Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California

Maestría en Ciencias en Ciencias de la Computación

Predicción de regiones de cromatina abierta a partir del perfil de activación de la histona H3K27ac.

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de Maestro en Ciencias

Presenta:

César Miguel Valdez Córdova

Ensenada, Baja California, México

2020

Tesis defendida por

César Miguel Valdez Córdova

y aprobada por el siguiente Comité

Dr. Carlos Alberto Brizuela Rodríguez
Codirector de tesis

Dra. Rosario Ivetth Corona de la Fuente
Codirector de tesis

Dr. Hugo Homero Hidalgo Silva

Dr. Pierrick Gerard Jean Fournier

Dr. Víctor Manuel Treviño Alvarado

Dr. Israel Marck Martínez Pérez
Coordinador del Posgrado en Ciencias de la Computación

Dra. Rufina Hernández Martínez
Directora de Estudios de Posgrado

César Miguel Valdez Córdova © 2020

Queda prohibida la reproducción parcial o total de esta obra sin el permiso formal y explícito del autor y director de la tesis

Resumen de la tesis que presenta César Miguel Valdez Córdova como requisito parcial para la obtención del grado de Maestro en Ciencias en Ciencias de la Computación.

Predicción de regiones de cromatina abierta a partir del perfil de activación de la histona H3K27ac.

Resumen aprobado por:

Dr. Carlos Alberto Brizuela Rodríguez
Codirector de tesis

Dra. Rosario Ivetth Corona de la Fuente
Codirector de tesis

En el contexto de la epigenética, la determinación de la estructura de la cromatina asociada a las modificaciones histonales y sus dinámicas asociadas es crucial para perfilar con precisión los fenómenos relacionados con procesos biológicos complejos, tales como, la diferenciación celular, el proceso de desarrollo celular y la aparición y progresión de diversas enfermedades. La predicción de la estructura de la cromatina, y en particular, la identificación de los sitios accesibles de cromatina son problemas biológicos fundamentales que están siendo activamente investigados. En esta tesis, presentamos un clasificador binario, basado en una máquina de soporte vectorial (SVM), entrenado con datos disponibles de ChIP-sec y ATAC-sec derivados de experimentos relacionados con la modificación histonal H3K27ac. Nuestro modelo utiliza características construidas a partir de señales que se extraen directamente de los datos genómicos de acuerdo con una abstracción computacional basada en un fenomeno biológico, con el objetivo de determinar si es posible predecir sitios de ATAC-seq a partir de datos de ChIP-seq. Nuestro modelo con mejor desempeño puede identificar exitosamente el 82,84 % de los intervalos de ATAC-seq que coinciden con intervalos de ChIP-seq. Para la realización de las pruebas se desarrolló histoneSig, un paquete de R que está disponible en github para uso público. Si bien los resultados iniciales son prometedores, nuestro modelo aún necesita ser capaz de cuantificar y discernir el número de regiones asociadas y sus diferentes grados de accesibilidad. Idealmente, esto ayudará en el desarrollo de abstracciones biológicas cada vez más robustas, lo cual resultará en características más informativas y altamente generalizables.

Palabras clave: Epigenética, ChIP-seq, ATAC-seq, Accesibilidad de la Cromatina, SVM, clasificador binario

Abstract of the thesis presented by César Miguel Valdez Córdova as a partial requirement to obtain the Master of Science degree in Computer Science.

Prediction of open chromatin regions from H3K27ac ChIP-Seq profiles

Abstract approved by:

Dr. Carlos Alberto Brizuela Rodríguez
Thesis Co-Director

Dra. Rosario Ivetth Corona de la Fuente
Thesis Co-Director

In the context of epigenetics, determining chromatin structure associated to histonal modifications and its associated dynamics is crucial for accurately profiling phenomena related to complex biological processes, such as cellular differentiation, development and disease onset and progression. Prediction of chromatin structure, and in particular, the identification of accessible chromatin sites are fundamental, actively researched biological problems. In this thesis, we present a binary Support Vector Machine (SVM) classifier trained on openly available ChIP-seq and ATAC-seq data derived from experiments related to the H3K27ac histonal modification. Our model utilizes features constructed from signals that are extracted directly from genomics data in accordance with a computational abstraction that is based on a biological phenomenon, with the goal of determining if it is possible to predict ATAC-seq sites from ChIP-seq data. Our top performing model can successfully identify 82.84 %. of the ATAC-seq intervals that overlap with ChIP-seq intervals. To perform the various tests contained in this thesis, an R package, histoneSig, was developed. It is available on github for public use. While initial results are promising, our model still needs to be able to discern the number of associated regions and different degrees of accessibility. This will ideally aid in the development of increasingly robust biological abstractions, which will result in more informative, highly generalizable features.

Keywords: Epigenetics, ChIP-seq, ATAC-seq, Chromatin accessibility, SVM, binary classifier

Dedicatoria

A tí, apreciable lector.

Agradecimientos

Al Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE) y en particular al Departamento de Ciencias de la Computación por haberme bienvenido a su programa de Maestría y proporcionarme todos los recursos necesarios, desde personal capacitado hasta instalaciones, para completar el programa exitosamente.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme el apoyo económico para realizar mis estudios de maestría. No. de becario: 634041.

A mis directores de Tesis: el Dr. Carlos Alberto Brizuela Rodríguez y la Dra. Rosario Ivetth Corona Fuentes. Por su excelencia, tanto técnica como humana, consejo, tiempo, dedicación tanto al proyecto de tesis como a mi desarrollo como investigador y sobre todo, paciencia. Estoy seguro que cada momento invertido contribuyó de manera importante. Llevaré conmigo todo el aprendizaje, científico y personal, que adquirí durante este tiempo por el resto de mi vida profesional.

A los miembros de mi comité de tesis: el Dr. Hugo Homero Hidalgo Silva, el Dr. Pierrick Gerard Jean Fournier y el Dr. Víctor Manuel Treviño Alvarado. Estoy seguro que tener a un grupo de profesionales provenientes de disciplinas tan variadas discutiendo y haciendo aportaciones en conjunto nutrió mi trabajo enormemente. Recordaré mis avances de tesis como momentos sumamente enriquecedores. También, a este ultimo en particular, por su atención durante mi desarrollo como profesional desde el inicio de mi trayectoria; de no ser por su ánimo, instrucción y recomendación, el CICESE probablemente no habría formado parte de la misma.

A todos mis amigos. Por los momentos buenos, malos y todo en medio de estos. Gracias por el tiempo que me dieron. Sé que con ustedes, el mio fue invertido de la mejor manera posible.

A toda mi familia y en particular a mi madre; por esa característica fe ciega e incansable, que me han tenido a lo largo de mi vida. Por haberme apoyado desde el primer momento a manos llenas y sin condición alguna. Nada de esto sería un hecho de no haber sido por ustedes.

Tabla de contenido

Página

Tabla de contenido (continuación)

Lista de figuras

Figura Página

Lista de figuras (continuación)

Lista de tablas

Tabla Página

Lista de tablas (continuación)

Capítulo 1. Introducción

La investigación biológica en su concepción moderna es una disciplina cada vez más intensiva en datos (Navarro *et al.*, 2019). Los datos de secuenciación, en sus diversos formatos, se duplican aproximadamente cada siete meses. Si las tendencias actuales se mantienen constantes, para el año 2025 se convertirá en el principal generador de datos en escala masiva entre dominios y disciplinas (Stephens *et al.*, 2015). Los paradigmas de investigación actuales en esta área son tan variados como los diferentes fenómenos biológicos subyacentes que se estudian, experimentos asociados a los mismos, los datos y los formatos en los que se capturan. No existe todavía una manera consistente y sistemática de analizar los datos de todos estos experimentos, y las relaciones funcionales a menudo no se entienden a profundidad. Si bien no hay garantía de llegar a una estandarización completa, es de gran interés desarrollar métodos novedosos para aprovechar las inmensas cantidades de datos biológicos disponibles en la actualidad.

El aprendizaje automático se ha utilizado ampliamente para resolver problemas en todos los niveles físicos del genoma, desde el mononucleótido hasta el nivel de proteínas y tejidos. Los modelos de aprendizaje tienen como objetivo principal proporcionar representaciones informativas y de baja dimensión para datos de alta dimensión. Dichas representaciones se obtienen como características que se perfeccionan durante el proceso de aprendizaje. Es posible capturar y resumir la información a través de varios niveles de abstracción dentro del genoma. Debido a los intrincados mecanismos celulares que se manifiestan a través de los fenómenos biológicos, a menudo existen relaciones significativas entre estos niveles. La integración de la información pertinente en un modelo único y completo sigue siendo una tarea difícil. Las características provenientes de diferentes regiones del genoma pueden ser estáticas, descriptivas o dinámicas; pueden variar dentro de una misma célula o entre múltiples condiciones, individuos y organismos, o ambos.

1.1. Antecedentes y Motivación

1.1.1. Aplicaciones generales de aprendizaje de máquinas en el ámbito biológico.

Actualmente, los métodos de aprendizaje de máquina aplicados varían tanto en su alcance como en su implementación biológica y computacional. Incluso cuando se trata de resolver el mismo problema biológico, la integración de datos no es homogénea. De manera similar, los paradigmas computacionales tienden a variar incluso cuando un modelo utiliza datos similares. Métodos considerados “clásicos“ en la comunidad de aprendizaje de máquina, que generalmente consisten en clasificadores como: Los discriminantes cuadráticos y lineales, los vecinos más cercanos (k-NN), las máquinas de soporte vectorial, los árboles de decisión y modelos asociados, se han utilizado exhaustivamente a lo largo de la historia de la biología “moderna“ basada en datos; se pueden apreciar ejemplos notables desde las últimas décadas del siglo pasado (Tarca *et al.*, 2007). Sin embargo, recientemente han surgido paradigmas modernos basados en otras subdisciplinas del aprendizaje de máquina. A continuación se presenta una lista no exhaustiva de ejemplos notables recientes, en un intento de proporcionar una síntesis variada del contexto general del aprendizaje de máquina biológico.

El método de Keilwagen et al. predice factores de transcripción de manera específica por tipo celular. Este modela la distribución conjunta de una amplia gama de características numéricas mediante distribuciones discretas. Su principio de aprendizaje consiste en utilizar una variante ponderada del principio de máxima verosimilitud condicional. La selección y la ingeniería de las características se realiza a partir de diversos protocolos de secuenciación profunda. Las nuevas estadísticas se calculan a partir de dichas características, las cuales se utilizan para la segmentación y agrupamiento de los datos. Posteriormente, estos datos se someten a un algoritmo de aprendizaje de máquina supervisado relacionado con la regresión logística. Finalmente, dicho modelo combina las predicciones de los clasificadores entrenados obtenidas del entrenamiento por diferentes tipos de células e iteraciones en un enfoque de conjunto (Keilwagen *et al.*, 2019).

Las redes neurales profundas han visto la adopción generalizada de una multitud de tareas relacionadas con la predicción. En particular, se busca encontrar la respuesta

al problema de predicción de la estructura y la función de las proteínas y otras estructuras relacionadas que interactúan con las mismas. Un ejemplo concreto es el de la predicción de afinidad de acoplamiento de los factores de transcripción. Para este problema, Deepbind (Alipanahi *et al.*, 2015), incorpora una red neuronal convolucional en dos pasos. Primero, aplica un módulo de convolución para “aprender“ los motivos de secuencia por medio de una representación de Matrices de Peso Posicionales (PWM). Posteriormente, utiliza una red neuronal no lineal para fabricar motivos de predicción de forma combinatoria. Dicho modelo superó consistentemente a todos los modelos conocidos hasta ese momento. Un esfuerzo similar fue empleado por DeepSEA, que utilizó una mezcla de capas convolucionales, agrupadas y completamente conectadas (Zhou y Troyanskaya, 2015). En este caso, para estudiar los efectos funcionales que tendrían variantes de secuencias en regiones no codificantes.

Se han utilizado arquitecturas profundas similares en el caso de modelos como Deep NF (Gligorijević *et al.*, 2017), un modelo basado en un autocodificador profundo multimodal. A través de una combinación de redes de la base de datos STRING [1], construye representaciones comunes de baja dimensión que contienen características proteicas de alto nivel. Se recrean redes regulatorias enteras y luego se hacen predicciones a partir de dichas redes construidas. Las métricas de evaluación para este estudio contienen validación cruzada y retención temporal sobre el desempeño predictivo.

Los estudios de especificidad celular han sido exitosos con los enfoques de Aprendizaje Profundo. Tal es el caso de TFImpute, cuyo objetivo principal es la imputación de datos sobre conjuntos de datos incompletos, como suele ser el caso de la información de especificidad de TF (Qin y Feng, 2017). En este caso, se modela el problema de TF-Binding en un entorno de aprendizaje multitarea que toma información de factores de transcripción y líneas celulares; también, dicho modelo es capaz de predecir las actividades potenciadoras específicas por línea celular. FactorNet (Quang y Xie, 2017) funciona de manera similar.

Los modelos de aprendizaje a partir de paradigmas no tradicionales de aprendizaje automatizado se han implementado entre disciplinas de forma exitosa. BindSpace(Yuan *et al.*, 2019) aprende a integrar secuencias de ADN basadas en etiquetas de

[1] `https://string-db.org/`

Familia/Clase de diversos factores de transcripción. StarSpace[2], un algoritmo reciente de Procesamiento del Lenguaje Natural (NLP), fue aplicadco al espacio de características mencionado anteriormente. Dicho modelo aprende una representación de palabras en un espacio semántico; en este caso, de *k*-meros a secuencias de sondas de ADN. Posteriormente, dicho algoritmo maximiza la similitud entre un ejemplo de secuencia y sus capas correspondientes.

Los enfoques de aprendizaje de transferencia modelados bajo marcos bayesianos pueden ayudar a modelar la relación predictiva entre los factores de transcripción y la expresión génica en casos de alta degeneración de secuencia (Zou *et al.*, 2015). Su idea fundamental es reutilizar muestras/instancias de dominio antiguas como datos auxiliares para un nuevo dominio (es decir, general para conjuntos de datos de histonas específicos). El modelo de Zou et al. presenta un modelo de transferencia de parámetros para sistemas degenerados, que puede ser útil para ciertos casos de cáncer.

1.1.2. Modificaciones histonales y accesibilidad de cromatina

Las marcas epigenéticas, en particular las modificaciones histonales y sus procesos y estados asociados, como la accesibilidad de cromatina, producen conjuntos de datos de alta dimensión que pueden ayudar a esclarecer información importante para los mecanismos biológicos fundamentales y la progresión de la enfermedad. Sin embargo, a menudo se caracterizan por una compleja interacción entre los fenómenos biológicos de cada experimento, lo que ha demostrado ser un obstáculo para una interpretación eficaz. Un ejemplo notable es el estudio del cáncer; aunque durante mucho tiempo se consideró como una enfermedad exclusivamente genética a nivel secuencia de carácter hereditario, estudios recientes muestran que las alteraciones en los procesos epigenéticos son clave en la activación o inhibición de los genes relacionados con la progresión del cáncer (Kagohara *et al.*, 2017). Ejemplos de estudios epigenéticos relacionados con el cáncer incluyen la anotación funcional del estado de cromatina, la identificación de los factores genéticos que impulsan las marcas epigenéticas, diversos estudios de causalidad y predicción de la estructura de la cromatina, entre otros

[2] https://arxiv.org/abs/1709.03856

(Cazaly *et al.*, 2019). Las relaciones funcionales también pueden determinarse dados los patrones de accesibilidad de la cromatina, como se muestra para los patrones de unión de los receptores de glucocorticoides (John *et al.*, 2011).

Los métodos que tratan con el estado de las modificaciones histonales generalmente toman un enfoque no supervisado debido a la baja disponibilidad de conjuntos de datos verificados de referencia. Un ejemplo notable son los datos para sitios de unión de factores de transcripción (TFBS), que son sumamente heterogéneos entre diferentes marcas histonales. Las predicciones pueden entonces agruparse de manera similar a los métodos supervisados, etiquetando las secuencias asociadas como “Acoplado“ o “No Acoplado“. Otros métodos pueden utilizar pruebas estadísticas o puntuaciones de predicción de afinidad vinculantes (Keilwagen *et al.*, 2019). En el caso del acoplamiento de Factores de Transcripción, los motivos de sequencia son la representación habitual para extraer dicho conocimiento en un marco computacional (Park y Kellis, 2015). Para entrenar un modelo supervisado, se deben proporcionar secuencias objetivo etiquetadas como “‘Acoplado“ o “No Acoplado“ derivadas de los datos de ChIP-Seq. Los modelos habituales utilizados en la clasificación incluyen máquinas de soporte vectorial, regresión logística, bosques aleatorios o, recientemente, enfoques basados en aprendizaje profundo (Quang y Xie, 2017). En el caso de la cromatina abierta, se puede hacer un modelo análogo etiquetando las regiones como “inaccesibles“ o “accesibles“, tomando en cuenta grados variables de accesibilidad.

La determinación de accesibilidad a la cromatina se considera generalmente como un subproblema de la determinación de la estructura de la misma. Dado que se trata de un problema complejo por sí mismo, las características del modelo varían considerablemente; estos generalmente consisten en información recopilada de marcas histonales específicas, ya que cada una de ellas tiende a presentar características específicas de modificación (Xu *et al.*, 2018). Por ejemplo, los elementos reguladores activos, como la modificación histónica H3K27ac, tienden a estar ubicados en regiones accesibles y sin nucleosomas; por lo tanto, la accesibilidad a la cromatina puede ser indicativa de la actividad reguladora específica del tipo de célula (Thurman *et al.*, 2012). Los protocolos de accesibilidad experimental tradicionales, como ATAC-seq o DNAse-seq, pueden utilizarse para determinar dichas regiones. Estos métodos suelen basarse en la búsqueda de regiones enriquecidas en las muestras resultantes. Las

particularidades específicas de los métodos experimentales y del análisis de datos que deben tenerse en cuenta para una interpretación adecuada de los datos. Sin embargo, la determinación del enriquecimiento no es específica del fenómeno de accesibilidad a la cromatina.

El enfoque tradicional en los análisis de secuenciación para la identificación de densidades de lectura marcadamente enriquecidas, o “picos“, en las muestras correspondientes se denomina extracción de picos o “peak calling“ (Bardet *et al.*, 2011). El objetivo de dicho proceso es filtrar el ruido de fondo e identificar con precisión las localizaciones de los “picos“ en las regiones genómicas asociadas. Dichos picos suelen corresponder a los sitios de acoplamiento de factor de transcripción, aunque no siempre es el caso; las proteínas no específicas e indirectas que se unen al ADN pueden estar asociadas a los picos obtenidos (Valouev *et al.*, 2008). Se sabe que la forma de los picos varía entre diferentes tipos de proteínas. Sin embargo, es posible caracterizarlos en tres categorías generales, cada una con sus correspondientes asociaciones biológicas. Existen picos “agudos“, correspondientes a posiciones genómicas específicas. Picos “amplios“ que se asocian con grandes dominios genómicos y picos “mixtos“, que implican una combinación de ambos tipos de picos mencionados anteriormente. Los experimentos ChIP-Seq de modificación de histona se asocian generalmente con picos amplios (Nakato y Shirahige, 2016). En algunos casos, la integración de métodos de extracción de picos tanto agudos como amplios puede ser útil para la caracterización adecuada de los fenómenos biológicos subyacentes ((Young *et al.*, 2011), (Starmer y Magnuson, 2016)). Una vez que los picos son encontrados, deben ser probados para determinar su significancia estadística y finalmente, efectuar análisis de enriquecimiento sobre los picos determinados como significativos. Sin embargo, este proceso es susceptible a artefactos biológicos a lo largo de todas sus etapas. Adicionalmente, la sintonización de parámetros en la identificación de picos a menudo produce resultados diferentes, incluso cuando se trata del mismo conjunto de datos experimentales. Además, en el caso particular de las marcas de histonas, los sitios de unión conocidos varían enormemente entre modificaciones, tipos de células y condiciones, lo que añade ruido adicional entre muestras a los métodos que incorporan regiones “conocidas“ en sus análisis. Por ejemplo, se ha observado que los tipos de células con promotores activos dan lugar a “picos fantasma“ con carácter de falso positivo en experimentos de ChIP-seq (Jain *et al.*, 2015).

