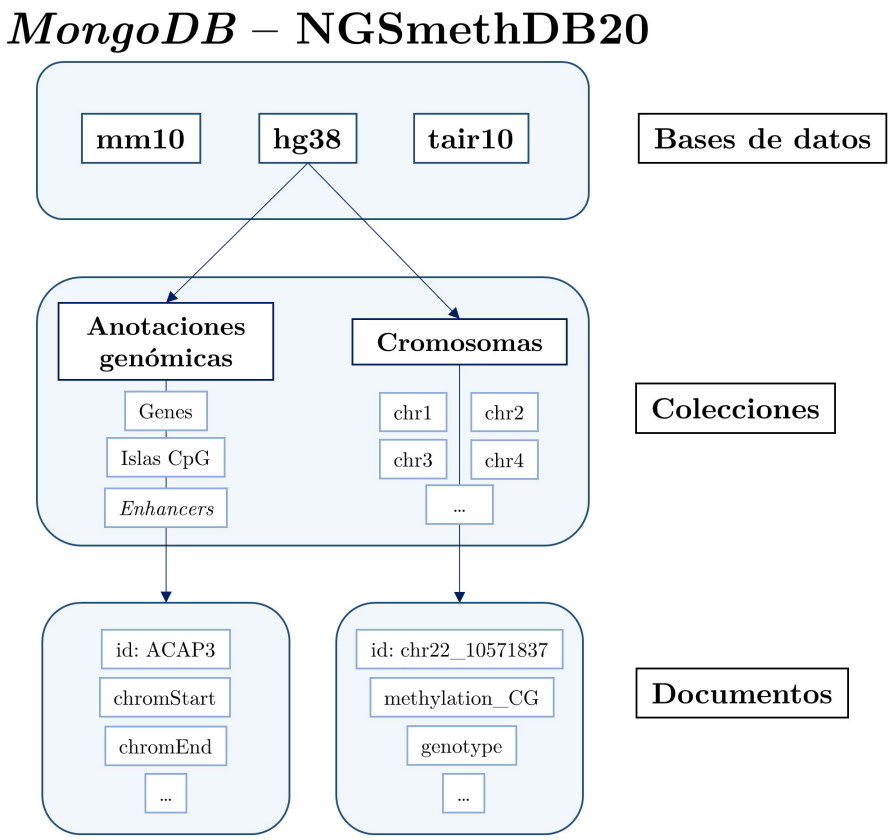
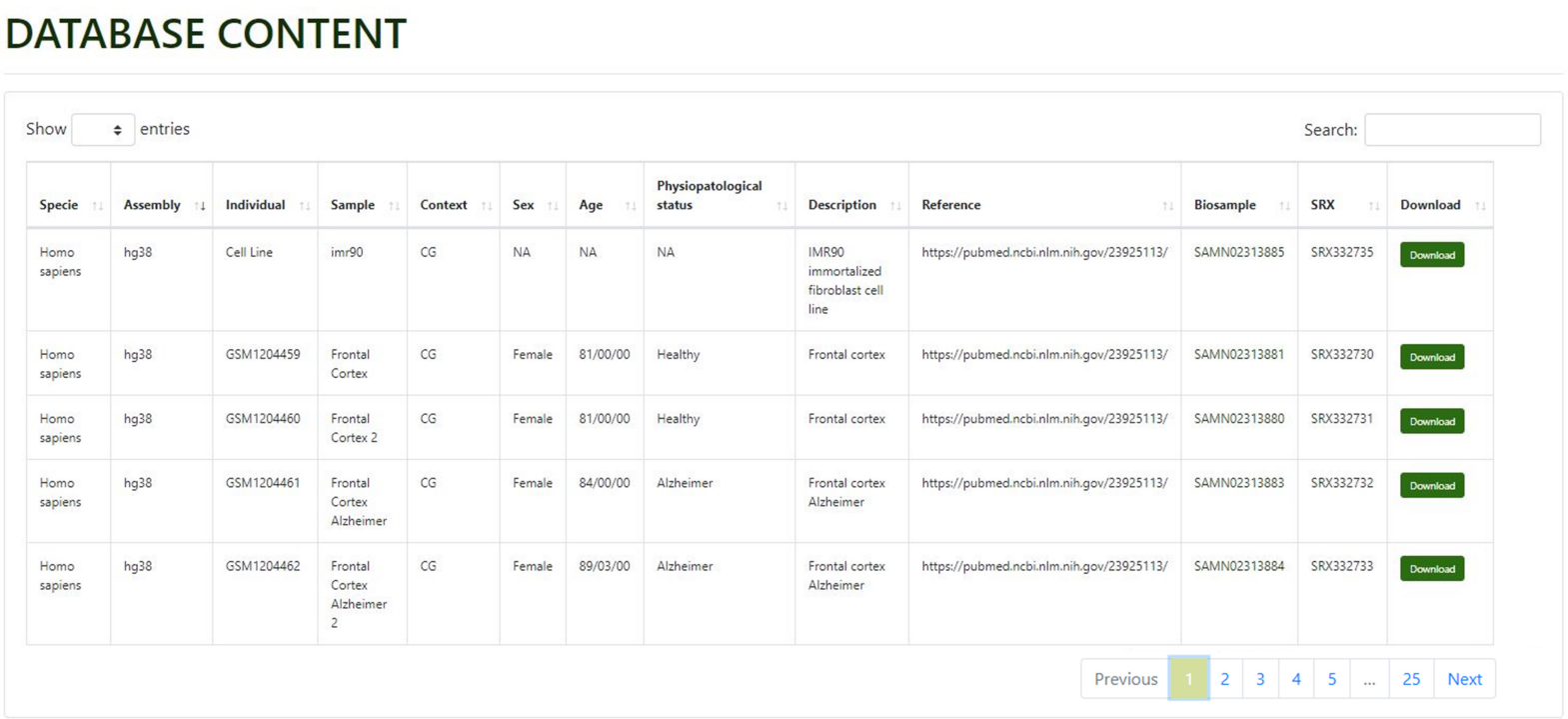


Figura 24: Esquema de la base de datos *MongoDB* implementada en *NGSmethDB20*.

4.3.2. Contenido de la base de datos

El contenido de la base de datos se puede encontrar en la sección *Database content* en forma de tabla dinámica (Figura 25). En esta, además de la información básica de cada una de las muestras, como el ensamblado, tejido del que proviene, o distintos identificadores como el *biosample*, se puede encontrar un botón de descarga del metiloma completo. En esta actualización se han añadido hasta el momento las muestras incluidas en el estudio de asociación (Tabla Suplementaria SI), y principalmente muestras provenientes de dos proyectos: i) *Cancer Cell Line Encyclopedia* (*CCLE*) [Li et al., 2019, Ghandi et al., 2019], compuesto por datos de metilación obtenidos por la técnica *RRBS* de 919 líneas celulares de 24 órganos distintos y 109 tipos de cáncer que se resumen en la Tabla Suplementaria SII; ii) El proyecto

PRJNA421218 que está compuesto de 55 muestras de *WGBS* provenientes de neuronas y oligodendrocitos de pacientes con esquizofrenia y pacientes control [Mendizabal et al., 2019]. Además de metilomas humanos, la base de datos dispone de metilomas de otras 5 especies: *Mus musculus*, *Pan troglodytes*, *Rhesus macaque*, *Solanum lycopersicum* y *Arabidopsis thaliana*.



DATABASE CONTENT

Show entries

Search:

Specie	Assembly	Individual	Sample	Context	Sex	Age	Physiopatological status	Description	Reference	Biosample	SRX	Download
Homo sapiens	hg38	Cell Line	imr90	CG	NA	NA	NA	IMR90 immortalized fibroblast cell line	https://pubmed.ncbi.nlm.nih.gov/23925113/	SAMN02313885	SRX332735	Download
Homo sapiens	hg38	GSM1204459	Frontal Cortex	CG	Female	81/00/00	Healthy	Frontal cortex	https://pubmed.ncbi.nlm.nih.gov/23925113/	SAMN02313881	SRX332730	Download
Homo sapiens	hg38	GSM1204460	Frontal Cortex 2	CG	Female	81/00/00	Healthy	Frontal cortex	https://pubmed.ncbi.nlm.nih.gov/23925113/	SAMN02313880	SRX332731	Download
Homo sapiens	hg38	GSM1204461	Frontal Cortex Alzheimer	CG	Female	84/00/00	Alzheimer	Frontal cortex Alzheimer	https://pubmed.ncbi.nlm.nih.gov/23925113/	SAMN02313883	SRX332732	Download
Homo sapiens	hg38	GSM1204462	Frontal Cortex Alzheimer 2	CG	Female	89/03/00	Alzheimer	Frontal cortex Alzheimer	https://pubmed.ncbi.nlm.nih.gov/23925113/	SAMN02313884	SRX332733	Download

Previous 1 2 3 4 5 ... 25 Next

Figura 25: Sección *Database content* de la base de datos *NGSmethDB20* (`https://arn.ugr.es/NGSmethDB/dbcontent`)

4.3.3. Métodos de búsqueda

La nueva base de datos implementará distintos métodos de búsqueda, entre los que por ahora se encuentran: Región e Islas CpG. En la Figura 26 se puede ver el formato preliminar del modo de búsqueda por región.

En todos ellos será posible seleccionar un ensamblado, la región genómica a analizar y a partir de ahí una o varias muestras de las que obtener los resultados. Una vez enviado el formulario con la consulta, la web sigue un sistema de colas, por lo que a cada usuario se le asigna un ID único. Este ID puede ser consultado durante 15 días de forma que el usuario pueda guardar en marcadores su consulta y volver a ella en este espacio de tiempo.

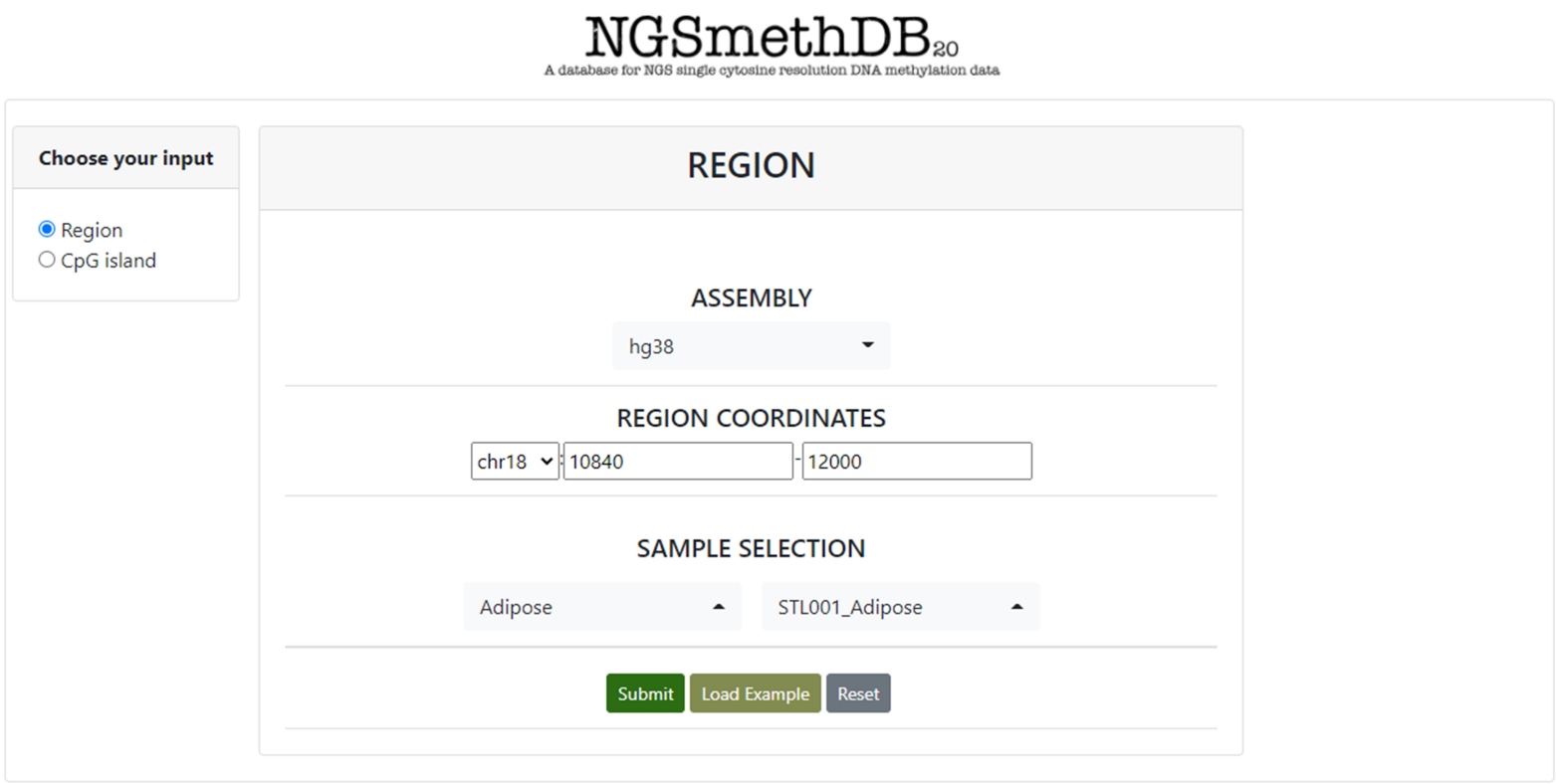


Figura 26: Método de consulta por región de *NGSmethDB20*. Imagen tomada de: `https://arn.ugr.es/NGSmethDB/browser`

En la Figura 27 se muestra una página de resultados para la consulta de la región *chr18:10840-12000*. Esta sección aún se encuentra en desarrollo y por el momento incluye la posibilidad de la descarga de los datos de metilación, su visualización en forma de tabla dinámica y un gráfico interactivo de la metilación de todos los CpGs que la componen.

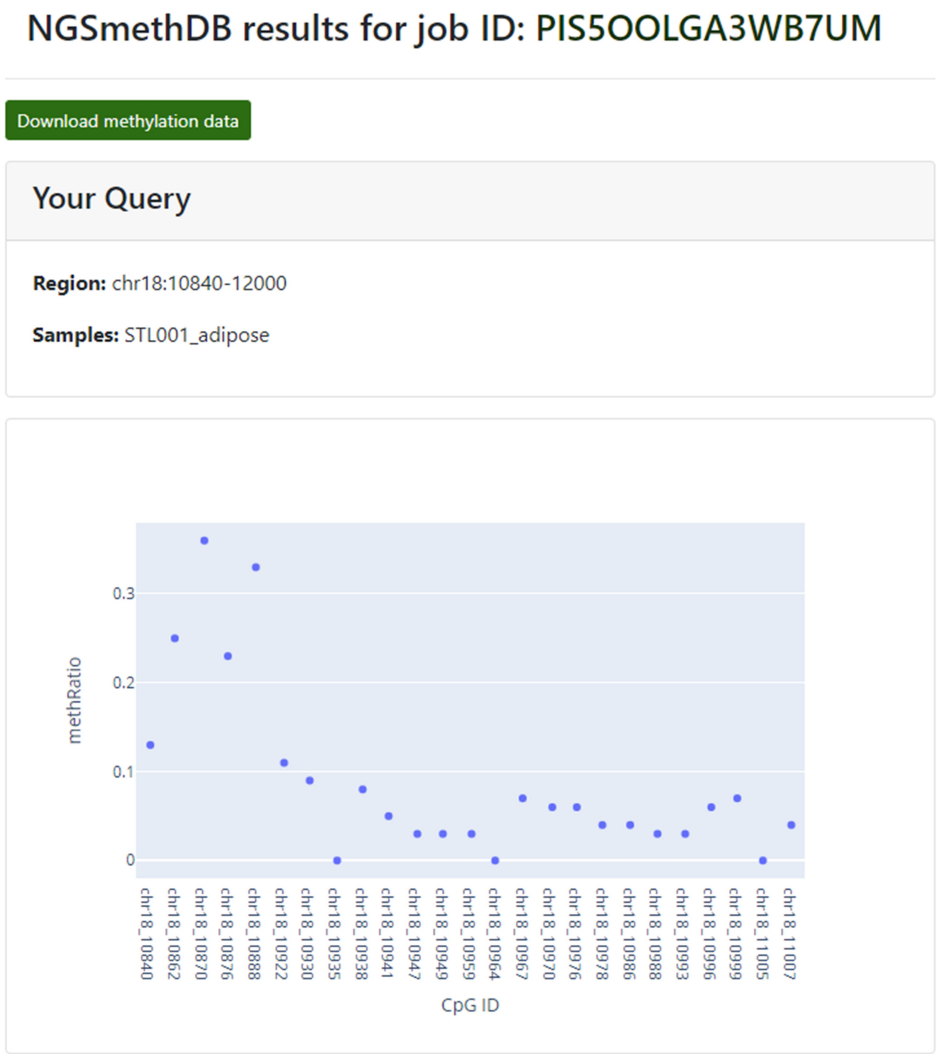


Figura 27: Ejemplo preliminar de la página de resultados de *NGSmethDB20*

4.4. *gCluster*

Durante esta Tesis Doctoral se desarrolló una mejora del método de predicción de islas mediante el software *gCluster* [Gómez-Martín et al., 2018], islas que luego han sido utilizadas como anotación genómica en la base de datos *NGSmethDB20*.

gCluster se basa en el mismo modelo de distancias utilizado por su predecesor *CpGcluster*, e incluye algunas mejoras en el modelo de distancias en el caso de *k-meros* solapantes y algunas funciones extra, como la posibilidad de obtener la distribución de distancias tanto observada como esperada (tanto global como por cromosoma) y el coeficiente de variación normalizado para cada cromosoma. Además incluye una serie de *helpers* o programas que ayudan a preparar la secuencia para su análisis por *gCluster* entre otras funciones. Todos los programas desarrollados se recogen en la Tabla 10.

Tabla 10: Software incluido con *gCluster*.