Idealmente, un modelo para la predicción de la accesibilidad de la cromatina no debe ser sensible a sesgos específicos de método o experimento. Se ha intentado mejorar los resultados de la identificación de picos para la accesibilidad de cromatina con clasificadores que han sido entrenados sobre marcas epigenómicas, expresión genética y modificaciones histonales: Epitensor captura las dependencias espaciales en la cromatina mediante la integración de datos de histonas, sitios hipersensibles a DNAsa 1 (DHS) y RNA-seq, que posteriormente se utilizan para la predicción de características estructurales (Zhu *et al.*, 2016). ChromAccPrediction utiliza un modelo de bosque aleatorio jerárquico entrenado en expresión genética y datos de accesibilidad de cromatina (Jung *et al.*, 2017). En este último, los datos de accesibilidad de cromatina se alinearon con el genoma conocido. El clasificador determinó entonces, de manera binaria, si la cromatina era accesible o no; si al menos una de las regiones de enriquecimiento obtenidas se superponía a un promotor conocido o a una región de código genético, se decía que estaba en cromatina accesible, inaccesible si no lo estaba.

La mayoría de los clasificadores actualmente implementados utilizan la expresión genética junto con los datos transcriptómicos asociados a la misma. Sin embargo, ninguno de ellos utiliza los datos de ChIP-seq exclusivamente para predecir la accesibilidad. Además, estos modelos de predicción no suelen ser específicos para un experimento en particular. Usualmente, las regiones de interés sólo se determinan como accesibles o no, a diferencia de efectuar la predicción de intervalos que pudieran resultar a partir de un experimento biológico en particular; en nuestro caso, ATAC-seq, experimento para el cual no se ha desarrollado un modelo de predicción específico. Cuando se trata de la detección de regiones de cromatina abierta, los picos obtenidos a partir de datos H3k27ac ChIP-seq se han considerado previamente como regiones “positivas“ o “abiertas“ (Kumar *et al.*, 2013a). En vista de lo anterior, hemos desarrollado un enfoque basado en la superposición.

Esta tesis tratará sobre la predicción de las regiones de cromatina abierta, tal y como se encuentran en un conjunto de datos ATAC-seq, a partir de datos específicos de ChIP-seq de la modificación histonal H3K27ac.

1.2. Objetivos

1.2.1. Objetivo general

Diseñar e implementar un encadenamiento de programas para predecir regiones de cromatina abierta específica de H3K27ac a partir de datos ChIP-Seq.

1.2.2. Objetivos específicos

- Establecer un conjunto de entrenamiento y prueba de datos de ChIP-seq y ATAC-seq para la predicción de regiones de cromatina abierta.
- Extraer y seleccionar características que conserven información relevante para la predicción.
- Determinar las características de señal más apropiadas para una correcta predicción de regiones de cromatina abierta.
- Desarrollar un paquete de R para facilitar el acceso y uso del modelo.

1.3. Organización de la Tesis

Esta tesis consta de 6 capítulos. El Capítulo 1 presenta una breve introducción para contextualizar al lector con el panorama actual del aprendizaje automático y el problema de predicción de cromatina abierta, así los objetivos de esta tesis. El Capítulo 2 tiene como objetivo proporcionar al lector un trasfondo general de conceptos relevantes para el desarrollo de esta tesis. Se divide en dos partes: la primera parte presenta conceptos biológicos, mientras que la segunda se centra en los computacionales. El Capítulo 3 delimita la metodología experimental, las herramientas asociadas, los modelos de aprendizaje y breves justificaciones de por qué se tomaron ciertas decisiones de diseño experimental en el desarrollo de esta tesis hasta la obtención de resultados relevantes para nuestros objetivos. El Capítulo 4 muestra los resultados obtenidos,

mientras que el Capítulo 5 los describe en detalle y proporciona al lector la perspectiva del autor con respecto a los hallazgos relevantes. El Capítulo 6 concluye la tesis, comenzando con un sumario de la misma, seguido de la presentación de los hallazgos más relevantes después de analizar los resultados de una manera concisa, finalizado con perspectivas adicionales de investigación.

Capítulo 2. Marco Teórico

2.1. Biológico

2.1.1. ADN y construcciones asociadas

Todos los organismos vivos actualmente documentados tienen como unidad fundamental las células, que están compuestas por los siguientes componentes: ADN, ARN, proteínas y moléculas pequeñas. Las células conservan toda la información esencial para la reproducción, funcionamiento y supervivencia en forma de ADN; su estructura consta de dos elementos complementarios de cuatro subunidades diferentes llamadas nucleótidos, que canónicamente forman pares de bases de Adenina-Timina (AT) y Citosina-Guanina (CG) (Chen, 2009). Las células de humano contienen grandes cantidades de ADN, aproximadamente 3.2×10^9 nucleótidos (Alberts, 2002). Sin embargo, no todo juega un papel activo en la célula; sólo alrededor de 1 % de las secuencias de nucleótidos en el genoma humano tienen un producto funcional asociado. Una secuencia que codifica para un producto de este tipo o que tiene una repercusión funcional dentro de la célula se conoce como gen (Gerstein *et al.*, 2007). Un genoma es el conjunto completo de ADN de un organismo. Este se encuentra empaquetado eficientemente dentro de la célula en forma de subestructuras llamadas cromosomas. Estos consisten en un complejo molecular llamado cromatina, que se compone de una sola molécula lineal de ADN y proteínas de empaque asociadas. La estructura central de la cromatina es el nucleosoma: un complejo de ocho proteínas que consiste en dos pares iguales de cuatro histonas diferentes, todas las cuales están envueltas con 147 pares de bases de ADN. Debido a su estado compacto, el ADN cromatinizado no siempre es fácilmente accesible. La accesibilidad de cromatina es el grado en el que otras moléculas del núcleo de la célula son capaces de interactuar físicamente con ADN superenrollado en forma de cromatina. Esto se determina por la topología de un nucleosoma y factores asociados que influyen en su ocupación. Normalmente se mide cuantificando la susceptibilidad de cromatina a modificaciones químicas específicas o escisión. Esta conformación tiene consecuencias funcionales para la regulación del proceso celular a nivel genético (Klemm *et al.*, 2019).

Si bien el ADN es esencialmente una molécula codificadora, existen estructuras asociadas que permiten la regulación de la información que está codificada dentro de

los genes. Esto asegura el funcionamiento correcto de las células. Dicha información se desenvuelve de forma dinámica y específica a través del proceso de expresión genética. Este proceso está influenciado por los Factores de Transcripción (TF), que son proteínas de unión al ADN específicas de una secuencia que ultimadamente tienen un efecto sobre el nivel de expresión genética específica de cada célula (Benveniste *et al.*, 2014). Además, la accesibilidad y la función de las características de la cromatina cambian cuando existen modificaciones covalentes en las histonas y cadenas de ADN asociadas. Los efectos de estas modificaciones dependen de la posición y los compuestos involucrados en la modificación; la expresión genética se ve afectada en patrones específicos, a menudo recurrentes, precisos y con cambios distribuidos de manera no uniforme en histonas particulares a través de diferentes regiones genómicas (Zhang y Li, 2017). En general, los elementos modificadores del estado de cromatina pueden regular la expresión genética por represión o activación, estados que son mutuamente excluyentes ((Charlet *et al.*, 2016), (Tie *et al.*, 2009)). Los patrones de modificación de cromatina pueden proporcionar información para elucidar los reguladores clave activos o reprimidos en una célula dada a través de un conjunto de condiciones (Creyghton *et al.*, 2010). Anotar dichos elementos reguladores e identificar los factores relacionados a los mismos son pasos clave para entender cómo funcionan las células, y de manera similar, cómo dejan de hacerlo cuando están afligidas por alguna enfermedad o condición que perturbe la estabilidad del estado celular (Slattery *et al.*, 2014). Un esquema que contiene algunos de estos elementos y sus posiciones relativas en el genoma se pueden visualizar en la Figura 1.

2.1.2. Epigenética y modificaciones histonales

El epigenoma de la célula consiste en las propiedades hereditarias independientes de la secuencia de ADN que pueden modular la producción funcional del genoma. Es un conjunto de alteraciones características del estado celular, dadas por cambios químicos dinámicos en los complejos de cromatina (C. David Allis, 2007). En términos generales, podemos dividir la regulación epigenética en tres mecanismos generales: modificaciones histonales específicas, interacciones de largo alcance entre los poten-

[1] https://haemgen.haem.cam.ac.uk/genomebrowser/encode/

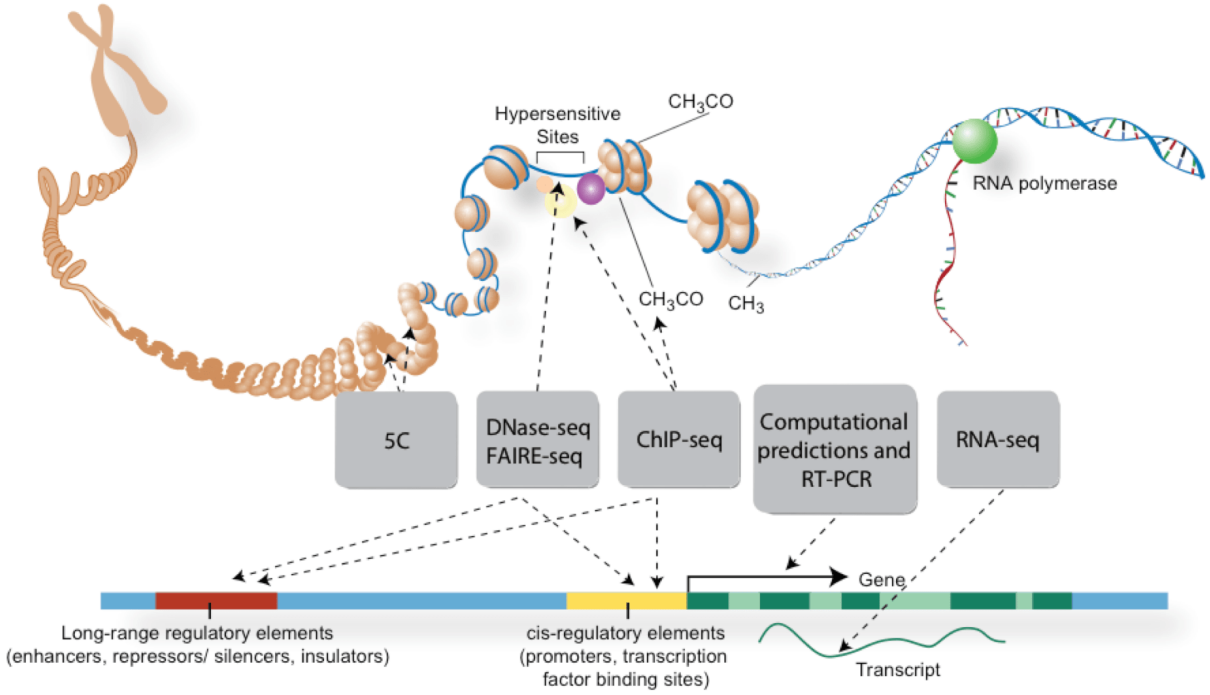

Figura 1. Construcciones biológicas relevantes, experimentos asociados para su determinación y sus posiciones relativas en el genoma. Obtenido de [1]

ciadores y sus objetivos, y el acceso selectivo, específico a cada región, de los elementos de regulación (Voss y Hager, 2013). Las modificaciones pueden manifestarse a través de múltiples mecanismos y exhibir repercusiones en diferentes niveles de granularidad dentro de la célula.

A nivel de secuencia, los elementos que son propensos a la modificación incluyen secuencias de ADN, proteínas histonales, y la localizacion de fragmentos de ARN no codificantes cortos y largos. A nivel del complejo de cromatina, podemos encontrar cambios en la organización espacial del ADN. Particularmente, en la ocupación y posicionamiento de los nucleosomas y las interacciones de cromatina en 3D; estas dependen del estado en particular de la accesibilidad de la cromatina de la célula y las proteínas modificadoras de la actividad que se unen a las regiones con niveles de exposición espacial variable (Sarda y Hannenhalli, 2014). La colección de estas características de regulación de la transcripción, las cuales son esenciales para el funcionamiento correcto de los cromosomas, se conocen comúnmente como "marcas epigenéticas".

Al analizar en conjunto las marcas mencionadas anteriormente, es posible deter-

minar el panorama general de la capacidad reguladora y funciones genómicas de una célula en particular y sus diferentes estados celulares (Allis y Jenuwein, 2016). En contraste con la secuencia de ADN, dicho panorama es altamente dinámico; debido al gran volumen de elementos interactivos que actualmente se consideran dentro del paisaje epigenético, el alcance total de la interacción entre los elementos y sus respectivas jerarquías funcionales sigue siendo ambigua (Lange y Schneider, 2010). Sin embargo, los mecanismos epigenéticos con dinámicas complementarias han sido vinculados a efectos celulares específicos. Un ejemplo notable es la regulación del silenciamiento transcripcional a través de la presunta interacción entre los efectos de la metilación del ADN y las modificaciones covalentes post-traduccionales de las proteínas histonales (Vaissiere *et al.*, 2008). Mientras que la comunicación entre los mecanismos no se entiende completamente, observar el comportamiento de la expresión genética a través del lente de las modificaciones histonales puede proporcionar una vista valiosa de los procesos y funciones regulatorias de una célula, incluyendo la dinámica de la cromatina.

En el contexto de la epigenética, las histonas están sujetas a múltiples modificaciones químicas post-traduccionales en residuos proteínicos específicos; se han identificado al menos ocho tipos distintos de modificaciones en 60 posiciones histonales conocidas. Estas modificaciones se pueden separar generalmente en dos categorías: las que regulan funciones celulares como la transcripción, replicación, condensación y diversos mecanismos de reparación, entre otras tareas biológicas basadas en el ADN y el establecimiento de entornos globales de cromatina en la célula (Kouzarides, 2007). Estas últimas logran dicho efecto sobre la estructura de la cromatina de orden mayor a través de la interacción de histonas modificadas con ADN o histonas en nucleosomas adyacentes. Mecánicamente hablando, la modificación con mayor potencial de desarrollo y remodelación local de la cromatina es la acetilación. Debido a esto, esta modificacion se asocia comúnmente a estados de transcripción activa (Zentner y Henikoff, 2013). Es posible establecer asociaciones similares entre los grados de accesibilidad de determinadas conformaciones de cromatina y su potencial regulador.

2.1.3. Accesibilidad de la cromatina

Los genomas de eucariotas se compactan en cromatina, un complejo de proteínas de ADN e histonas. Las histonas del núcleo empacan el ADN para formar un nucleosoma, la unidad basal de la cromatina. Dicho proceso reduce la cantidad de ADN expuesto a diversa maquinaria celular, o accesibilidad, que conduce a la inhibición de la transcripción (Fyodorov *et al.*, 2017). Dependiendo del nivel de compacidad, la estructura de la cromatina puede clasificarse en una de dos estructuras: heterocromatina o eucromatina; estos son los estados de cromatina condensados y sueltos, respectivamente (Brahmachari, 2013). La heterocromatina típicamente presenta patrones de expresión genética deficientes y no suele presentar alta actividad transcripcional. Por otra parte, la eucromatina, presenta patrones ricos de expresión y es más accesible a la transcripción (Tamaru, 2010). Estos estados de cromatina se encuentran frecuentemente asociados a modificaciones histonales específicas. Por ejemplo, la heterocromatina tiende a estar hipoacetilada e hipermetilada en la lisina 9 de la histona H3 (Wang *et al.*, 2016). Sin embargo, la cromatina no es una estructura inerte; está sujeta a una amplia gama de modificaciones histonales y cambios de conformación estructural, las cuales terminan por afectar la regulación del ADN (Bannister y Kouzarides, 2011). Dada la naturaleza dinámica del proceso de remodelación de la cromatina, la capacidad de las histonas nucleares para establecer los estados de cromatina depende en gran medida de las modificaciones covalentes presentes en ellas. Estas pueden permitir o negar modificaciones de estado adicionales, en una capa de regulación por encima del funcionamiento histonal intrínseco, durante el desarrollo celular o en respuesta a estímulos externos (Izzo y Schneider, 2016). La expresión celular depende de la activación y represión organizada de ciertos genes clave. Las células de cada organismo contienen en gran medida el mismo ADN; los factores transcripcionales y epigenéticos son los responsables de mantener el control sobre los patrones de expresión. La accesibilidad de la cromatina en las regiones asociadas con la regulación de genes afecta el resultado de los procesos transcripcionales en la célula (Miyamoto *et al.*, 2018). Se ha encontrado que el nivel de permisividad en los compuestos de cromatina en los sitios de regulación están fuertemente correlacionados con los niveles de actividad reguladora. El ADN regulatorio se expresa a menudo en sitios abiertos o accesibles de cromatina remodelada (Tsompana y Buck, 2014). Una ilustración que aproxima este fenómeno puede ser visualizada en la Figura 2. Los elementos de regulación suelen

estar situados en regiones de cromatina accesibles: el 16-21 % de los sitios de DNAsa I HS se encuentran en promotores o regiones exónicas de genes conocidos (Boyle *et al.*, 2008). Además, el estado de accesibilidad a la cromatina puede estar vinculado funcionalmente a un tipo particular de célula o tejido, y suele ser específico a una familia de proteínas (Xin y Rohs, 2018a).

La accesibilidad de la cromatina se utiliza de forma rutinaria como fuente de información para obtener conocimientos específicos sobre el tipo de célula y la condición en la que esta se encuentra. Recientemente, se ha demostrado que conocer el paisaje general de accesibilidad de determinadas regiones de cromatina en células de leucemia mieloide aguda (LMA) resultó crucial para determinar el tipo de origen celular de las células afectadas. Esto sugiere el uso potencial de patrones de loci de cromatina abierta como una firma pronóstica para la LMA humana (George *et al.*, 2016). Además, los paisajes de cromatina activa, en conjunción con otras mediciones a nivel de todo el genoma, tales como la expresión genética, la secuenciación de exoma entero, los perfiles de variación en el número de copias y los metilomas de ADN se han utilizado para capturar con éxito las características moleculares de la enfermedad, así como indicadores de dependencias de enfermedades no detectadas anteriormente. En conjunto, esta información puede allanar el camino para nuevos enfoques terapéuticos (Mack *et al.*, 2019). En términos generales, sondear la accesibilidad de la cromatina y sus factores epigenéticos asociados, puede ayudar a determinar la actividad de los elementos reguladores reprimidos y cuantificar los cambios en la expresión genética dentro y entre tipos de células. Esto se puede lograr de forma práctica con el uso de protocolos basados en tecnologías de secuenciación de próxima generación (Buenrostro *et al.*, 2015). Estos experimentos consisten en la separación del genoma, según un protocolo específico, para el aislamiento de las regiones accesibles del genoma. Los análisis posteriores asociados a menudo se centran en la identificación de paisajes de cromatina y los cambios epigenéticos asociados a procesos o condiciones de desarrollo particulares, o ambos.

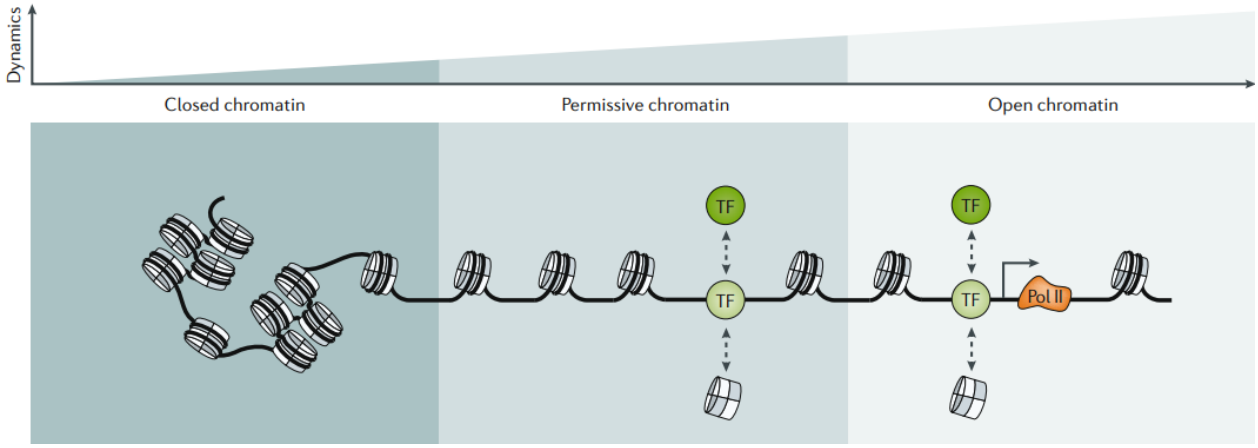

Figura 2. La variación en los niveles de accesibilidad de cromatina afecta la dinámica celular. Obtenido de(Klemm *et al.*, 2019).