Programa	Descripción
prepareAssembly.py	Obtiene la secuencia genómica y a continuación la divide en las secuencias canónicas (secuencias de referencia de los cromosomas) y las secuencias alternativas (ensamblados alternativos, secuencias sin ensamblar, etc). Normalmente las predicciones se realizan solo sobre las secuencias canónicas. Funciona tanto con genomas proporcionados por el usuario como con URLs o descargando los genomas de UCSC.
makeSeqObj.jar	Indexa los ficheros *fasta* del ensamblado, generando un archivo assembly.zip que será la entrada del resto de software.
randomizer.jar	Aleatoriza secuencias de ADN preservando la frecuencia de dinucleótidos (aunque no de forma estricta) y siguiendo el siguiente proceso: (1) Se concatenan los *contigs*; (2) se aleatoriza la secuencia; (3) los bloques de Ns se introducen de nuevo en la posición original.

gCluster.jar	Determina los clústers locales dada una palabra de ADN (en el caso de islas CpG la palabra CG) y sus propiedades globales de clusterización. Funciona para ambas hebras para palabras no palindrómicas y acepta cualquier combinación de palabras de ADN (CAG:CTG:CCG para representar el contexto CHH).
GenomeCluster.pl	Determina los clúster locales de elementos genómicos identificados por sus coordenadas.

4.4.1. Modelo de distancias

La distancia al vecino más próximo entre *k-meros* o palabras de ADN que pueden solapar es más difícil de definir que en el caso de palabras no solapantes. En la Figura 28 se ilustra este problema en el caso de los contextos de metilación CCG y CWG. CCG puede solapar consigo mismo si una de las copias está localizada en la hebra directa y otra en la hebra inversa (en rojo), mientras que CAG y CTG (abreviado CWG) son palíndromos el uno del otro y por tanto no pueden solapar.

```
5'- TCCGGTGCTACAGCTG -3' word start end strand
    ||||||||||||||||      CCG   2    4    +
3'- AGGCCACGATGTCGAC -5'  CCG   3    5    -
```

Figura 28: Ejemplo de *k-meros* solapantes (en rojo) y no solapantes (en verde y azul)

Teniendo en cuenta esto, la distancia entre palabras solapantes no se puede calcular de la forma habitual, inicio (palabra aguas abajo) – final (palabra aguas arriba) porque podría resultar en distancias negativas. En el caso de CCG (ilustrado en rojo, Figura 28) la distancia siguiendo ese modelo sería: 3 - 4 = -1.

Por otra parte, si definimos la distancia como inicio–inicio tendríamos que la distancia mínima depende del grado de solapamiento entre las dos palabras. En el caso de la Figura 28 tendríamos:

- CCG (+) - CCG(-) : 3-2 = 1 (Mínima distancia posible)

- CAG(+) – CTG(+): 14-11 = 3 (Mínima distancia entre dos palabras de ADN no solapantes de longitud 3).

Por tanto definimos finalmente la distancia como:

$$d_{j,i} = (SC_j - SC_i) - nf_i \times d_{j,i} = (SC_j - SC_i) - nf_i$$

Siendo SC_j y SC_i las coordenadas de inicio de la palabra aguas abajo y aguas arriba, respectivamente, y nf_i el número de bases no solapantes forzosamente de la palabra aguas arriba. En el caso de CAGCTG que se exponía en la Figura 28, nf_j seria 2 ya que CAG no puede ser solapado por CTG.

4.4.2. Distribución de distancias

La clusterización a nivel local se manifiesta por un aumento de la densidad local del elemento clusterizado. Esta densidad es inversamente proporcional a la media de distancias entre los elementos que se clusterizan, es decir, cuanto mayor sea la densidad, menores serán las distancias entre los elementos. Es por ello que si existe la clusterización, las distancias cortas entre los elementos estarán sobrerrepresentadas si las comparamos con una distribución aleatoria.

gCluster permite obtener la distribución de distancias tanto en genoma completo como por cromosoma. Sin embargo, dado que la estructura composicional es la misma en todos los cromosomas [Grantham et al., 1980, Bernardi, 1993], se usa por defecto la distribución genómica para la obtención de la distancia umbral. Además incluye la herramienta "*randomizer.jar*", que permite la aleatorización genómica preservando la frecuencia de dinucleótidos.

En la Figura 29 se muestra la distribución de distancias para cuatro genomas: *Homo sapiens*, *Caenorhabditis elegans*, *Solanum lycopersicum* y *Escherichia coli*. En el genoma humano y en el del tomate, las distancias cortas están claramente sobrerrepresentadas lo que es un indicador de la clusterización como se comentó anteriormente. En *C. elegans*, que no posee metilación

del ADN (al menos en contexto CpG) observamos una ligera sobrerrepresentación de las distancias cortas mientras que en *E. coli* no se muestra ninguna sobrerrepresentación a distancias cortas. La distribución geométrica ajusta perfectamente con la distribución observada en secuencias aleatorias, lo que prueba que el modelo de aleatorización es el adecuado. La intersección entre la curva de valores observados y la de valores esperados puede considerarse como la distancia umbral que separa los CpGs clusterizados de aquellos que no lo están.

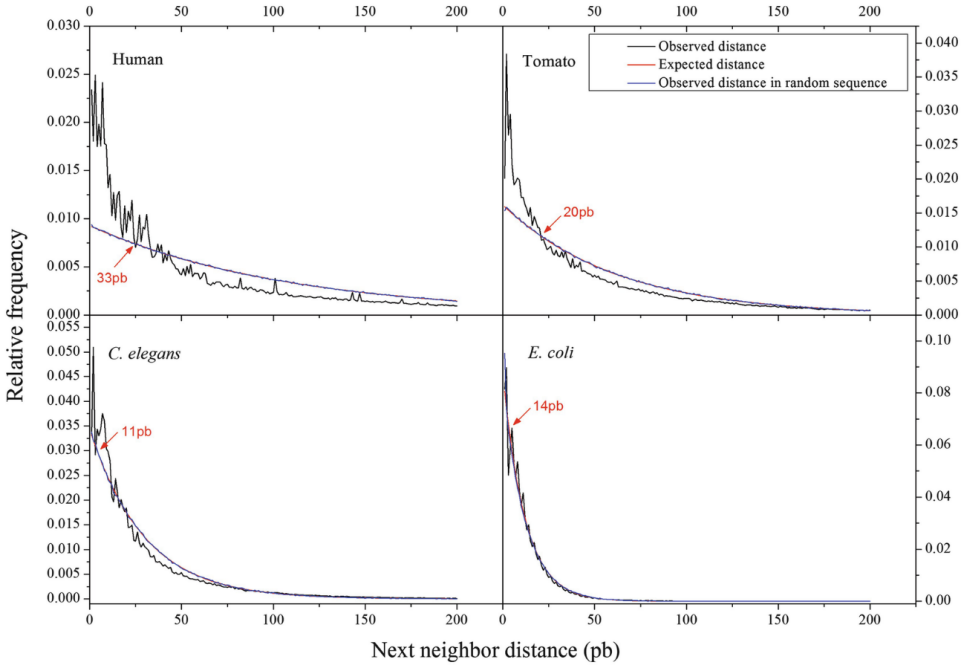


Figura 29: Distribución de distancias en cuatro especies. En cada plot se representan las distancias observadas (negro), las distancias esperadas (rojo) y las distancias observadas tras la aleatorización de la secuencia genómica (azul). La flecha indica la intersección genómica. Tomado de [Gómez-Martín et al., 2018].

4.4.3. Clusterización global de las palabras de ADN

La clusterización global de cualquier entidad en un espacio unidimensional se puede calcular por el coeficiente de variación de las distancias normalizadas [Bernaola-Galván et al., 2012, Hackenberg et al., 2012]. Este se calcula de la siguiente forma: (1) Para cada entidad, se calcula la distancia al vecino más cercano; (2) La distribución de distancias se normaliza a 1 dividiendo

cada distancia por la distancia media; (3) El coeficiente de variación (CV) se calcula como la desviación estándar dividida por la media; (4) Se corrige el CV para ciertos sesgos en las entidades con mucha frecuencia.

Dependiendo del valor que tenga el $CVcor$ (Coeficiente de variación corregido) nos indica la clusterización que existe:

- $CVcor = 1$: Los elementos se encuentran distribuidos aleatoriamente (siguiendo la distribución geométrica).
- $CVcor > 1$: Los elementos están clusterizados.
- $CVcor < 1$: Los elementos muestran repulsión, es decir, las distancias tienden a ser equidistantes entre ellos.

gCluster calcula el $CVcor$ aplicando el modelo de distancias que se ha explicado anteriormente en todos los contigs de un cromosoma. Análogamente a la distribución de distancias, este CV se puede calcular tanto en genoma completo como por cromosoma, aunque se usa el primero por defecto.

En la Figura 30A se muestra la clusterización global ($CVcor$) de seis especies frente a la de sus ensamblados aleatorizados, y en la Figura 30B las correspondientes proporciones O/E. Se puede observar como el dinucleótido CG está clusterizado en todos los genomas eucariotas analizados ($CVcor$ ¿1) incluido *C. elegans* que no muestra metilación de ADN. Además, la clusterización de los CpGs no está relacionada con la proporción O/E, ya que por ejemplo *C. elegans* muestra clusterización (Media $CVcor = 1{,}33$) pero una proporción O/E de 1 (el número de CpGs observados y esperados es el mismo), mientras que cuando aleatorizamos el genoma humano, la proporción O/E sigue siendo de 0,2, ya que durante el proceso se conserva la frecuencia de dinucleótidos, pero los coeficientes de $CVcor$ son virtualmente 1, lo que indica una distribución aleatoria. Incluso para secuencias de ADN pequeñas como es el caso del grupo de bacterias observamos solo una pequeña fluctuación del $CVcor$ en torno a 1 en las secuencias aleatorizadas, en contraste con la distribución sin aleatorizar.

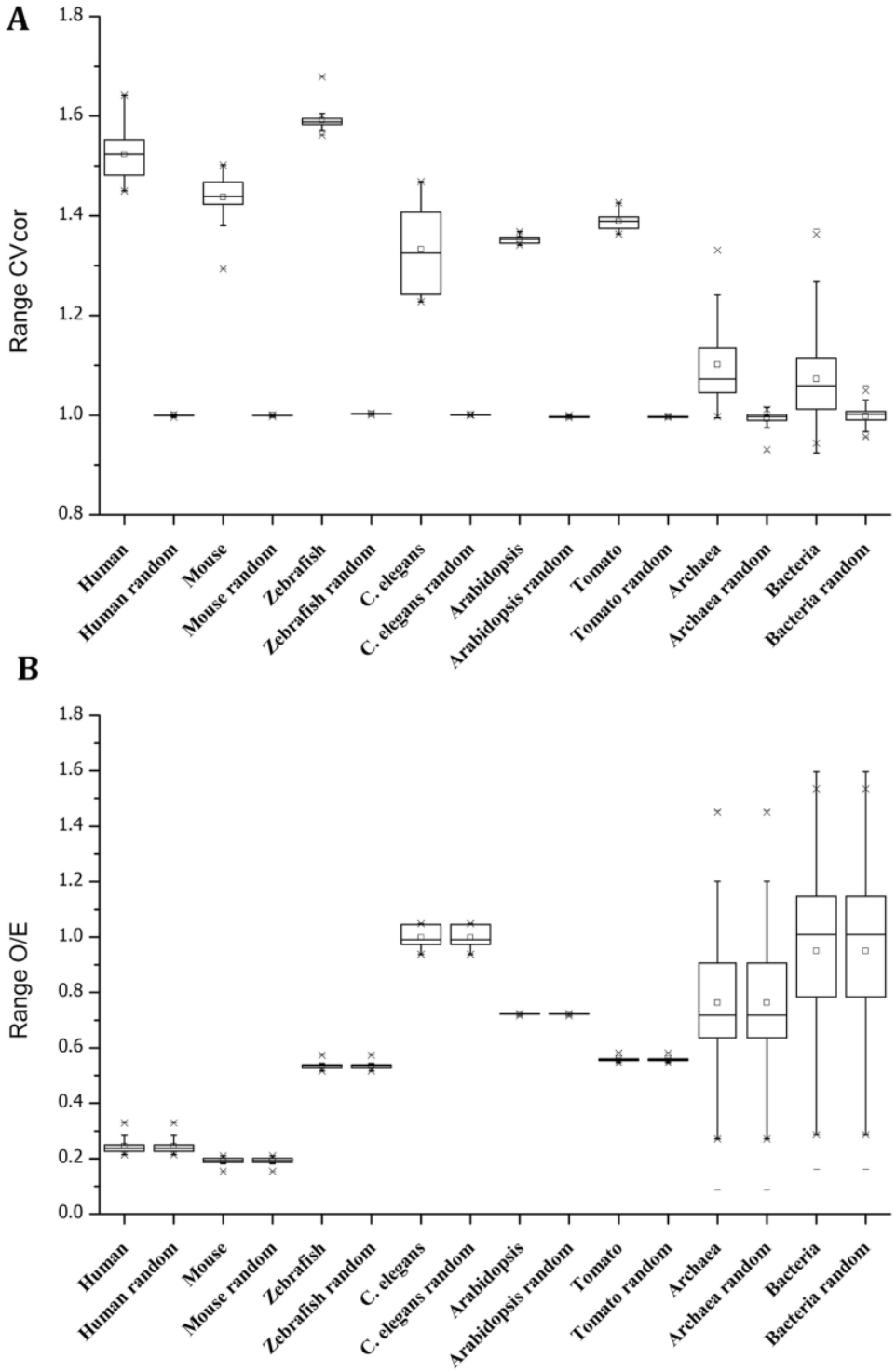


Figura 30: A) Clusterización global (*CVcor*) para seis especies y sus ensamblados aleatorizados (humano, ratón, pez zebra, *C. elegans*, Arabidopsis, tomate y las colecciones de 163 archaea y la de 124 bacterias. B) Proporciones O/E (frecuencia observada/esperada de dinucleótidos CpG). Ambas medidas se han calculado por cromosoma y el plot muestra su distribución en los distintos cromosomas mediante un *boxplot*. Tomado de [Gómez-Martín et al., 2018].

Capítulo 5

Discusión

La variación genética, la metilación del ADN y la expresión génica interactúan de una forma compleja y bidireccional. En los últimos años, y gracias al desarrollo de los estudios de asociación en genoma completo, se ha profundizado en algunos aspectos concretos de esta interacción mediante los llamados *eQTLs* y *meQTLs*. Estos se centran en la interacción entre variación genética y expresión génica (*eQTLs*) y en la interacción entre variación genética y metilación (*meQLTs*). Además suelen estar focalizados en un fenotipo concreto, realizando usualmente la comparación de pacientes de alguna enfermedad compleja contra controles sanos, y vinculando los resultados de estudios *GWAS* anteriores con la expresión génica o la metilación respectivamente. Indudablemente, estos estudios han aumentado el conocimiento sobre la interacción entre variación genética y expresión génica, por ejemplo analizando diferencias de expresión entre diferentes poblaciones [Nica and Dermitzakis, 2013], pero tienen ciertas desventajas:

- Son muy caros ya que requieren tamaños muestrales muy grandes y son necesarios cientos de individuos para cada estudio.
- Frecuentemente no puede estudiarse el tejido más relevante en el fenotipo. Un ejemplo claro de esto sería el caso del Alzheimer.
- No es posible obtener resultados generales acerca de la interacción en un loci dado ya que se suele emplear solamente un tipo de tejido por estudio.

Para solventar algunos de estos problemas y complementar los estudios *eQTL*/*meQTL*, en esta Tesis se planteó usar datos de secuenciación masiva ya disponibles en los repositorios públicos para analizar el impacto de la variación genética sobre la metilación y la expresión génica de una forma genérica y que no estuviese vinculada a un tipo de tejido o fenotipo. Para ello se han analizado 58 muestras de secuenciación masiva (*WGBS*) detectando 51.585 SNPs (1,3 %) asociados con al menos un CpG ($FDR \leq 0{,}05$).

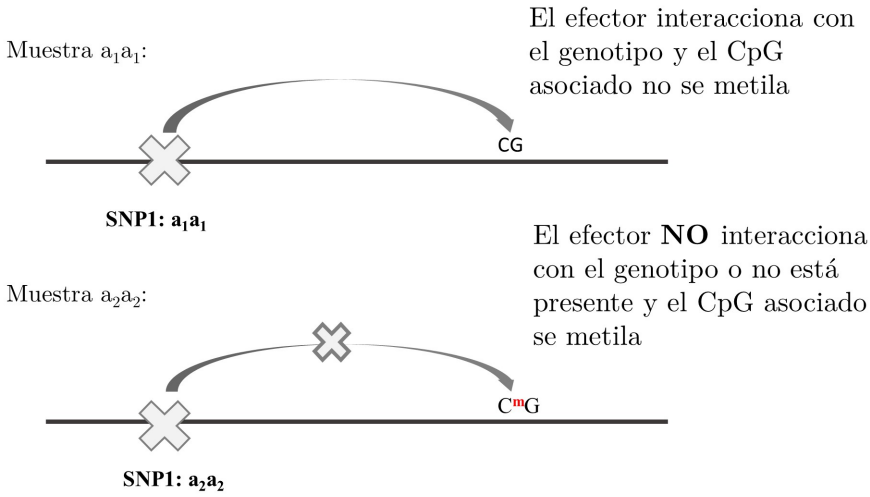
Una de las aplicaciones más prometedoras de esta aproximación es que permite la búsqueda de enlaces funcionales de *SNPs* asociados a un fenotipo mediante *GWAS* cuyo vínculo funcional no se conozca hasta el momento. Frecuentemente estos *SNPs* no solapan con ninguna región reguladora o codificante lo que dificulta detectar posibles mecanismos moleculares que vinculen el *SNP* con el fenotipo a nivel funcional. Sin embargo, a través de la interacción entre variación genética y metilación del ADN, ésta puede influir en la expresión génica de forma muy distal.

Para poder obtener datos de asociación estadística entre variación genética y metilación, se requiere procesar y analizar una gran cantidad de datos de diferentes tipos. Es por ello que, dentro del marco de esta Tesis Doctoral, se han desarrollado y puesto a disposición de la comunidad científica los siguientes recursos generados: *MethylExtract2*, *NGSmethDB*, *gCluster* y *geno*^{5}*mC*.