2.1.4. Secuenciación de nueva generación

Es posible extraer y analizar información asociada al estado particular de una célula mediante el uso de tecnologías de secuenciación genómica. Cierta información específica de los fenómenos, como la determinación de la unión proteína-ADN puede obtenerse mediante la realización de experimentos que se centran en la bioquímica subyacente de cada fenómeno. Después de realizar un experimento, el material fuente puede ser analizado por secuenciación de nueva generación, en inglés, Next Generation Sequencing (NGS). Un flujo de trabajo estándar de NGS generalmente implica los siguientes pasos: Extracción de ADN o ARN, preparación de una biblioteca genómica, preparación de una plantilla de ADN, el protocolo de secuenciación y finalmente, el análisis de los datos obtenidos (Besser *et al.*, 2018).

Brevemente, una vez terminado el experimento biológico deseado, se extrae su ADN o ARN correspondiente y es transformado en una biblioteca genómica para el análisis de NGS. La construcción de bibliotecas implica la fragmentación de la información obtenida de ADN o ARN según una longitud previamente especificada, convertir dichos fragmentos en ADN de doble cadena, y unir los adaptadores de oligonucleótidos a los extremos de los fragmentos de destino con el objetivo de tornar estas secuencias reconocibles por un instrumento de secuenciación (Head *et al.*, 2014). Después de que los adaptadores se insertan exitosamente, una plantilla de ADN para secuenciación es generada amplificando las secuencias mencionadas anteriormente, lo cual se puede lograr con una amplia gama de técnicas, a menudo específicas por el fabricante (Metz-

ker, 2009). Este proceso genera una alta densidad de conjuntos de secuencias de ADN comunes al experimento realizado, que luego se unen a una superficie sólida, la cual posteriormente se somete a secuenciación (Linnarsson, 2010). La secuenciación consiste en analizar estos conjuntos anclados a la plantilla en paralelo. Esto generalmente se hace a través de un proceso de "lavado y escaneo", el cual involucra lavar la plantilla de ADN con nucleótidos etiquetados, incorporándolos en las cadenas de ADN de la plantilla, y escaneándolos para identificar qué bases se incorporaron en cada cadena. Esto permite la cuantificación y el análisis posterior del ADN o ARN extraído específico del experimento (Schadt *et al.*, 2010).

En tiempos recientes, las aplicaciones de secuenciación de nueva generación han experimentado un cambio de paradigma. La primera generación de los métodos de secuenciación está constituida por la secuenciación automatizada tipo Sanger. Estos se utilizan en gran medida para reconstruir la secuencia completa de ADN de un organismo en un proceso conocido como Secuenciación de Genoma Completo (WGS). Sin embargo, la capacidad de NGS para crear datos rápidamente en paralelo, en el orden de miles de millones de lecturas cortas por día, ha permitido a los investigadores ampliar sus objetivos de investigación más allá de la determinación de secuencias canónicas en el genoma (Metzker, 2009). Los métodos basados en secuencia nos permiten realizar búsquedas específicas y consultas con orientación biológica; pueden ser utilizadas para identificar y cuantificar transcritos raros dadas condiciones biológicas particulares. Los sectores de investigación y sus aplicaciones incluyen, pero no se limitan a: genómica comparativa, evolución, medicina forense, epidemiología, medicina aplicada y diagnóstico. Llevar a cabo investigaciones en los campos antes mencionados requiere de la integración de múltiples fuentes de datos biológicos, que normalmente son generados a través de múltiples experimentos de enriquecimiento acoplados a NGS. Los experimentos más comunes pueden ser visualizados en la Figura 3. Esto permite cambiar el enfoque de los análisis de la vista alguna vez considerada como estática y limitada a nivel de secuencia del genoma hacia la consideración de propiedades hereditarias independientes de dicha secuencia. Tal es el caso del dominio de la biología de la cromatina, específicamente, a través de la observación de las interacciones proteína-ADN (Goodwin *et al.*, 2016). Actualmente se puede observar dicho fenómeno mediante la aplicación de protocolos que enriquecen los fragmentos de ADN que interactuan con proteínas dentro de una muestra dada. Ejemplos notables

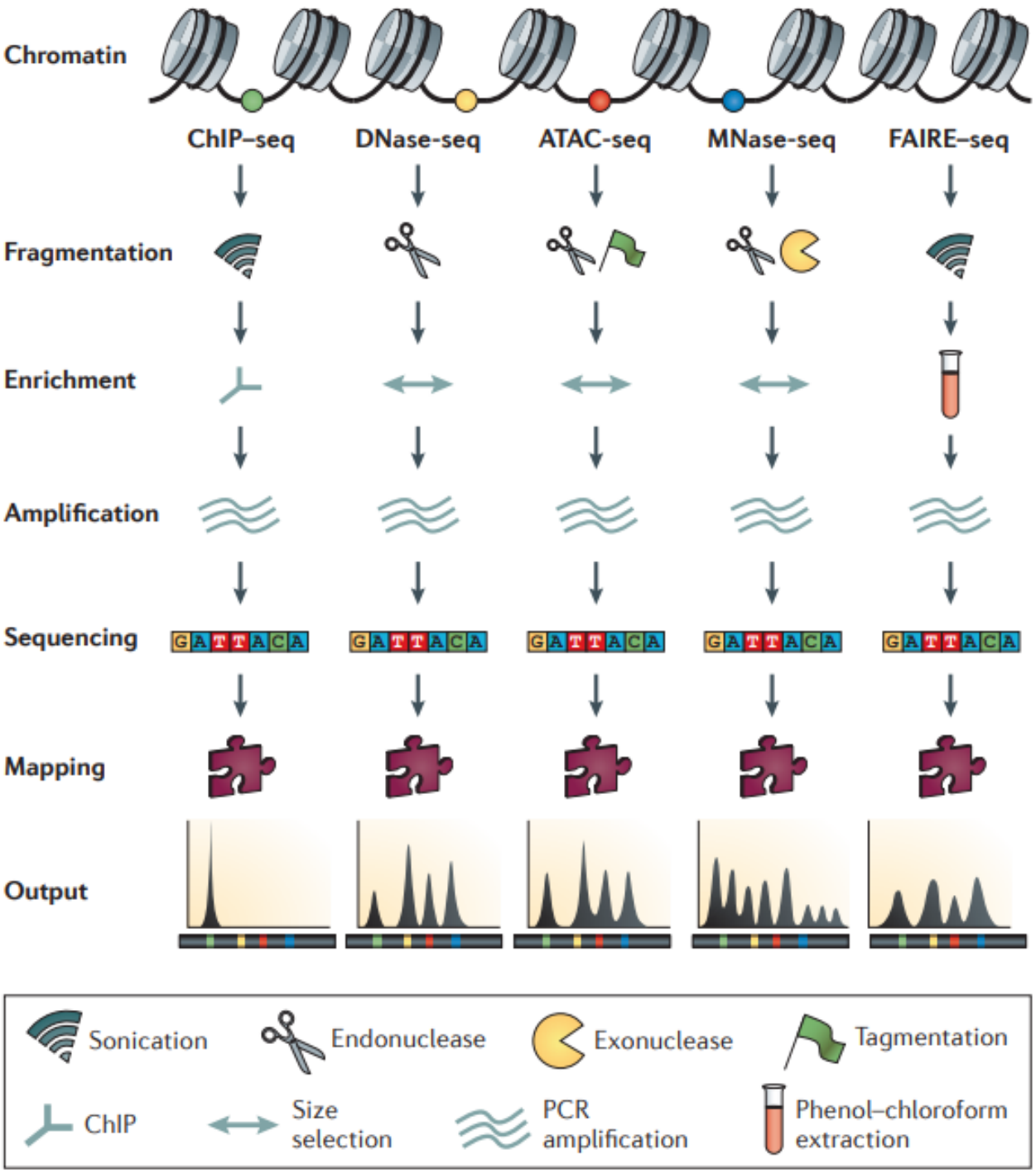

Figura 3. Contexto general y comparación de protocolos experimentales comunes acoplados a NGS para la extracción de información específica de ciertos fenómenos biológicos. Obtenido de (Meyer y Liu, 2014)

incluyen la Inmunoprecipitación de la Cromatina y el Ensayo para Cromatina Accesible por Transposasa seguido de secuenciación de nueva generación (ChIP-seq y ATAC-seq, por sus siglas en inglés, respectivamente), las cuales se discutirán en breve.

2.1.5. ChIP-Seq

La inmunoprecipitación de cromatina seguida de secuenciación de nueva generación (ChIP-seq) identifica las localizaciones genómicas que están ligadas a proteínas. Esto se logra mediante la unión química, o reticulación, de las proteínas y la cromatina, seguido de la separación de las proteínas ligadas del resto de las secuencias de ADN por sonicación, aislando dichos pares de proteína-cromatina. Posteriormente se capturan los fragmentos de ADN unidos a las proteínas utilizando anticuerpos específicos correspondientes a la proteína que se intenta precipitar y, por último, se efectua un protocolo de secuenciación de nueva generación (Bailey *et al.*, 2013). Una vez que el experimento de NGS es ejecutado exitosamente, el análisis consiguiente puede ser dividido en tres pasos: filtrar las lecturas que caen por debajo de algún umbral de calidad, alinear las lecturas a un genoma o transcriptoma de referencia y la cuantificación de los niveles de expresión genética, o en el caso de ChIP-Seq, identificación de picos (Fonseca *et al.*, 2014). ChIP-seq, a diferencia de ATAC-seq, es una tecnología bien establecida. Sus limitaciones comunes, incluyendo fuentes de sesgo, consideraciones experimentales como el requisito de un número elevado de células en la muestra y la calidad y selección limitada de proteínas reticulantes y proteínas precipitantes, o ambas, han sido en gran medida abordadas o se toman en cuenta de forma activa en el desarrollo de protocolos para evitar los efectos de dichas limitaciones en el análisis posterior.

El ChIP-seq se utiliza para estudiar las interacciones de ADN-Proteína en todo el genoma. Esta técnica se utiliza usualmente para la identificación de Factores de Transcripción y proteínas asociadas; se ha demostrado que estas proteínas de unión funcionan predominantemente en regiones de cromatina accesible: 94,4 % de los picos de ChIP-seq de los factores de transcripción de la Enciclopedia de los Elementos del ADN (ENCODE) se encuentran dentro de regiones de cromatina accesible, a diferentes niveles de intensidad (Thurman *et al.*, 2012). Cabe mencionar que los eventos de acoplamiento en ChIP-seq y experimentos relacionados tienden a ser específico al tipo celular y tejido en el que se realizan. Por lo tanto, obtener regiones de cromatina abierta concurrentes a un conjunto de perfiles de acoplamiento de ChIP-seq podría ayudar a la identificación de regiones de regulación comunes y esenciales, sus conformacio-

nes de cromatina, reguladores clave y sus programas de transcripción asociados en diferentes tipos de células y condiciones.

El ChIP-seq normalmente se realiza precipitando una proteína de interés a la vez. Estas proteínas, generalmente TFs o histonas, están usualmente bajo efectos de regulación compuesta que se llevan a cabo simultáneamente dentro de la célula. Por lo tanto, realizar ChIP-seq como el único experimento para inferir los efectos funcionales de los genes que no se encuentran direcatamente asociados a la proteína objetivo suele ser insuficiente. Sin embargo, el ChIP-seq puede ser usado en conjunto con otras metodologías basadas en ChIP-seq para obtener resultados favorables. Un ejemplo notable es cuando la obtención del perfil de acoplamiento del genoma por medio de ATAC-seq falla debido a la existencia de múltiples regiones enriquecidas de ATAC con motivos similares que no pueden ser distinguidos solamente por virtud de su secuencia. En instancias como esta, realizar análisis de ChIP-seq para factores de transcripción específicos es una herramienta valiosa para la validacion de otros experimentos. En el caso específico de las modificaciones histonales, los datos de ChIP-seq pueden ser “mapeados“ a picos derivados de ATAC-seq para confirmar el estado particular de cromatina de un elemento regulador dado, ya que se ha demostrado que estos tienen diferentes niveles de actividad en función de su etapa de desarrollo actual (Jiang y Mortazavi, 2018b).

2.1.6. ATAC-Seq

El ensayo para cromatina accesible por transposasa seguido de secuenciación de nueva generación (ATAC-seq) tiene como objetivo medir el conjunto de cromatina accesible en el genoma a través de la detección de eventos de clivado químico. Esto es logrado mediante la utilizacion de una proteina de “cortado y pegado“, específicamente, una enzima transposasa hiperactiva conocida como Tn5. Dicha enzima reconoce y “marca“ repeticiones invertidas específicas de 19 pares de bases; estas son secuencias idénticas que se encuentran en las dos hebras correspondientes del ADN marcado, una invertida con respecto a la otra para preservar la complementariedad de pares de base. Es posible insertar dichas secuencias repetidas como adaptadores de secuencia

en regiones de cromatina accesible para hacerlas susceptibles a la actividad enzimática de Tn5 (Buenrostro *et al.*, 2013). Un flujo de trabajo computacional básico después de ejecutar el protocolo experimental implica alinear las lecturas obtenidas que fueron seleccionadas por los adaptadores de Tn5, realizar un control de calidad, cuantificar el enriquecimiento o bien, efectuar una identificación de picos y posteriormente analizar las regiones asociadas y la huella general de acoplamiento. Aunque dicho experimento es utilizado principalmente para evaluar la cromatina accesible, el ATAC-seq también permite realizar un perfil general de factores de transcripción en todo el genoma y sus rutinas de expresión génica asociadas. Esto se debe a que el acoplamiento de los factores de transcripción protege las regiones unidas contra clivado químico por nucleasa y la actividad de Tn5, fragmentando regiones que de otro modo serían largas y abiertas en fragmentos más pequeños. Dichas regiones fragmentadas pueden posteriormente identificarse como factores de transcripción, contrastando las secuencias de cada región contra motivos de TF conocidos (Corces *et al.*, 2018).

Mientras que existen otros protocolos para evaluar la cromatina accesible, ATAC-seq tiene claras ventajas. Se requiere una baja cantidad de células y presenta una alta relación señal/ruido, en comparación con otros métodos existentes (Klemm *et al.*, 2019). No es tan propenso a sesgos específicos del método, a diferencia de la alternativa frecuentemente utilizada, DNAseq. La contabilización del sesgo del ADN mitocondrial sigue siendo un reto debido a la digestión excesiva o insuficiente de dicho ADN en la muestra. Las ventajas prácticas incluyen el desarrollo y uso de librerías de ADN generadas con este protocolo, ya que pueden ser empleadas para análisis de NGS directamente después del aislamiento y la amplificación por PCR. Otros rasgos distintivos son la alta sensibilidad, simplicidad técnica, alto rendimiento y la capacidad de procesar grandes volumenes de muestras, lo que resulta útil en estudios clínicos de cohorte de gran tamaño. Además, esta técnica muestra mayor sensibilidad para la detección de conjuntos específicos de regiones libres de nucleosomas fuera de los sitios anotados de inicio de la transcripción (Nordström *et al.*, 2019).

Determinar los niveles de permisividad de la cromatina de una célula a través de los loci conocidos puede resultar crucial a la hora de hacer análisis río abajo que están ligados a condiciones particulares. Esto se debe a que la conformación tridimensional del genoma influye en los niveles y patrones de actividad de los reguladores cono-

cidos, específicamente, factores de transcripción. Como estos últimos son capaces de remodelar la cromatina y acoplarse de forma específica por secuencia, existe una interacción considerable entre sus patrones de unión y su efecto en el paisaje de cromatina tridimensional resultante (Kim y Shendure, 2019). El estado de la organización de la cromatina puede impactar el efecto y la distribución de las mutaciones; esto se observa a menudo en el cáncer, donde la organización de la cromatina se altera con frecuencia y en consecuencia, afecta a las tasas y patrones de mutación, a menudo de forma altamente específica dependiendo del contexto celular, por tipo de cáncer (Makova y Hardison, 2015). La utilización de ATAC-seq en conjunto con otras tecnologías basadas en secuenciación de nueva generación puede ser de ayuda en determinar qué zonas accesibles de la cromatina en nuestros análisis están vinculadas a mutaciones funcionales y regiones de actividad de interés.

2.2. Computacional

2.2.1. Aprendizaje automático

El Aprendizaje Automático se refiere al conjunto de herramientas computacionales para la detección automatizada de patrones en los datos (Shalev-Shwartz y Ben-David, 2014). Estos métodos se aplican en la Biología Computacional como enfoques de propósito general para aprender relaciones funcionales a partir de los datos; esto permite la obtención de modelos predictivos sin necesidad de suposiciones previas sobre los mecanismos biológicos subyacentes (Angermueller *et al.*, 2016). Un flujo de trabajo típico se puede observar en la Figura 4. Los algoritmos pertenecientes a los paradigmas de aprendizaje supervisado y no supervisado han demostrado tener mayor representación en el panorama de la investigación contemporánea. En resumen, el primero tiene por objeto predecir correctamente el valor de una medida de resultado basada en una serie de medidas de entrada, mientras se conocen las etiquetas de salida correspondientes. Este último tiene como objetivo descubrir patrones y asociaciones a partir de muestras de datos sin una necesidad estricta de etiquetas de salida. Dada la naturaleza de nuestro problema, la discusión se centrará en el primero.

El problema de aprendizaje se constituye de la siguiente manera: dado un espacio

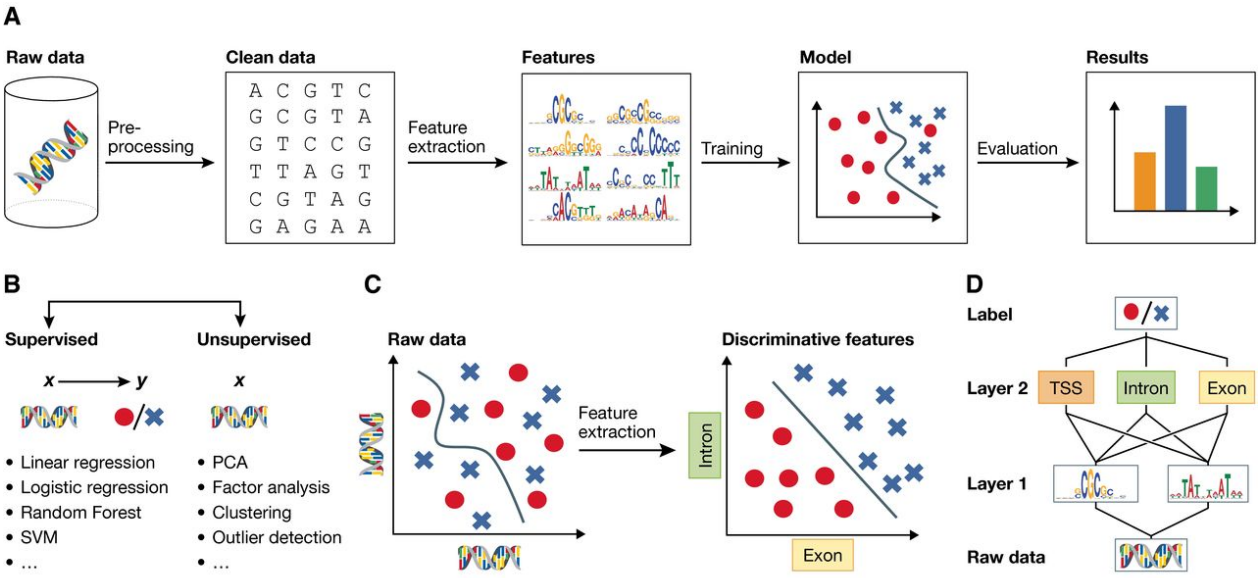

Figura 4. Flujo de trabajo típico de Aprendizaje Automatico. El inciso A contiene un esquema que detalla el proceso desde la obtención, exploración y manejo de los datos hasta la obtención de resultados. El inciso B contiene algoritmos populares comunmente utilizados en paradigmas tanto supervisados como no supervisados. El inciso C ilustra como los datos biólogicos habitualmente se prestan para errores de clasificación, por lo que utilizar representaciones que separen los datos con menor error de clasificación suele prestarse para mejores resultados. El inciso D muestra una red neuronal extrayendo representaciones abstractas de los datos sin procesamiento. Obtenido de (Angermueller *et al.*, 2016).

de entrada X con una serie de entradas x, se debe encontrar una función objetivo f para la asignación de valores de la forma $f : X \rightarrow Y$, donde Y es el espacio de salida, el cual se compone de un conjunto de elementos y.

En un proceso conocido como entrenamiento, un algoritmo de aprendizaje A elige entre un conjunto de funciones de destino candidatas H mediante pruebas con combinaciones de x, y hasta que se encuentra una función $g : X \rightarrow Y$ que mejor se aproxima a la función objetivo real, f (Abu-Mostafa *et al.*, 2012). Una vez que se encuentra una función apropiada g, es posible probar nuevas entradas x en el modelo obtenido para efectuar predicciones de valores de salida y . Si el modelo es capaz de categorizar ejemplos nuevos correctamente, nunca antes vistos por el modelo, se dice que este generaliza correctamente. La comprensión adecuada de los fenómenos subyacentes es crucial para la construcción apropiada del modelo. Para evaluar adecuadamente nuestro modelo de aprendizaje, una medida de evaluación que esté de acuerdo con el comportamiento de predicción deseado por el modelo debe ser establecida. Para la clasificación, ejemplos ilustrativos incluyen pero no se limitan a: precisión, el Coeficiente de Correlación de Matthew (MCC), la especificidad, la medida de puntuación F1, la sensibilidad y la curva ROC.

Para asegurar el funcionamiento adecuado de un modelo de aprendizaje con mayor certeza, se deben tener en cuenta los aspectos computacionales prácticos durante la fase de construcción del modelo. De forma prioritaria, se desea aumentar la capacidad de generalizar mediante la reducción de la tasa de error de clasificación en todas las pruebas futuras. Como los datos futuros no suelen estar disponibles de antemano, los datos deben ser particionados para maximizar la eficacia durante la fase de entrenamiento. Usualmente se aconseja la división proporcional de los datos disponibles en tres conjuntos diferentes de la siguiente manera: un conjunto de entrenamiento, para la construcción básica del modelo, un conjunto de validación, para ajustar correctamente el modelo y finalmente, un conjunto de prueba, para medir el rendimiento del modelo construido. Esta es una de las precauciones adicionales que se deben tomar al ajustar los modelos. Utilizar datos ruidosos o de otro modo poco informativos, o una partición de datos incorrecta cuando se entrena un modelo puede resultar en sobre o sub-ajuste del modelo. Concisamente, los modelos deben estar diseñados para evitar modelar fuentes de variación excesivamente minúsculas o extensas. De lo contrario, pueden aumentar las tasas de error debido al alto sesgo del modelo, forzando patrones que son irrelevantes para los fenómenos subyacentes de los datos de prueba proporcionados, o fallar en la captura de la estructura subyacente de los datos debido a la alta variación del modelo. Idealmente, en un modelo bien entrenado habrá separación mínima entre las tasas de error observadas entre los conjuntos de entrenamiento y prueba (Murphy, 2012).

2.2.2. Extracción de características

Cuando un conjunto de datos crece en magnitud, la cantidad de datos necesarios para proporcionar un análisis fiable crece exponencialmente (Hira y Gillies, 2015). Una solución es proyectar los datos a un espacio dimensional más bajo, es decir, reducir los datos originales a un número determinado de variables o características mediante el uso de transformaciones, con el fin de eliminar las redundancias latentes en forma de ruido de clase o atributo. La extracción de características se refiere al uso de las transformaciones de las características de entrada para producir características nuevas (Jain *et al.*, 1999). Esta técnica a menudo se utiliza para obtener conjuntos

de características que sean apropiados para que estas sean utilizadas por un algoritmo de aprendizaje para predecir exitosamente. La extracción de representaciones simples y compactas de conjuntos de datos es altamente deseable, ya que a menudo estas reducen la complejidad computacional del problema de aprendizaje, así como la variabilidad entre los predictores y puede proporcionar nuevos conocimientos sobre los parámetros del conjunto de datos relevantes para la predicción de un problema determinado. Cuando un conjunto de datos es lo suficientemente grande y contiene subcaracterísticas informativas, la extracción de características tiene un efecto intrínseco de reducción de dimensionalidad (Bishop, 2006). Extraer apropiadamente dichas características es clave para la construcción efectiva de modelos de predicción.

La elección de un algoritmo de extracción de características apropiado para un problema determinado no es una tarea trivial. Diversas cuestiones deben tenerse en cuenta antes de tomar una decisión: ¿Qué tarea de predicción está siendo realizada? ¿El método es de naturaleza supervisada o no supervisada? ¿Se requieren soluciones económicas en recursos y aproximadas o exactas? Adicionalmente, después del proceso de extracción, se debe evaluar la calidad de las características obtenidas. Idealmente, deben ser significativamente diferentes entre clases, mostrar valores similares para un patrón dado por clase y no deben estar correlacionados entre sí. Aunque aplicar eficientemente los métodos de extracción de características es una tarea específica de un dominio, existen métodos de aplicación general para diversas tareas de extracción de características. Ejemplos notables incluyen pero no se limitan a: estandarización, normalización, mejora de la señal, extracción de características locales, métodos de encajes matemáticos, expansiones no lineales y discretización de características (Guyon, 2006).

La extracción óptima de características puede describirse como la construcción de una función f que es capaz de comprimir un vector de entrada x de dimensiones d en un vector de salida $y = F(x)$ con dimensiones p tales que $p < d$ (Hornik y Kuan, 1992). Algunos de los métodos estadísticos más representativos para esta tarea son: Análisis de Componentes Prinicipales (PCA), Análisis Discriminante Lineal (LDA), Análisis Factorial (FA), Mínimos Cuadrados Ordinarios (OLS), entre otros (Ding *et al.*, 2011). Aunque no todos los métodos reducen directamente la dimensionalidad de un problema, la extracción de características tiene como objetivo reducir la varianza en el desempeño de

la clasificación para aumentar la generalización del modelo. La reducción del número de funciones a menudo reduce el número de parámetros necesario para la clasificación exitosa. Sin embargo, la reducción de las características antes mencionadas no mejora indefinidamente el rendimiento de la clasificación. La reducción indiscriminada de la dimensionalidad a menudo tiene el efecto contrario; es posible perder información relevante para el problema de la predicción si se filtran características informativas (Hild *et al.*, 2006). Es necesario trabajar con combinaciones de características construidas y crudas y analizarlas tomando en cuenta el contexto del desempeño específico del clasificador y elegir el conjunto final de características en consecuencia.

2.2.3. Filtros

Fundamentalmente, el filtrado es el procesamiento de una señal en el dominio del tiempo como entrada que resulta en una modificación del contenido de dicha señal como salida. A menudo se aplican para reducir o filtrar el contenido no deseado en la señal. Esto se consigue normalmente estableciendo un umbral que capta ciertas frecuencias en una señal, ignorando o contabilizando aquellas que no son de interés. Los filtros se pueden clasificar ampliamente en dos categorías: analógicos y digitales; el primero funciona con señales continuas, mientras que el segundo funciona con una secuencia de valores de muestra discretos. Dada la naturaleza de nuestros datos, centraremos nuestra discusión en este último tipo.

El tipo más simple de filtro digital es el filtro de respuesta finita al impulso (FIR). Este utiliza exclusivamente valores presentes y pasados de entrada en sus cálculos, sin tomar en cuenta los valores calculados por el mismo filtro durante el proceso de filtrado. Una versión comúnmente utilizada es el filtro de media móvil (MA), que funciona calculando promedios sobre intervalos de longitud definidos, o ventanas, secuencialmente a través de la señal de entrada. La ecuación 1 representa un filtro de media móvil de L puntos en tiempo discreto (Alan V. Oppenheim, 1982).

$$y[n] = \frac{1}{L} \sum_{k=0}^{L-1} x[n-k]. \tag{1}$$

Donde x es la señal de entrada, y el valor de salida promediado, L es la longitud de la ventana, n es el paso en el tiempo actual en la señal y k es el paso de tiempo anterior. Dicho filtro tiene el efecto de suavizar los cambios repentinos en la entrada, donde la suavidad aumenta con la longitud de la ventana. Además, este filtro se comporta como un filtro pasa bajas, donde se promedian las frecuencias bajas o ruidosas y sólo se mantienen los componentes con un valor de señal alto.

2.2.4. Selección de características

Después de extraer las características relevantes para el modelo de predicción, la siguiente tarea es ser capaz de discernir cuáles de estas características son las más relevantes para el modelo de predicción. Encontrar un subconjunto de características óptimo es un problema generalmente intratable, por lo tanto, la selección de características apunta a elegir un pequeño subconjunto de características entre combinaciones de las originales y extraídas según cada modelo, con una evaluación previamente establecida, hasta que se cumpla una condición de parada; dichas medidas pueden basarse en la información, la distancia, consistencia o dependencia de las variables (Liu y Yu, 2005). En el contexto de la clasificación, la selección de características se emplea a menudo con el objetivo de mejorar el rendimiento de un modelo en un problema de aprendizaje supervisado. Por lo tanto, el objetivo del proceso es eliminar las características que son redundantes para el objetivo de clasificación. Con ello se pretende mantener la mayor precisión de clasificación posible con la mínima cantidad de características y tener una distribución de clases de característica resultante lo más parecida posible a la del problema original, dadas todas sus características. Este proceso afecta principalmente a la fase de formación de un modelo y puede hacerse en conjunción con o de forma independiente del algoritmo de aprendizaje que se va a aplicar sobre las características obtenidas (Tang *et al.*, 2014). Además, la selección de características no está necesariamente limitada a los problemas de clasificación; puede extenderse a problemas como la agrupación y la regresión. Un procedimiento típico de selección de características consta de tres partes principales: un algoritmo de búsqueda principal para buscar subconjuntos candidatos de solución, una función objetivo para guiar el proceso de búsqueda con una meta de optimización, minimizar

o maximizar, y una función de rendimiento para evaluar los resultados del algoritmo.

Los métodos de selección de características pueden ser separados por su perspectiva de supervisión o estrategia de selección. En el contexto de un paradigma de aprendizaje supervisado, existen tres amplias categorías de estrategias de selección: Filtros, métodos de envoltura y métodos embebidos (Li *et al.*, 2017). Los métodos de filtrado funcionan de manera completamente independiente del algoritmo de aprendizaje que se está empleando. Estas clasificaciones calculan puntajes entre las características y una métrica objetivo, para cada característica o conjunto de características, según un umbral establecido para un conjunto de datos determinado. El conjunto con puntaje más alto se guarda para modelado posterior. Las métricas para calcular los puntajes con frecuencia incluyen la correlación de Pearson y el criterio de información mutua (MI) (Chandrashekar y Sahin, 2014). Las ventajas de las métricas basadas en filtro son su simplicidad y robustez computacional contra el sobreajuste. Sin embargo, el subconjunto resultante de esta operación a menudo no es el óptimo, lo cual puede complicar la búsqueda de un algoritmo de aprendizaje adecuado. En respuesta, los métodos de envoltura utilizan iterativamente la predicción del algoritmo de aprendizaje, dado un subconjunto de variables, como función objetivo para la evaluación de las características. Los algoritmos de búsqueda heurísticos o de selección secuencial son frecuentemente utilizados. Estos métodos no incorporan conocimientos sobre la estructura específica de la función de clasificación, lo cual permite una mayor flexibilidad a la hora de seleccionar una máquina de aprendizaje a expensas de un óptimo rendimiento para discriminación de subconjuntos y un sobreajuste. Debido a esto último y contemplando una mayor eficiencia computacional, las estrategias de búsqueda voraces se utilizan a menudo en conjunción con los métodos de envoltura. A diferencia de los dos métodos mencionados anteriormente, los métodos integrados directamente incorporan la selección de características en el proceso de entrenamiento de un algoritmo. En esta estrategia de selección, el tratamiento de características difiere en tanto que hay parámetros asignados por el modelo que están directamente asociados a las características proporcionadas, como los pesos asignados a cada característica por el modelo. Las funciones objetivo relacionadas deben ser diseñadas para maximizar la cantidad de la información suministrada al modelo, cuantificada a través de una métrica como el MI, mediante parámetros suministrados por el modelo, en conjunción con las características proporcionadas para obtener la salida de la clase deseada.

Idealmente, se mantienen las características altamente clasificadas, mientras que las menos informativas se eliminan. Métodos populares de este tipo incluyen la regularización L1 y L2 y árboles de decisión (Lal *et al.*, 2006).

Todos los enfoques mencionados anteriormente a menudo se refieren a objetivos de minimización o maximización que no tienen una solución óptima que se puede tratar de forma fiable. Se pueden aplicar métodos heurísticos para obtener resultados favorables. La utilización de estrategias de búsqueda voraces para navegar el posible espacio de características puede reducir enormemente la complejidad del problema, reducir el sobresajuste y los recursos computacionales necesarios para implementaciones exitosas. Un ejemplo rutinario es el uso de algoritmos secuenciales, que a menudo no son soluciones óptimas, aunque fiables y rentables. Existen dos ejemplos generales: selección hacia adelante y hacia atrás. En la selección hacia adelante se comienza con un modelo sin características y se añaden variables secuencialmente, es decir, se evaluan todos los modelos con características individuales y se mantiene el de mayor puntaje. Posteriormente, se añaden todas las características restantes para un modelo con un total de dos características y así sucesivamente, hasta que todas las características hayan sido evaluadas y se mantenga el modelo de mejor rendimiento (Reunanen, 2003). La selección hacia atrás funciona de manera similar; sin embargo, comenzamos con un conjunto completo de características y se eliminan las redundantes hasta que sólo se mantengan las más informativas (Fisher, 1996). Las variantes asociadas, como la selección por etapas o la selección flotante hacia delante, tienen limitaciones específicas en sus funciones objetivo que eliminan o añaden dichas variables de forma dinámica en vez de por etapas, pero el comportamiento general permanece similar (Pudil *et al.*, 1994). Los procedimientos de eliminación hacia atrás generalmente pueden producir mejores resultados para clasificación binaria, a expensas de conjuntos de características más grandes, en contraste con los métodos de selección hacia adelante. Este enfoque depende en gran medida de la elección del algoritmo de aprendizaje que se va a utilizar posterior al del procedimiento de selección de características (Borboudakis y Tsamardinos, 2019).

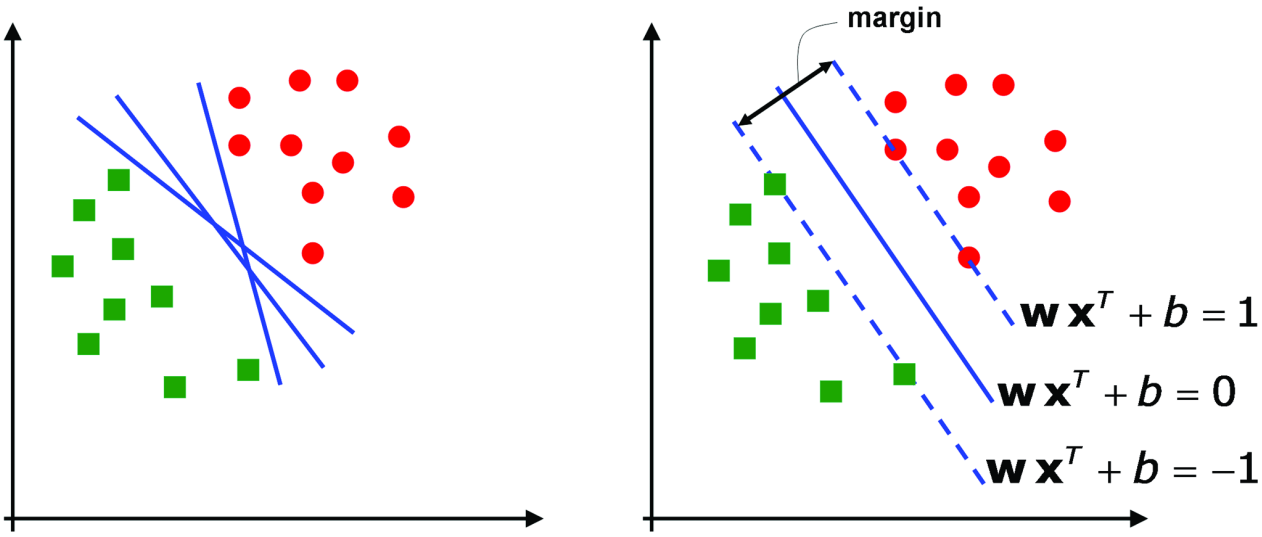

Figura 5. Un clasificador de máquina de soporte vectoriale binario. La izquierda ilustra posibles hiperplanos mientras que la derecha muestra el límite de decisión óptimo, lo que maximiza la distancia entre ambas clases. Obtenido de (Tarca *et al.*, 2007).

2.2.5. Máquina de soporte vectorial

La máquina de soporte vectorial (SVM) es un algoritmo de aprendizaje que se implementó originalmente para problemas de clasificación de dos grupos. Fundamentalmente, los vectores de las características de entrada son mapeados de manera no lineal a un espacio de características de alta dimensión. Posteriormente se construye una frontera de decisión sobre este espacio de características para separar los puntos de los datos entre clases (Cortes y Vapnik, 1995). Una explicación simple es la siguiente: en el caso de la clasificación binaria, donde hay n ejemplos totales de entrenamiento compuestos de vectores x_i de dimensión semejante p, nuestra meta es categorizar cada punto en todos los vectores de entrada como perteneciente a una de dos clases $y_i = -1$ o 1 (Fletcher, 2009). Suponiendo que nuestros datos son linealmente separables, todos los puntos de datos pueden ser separados en sus respectivas clases por una línea en un plano dado por los puntos x_1 y x_2. Dicha línea funciona como un hiperplano de separación, que tiene dimensiones $p-1$. Si extendemos este ejemplo a un espacio dimensional de p dimensiones, un hiperplano se definiría como el correspondiente subespacio $p-1$ dimensional definido por la ecuación 2. El hiperplano es entonces capaz de dividir nuestras observaciones en las que son mayores y menores que 0, que corresponden a las clases $y_i = 1$ y y_i = -1, respectivamente.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0. \quad (2)$$

Se pueden calcular múltiples hiperplanos; el problema radica en encontrar el que mejor separe las categorías. Encontrar un hiperplano óptimo depende de la maximización de la distancia del plano de separación con respecto a ambas categorías. Sin embargo, a menudo es el caso que las categorías no pueden ser completamente separadas. Por lo tanto, el problema de clasificación puede enmarcarse como el de encontrar un hiperplano, dado por la ecuación 3, que puede separar clases de manera óptima, permitiendo una clasificación con error mínimo, donde **w** es un vector *p* dimensional perpendicular al hiperplano y *b* es un término de sesgo. Una conceptualización de dicho hiperplano y los márgenes de separación asociados se pueden visualizar en la Figura 5. En el caso de que diferentes hiperplanos de separación tengan el mismo error dentro de la muestra, un margen de separación más amplio es beneficioso, de modo que permite que la máquina de aprendizaje tenga un mayor potencial de generalización. Tener un margen de separación estrecho hace que el modelo sea propenso a errores de clasificación.

$$\mathbf{w}\mathbf{x}^{\mathbf{T}} + b = 0. \quad (3)$$

El hiperplano óptimo puede ser construido encontrando valores de **w** y *b* que maximicen el margen de separación del hiperplano, dado por $\frac{1}{\|\mathbf{w}\|^2}$. La solución para el problema del clasificador de soporte vectorial para el caso lineal puede ser expresada por medio de los productos internos de los datos asociados, indicados por $(\mathbf{x_i} \cdot \mathbf{x_i^T})$ (Hastie *et al.*, 2009). Si extendemos el producto interno a un caso general, puede expresarse como $K(x_i, x_i')$, donde K es una función que transforma el producto interno, llamado núcleo. El clasificador vectorial de soporte del núcleo lineal puede representarse con la ecuación 4.

$$K(x_i, x_i') = \sum_{j=1}^{p} x_{ij} x_{ij}'. \quad (4)$$

Cuando los datos no son linealmente separables, o se requiere de fronteras de

decisión sofisticados para discernir entre clases dada una distribución específica de datos, el clasificador de soporte vectorial puede extenderse más allá del caso lineal. Las máquinas de soporte vectorial consisten entonces, en aplicar una transformación del núcleo a los datos del clasificador mediante la sustitución de productos de matriz lineal en el clasificador de soporte vectorial lineal por varias transformaciones de mayor dimensión. Los núcleos populares incluyen el núcleo polinomial y el núcleo de la función de base radial (RBF), el último de los cuales está dado por la ecuación 5 (Tarca *et al.*, 2007).

$$K(\mathbf{x}, \mathbf{z}) = exp(-\gamma \|\mathbf{x} - \mathbf{z}\|^2). \tag{5}$$

Capítulo 3. Metodología

3.1. Datos y métodos

Proporcionarle información a un modelo sobre las propiedades subyacentes del fenómeno biológico que se intenta predecir puede mejorar su capacidad de predicción. Los datos que provienen de los experimentos biológicos que se realizan para desentrañar dichos fenómenos suelen estar representados en una variedad de formatos de datos predefinidos; realizar un análisis exploratorio de datos puede ayudar a revelar características en común entre los experimentos o tipos de datos, con el objetivo de incorporarlos como información relevante para la predicción dentro del modelo. A continuación se detallan los pasos que se han dado para la adquisición y adecuación de datos y una serie de medidas realizadas tanto en los experimentos de ChIP-seq como en los de accesibilidad de cromatina relacionados, obtenidos a partir de un tipo de célula conocida por mostrar actividad representativa de H3K27ac: A549, o células humanas epiteliales basales alveolares adenocarcinómicas (Chang *et al.*, 2016).