No obstante, la base de datos *geno*^{5}*mC* es sin lugar a dudas la herramienta en la que confluyen todos los datos, resultados y herramientas desarrollados. Esta base de datos permite explorar las asociaciones *SNP-CpG* obtenidas en el estudio de asociación así como su co-localización con otros elementos genómicos, permitiendo la obtención de conocimiento biológico. Esta utilidad y su concepto subyacente se han ilustrado mediante algunos ejemplos concretos, vinculando *SNPs* asociados a un fenotipo con genes relevantes.

Al igual que los estudios *eQTL* (*SNP* $\rightarrow$ expresión génica), la aproximación seguida en esta tesis doctoral (*SNP* $\rightarrow$ metilación del ADN $\rightarrow$ expresión génica) revela asociaciones estadísticamente significativas, pero no los mecanismos moleculares subyacentes. Parece razonable suponer que algún tipo de molécula tiene que interactuar con el ADN en la posición del *SNP* pa-

ra que los genotipos de este puedan influir en los estados de metilación o la expresión génica. Algunos factores de transcripción cambian su nivel de afinidad por algunos sitios de unión dependiendo del genotipo que haya en ellos (*SNPs*), por lo que estos podrían ser los mejores candidatos, aunque, a modo general se puede hablar de un efector que tiene que estar presente en la célula dada. En caso contrario el genotipo del *SNP* no podría afectar a la metilación del ADN (o la expresión génica). La Figura 31 muestra este modelo e ilustra las consecuencias que supone el hecho de usar muestras de diferentes tejidos o tipos celulares.



Efector presente en todas las muestras

	No metilado	Metilado
a_1a_1	N1	0
a_2a_2	0	N2

Efector NO presente en NINGUNA muestra

	No metilado	Metilado
a_1a_1	0	N1
a_2a_2	0	N2

Efector presente en ALGUNAS muestras

	No metilado	Metilado
a_1a_1	a	b
a_2a_2	0	N2

a = número de muestras con genotipo a_1a_1 que expresan el efector
b = número de muestras con genotipo a_2a_2 que no expresan el efector

Figura 31: Modelo de interacción entre el efector y los *SNPs* asociados.

Al existir la posibilidad de que el efector esté presente en algunas muestras y ausente en otras, en aquellos casos en que el efector no esté presente pero sí que exista una asociación, el valor-p será más alto y por lo tanto la asociación existente probablemente no será detectada (véase figura 31, tabla de contingencia). Este hecho no aumenta el número de falsos positivos, sino los falsos negativos ya que se trataría de asociaciones existentes no detectadas. A primera vista, eso parece una clara desventaja de la aproximación seguida en esta Tesis comparada con los estudios *meQTL* que generalmente emplean solo un tipo de célula o tejido. Sin embargo, empleando un único tipo de tejido solamente se podrán desvelar las asociaciones mediadas por los efectores presentes en él. Utilizando la aproximación multi-tejido, si el subconjunto de muestras en las que el efector no está presente no es demasiado grande, sí que podremos detectar la asociación.

Además, el modelo y la aproximación elegida en esta Tesis abren la puerta para mejorar el cálculo del valor-p eliminando las muestras en las que el efector se encuentre ausente. Un ejemplo de ello sería realizar de nuevo el cálculo del valor-p para aquellas asociaciones en las que todas las muestras en las que no se observa un impacto del genotipo sobre el estado de metilación derivan de un mismo tipo celular. En la misma línea se puede intentar detectar, a través de datos de expresión, el posible gen del efector o motivo en el ADN que solapa con los SNP asociados que sean compartidos entre varios *loci*. De esta forma se podrían detectar tanto los genes como sus sitios de unión aumentando así el conocimiento sobre los vínculos moleculares entre variación genética y expresión génica; algo que parece solamente posible con un diseño multi-tejido.

Capítulo 6

Conclusiones

1. En esta Tesis Doctoral se presenta un modelo genérico para el análisis de la relación entre metilación del ADN, variación genética y expresión génica. Para ello se han desarrollado: 1) un flujo de datos automatizado para la obtención de mapas de metilación; 2) un protocolo de asociación estadística entre metilación del ADN y variación genética; 3) una base de datos *geno*^{5}*mC* para almacenar los datos de asociación; 4) una base de datos *NGSmethDB* para almacenar los metilomas obtenidos; 5) un nuevo software de predicción de clústers de elementos genómicos *gCluster*.

2. El protocolo automatizado *MethylExtract2* integra todas las etapas del proceso de obtención de mapas de metilación de muestras a genoma completo (*WGBS* y *RRBS*), permitiendo la obtención simultanea (en la misma muestra) de los niveles de metilación y la variación genética. Incluye distintos controles de calidad asi como la posibilidad de personalizar los parámetros según la necesidad del usuario. Además, su implementación en Docker lo convierte en un software portable e independiente del sistema operativo en que se ejecute.

3. La base de datos *NGSmethDB* almacena y pone a disposición de la comunidad científica los metilomas de alta resolución procesados por *MethylExtract2*. En su última versión, *NGSmethDB20*, dispone de una nueva e intuitiva interfaz que permite distintos métodos de búsqueda, ofreciendo también gráficos interactivos para facilitar el análisis de los

metilomas.

4. *gCluster*, un nuevo software basado en el modelo de distancias de *CpG-cluster*, presenta una mejora de dicho modelo en el caso de *k-meros* solapantes. Además implementa funciones extra como la posibilidad de obtener la distribución de distancias observada y esperada o el Coeficiente de Variación corregido.

5. Tras su comparación con otros métodos, el modelo basado en el Test Exacto de Fisher ha resultado ser el más adecuado para evaluar la asociación estadística entre metilación y variación genética, ya que se basa en los tres estados biológicos de la metilación: M (Metilado), I (Metilación alelo específica) y U (No metilado).

6. Mediante el protocolo de asociación estadística entre metilación del ADN y variación genética hemos encontrado 51.585 *SNPs* (1,3 %) asociados con al menos un CpG ($FDR \leq 0{,}05$), usando datos de *WGBS* de 58 individuos distintos.

7. Tras la co-localización genómica de los pares *SNP-CpG* asociados se encontró una gran cantidad de *SNPs* (37.192 (72,10 %)) con CpGs asociados que también son *TL-CpGs*, lo que indica la presencia de vínculos, mediados por metilación, entre la variación genética y la expresión génica.

8. Una de las aplicaciones más prometedoras de la aproximación que se presenta en esta Tesis es que permite la búsqueda de enlaces funcionales de *SNPs* asociados a un fenotipo mediante *GWAS* cuyo vínculo funcional no se conozca hasta el momento. Esta utilidad se ha implementado en la base de datos $geno^5mC$, una aplicación web interactiva, provista de diversos métodos de búsqueda jerarquizada.

9. La utilidad de $geno^5mC$ se muestra mediante varios ejemplos como es el *SNP rs727563* que se había asociado previamente a la enfermedad inflamatoria intestinal mediante *GWAS*, pero se desconocía la conexión molecular con la enfermedad.

Capítulo 7

Perspectivas de futuro

Existen numerosas mejoras y experimentos que podrían ayudar en el futuro a mejorar el análisis de asociación llevado a cabo en esta Tesis Doctoral, algunos de los cuales se detallan a continuación.

En primer lugar, sería necesario realizar experimentos de confirmación de las asociaciones encontradas, con el objetivo de determinar la reproducibilidad de nuestro método. Esto debería realizarse utilizando un conjunto independiente, y la intersección entre el conjunto actual de asociaciones y el nuevo debe ser significativamente más alta de lo esperado por azar. Por otra parte podría ser interesante confirmar las asociaciones mediante el análisis de los heterocigotos, evaluando si existe una relación entre este genotipo y la metilación parcial.

En cuanto a la base de datos $geno^5mC$, se deberán actualizar los datos de asociación que contiene tras el nuevo cálculo de las asociaciones *SNP-CpG* con los datos de validación. Además, para algunas muestras se dispone de los valores de expresión, por lo que sería interesante poder mostrar los valores de expresión génica en las mismas muestras en las que se ha calculado la asociación en la base de datos. Por otra parte se podrían calcular los semáforos CpG (*TL-CpGs*) en estas mismas muestras e incorporar los resultados de co-localización con estos.

Como se ha demostrado con los ejemplos descritos en anteriores apartados, existen multitud de *SNPs* que se encuentran asociados a un fenotipo pero que aún carecen de un vínculo funcional entre ellos. Una mejora impor-

tante de nuestro método sería desarrollar un protocolo automatizado para encontrar vínculos funcionales a estas asociaciones basándonos en el modelo descrito en el Capítulo 6. Experimentos futuros podrían encaminarse en esta dirección, intentando encontrar de una forma automatizada si la presencia/ausencia de determinado factor de transcripción está afectando a las asociaciones *SNP-CpG* con el objetivo de encontrar un vínculo funcional para estas.

Material suplementario

Tabla SI: Muestras utilizadas en el estudio de asociación entre variación genética y metilación del ADN.