A lo largo de este proyecto de tesis, se utilizaron dos conjuntos de datos: Datos de experimentos fijados en H3K27ac de ChIP-seq de y ATAC-seq, correspondientes al tipo de célula antes mencionado, tratada con 100 nm de dexametasona, se obtuvieron del Portal ENCODE con números de resumen de los experimentos ENCSR778NQS[1] y ENCSR220ASC[2], respectivamente. Ambos están alineados con el ensamblaje del genoma hg38/GRCh38 [3]. Los datos alojados en el portal ENCODE han sido procesados previamente de acuerdo a las especificaciones de ENCODE para cada experimento [4]. Se descargaron dos tipos de archivos por experimento: archivos de bigwig, que contienen todas las regiones genómicas del experimento con los valores “crudos“ o sin modificar de las señales después del tratamiento inicial de los datos NGS y archivos BED, considerados como “narrowpeak“ (o de “pico estrecho“) en este caso. Dichos archivos contienen las regiones más relevantes o “picos“ determinados por una herramienta de “identificación de picos“ en particular.

Los formatos predefinidos que contienen información genómica pueden ser accedidos y manipulados con una variedad de herramientas de software. El proyecto Bio-

[1] https://www.encodeproject.org/experiments/ENCSR778NQS/
[2] https://www.encodeproject.org/experiments/ENCSR220ASC/
[3] https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/
[4] https://www.encodeproject.org/pipelines/

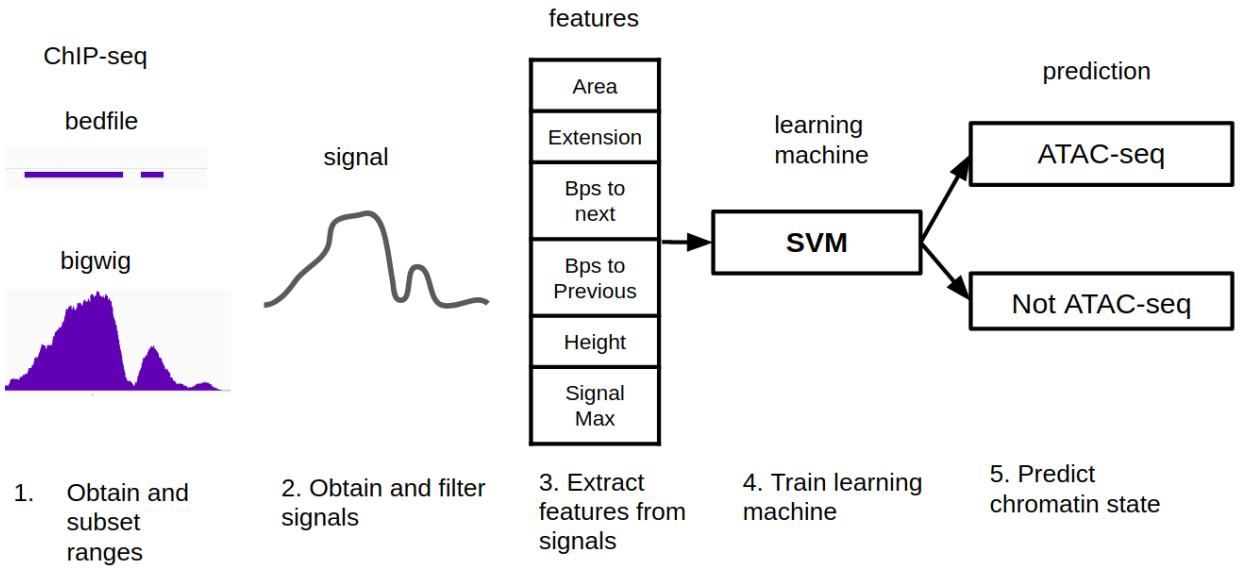

Figura 6. Esquema general secuencial del funcionamiento del modelo propuesto

conductor del lenguaje R se basa en el paquete GenomicRanges y paquetes asociados para acceder, representar y realizar análisis sobre rangos genómicos anotados ((Gentleman *et al.*, 2004), (Lawrence *et al.*, 2013)). Dado nuestro enfoque, nos basamos en esta infraestructura para permitir la adquisición del nivel de la señal por par de bases de un pico dado. Proponemos el paquete de R histoneSig [5] para leer, manipular y envolver pares de archivos bed/narrowpeak y bigwig en prácticos objetos *signalSet*, que facilitan las tareas posteriores aquí descritas. Un esquema general secuencial del proceso que se siguió para el modelo desarrollado puede ser visualizado en la Figura 6. Se utilizaron archivos “narrowpeak“ como punto de partida del modelo tanto para ChIP-seq como para ATAC-seq. Todos los intervalos genómicos del archivo Narrowpeak suministrados por este último se leyeron y envolvieron en un archivo *GRanges*, usando *import.np*, una función incluida en histoneSig, que es una extensión de Función *import* de GenomicRanges, publicada originalmente de forma no oficial por Beauparlant, C[6]. Los objetos *GRanges* resultantes fueron filtrados con la función de histoneSig *granges_to_chr* para conservar todos los cromosomas canónicos, con la excepción del cromosoma Y. Los intervalos provenientes del archivo BED o “narrowpeak“ de ATAC-seq fueron utilizados como referencia para etiquetar las observaciones del modelo como “Verdaderamente Positivo“ y “Verdaderamente Negativo“. El conjunto de datos ChIP-seq se dividió en dos clases: las que se traslapaban con los intervalos ATAC-seq y

[5] https://github.com/semibah/histoneSig
[6] https://charlesjb.github.io/How_to_import_narrowPeak/

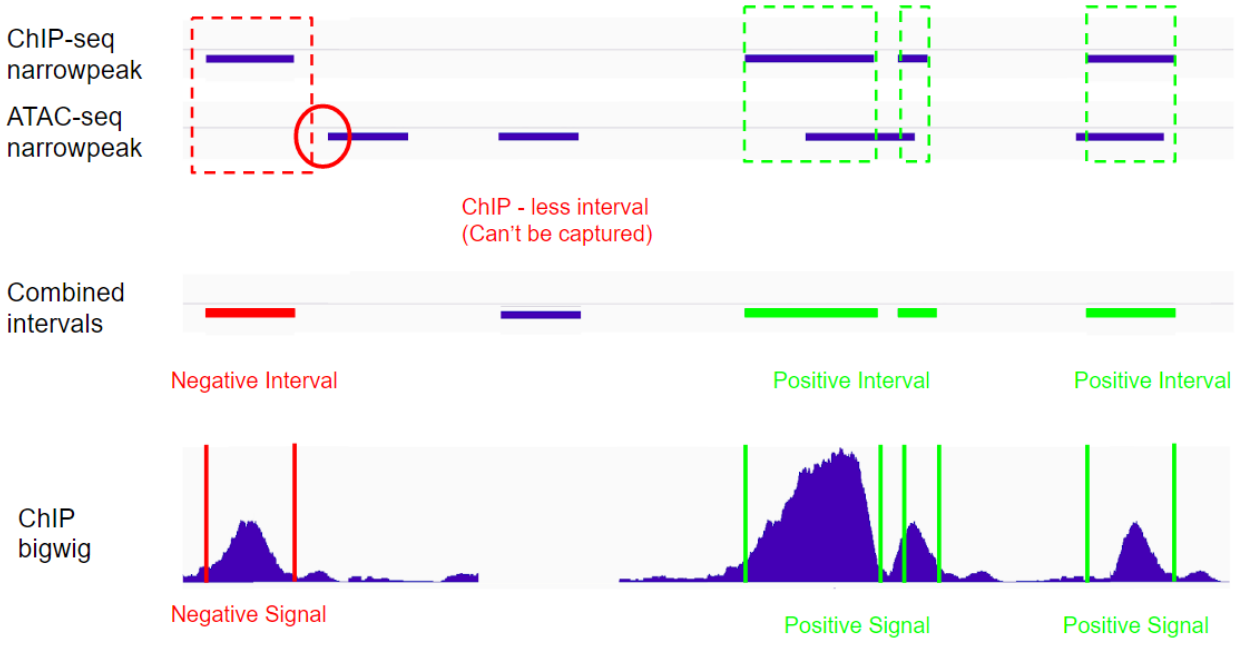

Figura 7. Esquema de la determinación de la etiqueta para los datos de señal.

las que no. Estos fueron etiquetados respectivamente como intervalos de ChIP ATAC-Positivo y ATAC-Negativo. Un esquema que detalla la determinación de la etiqueta puede ser visualizado en la Figura 7. Esta operación de traslape de rangos dio como resultado la generación del conjunto de intervalos regular.

Una vez obtenido el conjunto de intervalos regular, compuesto por los rangos de ChIP base ATAC-positivo y ATAC-negativo se extendieron los intervalos contenidos en el conjunto como prueba de concepto. Los intervalos existentes se ampliaron en 250 pares de base río arriba y río abajo; si se producía un traslape con un intervalo aledaño en menos de 250 pares de base, el intervalo compartido por ambos segmentos vecinos se dividía por la mitad. Cada una de estas mitades se añadieron a los intervalos vecinos. Las secciones añadidas río arriba y río abajo a cada intervalo se denominan “adaptadores“. Se obtuvieron las distribuciones de tamaño de los adaptadores añadidos. Después de añadir adaptadores a todos los intervalos de ChIP existentes, se volvieron a someter a prueba los intervalos ampliados para detectar superposiciones con intervalos ATAC. Si la adición de adaptadores resultó en una superposición previamente no detectada entre las regiones de ChIP y ATAC recientemente extendidas, las observaciones que anteriormente no mostraban traslapes, etiquetadas como “ATAC-negativas“ fueron re-etiquetadas como “ATAC-positivas“. La Figura 8 muestra un traslape recién generado, previamente ausente, en contraste con la Figura 7, que muestra el mismo

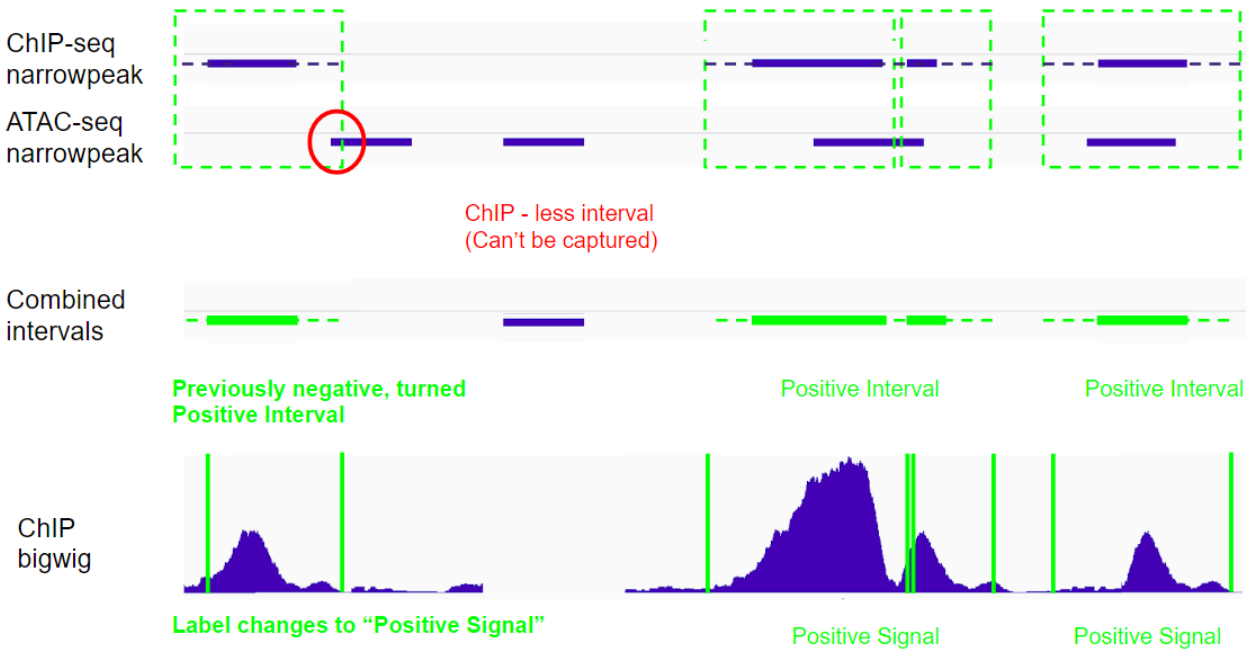

Figura 8. Reetiquetado de señales de negativo a positivo a medida que se genera un nuevo traslape después de la extensión.

segmento de señal sin extensión. Este proceso de extensión y traslape resultó en la generación del conjunto de intervalos extendido.

Una vez determinados ambos conjuntos de intervalos, la función de histoneSig *np_signals_from_bigwig*, se utilizó para acceder directamente al archivo bigwig de ChIP-seq asociado a los conjuntos de picos de ChIP-seq etiquetados como Positivo y Negativo para obtener valores específicos por posición de cada intervalo del conjunto en un formato continuo, generando lo que de ahora en adelante será designado como una "señal". Las señales, sus posiciones inicial y final, la anchura del intervalo analizado, la distancia hasta el pico anterior y el siguiente en la muestra, así como el cromosoma de origen fueron extraídos y envueltos en un objeto de tipo *signalSet*.

3.2. Preprocesamiento de señales

3.2.1. Filtrado

Dada la naturaleza ruidosa de las señales analizadas y debido a las restricciones de diseño y eficiencia computacional a nivel de extracción de características, las señales fueron filtradas antes de continuar su manipulación. La función *filter_signalSet* de histoneSig se utilizó para aplicar un filtro pasa bajas a todos los objetos *signalSet* correspondientes a los conjuntos de intervalos regulares y extendidos. El filtro subyacente utilizado por la función se encuentra en el paquete de R *stats* [7], que utiliza la ecuación 6. Donde y corresponde al valor resultante de la señal filtrada, x al valor de la señal, f al valor del filtro, i al paso de tiempo actual y p al tamaño de la ventana de tiempo utilizada.

$$y[i] = x[i] + f[1] * y[i-1] + \ldots + f[p] * y[i-p]. \tag{6}$$

Brevemente, la función de filtro base aplica un filtro de media móvil unilateral, en el que los coeficientes que se aplican por paso de tiempo como filtro se basan en la observación anterior en la señal que se va a filtrar. Dicha función de filtro toma dos argumentos, la señal y una ventana de filtro de tamaño p. Dado el tamaño variable de las señales que se encuentran en nuestros intervalos genómicos, se tomó un enfoque de tamaño de ventana fraccionaria. Todas las señales en los conjuntos de intervalos se dividieron en n partes iguales, donde n es el número suministrado a la opción *fractional* de la función *filter_signalSet*. El tamaño resultante por señal de cada pieza después de su segmentación se pasó al filtro como el tamaño de la ventana para calcular la media móvil, es decir, para una señal de tamaño 100, el tamaño de la ventana después de pasar un valor fraccionario de 25 sería 4. Nuestros análisis se llevaron a cabo con un valor fraccionario de 25. Un ejemplo de la señal antes y después del filtrado puede ser observado en la Figura 9.

[7] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/filter.html

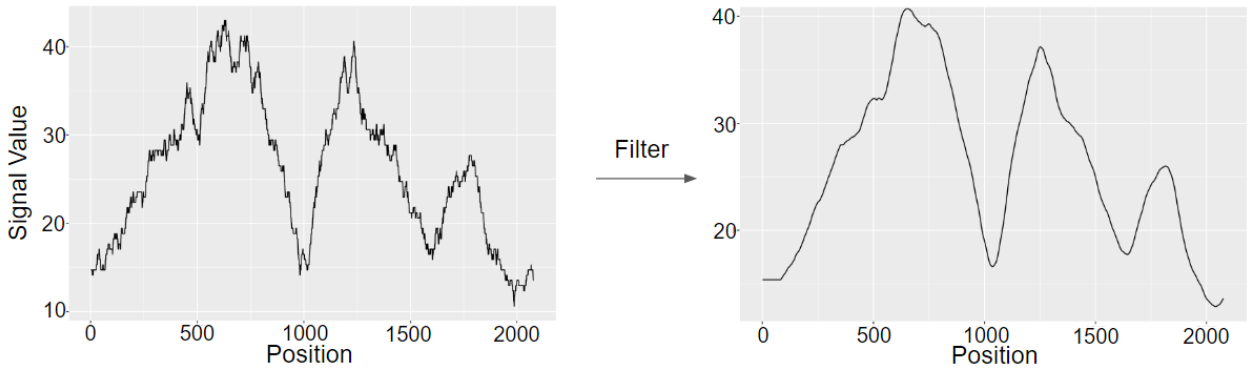

Figura 9. Representación funcional de la señal antes y después del filtrado con una señal fraccionaria con valor de filtro de 25.

3.3. Extracción de características

Diversas características específicas de señal fueron extraídas de los conjuntos de señales regulares y extendidos. Para cada intervalo de señal, se calcularon seis características. Las características descriptivas incluyen pares de base hasta el siguiente pico y pares de base hasta el pico anterior, nombradas *base_pairs_to_next_peak* y *base_pairs_to_previous_peak* en *histoneSig*, respectivamente. Ambas pueden ser vistas como una medida del aislamiento genómico de un intervalo dado; este se obtuvo calculando la distancia genómica en pares de base al intervalo de señal siguiente y anterior en cada conjunto, por cromosoma. La característica de máximo de señal o *signal_max* fue generada al obtener el valor máximo observado en cada intervalo de señal, que puede considerarse como una medida de enriquecimiento máximo por intervalo. Las últimas tres características, *extension*, *height* y *area* fueron extraídas de un cálculo de “valle“ con fundamento biológico, el cual será explicado en breve. Estos cálculos se realizan automáticamente suministrando a la función de histoneSig *base_features_from_signalSet* con un *signalSet*.