Tejido/Línea Celular	SRX	BioSample
Adiposo	SRX388732	SAMN02028525
Adiposo	SRX388741	SAMN02028527
Adiposo	SRX190155	SAMN00847536
Hígado	SRX213280	SAMN01881963
cd34	SRX142784	SAMN00113441
cd14	SRX323152	SAMN02252540
h1 mesenchymal	SRX116604	SAMN00774088
IPS DF 6.9	SRX056691	SAMN00255370
h9	SRX056693	SAMN00255371
hues64	SRX259086	SAMN01997279
h1 BMP4 Mesendoderm	SRX101207	SAMN00739304
Endoderm hESC	SRX142783	SAMN00857854
h1 BMP4	SRX027688	SAMN00114947
h1	SRX006241	SAMN00004462
h1 neuronal	SRX043405	SAMN00215949
IPS DF 19.11	SRX056688	SAMN00255367
imr90	SRX006785	SAMN00004463
Placenta	SRX323156	SAMN02252539
Fetal Adrenal Gland	SRX312981	SAMN02211822
Fetal Muscle	SRX312976	SAMN02211820
Fetal Large Intestine	SRX248145	SAMN00254261
Fetal Muscle	SRX323153	SAMN02252544

Frontal Cortex	SRX332731	SAMN02313880
Frontal Cortex	SRX332730	SAMN02313881
hepg2	SRX332734	SAMN02313882
Colon Tumor	SRX332736	SAMN02313886
Colon	SRX332737	SAMN02313887
Frontal Cortex Alzheimer	SRX332732	SAMN02313883
Frontal Cortex Alzheimer	SRX332733	SAMN02313884
Prefrontal Cortex	SRX140476	SAMN00854881
Prefrontal Cortex	SRX140477	SAMN00854882
Prefrontal Cortex	SRX140478	SAMN00854883
h9	SRX015763	SAMN00007497
	SRX015765	
h9 fibroblast	SRX015766	SAMN00007498
	SRX015768	
Foreskin fibroblast	SRX015769	SAMN00007499
	SRX015772	
gm12878	SRX186543	SAMN01173960
h1 hESC	SRX186542	SAMN01173959
Sperm	SRX081759	SAMN00632013
	SRX081760	
Sperm	SRX081761	SAMN00632014
	SRX081762	
Colon	SRX1631736	SAMN04558104
Skin fibroblast	SRX448680	SAMN02595800
Skin fibroblast	SRX448681	SAMN02595798
Skin fibroblast	SRX448682	SAMN02595801
Skin fibroblast	SRX448683	SAMN02595799
Brain	SRX306583	SAMN02204644
Brain	SRX306584	SAMN02204642
Brain	SRX306585	SAMN02204643
Brain	SRX308340	SAMN02206255
Brain	SRX309596	SAMN02206500
Brain	SRX309597	SAMN02206499

Brain cortex	SRX314937	SAMN02213489
Brain DPC Nn	SRX314938	SAMN02213482
Brain DPC	SRX314940	SAMN02213487
hues6	SRX314943	SAMN02213481
tCell	SRX091574	SAMN00709114
Mononuclear	SRX111392	SAMN00765484
tCell	SRX091573	SAMN00709113
Prefrontal cortex	SRX275878	SAMN02141269

Tabla SII: Resumen de las muestras analizadas y añadidas a la base de datos NGSmethDB20 del proyecto *Cancer Cell Line Encyclopedia* (*CCLE*) [Li et al., 2019, Ghandi et al., 2019].

Órgano	Tipo de cáncer	Número de muestras
Endometrio	Carcinoma mixto adenoescamoso	1
	Adenocarcinoma	21
	Carcinosarcoma maligno, tumor mesodérmico mixto	3
	Carcinoma de células claras	1
	No especificado	2
Esófago	Adenocarcinoma	1
	Carcinoma de células escamosas	22
	Metaplasía	1
	No especificado	2
Endometrio	Carcinoma mixto adenoescamoso	1
	Adenocarcinoma	21
	Carcinosarcoma maligno, tumor mesodérmico mixto	3
	Carcinoma de células claras	1
	No especificado	2
Ganglios autonómicos	Neuroblastoma	16
Glándulas salivares	Carcinoma mucoepidermoide	2
Hígado	Adenocarcinoma	1

	Carcinoma hepatocelular	22
	Hepatoblastoma	2
Hueso	Condrosarcoma	1
	Sarcoma Ewings, tumor periférico primitivo neuroectodérmico	7
	Tumor células gigantes	2
	Osteosarcoma	6
Intestino Delgado	Adenocarcinoma	1
Intestino Grueso	Adenocarcinoma	42
	No especificado	15
Ovario	Tumor cordones sexuales-estromal	1
	Adenocarcinoma	16
	Tumor Brenner	1
	Carcinoma de células claras	7
	Carcinoma endometrioide	2
	Carcinoma mixto	1
	Carcinoma mucinoso	1
	Carcinoma seroso	3
	Carcinoma indiferenciado	2

Índice de figuras

Índice de tablas

.

Bibliografía

[Abascal et al., 2020] Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Aken, B., Akiyama, J. A., Jammal, O. A., Amrhein, H., Anderson, S. M., Andrews, G. R., Antoshechkin, I., Ardlie, K. G., Armstrong, J., Astley, M., Banerjee, et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710.

[Adams et al., 2012] Adams, D., Altucci, L., Antonarakis, S. E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., Dahl, F., Dermitzakis, E. T., Enver, et al. (2012). BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotechnology*, 30(3):224–226.

[Andrews, 2018] Andrews, S. (2018). FastQC: A quality control tool for high throughput sequence data.

[Barturen et al., 2013] Barturen, G., Rueda, A., Oliver, J. L., and Hackenberg, M. (2013). MethylExtract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000Research*, 2(1):217.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

[Bernaola-Galván et al., 2012] Bernaola-Galván, P., Oliver, J. L., Hackenberg, M., Coronado, A. V., Ch. Ivanov, P., and Carpena, P. (2012). Segmentation of time series with long-range fractal correlations. *European Physical Journal B*, 85(6).

[Bernardi, 1993] Bernardi, G. (1993). Genome organization and species formation in vertebrates. *Journal of Molecular Evolution*, 37(4):331–337.

[Bird, 2002] Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes and Development*, 16(1):6–21.

[Bird, 2008] Bird, T. D. (2008). Genetic aspects of Alzheimer disease. *Genetics in Medicine*, 10(4):231–239.

[Broad Institute, 2019] Broad Institute (2019). Picard toolkit. \url{http://broadinstitute.github.io/picard/}.

[Brzozowski et al., 2016] Brzozowski, B., Mazur-Bialy, A., Pajdo, R., Kwiecien, S., Bilski, J., Zwolinska-Wcislo, M., Mach, T., and Brzozowski, T. (2016). Mechanisms by which Stress Affects the Experimental and Clinical Inflammatory Bowel Disease (IBD): Role of Brain-Gut Axis. *Current Neuropharmacology*, 14(8):892–900.

[Clark et al., 2006] Clark, S. J., Statham, A., Stirzaker, C., Molloy, P. L., and Frommer, M. (2006). DNA methylation: Bisulphite modification and analysis. *Nature Protocols*, 1(5):2353–2364.

[Corradin and Scacheri, 2014] Corradin, O. and Scacheri, P. C. (2014). Enhancer variants: Evaluating functions in common disease. *Genome Medicine*, 6(10):85.

[De Mendoza et al., 2018] De Mendoza, A., Bonnet, A., Vargas-Landin, D. B., Ji, N., Hong, F., Yang, F., Li, L., Hori, K., Pflueger, J., Buckberry, S., Ohta, H., Rosic, N., Lesage, P., Lin, S., and Lister, R. (2018). Recurrent acquisition of cytosine methyltransferases into eukaryotic retrotransposons. *Nature Communications*, 9(1):1–11.

[Deaton and Bird, 2011] Deaton, A. M. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes and Development*, 25(10):1010–1022.

[Depristo et al., 2011] Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M.,

Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–501.

[Dios et al., 2014] Dios, F., Barturen, G., Lebrón, R., Rueda, A., Hackenberg, M., and Oliver, J. L. (2014). DNA clustering and genome complexity. *Computational Biology and Chemistry*, 53(PA):71–78.

[Django Software Foundation, 2020] Django Software Foundation (2020). Django.

[Dreos et al., 2017] Dreos, R., Ambrosini, G., Groux, R., Perier, R. C., and Bucher, P. (2017). The eukaryotic promoter database in its 30th year: Focus on non-vertebrate organisms. *Nucleic Acids Research*, 45(D1):D51–D55.

[Duncan and Miller, 1980] Duncan, B. K. and Miller, J. H. (1980). Mutagenic deamination of cytosine residues in DNA. *Nature*, 287(5782):560–561.

[Feingold et al., 2004] Feingold, E. A., Good, P. J., Guyer, M. S., Kamholz, S., Liefer, L., Wetterstrand, K., Collins, F. S., Gingeras, T. R., Kampa, D., Sekinger, E. A., Cheng, J., Hirsch, H., Ghosh, S., Zhu, Z., Patel, S., Piccolboni, A., Yang, A., Tammana, H., Bekiranov, S., Kapranov, et al. (2004). The ENCODE (ENCyclopedia of DNA Elements) Project. *Science*, 306(5696):636–640.