3.3.1. Cálculo del Valle

Al analizar diversos conjuntos de datos transcriptómicos de modificaciones histonales, se ha observado empíricamente que las regiones asociadas a modificaciones histonales vinculadas a la regulación activa, como es el caso del h3k27ac, muestran distintivos patrones de acoplamiento en regiones enriquecidas que se asemejan a “va-

lles“ capturados entre dos “picos“ en una señal dada (Kumar *et al.*, 2013a). Esto se debe al hecho de que los elementos reguladores, tales como los potenciadores y promotores, se vinculan a regiones que no están ligadas por nucleosomas, y por lo tanto se consideran “abiertas“ para el acoplamiento, lo cual resulta en un patrón distintivo de “pico-valle-pico“ al analizar las señales asociadas (Pundhir *et al.*, 2016). Por lo tanto, introducimos el concepto de calcular un “área“ entre dos picos de alto enriquecimiento o “señal“ para aproximar un evento de acoplamiento en cromatina previamente abierta; hipotetizamos que los intervalos de señal con un área pronunciada entre dos picos deben caracterizar correctamente un patrón de “pico-valle-pico“, indicando un evento de acoplamiento y, por asociación, cromatina abierta que fue acoplada éxitosamente. Dado el contexto anterior, después de filtrar todas las señales, se realizó un paso adicional para capturar los puntos más representativos de la señal “valle“ que se encontraban entre uno o dos “picos“ adyacentes. Todos los valles y picos dentro de la señal filtrada se calcularon con versiones modificadas de las funciones *findValleys* y *findPeaks* del paquete de R quantmod[8], las cuales esencialmente capturan todos los mínimos y máximos locales en una serie temporal, respectivamente. Dichas versiones modificadas se incluyen en histoneSig como un cálculo interno en la función *positions_from_signalSet*; estas modificaciones se implementaron para eliminar una restricción de las funciones originales, las cuales retornan el primer punto posterior a un pico o valle, con motivo de devolver la posición exacta del pico o valle. Los valles y picos se calcularon llamando a la función *positions_from_signalSet* dos veces y estableciendo su argumento *points* en *valley* y *peak*, respectivamente. Después, se hizo un cálculo simple y rectangular del área: se calculó la variable *extension* para todos los pares pico-valle como la distancia desde el valle detectado hasta el(los) pico(s) más cercano(s) en pares de base y la variable *length* se obtuvo restando el valor de la señal del (los) pico(s) adyacente(s) al valle desde el valor de la señal del valle. Una vez obtenidas ambas variables, se multiplicaron para obtener una ventana de área que contiene la mitad del valor del área del valle correspondiente, y el área bajo la curva entre el pico capturado y el valle más cercano. Las áreas obtenidas se redujeron a la mitad para aproximarse sólo al valle formado entre picos. En el caso de que un valle tuviera dos picos adyacentes, los valores calculados de *area* y *extension* fueron agregados para representar un solo valle. Los componentes relevantes del cálculo se

[8] https://cran.r-project.org/web/packages/quantmod/quantmod.pdf

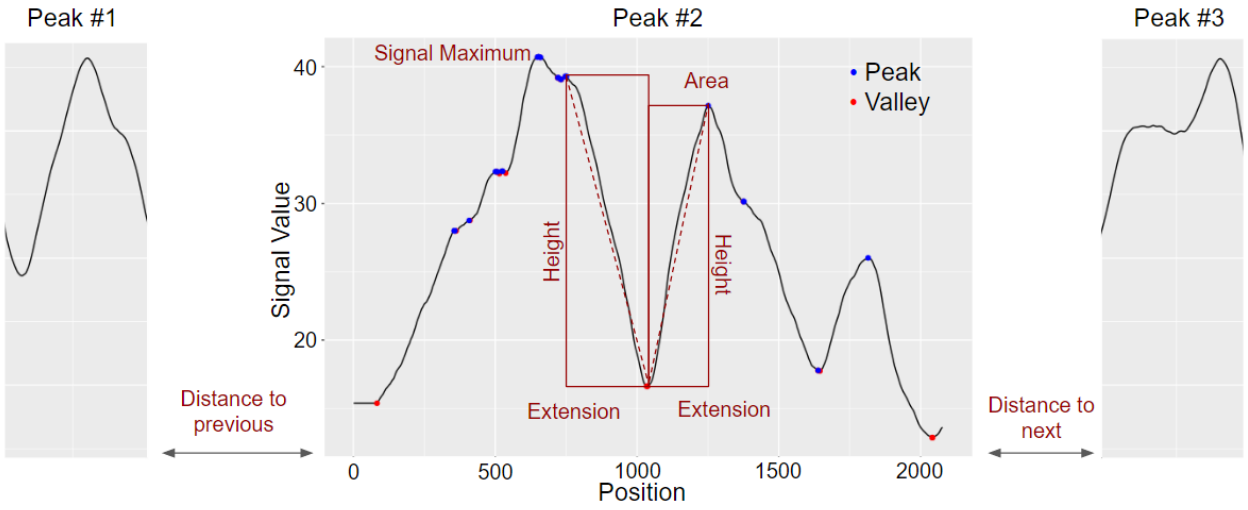

Figura 10. Diagrama que ilustra las características básicas actuales calculadas por la función *base_features_from_signalSet*

pueden visualizar en la Figura 10.

Una serie de parámetros de control fueron suministrados a *base_features_from_signalSet* para refinar aún más la abstracción del cálculo del valle. Dado que se suelen detectar múltiples valles por intervalo, se considera que el que tiene el área mayor es el valle más representativo de ese intervalo, lo cual se especificó configurando la opción *max_area_filter* con un valor de *TRUE*. En el caso dado de que la posición genómica del valle fuera necesaria para análisis río abajo, el parámetro *section* se estableció como *"valley"*, para mantener el punto exacto en el que se detectó el valle, a diferencia de mostrar el intervalo que abarca el valle detectado y su(s) pico(s) adyacente(s).

3.4. Selección de características

Las características se evaluaron en función de su capacidad para discernir entre observaciones positivas y negativas para ATAC tanto para los conjuntos regulares como extendidos. Se realizaron análisis exploratorios de cromosomas con las características extraídas para determinar si alguna característica en particular era suficiente para distinguir entre los intervalos de ChIP ATAC-positivo y ATAC-negativo. Para probar la significancia estadística de la diferencia entre las medianas para los grupos positivos y negativos, se realizó una prueba de suma de clasificación de Wilcoxon en cada una de las variables extraídas utilizando la función *wilcox.test* del paquete stats de R. Se realizaron pruebas de dos lados, no emparejadas y se calcularon intervalos de confian-

za de 0,95 por conjunto.

Debido a las limitaciones computacionales, los análisis de la selección de características se limitaron a los cromosomas 1, 8 y 21. Los resultados de estos análisis fueron luego extrapolados a las variables incluidas en la máquina de aprendizaje final la cual utilizó los cromosomas canónicos humanos completos, a excepción del Y. Dos variantes de eliminación recursiva de características (RFE) fueron empleados: RFE iterativo y RFE combinado con una SVM lineal (SVM-RFE). La función RFE del paquete R *caret* se utilizó para implementar ambos (Kuhn *et al.*, 2008). Dicho paquete funciona asignando la función *rfeControl* a una variable con parámetros que especifican el tipo y el tamaño de la variable, además de características del método de remuestreo que se empleará, entre otras. Se seleccionó la validación cruzada de 5 iteraciones tanto para la versión iterativa como para la versión RFE-SVM. Un parámetro adicional que se debe especificar es el argumento *functions*, que proporciona a cualquier modelo posterior basado en el control previamente definido con funciones de ayuda para definir cómo se deben implementar el ajuste del modelo, la predicción y la importancia de las variables; esta variable se estableció a *caretFuncs*, que es un conjunto de funciones que se utilizan cada vez que se realiza un procedimiento de forma iterativa, como es el caso en la eliminación recursiva de características [9].

La selección iterativa de características se empleó para determinar la variable predictiva independiente con mayor relevancia. Esto se logra evaluando variables sobre múltiples modelos, eliminando variables secuencialmente, comenzando con el subconjunto completo del predictor. El paquete *caret* implementa este procedimiento al llamar a la función *rfeIter*. Las variables fueron evaluadas con un modelo de bosque aleatorio construido para cada uno de los subconjuntos generados durante el proceso de eliminación de variable. *caret* llama automáticamente al paquete de R *randomForest* para este procedimiento de evaluación. El procedimiento de clasificación se puede encontrar en la documentación del paquete R *randomForest* [10]. En resumen, para cada modelo de clasificación de bosque aleatorio, la precisión de la predicción se registra en los datos que están fuera de la muestra validada cruzada actual. En los datos de la muestra, la precisión se registra después de generar nuevas

[9] https://topepo.github.io/caret/recursive-feature-elimination.html

[10] http://math.furman.edu/~dcs/courses/math47/R/library/randomForest/html/importance.html

permutaciones con las variables predictoras y se obtienen las diferencias de éstas dos precisiones y luego se promedian sobre todos los modelos de árbol generados y se normalizan mediante el error estándar.

Posteriormente, se empleó RFE-SVM para determinar cuáles de las características dadas en conjunto conducirían al modelo de predicción óptimo. Este algoritmo emplea directamente la selección voraz hacia atrás, tratando iterativamente de elegir las características que llevarán al mayor margen de separación de clases en el clasificador SVM vinculado (Guyon *et al.*, 2002). Este problema combinatorio se resuelve en cada iteración del entrenamiento eliminando de forma voraz la dimensión de entrada que disminuye el margen de separación de clases, al menos hasta que sólo queden las dimensiones de entrada que llevarían a una separación óptima alcanzable. Esto se implementó llamando a la función *rfe* de R *caret* con la variable *rfeControl* previamente definida, mientras que la opción *method* de *rfe* se estableció en *svmLinear*. Los subconjuntos de variables se clasificaron con base en el área bajo la curva ROC (AuROC) para cada uno de los modelos lineales de SVM generados.

3.5. Clasificación

3.5.1. Preprocesamiento de datos

Una vez determinadas las variables relevantes, se prepararon los datos para su clasificación. Para facilitar el cálculo y mejorar el rendimiento en el caso de los clasificadores lineales, los datos fueron estandarizados según escala antes del entrenamiento con la función *scale* de R *base* (Hsu *et al.*, 2003). Dicha función calcula la media y desviación estándar de cada uno de los vectores de características y luego procede a restar la media y dividir cada uno de los elementos del vector por la desviación estándar asociada del vector. Los datos de los conjuntos extendidos y regulares se dividieron con la función *createDataPartition* del paquete *caret* en conjuntos de entrenamiento y prueba, que comprenden el 80 % y el 20 % del total de las observaciones, respectivamente. La función anterior genera las particiones mediante un muestreo aleatorio que intenta conservar la distribución y proporción de los diferentes factores de respuesta

y de la muestra original en la nueva partición generada; en nuestro caso, positivo y negativo.

3.5.2. SVM

Después del preprocesamiento, se entrenaron múltiples máquinas de aprendizaje sobre los datos resultantes. El paquete R *e*1071 fue usado para entrenar todas las máquinas de soporte vectorial (Meyer *et al.*, 2019). Los SVMs de funcion lineal y radial fueron entrenados en los datos obtenidos de los conjuntos de datos de señales regulares y extendidas. Se mantuvieron todos los parámetros predeterminados. Una vez que se obtuvieron los objetos de SVM correspondientes, las predicciones se llevaron a cabo utilizando la función *predict* de R *stats*. Se realizaron dos tipos de predicciones. En el primer conjunto, tanto los SVMs entrenados en los conjuntos regulares y extendidos fueron suministrados con su correspondiente conjunto de prueba para predicciones. Un segundo experimento se llevó a cabo para probar si predecir con una máquina de aprendizaje entrenada en las señales extendidas aumentaba el rendimiento del clasificador cuando se utilizaban datos extendidos. Una predicción adicional fue hecha usando el conjunto de prueba de señal extendida con el SVM entrenado en señales regulares. Se calcularon entonces múltiples métricas de evaluación sobre las predicciones obtenidas. Las matrices de confusión se obtuvieron con la función *confusionMatrix* de R *caret*, así como métricas asociadas como precisión, sensibilidad y especificidad. Además, la función *MCC* del paquete R *mltools* se utilizó para calcular el coeficiente de correlación de Matthew (MCC) [11].

[11] https://cran.r-project.org/web/packages/mltools/index.html

Capítulo 4. Resultados

La serie de experimentos aquí propuestos tiene el objetivo de responder a las siguientes preguntas:

1. ¿Cuántos de los segmentos de ATAC-seq pueden ser capturados por segmentos de ChIP-seq, tanto en su versión original y extendiendo los segmentos originales de ChIP-seq un máximo de 250 pares de base en cada extremo?
2. ¿Cuál es el porcentaje de señales de ChIP-seq que no coinciden con intervalos de ATAC-seq?
3. ¿Cuál es la distribución de las extensiones agregadas a ambos extremos de segmentos de ChIP-seq, en pares de base?
4. ¿Qué tan efectivas son las características propuestas en la sección 3.3 para identificar los segmentos de ATAC-seq positivos, individualmente y en conjunto?
5. ¿Existe una diferencia en desempeño entre usar solo los segmentos regulares de ChIP-seq, versus el uso de los intervalos extendidos?

A continuación se describen los resultados dirigidos a responder secuencialmente cada una de estas preguntas.

4.1. Datos de los Intervalos

Posterior al análisis de los conjuntos de datos de H3K27ac ChIP-seq y ATAC-seq, con números de acceso del portal ENCODE ENCSR778NQS y ENCSR220ASC, se obtuvieron 115,352 y 66,883, respectivamente. Con excepción de la cromosoma Y, se contemplaron todas las cromosomas canónicas: los pares 1-22 y la X. El número de picos en las muestras tanto positivas como negativas con ATAC para los conjuntos de datos regulares y extendidos que resultaron de las operaciones de traslape positivas y negativas se pueden observar en la Tabla 1.

Tabla 1. Cuantificación basal de picos

	Regular	Extendido
ATAC-positivo	41,284	45,743
ATAC-negativo	74,068	69,609

La Tabla 2 contiene dicha cuantificación de picos en porcentajes con respecto a la muestra completa. Adicionalmente, dicha tabla ilustra la relación entre el total de los picos de ATAC que traslapan con nuestros picos de ChIP ATAC-positivos y el total de los picos de ATAC.

Tabla 2. Cuantificación de picos y máximo teórico de intervalos recuperables por el clasificador en porcentaje de muestra

	Regular	Extendido
ATAC-positivo	36.57 %	39.66 %
ATAC-negativo	64.21 %	60.35 %
% de Traslape ATAC-en-P-ChIP/ATAC (Cota Superior)	62.48 %	65.16 %

4.2. Análisis de extensiones

Se cuantificaron los segmentos, o adaptadores, que se añadieron a ambos extremos de los intervalos de ChIP por cada extensión. La Figura 11 ilustra un histograma de la distribución de tamaño de los adaptadores, hasta un máximo de 250 pares de bases por lado.

Un histograma similar que detalla los adaptadores con longitudes menores a 250 pares de base puede observarse en la Figura 12.

Se hicieron extensiones por ambos lados de los intervalos de ChIP analizados. La Figura 13 muestra las densidades correspondientes para las extensiones realizadas tanto del lado izquierdo como en el derecho de los intervalos extendidos de ChIP positivos y negativos con ATAC en la muestra.

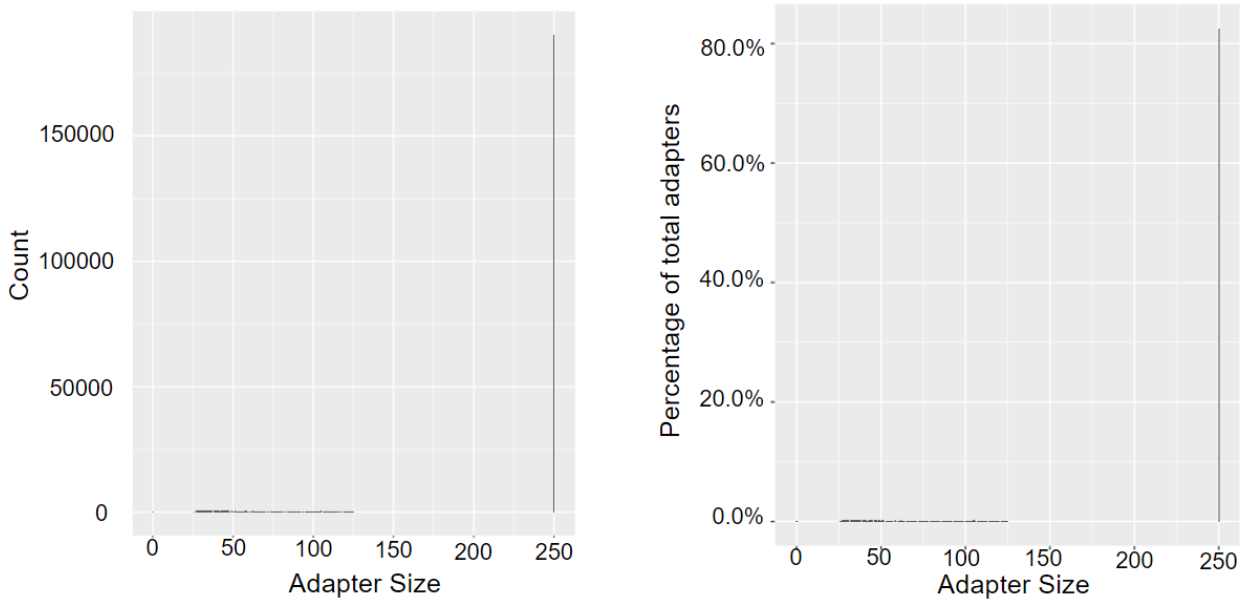

Figura 11. Histogramas que muestran la distribución del tamaño del adaptador agregado en los intervalos de ChIP extendidos en frecuencia y porcentaje de muestra.

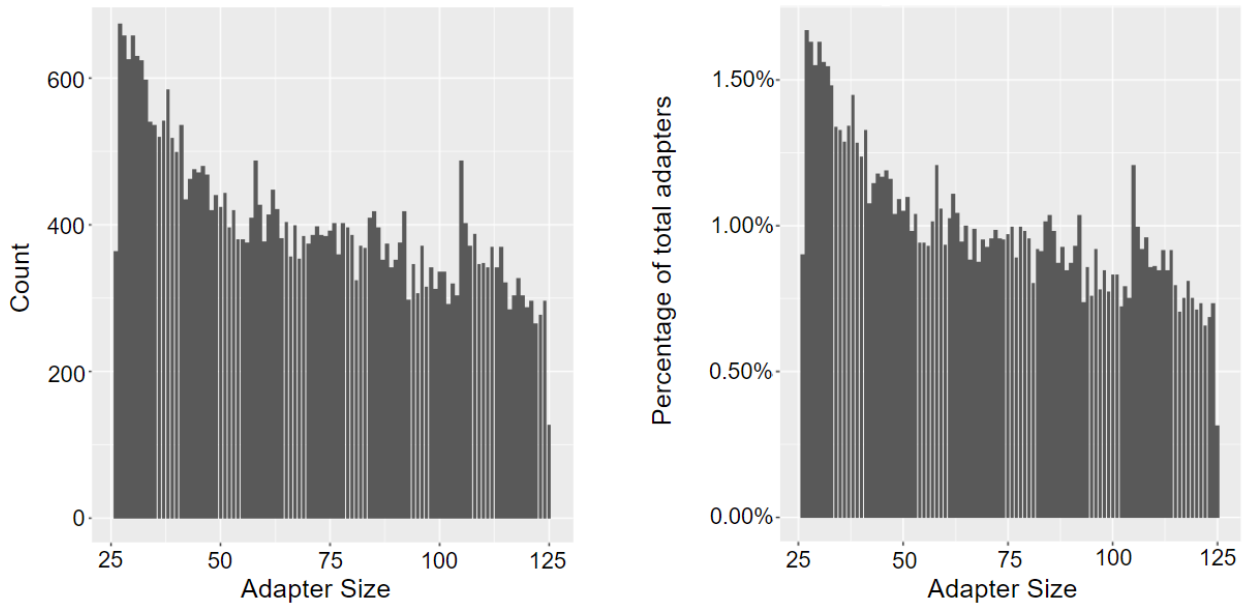

Figura 12. Histogramas que muestran la distribución del tamaño del adaptador agregado para adaptadores con longitud menor a 250 pares de base en los intervalos ChIP extendidos en frecuencia y porcentaje de muestra.

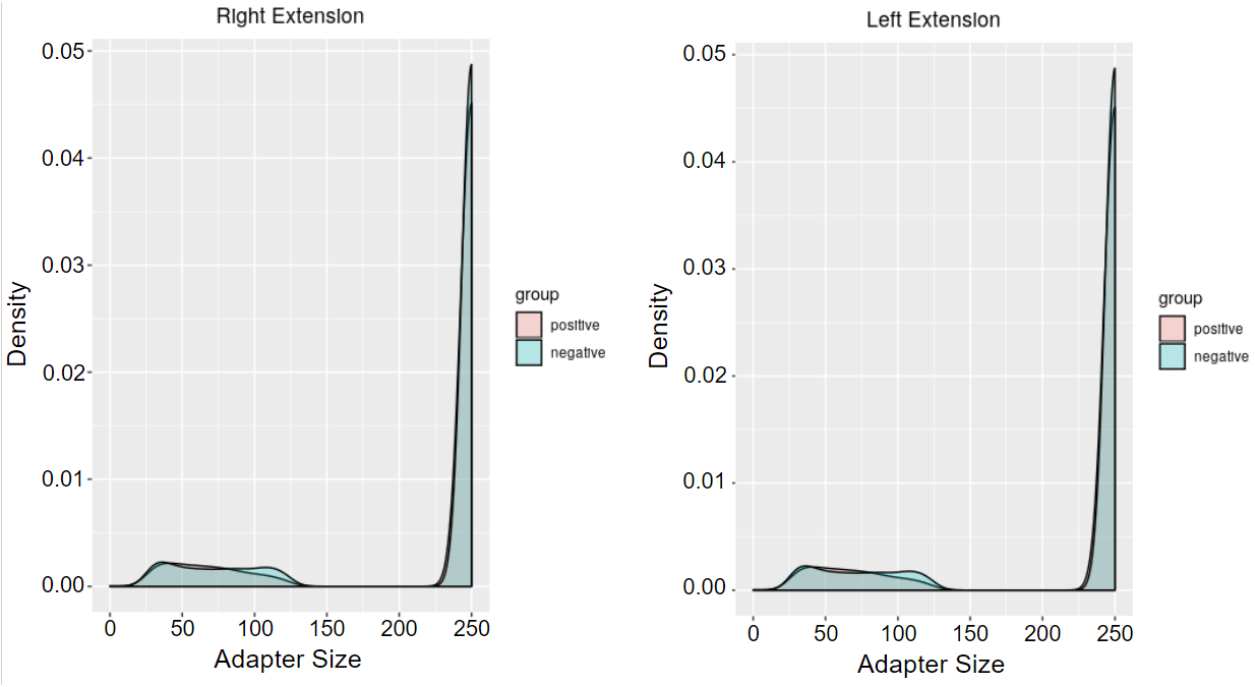

Figura 13. Gráficos de densidad de los adaptadores que contrastan por orientación a los intervalos extendidos de ChIP con traslapes positivos y negativos con ATAC.

4.3. Selección de características

El análisis exploratorio por diagrama de caja realizado con los conjuntos negativos y positivos de las características: máximo de señal, altura, extensión, área, pares de base al siguiente pico y pares de base al pico previo para los cromosomas 1, 11 y 21 puede ser visualizado en la Figura 14. Se aplicó una transformada logarítimica en el eje vertical a las variables área, pares de base al siguiente pico y pares de base al pico previo, mientras que para las variables altura, extensión y máximo de señal, los ejes permanecieron sin modificación. Esto se hizo debido a que las variables que fueron transformadas mostraron una diferencia de gran magnitud entre los valores obtenidos. Adicionalmente, se realizó un análisis preliminar de componentes principales. Las gráficas de pares y composiciones asociadas de los primeros tres componentes obtenidos pueden ser visualizados en la Figura 15.

Las pruebas de Wilcoxon presentaron un valor p inferior a 2.2^{-16} para el conjunto completo de las seis características, tanto para los conjuntos regulares como para los extendidos, al probar las diferencias de grupo entre los intervalos ChIP negativos y positivos. La Tabla 3 contiene los intervalos de confianza calculados para la prueba.

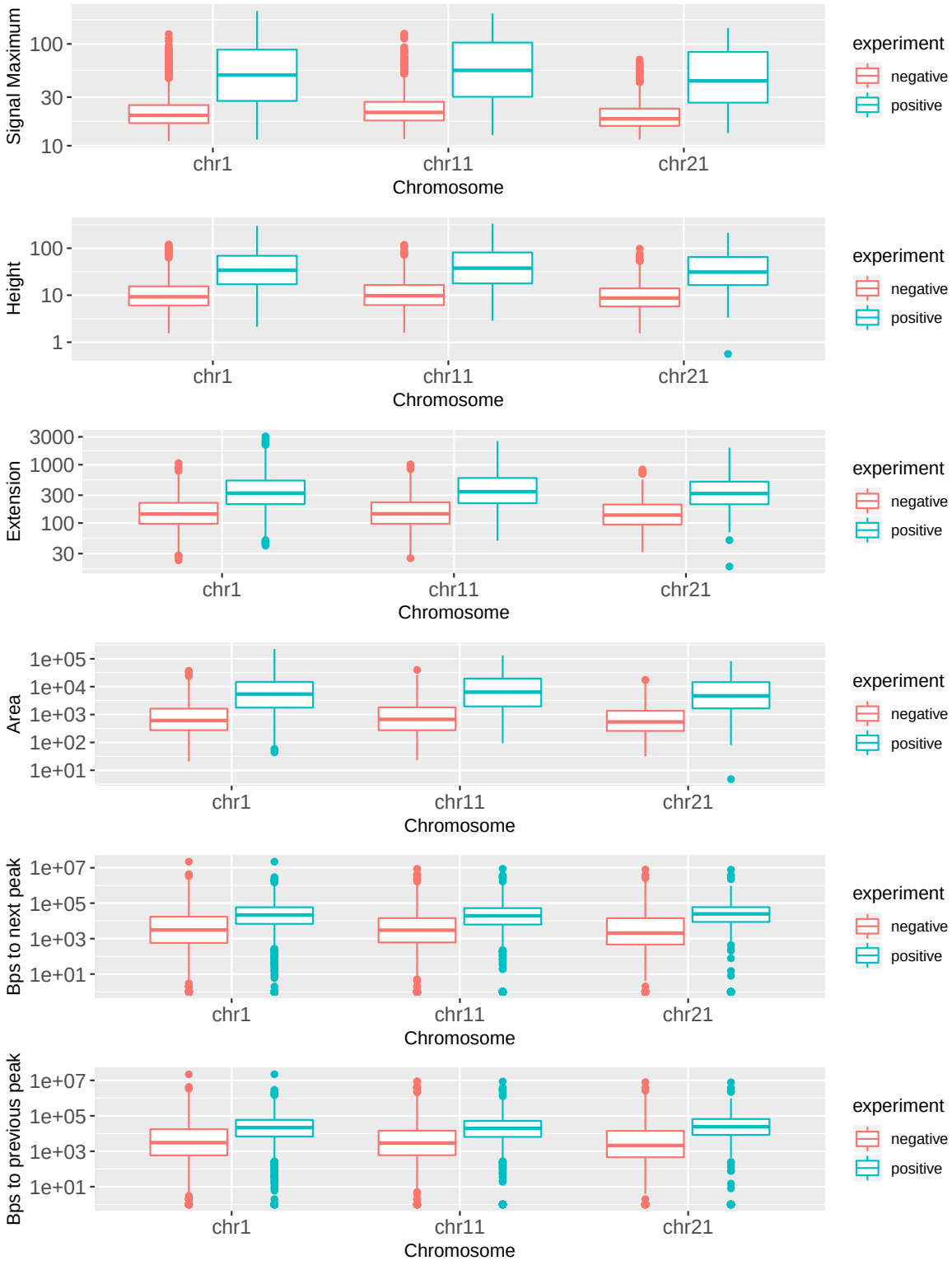

Figura 14. Diagramas de caja que contrastan los conjuntos de intervalos de ChIP positivos y negativos con ATAC de las características máximo de señal, altura, extensión, área, pares de base al siguiente pico y pares de base al pico previo para los cromosomas 1, 11 y 21.

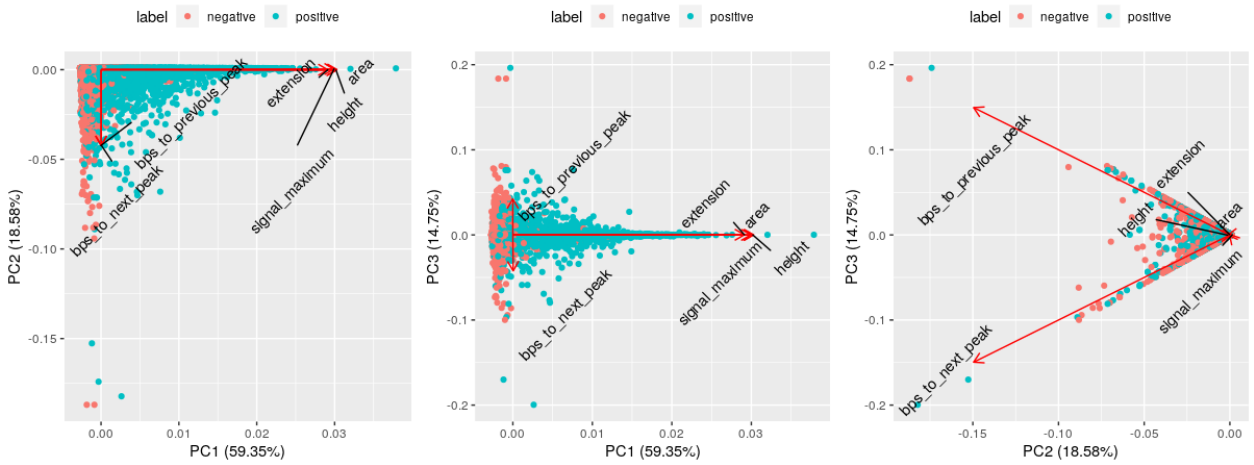

Figura 15. Gráficas de pares de los primeros tres componentes principales y sus composiciones obtenidos posterior a un análisis de componentes principales del conjunto de datos de seis características extraídas del conjunto de señales regulares

Tabla 3. Intervalos de confianza del 95 % obtenidos después de realizar una prueba de Wilcoxon entre los conjuntos de ChIP positivos y negativos con ATAC para las seis características.

	Regular	Extendido
Extensión	[168, 172]	[168, 173]
Máximo de Señal	[27.50,28.31]	[33.32, 34.24]
Altura	[1.22, 1.24]	[1.74, 1.77]
Área	[1.99, 2.02]	[3.21, 3.26]
Bps al siguiente	[1.81, 1.85]	[1.67, 1.74]
Bps al previo	[1.81, 1.86]	[1.67, 1.74]

Los resultados del procedimiento de RFE iterativo realizado en el cromosoma 1 para los conjuntos de señales regulares y extendidas se pueden observar en las tablas 4 y 5, respectivamente.

La validación cruzada de 5 pliegues RFE-SVM lineal realizada en el cromosoma 1 dio como resultado las AuROC reportadas en la Tabla 6.

4.4. Clasificación binaria con máquinas de soporte vectorial

Las matrices de confusión para los SVMs Lineales y de RBF para los conjuntos de entrenamiento y prueba se pueden ver en las figuras 16 y 17, respectivamente.

Tabla 4. Orden de eliminación de variable e importancia de la característica de permutación (PIR) obtenida después de realizar RFE iterativo en señales regulares del cromosoma 1.

	Orden de eliminación de variable	PIR General
Máximo de Señal	6	1052.06
Altura	5	761.22
Área	4	621.76
Extensión	1	437.54
Bps al siguiente	2	512.45
Bps al previo	3	521.17

Tabla 5. Orden de eliminación de variable e importancia de la característica de permutación (PIR) obtenida después de realizar RFE iterativo en señales extendidas del cromosoma 1.

	Orden de eliminación de variable	PIR General
Máximo de Señal	6	1060.31
Altura	5	778.20
Área	4	640.24
Extensión	1	426.21
Bps al siguiente	3	581.16
Bps al previo	2	565.70

4.4.1. Predicciones de señal de los conjuntos regulares y extendidos

Las tablas 7, 8, 9 y 10 contienen las métricas de rendimiento de precisión, MCC, sensibilidad y especificidad obtenidas a partir de los SVMs Lineales y de función de base radial tanto para los conjuntos de entrenamiento como para los de prueba.

Los conjuntos de entrenamiento y prueba para los modelos de predicción se obtuvieron mediante la segmentación aleatoria de los conjuntos de datos que se obtuvieron al extraer el total de picos de los experimentos ENCSR778NQS y ENCSR220ASC, y seleccionando los cromosomas 1-22 y X. Las pruebas realizadas de selección de variable fueron efectuadas solamente para los cromosomas 1, 8 y 22. No obstante, para los procedimientos asociados a la selección de variables, se muestran solamente los resultados del cromosoma 1. Los resultados de las cromosomas 8 y 21 pueden consultarse en el Anexo.

Tabla 6. AuROC obtenida utilizando el método lineal RFE-SVM en señales regulares y extendidas del Cromosoma 1.

Número de variables	Regular	Extendido
1 (MS)	0.83	0.80
2 (MS, AL)	0.83	0.80
3 (MS, AL, AR)	0.83	0.80
4 (MS, AL, AR, BpsS)	0.83	0.80
5 (MS, AL, AR, BpsS, BpsN)	0.83	0.80
6 (MS, AL, AR, BpsS, BpsN, EX)	0.83	0.80

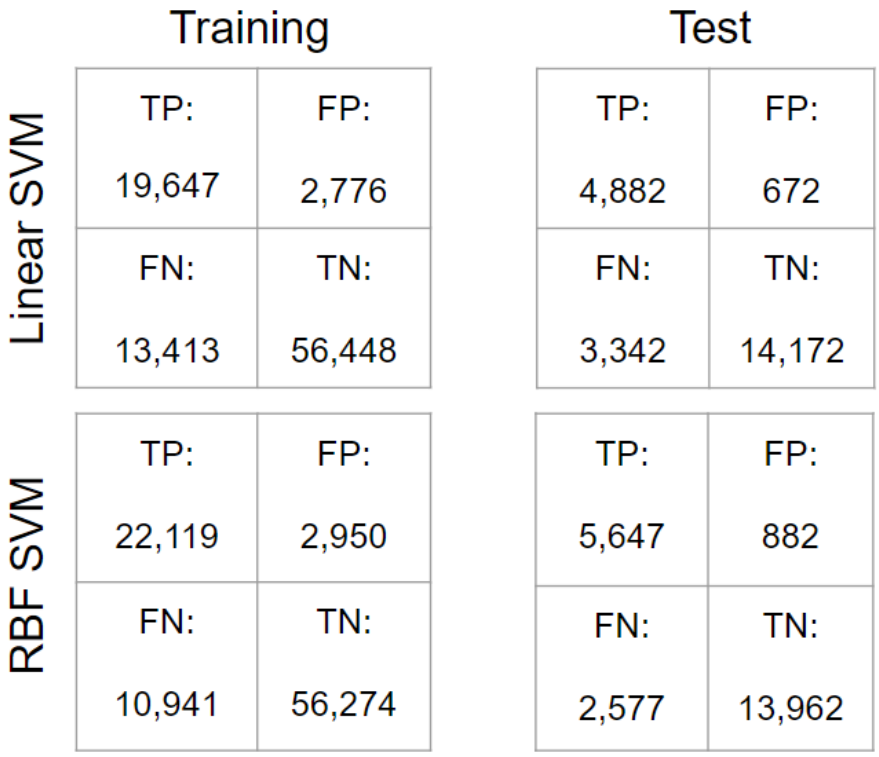

Figura 16. Matrices de confusión del conjunto de entrenamiento y de prueba para los SVMs Lineales y de función de base radial entrenados en el conjunto de señales regular.

4.4.2. Predicciones Cruzadas

La Tabla 11 y la Tabla 12 contienen las cuatro métricas de desempeño antes mencionadas que se obtuvieron al utilizar el SVM entrenado en el conjunto de señales extendidas para comparar el rendimiento de dicha máquina entre las predicciones realizadas cuando se utiliza el conjunto regular o el extendido.

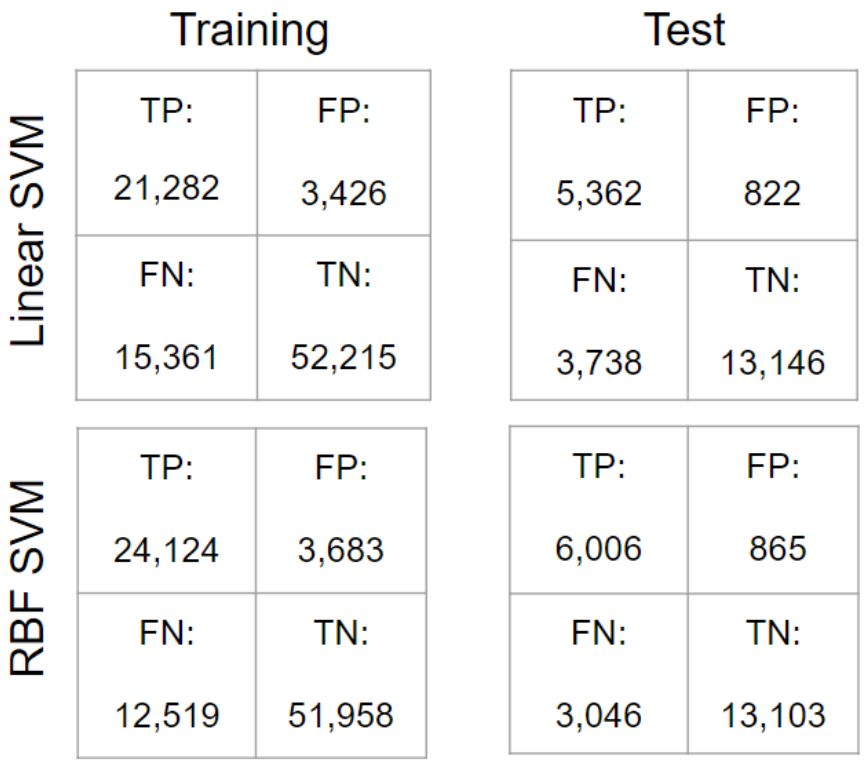

Figura 17. Matrices de confusión del conjunto de entrenamiento y de prueba para los SVMs Lineales y de función de base radial entrenados en el conjunto de señales extendido.

Tabla 7. Precisión de las predicciones efectuadas por los SVM lineales y de función de base radial de los conjuntos de entrenamiento y de prueba

Señal	SVM Lineal (Entrenamiento)	SVM Lineal (Prueba)	SVM RBF (Entrenamiento)	SVM RBF (Prueba)
Regular	82.46 %	82.60 %	84.95 %	85.01 %
Extendido	79.64 %	80.23 %	82.44 %	82.84 %

Tabla 8. Coeficiente de Correlación de Matthew de las predicciones efectuadas por los SVM lineales y de función de base radial de los conjuntos de entrenamiento y de prueba

Señal	SVM Lineal (Entrenamiento)	SVM Lineal (Prueba)	SVM RBF (Entrenamiento)	SVM RBF (Prueba)
Regular	0.62	0.61	0.67	0.67
Extendido	0.57	0.59	0.63	0.64

Tabla 9. Sensibilidad de las predicciones efectuadas por los SVM lineales y de función de base radial de los conjuntos de entrenamiento y de prueba. Las filas adicionales indican la cantidad de nuevos “Verdaderos Positivos“ incorporados por cada clasificador en los casos regulares y extendidos.

Señal	SVM Lineal (Entrenamiento)	SVM Lineal (Prueba)	SVM RBF (Entrenamiento)	SVM RBF (Prueba)
Regular	59.43 %	59.36 %	66.91 %	68.66 %
Extendido	58.08 %	58.92 %	65.84 %	66.0 %
Regular / TP	19,647	4,882	22,119	5,647
Extendido / TP	21,282	5,362	24,124	6,006

Tabla 10. Especificidad de las predicciones efectuadas por los SVM lineales y de función de base radial de los conjuntos de entrenamiento y de prueba.