[Fisher, 1992] Fisher, R. A. (1992). Statistical Methods for Research Workers. In *Statistical Methods for Research Workers*, pages 66–70. Springer, New York, NY.

[Fisher et al., 2018] Fisher, V. A., Wang, L., Deng, X., Sarnowski, C., Cupples, L. A., and Liu, C. T. (2018). Do changes in DNA methylation mediate or interact with SNP variation? A pharmacoepigenetic analysis 06 Biological Sciences 0604 Genetics. *BMC Genetics*, 19(Suppl 1).

[Fishilevich et al., 2017] Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., Lancet, D., and Cohen, D. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database*, 2017(1).

[Frommer et al., 1992] Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5- methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, 89(5):1827–1831.

[Gardiner-Garden and Frommer, 1987] Gardiner-Garden, M. and Frommer, M. (1987). CpG Islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2):261–282.

[Geisen et al., 2014] Geisen, S., Barturen, G., Alganza, Á. M., Hackenberg, M., and Oliver, J. L. (2014). NGSmethDB: An updated genome resource for high quality, single-cytosine resolution methylomes. *Nucleic Acids Research*, 42(D1):D53–D59.

[Ghandi et al., 2019] Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., Barretina, J., Gelfand, E. T., Bielski, C. M., Li, H., Hu, K., Andreev-Drakhlin, A. Y., Kim, J., Hess, J. M., Haas, B. J., Aguet, F., Weir, B. A., Rothberg, M. V., Paolella, B. R., Lawrence, et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, 569(7757):503–508.

[Goddard et al., 2016] Goddard, M. E., Kemper, K. E., MacLeod, I. M., Chamberlain, A. J., and Hayes, B. J. (2016). Genetics of complex traits: Prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proceedings of the Royal Society B: Biological Sciences*, 283(1835).

[Gómez-Martín et al., 2020] Gómez-Martín, C., Aparicio-Puerta, E., Medina, J., Barturen, G., Oliver, J., and Hackenberg, M. (2020). geno5mC: a database to explore the association between genetic variation (SNPs) and CpG methylation in the human genome. *Journal of Molecular Biology*.

[Gómez-Martín et al., 2018] Gómez-Martín, C., Lebrón, R., Oliver, J. L., and Hackenberg, M. (2018). Prediction of CpG Islands as an intrinsic clustering property found in many Eukaryotic DNA sequences and its relation to DNA methylation. In *Methods in Molecular Biology*, volume 1766, pages 31–47. Humana Press Inc.

[Grantham et al., 1980] Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pavé, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, 8(1):197.

[Grimm et al., 2003] Grimm, C. H., Rogner, U. C., and Avner, P. (2003). Lrmp and Bcat1 are candidates for the type I diabetes susceptibility locus Idd6. *Autoimmunity*, 36(4):241–246.

[Hackenberg et al., 2011a] Hackenberg, M., Barturen, G., and Oliver, J. L. (2011a). NGSmethDB: A database for next-generation sequencing single-cytosine- resolution DNA methylation data. *Nucleic Acids Research*, 39(SUPPL. 1):D75–D79.

[Hackenberg et al., 2011b] Hackenberg, M., Carpena, P., Bernaola-galván, P., Barturen, G., Alganza, Á. M., and Oliver, J. L. (2011b). WordCluster : detecting clusters of DNA words and genomic elements. *Algorithms for Molecular Biology*, 6(1):2.

[Hackenberg et al., 2006] Hackenberg, M., Previti, C., Luque-escamilla, P. L., Carpena, P., Martínez-aroza, J., and Oliver, J. L. (2006). CpG-cluster : a distance-based algorithm for CpG-island detection. *BMC bioinformatics*, 13:1–13.

[Hackenberg et al., 2012] Hackenberg, M., Rueda, A., Carpena, P., Bernaola-Galván, P., Barturen, G., and Oliver, J. L. (2012). Clustering of DNA words and biological function: A proof of principle. *Journal of Theoretical Biology*, 297:127–136.

[Hon et al., 2013] Hon, G. C., Rajagopal, N., Shen, Y., McCleary, D. F., Yue, F., Dang, M. D., and Ren, B. (2013). Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nature Genetics*, 45(10):1198–1206.

[Illingworth et al., 2010] Illingworth, R. S., Gruenewald-Schneider, U., Webb, S., Kerr, A. R. W., and James, K. D. (2010). Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. *PLoS Genet*, 6(9):1001134.

[Jones, 2012] Jones, P. A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492.

[Koboldt et al., 2012] Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576.

[Krueger, 2019] Krueger, F. (2019). 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 libraries.

[Krueger and Andrews, 2011] Krueger, F. and Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572.

[Laird, 2010] Laird, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3):191–203.

[Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.

[Laurent et al., 2010] Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., Low, H. M., Sung, K. W. K., Rigoutsos, I., Loring, J., and Wei, C. L. (2010). Dynamic changes in the human methylome during differentiation. *Genome Research*, 20(3):320–331.

[Lebrón et al., 2017] Lebrón, R., Gómez-Martín, C., Carpena, P., Bernaola-Galván, P., Barturen, G., Hackenberg, M., and Oliver, J. L. (2017). NGSmethDB 2017: Enhanced methylomes and differential methylation. *Nucleic Acids Research*, 45(D1):D97–D103.

[Lee, 2013] Lee, C. Y. (2013). Chronic restraint stress induces intestinal inflammation and alters the expression of hexose and lipid transporters. *Clinical and Experimental Pharmacology and Physiology*, 40(6):385–391.

[Leinonen et al., 2011] Leinonen, R., Sugawara, H., and Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*, 39(SUPPL. 1):D19.

[Lev Maor et al., 2015] Lev Maor, G., Yearim, A., and Ast, G. (2015). The alternative role of DNA methylation in splicing regulation. *Trends in Genetics*, 31(5):274–280.

[Li et al., 1993] Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature*, 366(6453):362–365.

[Li et al., 2019] Li, H., Ning, S., Ghandi, M., Kryukov, G. V., Gopal, S., Deik, A., Souza, A., Pierce, K., Keskula, P., Hernandez, D., Ann, J., Shkoza, D., Apfel, V., Zou, Y., Vazquez, F., Barretina, J., Pagliarini, R. A., Galli, G. G., Root, D. E., Hahn, W. C., Tsherniak, A., Giannakis, M., Schreiber, S. L., Clish, C. B., Garraway, L. A., and Sellers, W. R. (2019). The landscape of cancer cell line metabolism. *Nature Medicine*, 25(5):850–860.

[Lin et al., 2013] Lin, X., Sun, D., Rodriguez, B., Zhao, Q., Sun, H., Zhang, Y., Li, W., and Bishop, M. (2013). BSeQC: Quality control of bisulfite sequencing experiments. *Bioinformatics*, 29(24):3227–3229.

[Lioznova et al., 2019] Lioznova, A. V., Khamis, A. M., Artemov, A. V., Besedina, E., Ramensky, V., Bajic, V. B., Kulakovskiy, I. V., and Medvedeva, Y. A. (2019). CpG traffic lights are markers of regulatory regions in human genome. *BMC Genomics*, 20(1):102.

[Lister et al., 2009] 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., and Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322.

[Liu et al., 2015] Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., Ripke, S., Lee, J. C., Jostins, L., Shah, T., Abedian, S., Cheon, J. H., Cho, J., Dayani, N. E., Franke, L., Fuyuno, Y., Hart, A., Juyal, R. C., Juyal, G., Kim, et al. (2015). Association analyses

identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics*, 47(9):979–986.

[Liu et al., 2012] Liu, Y., Siegmund, K. D., Laird, P. W., and Berman, B. P. (2012). Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biology*, 13(7).

[Maunakea et al., 2013] Maunakea, A. K., Chepelev, I., Cui, K., and Zhao, K. (2013). Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Research*, 23(11):1256–1269.

[McKenna et al., 2010] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

[McRae et al., 2018] McRae, A. F., Marioni, R. E., Shah, S., Yang, J., Powell, J. E., Harris, S. E., Gibson, J., Henders, A. K., Bowdler, L., Painter, J. N., Murphy, L., Martin, N. G., Starr, J. M., Wray, N. R., Deary, I. J., Visscher, P. M., and Montgomery, G. W. (2018). Identification of 55,000 Replicated DNA Methylation QTL. *Scientific Reports*, 8(1):1–9.

[Medvedeva et al., 2014] Medvedeva, Y. A., Khamis, A. M., Kulakovskiy, I. V., Ba-Alawi, W., Bhuyan, M. S. I., Kawaji, H., Lassmann, T., Harbers, M., Forrest, A. R. R., and Bajic, V. B. (2014). Effects of cytosine methylation on transcription factor binding sites. *BMC genomics*, 15(1):119.

[Meissner et al., 2005] Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18):5868–5877.

[Mendizabal et al., 2019] Mendizabal, I., Berto, S., Usui, N., Toriumi, K., Chatterjee, P., Douglas, C., Huh, I., Jeong, H., Layman, T., Tamminga, C. A., Preuss, T. M., Konopka, G., and Yi, S. V. (2019). Cell type-specific epigenetic links to schizophrenia risk in the brain. *Genome Biology*, 20(1).

[Merkel et al., 2019] Merkel, A., Fernández-Callejo, M., Casals, E., Marco-Sola, S., Schuyler, R., Gut, I. G., and Heath, S. C. (2019). GemBS: High throughput processing for DNA methylation data from bisulfite sequencing. *Bioinformatics*, 35(5):737–742.

[Merkel, 2014] Merkel, D. (2014). Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux Journal*, 2014(239):2.

[Metodiev et al., 2014] Metodiev, M. D., Gerber, S., Hubert, L., Delahodde, A., Chretien, D., Gérard, X., Amati-Bonneau, P., Giacomotto, M. C., Boddaert, N., Kaminska, A., Desguerre, I., Amiel, J., Rio, M., Kaplan, J., Munnich, A., Rötig, A., Rozet, J. M., and Besmond, C. (2014). Mutations in the tricarboxylic acid cycle enzyme, aconitase 2, cause either isolated or syndromic optic neuropathy with encephalopathy and cerebellar atrophy. *Journal of Medical Genetics*, 51(12):834–838.

[Michael Bayer, 2012] Michael Bayer (2012). SQLAlchemy. In Wilson, A. B. and Greg, editors, *The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks*, page 20.

[NCBI, 2017] NCBI (2017). SRA Toolkit.

[Nica and Dermitzakis, 2013] Nica, A. C. and Dermitzakis, E. T. (2013). Expression quantitative trait loci: Present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620).

[Okada et al., 2014] Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., Graham, R. R., Manoharan, A., Ortmann, W., Bhangale, T., Denny, J. C., Carroll, R. J., Eyler, A. E., Greenberg, J. D., Kremer, J. M., Pappas, et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381.

[Peters et al., 2017] Peters, L. A., Perrigoue, J., Mortha, A., Iuga, A., Song, W. M., Neiman, E. M., Llewellyn, S. R., Di Narzo, A., Kidd, B. A., Telesco, S. E., Zhao, Y., Stojmirovic, A., Sendecki, J., Shameer, K., Miotto, R., Losic, B., Shah, H., Lee, E., Wang, M., Faith, J. J., Kasarskis, A., Brodmerkel, C., Curran, M., Das, A., Friedman, J. R., Fukui, Y., Humphrey, M. B., Iritani, B. M., Sibinga, N., Tarrant, T. K., Argmann, C., Hao, K.,

Roussos, P., Zhu, J., Zhang, B., Dobrin, R., Mayer, L. F., and Schadt, E. E. (2017). A functional genomics predictive network model identifies regulators of inflammatory bowel disease. *Nature Genetics*, 49(10):1437–1449.

[Pierce et al., 2018] Pierce, B. L., Tong, L., Argos, M., Demanelis, K., Jasmine, F., Rakibuz-Zaman, M., Sarwar, G., Islam, M. T., Shahriar, H., Islam, T., Rahman, M., Yunus, M., Kibriya, M. G., Chen, L. S., and Ahsan, H. (2018). Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. *Nature Communications*, 9(1):1–12.

[Quinlan and Hall, 2010] Quinlan, A. R. and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

[Ramos et al., 2014] Ramos, E. M., Hoffman, D., Junkins, H. A., Maglott, D., Phan, L., Sherry, S. T., Feolo, M., and Hindorff, L. A. (2014). Phenotype-genotype integrator (PheGenI): Synthesizing genome-wide association study (GWAS) data with existing genomic resources. *European Journal of Human Genetics*, 22(1):144–147.

[Roadmap Epigenomics Consortium et al., 2015] Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–329.

[Sardana et al., 2017] Sardana, M., Vasim, I., Varakantam, S., Kewan, U., Tariq, A., Koppula, M. R., Syed, A. A., Beraun, M., Drummen, N. E., Vermeer, C., Akers, S. R., and Chirinos, J. A. (2017). Inactive matrix gla-protein and arterial stiffness in type 2 diabetes mellitus. *American Journal of Hypertension*, 30(2):196–201.

[Shabalin, 2012] Shabalin, A. A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358.

[Sharp et al., 2011] Sharp, A. J., Stathaki, E., Migliavacca, E., Brahmachary, M., Montgomery, S. B., Dupre, Y., and Antonarakis, S. E. (2011). DNA methylation profiles of human active and inactive X chromosomes. *Genome Research*, 21(10):1592–1600.

[Sherry et al., 2001] Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). DbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311.

[Shukla et al., 2011] Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, 479(7371):74–82.

[Singer et al., 2015] Singer, M., Kosti, I., Pachter, L., and Mandel-Gutfreund, Y. (2015). A diverse epigenetic landscape at human exons with implication for expression. *Nucleic Acids Research*, 43(7):3498–3508.

[Spiegel et al., 2012] Spiegel, R., Pines, O., Ta-Shma, A., Burak, E., Shaag, A., Halvardson, J., Edvardson, S., Mahajna, M., Zenvirt, S., Saada, A., Shalev, S., Feuk, L., and Elpeleg, O. (2012). Infantile cerebellar-retinal degeneration associated with a mutation in mitochondrial aconitase, ACO2. *American journal of human genetics*, 90(3):518–23.

[Strunz et al., 2018] Strunz, T., Grassmann, F., Gayán, J., Nahkuri, S., Souza-Costa, D., Maugeais, C., Fauser, S., Nogoceke, E., and Weber, B. H. (2018). A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. *Scientific Reports*, 8(1):1–11.

[Suzuki and Bird, 2008] Suzuki, M. M. and Bird, A. (2008). DNA methylation landscapes: Provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6):465–476.

[Tak and Farnham, 2015] Tak, Y. G. and Farnham, P. J. (2015). Making sense of GWAS: Using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics and Chromatin*, 8(1).

[Takai and Jones, 2002] Takai, D. and Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6):3740–5.

[Tam et al., 2019] Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484.

[Urich et al., 2015] Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J., and Ecker, J. R. (2015). MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nature Protocols*, 10(3):475–483.

[Wang et al., 2012] Wang, H., Maurano, M. T., Qu, H., Varley, K. E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., Thurman, R. E., Kaul, R., Myers, R. M., and Stamatoyannopoulos, J. A. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Research*, 22(9):1680–1688.

[Wang et al., 2020] Wang, X., Gessier, F., Perozzo, R., Stojkov, D., Hosseini, A., Amirshahrokhi, K., Kuchen, S., Yousefi, S., Lötscher, P., and Simon, H.-U. (2020). RIPK3–MLKL–Mediated Neutrophil Death Requires Concurrent Activation of Fibroblast Activation Protein-α. *The Journal of Immunology*, 205(6):1653–1663.

[Wu et al., 2016a] Wu, S., Powers, S., Zhu, W., and Hannun, Y. A. (2016a). Substantial contribution of extrinsic risk factors to cancer development. *Nature*, 529(7584):43–47.

[Wu et al., 2016b] Wu, T. P., Wang, T., Seetin, M. G., Lai, Y., Zhu, S., Lin, K., Liu, Y., Byrum, S. D., Mackintosh, S. G., Zhong, M., Tackett, A., Wang, G., Hon, L. S., Fang, G., Swenberg, J. A., and Xiao, A. Z. (2016b). DNA methylation on N 6-adenine in mammalian embryonic stem cells HHS Public Access. *Nature*, 532(7599):329–333.

[Xi and Li, 2009] Xi, Y. and Li, W. (2009). BSMAP: Whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, 10.

[Yang et al., 2020] Yang, L., Chen, Z., Stout, E. S., Delerue, F., Ittner, L. M., Wilkins, M. R., Quinlan, K. G., and Crossley, M. (2020). Methylation of a CGATA element inhibits binding and regulation by GATA-1. *Nature Communications*, 11(1):1–10.

[Yearim et al., 2015] Yearim, A., Gelfman, S., Shayevitch, R., Melcer, S., Glaich, O., Mallm, J. P., Nissim-Rafinia, M., Cohen, A. H. S., Rippe, K., Meshorer, E., and Ast, G. (2015). HP1 Is Involved in Regulating the Global Impact of DNA Methylation on Alternative Splicing. *Cell Reports*, 10(7):1122–1134.

[Yoder et al., 1997] Yoder, J. A., Walsh, C. P., and Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, 13(8):335–340.

[Yu et al., 2013] Yu, D.-H., Ware, C., Waterland, R. A., Zhang, J., Chen, M.-H., Gadkari, M., Kunde-Ramamoorthy, G., Nosavanh, L. M., and Shen, L. (2013). Developmentally Programmed 3' CpG Island Methylation Confers Tissue- and Cell-Type-Specific Transcriptional Activation. *Molecular and Cellular Biology*, 33(9):1845–1858.

[Zhang et al., 2020] Zhang, X., Jeong, M., Huang, X., Wang, X. Q., Wang, X., Zhou, W., Shamim, M. S., Gore, H., Himadewi, P., Liu, Y., Bochkov, I. D., Reyes, J., Doty, M., Huang, Y. H., Jung, H., Heikamp, E., Aiden, A. P., Li, W., Su, J., Aiden, E. L., and Goodell, M. A. (2020). Large DNA Methylation Nadirs Anchor Chromatin Loops Maintaining Hematopoietic Stem Cell Identity. *Molecular Cell*, 78(3):506–521.e6.

[Ziller et al., 2013] Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., Gnirke, A., and Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463):477–481.