Señal	SVM Lineal (Entrenamiento)	SVM Lineal (Prueba)	SVM RBF (Entrenamiento)	SVM RBF (Prueba)
Regular	95.31 %	95.47 %	95.02 %	94.06 %
Extendido	93.84 %	94.12 %	93.38 %	93.81 %

Tabla 11. Métricas de rendimiento obtenidas después de utilizar un conjunto de prueba de señal regular para predicciones en SVMs entrenados en señales extendidas.

Núcleo de la SVM	Lineal (Prueba)	RBF (Prueba)
Precisión	79.14 %	81.45 %
MCC	0.56	0.61
Sensibilidad	55.76 %	60.76 %
Especificidad	94.12 %	94.72 %

Tabla 12. Métricas de rendimiento obtenidas después de utilizar un conjunto de prueba de señal extendida para predicciones en SVMs entrenados en señales extendidas.

Núcleo de la SVM	Lineal (Prueba)	RBF (Prueba)
Precisión	80.23 %	82.84 %
MCC	0.59	0.64
Sensibilidad	58.92 %	66.0 %
Especificidad	94.12 %	93.81 %

Capítulo 5. Discusión

El primer paso en la creación del modelo fue desarrollar una abstracción que pudiera ser utilizada de manera confiable para aproximar la predicción de cromatina abierta. Se eligieron datos de marca histonal porque el acoplamiento dentro de las regiones abiertas depende en gran medida de la accesibilidad cuantitativa del ADN y del enriquecimiento de las marcas de cromatina, como H3K27ac (Grossman *et al.*, 2017). Extender los picos de ChIP-seq podría permitir la detección de regiones de ATAC-seq que previamente no traslapaban, si los eventos de acoplamiento en el intervalo original de ChIP-seq fueran “centrados“ de forma relativa a las posiciones de los nucleosomas. Posteriormente a la extensión, se encontraron 4,459 superposiciones del ATAC-seq no detectadas previamente con los intervalos de ChIP-seq, lo que justifica la extensión.

La segmentación de nuestro conjunto de datos ChIP-seq original en ATAC-Positivo y ATAC-Negativo produjo un desequilibrio entre la muestra positiva y la negativa: para la muestra regular sin extensiones, los conjuntos positivos y negativos representaron el 36,57 % y el 64,21 % de la muestra total, respectivamente. Extender todos los intervalos de secuencias de ChIP a una distancia de 250 pares de bases o hasta que los intervalos “se unieron“ entre sí, tanto río arriba como río abajo resultó en conjuntos positivos y negativos de 39.66 % y 60.35 % del total de la muestra, respectivamente. Esto puede verse como un desplazamiento del balance de la muestra del 3.1 % hacia la muestra positiva. Hablando meramente de porcentajes esto puede parecer no significativo; no obstante, cuando se toma dentro del contexto de picos únicos capturados, el resultado es favorable. Los cromosomas 1 a 22 y X fueron contemplados para los siguientes análisis. La muestra regular de 41,284 intervalos ChIP ATAC-positivos traslapó con 41,792 intervalos únicos de ATAC, mientras que la muestra extendida de 45,743 ChIP ATAC-positivos traslapó con 43,581 intervalos únicos de ATAC. Esto corresponde a 1,789 traslapes únicos con intervalos de ATAC-seq que antes no estaban contemplados por nuestro clasificador. Consulte el segundo intervalo ATAC-seq de la Figura 7 para un ejemplo visual. Además, si un intervalo de ATAC-seq no tiene un intervalo ChIP-seq asociado con cual traslaparse, nuestro clasificador no será capaz de predecir dicho intervalo, ya que está entrenado con datos ChIP-seq. Esto impone una cota superior a la cantidad de intervalos ATAC-seq que pueden predecirse a partir de un conjunto de

intervalos ChIP-seq. La última fila de la Tabla 2 ilustra el porcentaje total de picos de ATAC que pueden capturarse con éxito en un caso ideal, 62.48 % cuando se utiliza el ChIP-seq regular y 65.16 % cuando se utiliza el conjunto de ChIP-seq extendido para entrenamiento. Un clasificador libre de errores sería capaz de capturar como máximo dicho porcentaje de los picos de ATAC en la muestra. El hecho de que los experimentos de ChIP-seq no tengan una correspondencia del 100 % con un experimento de ATAC-seq, también puede interpretarse como un acoplamiento que no ocurre en regiones de cromatina abierta o la antes mencionada asincronía de eventos de acoplamiento o maquinaria biológica no contabilizada entre los experimentos.

La extensión de los intervalos de ChIP-seq fue en su mayor parte homogénea. Aproximadamente el 82.5 % de todos los intervalos de ChIP-seq no tenían otras regiones de ChIP en proximidad y se sometieron a una extensión de 250 pares de bases tanto río arriba como río abajo. El resto de los tamaños de subconjunto se distribuyeron uniformemente en el intervalo de 25 a 125 pares de bases, como se ilustra en la Figura 12. Al considerar la orientación de las extensiones en muestras positivas y negativas tanto para las extensiones ascendentes como para las descendentes, o a la derecha y a la izquierda, como se considera en las figuras, no hubo un cambio aparente en la distribución de tamaños de par de base. Sin embargo, los intervalos positivos muestran una representación ligeramente más alta que sus contrapartes negativas en el intervalo de 25-75 pares de base como se observa en la Figura 13. Esto es consistente con la distribución de los sitios de unión de factores de transcripción (TFBS) en eucariontes; la longitud promedio en las regiones potenciadoras (enhancers) es de 6-12 pares de bases (Tugrul *et al.*, 2015). Los eventos sucesivos de acoplamiento en un intervalo de cromatina abierta explicarían extensiones de pares de bases inferiores a 125 para el caso ATAC positivo, si hubiera múltiples eventos de acoplamiento presentes en un único intervalo de cromatina abierta, y cada uno de ellos estuviera representado por los intervalos ChIP correspondientes, podría esperarse razonablemente un espaciamiento de pares de bases inferior a 125 entre estos intervalos.

Una vez obtenidos y ampliados los intervalos, se realizó la extracción de características. Las características extraídas se pueden separar en dos grupos: métricas de intervalo asociadas al enriquecimiento en la muestra y métricas descriptivas de distancia genómica. Las métricas de distancia genómica pueden ser vistas como una

medida del aislamiento de un intervalo en el genoma y la extensión del evento de acoplamiento. El primero se compone de dos características, pares de bases al siguiente pico y pares de bases al pico anterior, mientras que el segundo, la extensión, es la longitud del par de bases de un valle detectado en cada intervalo. Las métricas asociadas al enriquecimiento se componen de máximo de señal, área y altura. El primero es el valor máximo observado de la señal por intervalo. La altura corresponde al valor más alto de la señal en el valle detectado y el área es un cálculo de área basado en la altura y la extensión, como se describe en la sección 3.3.1.

Partiendo de la noción anterior de que las muestras positivas y negativas tienen valores suficientemente diferentes en las distintas características para hacer posible el discernimiento entre estos dos grupos, se realizó un análisis exploratorio de los datos con las características obtenidas. Se realizaron análisis superficiales con diagramas de caja, de los que se puede observar un ejemplo en la Figura 14. Dichos análisis mostraron que las muestras positivas y negativas parecen estar separadas de forma fiable tanto por las métricas de distancia genómica como por las métricas de enriquecimiento. Sin embargo, estas últimas mostraron una separación consistentemente mayor entre los valores medios. Una prueba de Wilcoxon, además de un análisis preliminar de PCA, el cual puede ser visualizado en la Figura 15, confirmaron estas nociones. Mientras que todos los valores *p* estaban muy por debajo del nivel de umbral de significancia, los intervalos de confianza demostraron ser informativos con respecto a la separación de grupos. Los valores relacionados con el enriquecimiento demostraron tener intervalos de confianza con una escala apropiada para la separación de grupos, en particular, el máximo de la señal y el área. Estos difirieron significativamente entre los grupos Negativo y Positivo, lo que se reflejó en los intervalos de confianza obtenidos. Con la excepción de la extensión, las métricas de distancia genómica mostraron un cambio en el intervalo de 1 ó 2 pares de bases entre positivo y negativo, lo cual no resulta significativo dada la variabilidad del tamaño del intervalo, incluso dentro de los grupos negativo y positivo, que pueden ser de hasta cuatro órdenes de magnitud en diferencia entre el ejemplo más corto y el más largo.

Se estableció una clara división durante la selección de características entre las métricas de enriquecimiento y distancia genómica. Las tres características asociadas a esta última, pares de bases a la siguiente y anterior, así como la extensión, fueron

consistentemente entre las primeras en ser eliminadas durante la selección de características. Para defender aún más las medidas relacionadas con el enriquecimiento por tener un mayor potencial de discernimiento entre ejemplos negativos y positivos, el máximo de señal, área y altura, tuvieron una clasificación de importancia de la permutación (PIR) más alta que las métricas de distancia genómica para las dos estrategias de selección de características empleadas, siendo el máximo de señal la última variable que se eliminó en ambos casos. La variable antes mencionada era robusta, en tanto que al emplear la estrategia de selección de características de RFE-SVM, un SVM de un solo predictor producía el AuROC más alto, aunque otros modelos multicaracterística fueron muy competitivos. Aun así, no se obtuvo un aumento evidente en el rendimiento con la adición o sustracción de variables adicionales. Incluso, cuando no se esperaba lograr un aumento claro del rendimiento con una mayor complejidad del modelo para las variables observadas, las máquinas de aprendizaje subsiguientes fueron entrenadas con las seis características obtenidas.

Las métricas de rendimiento se comportaron de forma casi homogénea para las SVMs de seis características tanto para los casos de núcleo lineal como para los de núcleo RBF. No se detectaron fuertes discrepancias en ninguna de las métricas tradicionales: precisión, especificidad o sensibilidad al comparar los conjuntos de entrenamiento y los conjuntos de prueba, lo que indica un procedimiento de entrenamiento que permite que un SVM pueda generalizar adecuadamente, independientemente del núcleo utilizado. Debido a la presencia de una muestra desequilibrada, con el grupo mayoritario en la clase “negativa“, se observó una alta especificidad; las señales “regulares“ con un núcleo lineal mostraron el mayor valor en el conjunto de prueba, con un 95.47 %. Cuando se contrasta con la especificidad, la sensibilidad parece pobre ya que los SVMs negativamente desequilibrados tienden a clasificar todo como negativo. Cualquier algoritmo que intente mejorar esta restricción de rendimiento inevitablemente sacrifica especificidad para mejorar la sensibilidad (Akbani *et al.*, 2004). Esta situación se observó al cambiar de una SVM lineal a una SVM con núcleo RBF. El RBF mostró un aumento de la sensibilidad a expensas de disminuir la especificidad: al pasar de núcleo lineal a RBF en el caso de las señales regulares, se observó una ganancia de sensibilidad del 9,3 % a expensas de una pérdida de especificidad del 1,41 %. En el caso de la señal extendida se obtuvo una ganancia de sensibilidad similar del 7,09 % a expensas de un 0,31 % de especificidad. No parece ser que los nuevos positivos aña-

didos después de la extensión ayudaran al valor de sensibilidad, ya que la sensibilidad del conjunto extendido es menor que la sensibilidad del conjunto regular en un 2.66 %. La utilización de un núcleo RBF, a diferencia de uno lineal, permitió una ligera ganancia de precisión de 2,41 % y 2,61 % para los conjuntos de prueba regulares y extendidos, respectivamente. Examinar el cambio en los verdaderos positivos capturados cuando se contrasta entre SVMs extendidos y regulares es coherente con los cambios de sensibilidad antes mencionados. El conjunto de prueba extendido mostró un aumento en el número total de verdaderos positivos detectados con respecto al conjunto regular de 480 y 359 para las SVMs de núcleo lineal y RBF, respectivamente. El conjunto de prueba extendido, consta de 23,068 picos, de los cuales 9,159 son positivos. Dentro de la muestra positiva extendida, 284 picos que anteriormente eran negativos en el conjunto regular fueron etiquetados como positivos. Es decir, 284 intervalos ChIP-seq que anteriormente no tenían traslapes con los intervalos ATAC-seq adquirieron un traslape con ATAC-seq después de la extensión. De estos 284 picos, el clasificador lineal SVM etiquetó 62 como positivos y 222 como negativos, mientras que el RBF etiquetó 69 como positivos y 215 como negativos. El RBF-SVM puede capturar más de los nuevos positivos re-etiquetados, pero su especificidad sufre ya que esto lo priva de muestras negativas adicionales “verdaderas“ y, por lo tanto, de potencial de discernimiento. Incluso si se trata de un subconjunto reducido para el análisis. Esto puede explicar en parte la especificidad comprometida mostrada por los SVMs de RBF después de la extensión.

Además de las ya mencionadas métricas, decidimos incorporar el Coeficiente de Correlación de Matthew (MCC) como una métrica de evaluación adicional debido a que es capaz de discernir favorablemente cuando se cuenta con clases desequilibradas. Debido a que el MCC es un caso especial del Coeficiente de Correlación de Pearson, se puede utilizar una interpretación similar (Powers, 2011). Todos los resultados del MCC tendieron hacia un valor positivo de 1, lo que indica una fuerte correspondencia entre las etiquetas “verdaderas“ y las predicciones efectuadas por nuestro clasificador. Cambiar el núcleo permitió un aumento, en MCC de 0.0526 y 0.0571 para los conjuntos regulares y extendidos, respectivamente. Mientras que el aumento de rendimiento fue mayor en el conjunto extendido cuando se utilizó el núcleo RBF, se observó un descenso de rendimiento similar al mencionado anteriormente cuando se compararon conjuntos regulares y extendidos tanto en los casos de núcleo lineal como en los de

núcleo RBF. Esto solidifica la posición de que los positivos re-etiquetados después de la extensión pueden estar comprometiendo la capacidad del clasificador para identificar correctamente la etiqueta de una señal extendida, de acuerdo con el carácter biológico “verdadero“ o de señal general. En términos generales, la incorporación de nuevos positivos “reales“ basados en traslape en la muestra y la exploración de nuevos núcleos para límites de decisión pueden permitir ligeras ganancias de rendimiento, con el riesgo de perder el potencial de generalización. Sin embargo, es de suma importancia añadir criterios adicionales para discernir si una señal debe conservar una etiqueta positiva o negativa después de la extensión, incluso cuando se consideran nuevas superposiciones, ya que las características que utilizamos actualmente pueden ser similares incluso dentro de diferentes grupos de etiquetas. A continuación se describen algunos ejemplos representativos de cada uno de los casos.

Para profundizar en la noción de una etiqueta basalmente “verdadera“ que correspondiera al fenómeno biológico subyacente, se realizó un análisis empírico de varias señales y sus correspondientes etiquetas antes y después de la extensión tanto para los SVMs lineales como para los RBFs. Esto resultó en tres casos diferentes en los que los intervalos de señal mostraron una de las tres siguientes tendencias durante la clasificación. Las predicciones de un clasificador sobre intervalos de señales podían re-etiquetar una señal previamente negativa, convertida en positiva posterior al traslape como positiva, mantener una señal previamente negativa etiquetada como negativa incluso después de un re-etiquetado inducido por extensión y traslape, y los casos en los que había una decisión dividida sobre el re-etiquetado entre clasificadores con diferentes núcleos.

Las gráficas contenidas en las figuras mostradas a continuación muestran la misma señal en sus formas regulares y extendidas. El eje vertical indica la intensidad de la senal, mientras que el eje horizontal es una medida de extensión de señal. Las posiciones del eje horizontal son relativas a cada señal. Es decir, mientras se contempla el mismo intervalo genómico en ambas señales, la posición mostrada en el eje horizontal en el caso extendido no corresponde a la posición original en el caso regular, puesto que el caso extendido contempla pares de bases adicionales que fueron añadidos en el proceso de extensión. Las siguientes tres figuras muestran extensiones de 250 pares de bases en ambos extremos.

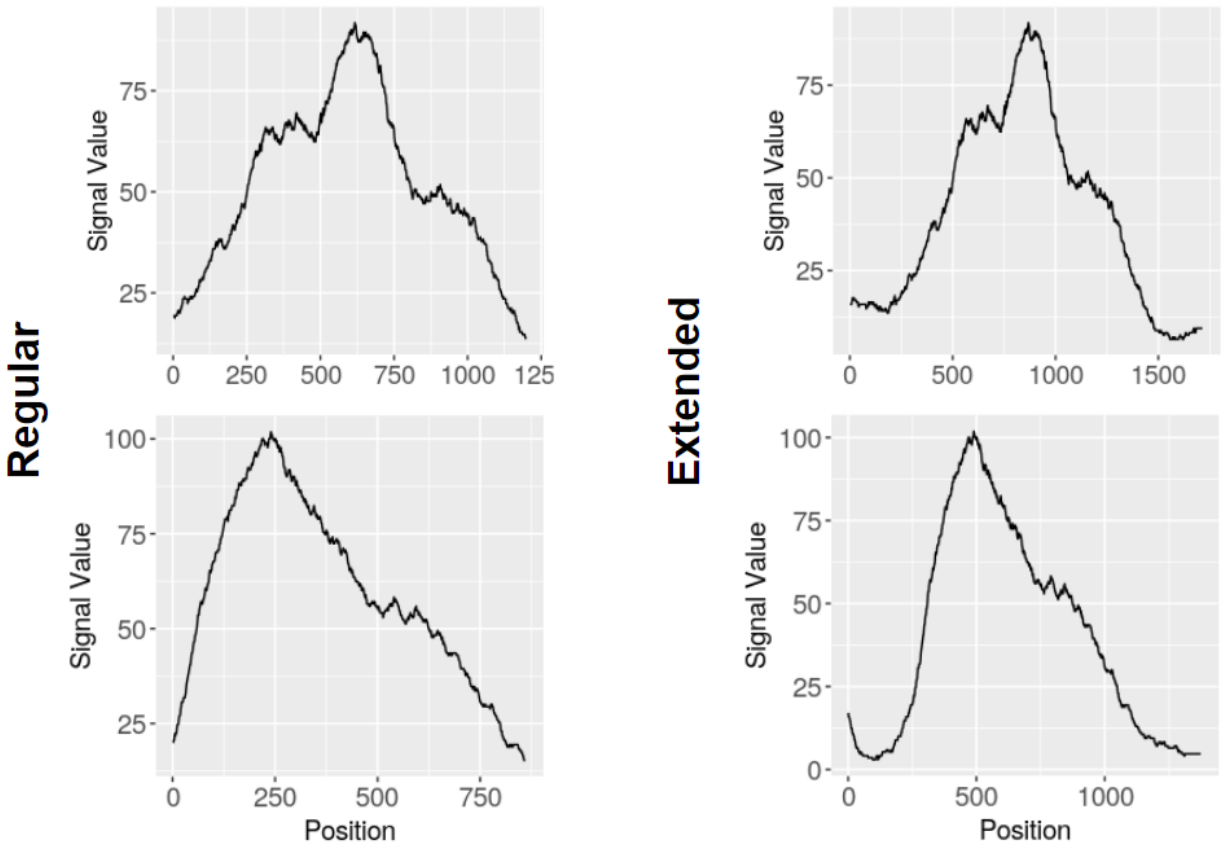

Figura 18. Ejemplos de señales anteriormente negativas clasificadas como positivas por los SVMs tanto lineales como de RBF, después de haber sido re-etiquetadas como positivas después del proceso de extensión.

Una señal que ilustra el primer caso puede observarse en la Figura 18. Señales previamente negativas que son re-etiquetadas como positivas muestran que después de la extensión, un valle pronunciado parece formarse consistentemente. En ambos ejemplos, se pueden ver curvas que podrían parecerse a la arquitectura típica del “valle“ en ambos extremos. Incluso si la longitud de la señal no es notablemente larga, 250 extensiones de pares de bases en cada extremo son suficientes para formar un valle que de otra manera no estaría allí, y parece ser más pronunciado que lo que una señal extendida pudiera contener. Biológicamente, esto podría llevar a un clasificador a determinar que dos valles en un intervalo son altamente indicativos de un evento de acoplamiento.

En la Figura 19 se puede ver un ejemplo de señales negativas que mantuvieron su carácter negativo después de la extensión. Estas señales son generalmente más largas y una extensión de 250 pares de bases no parece ser suficiente para cambiar el carácter general de la señal, ya que no se puede observar la formación de nuevos

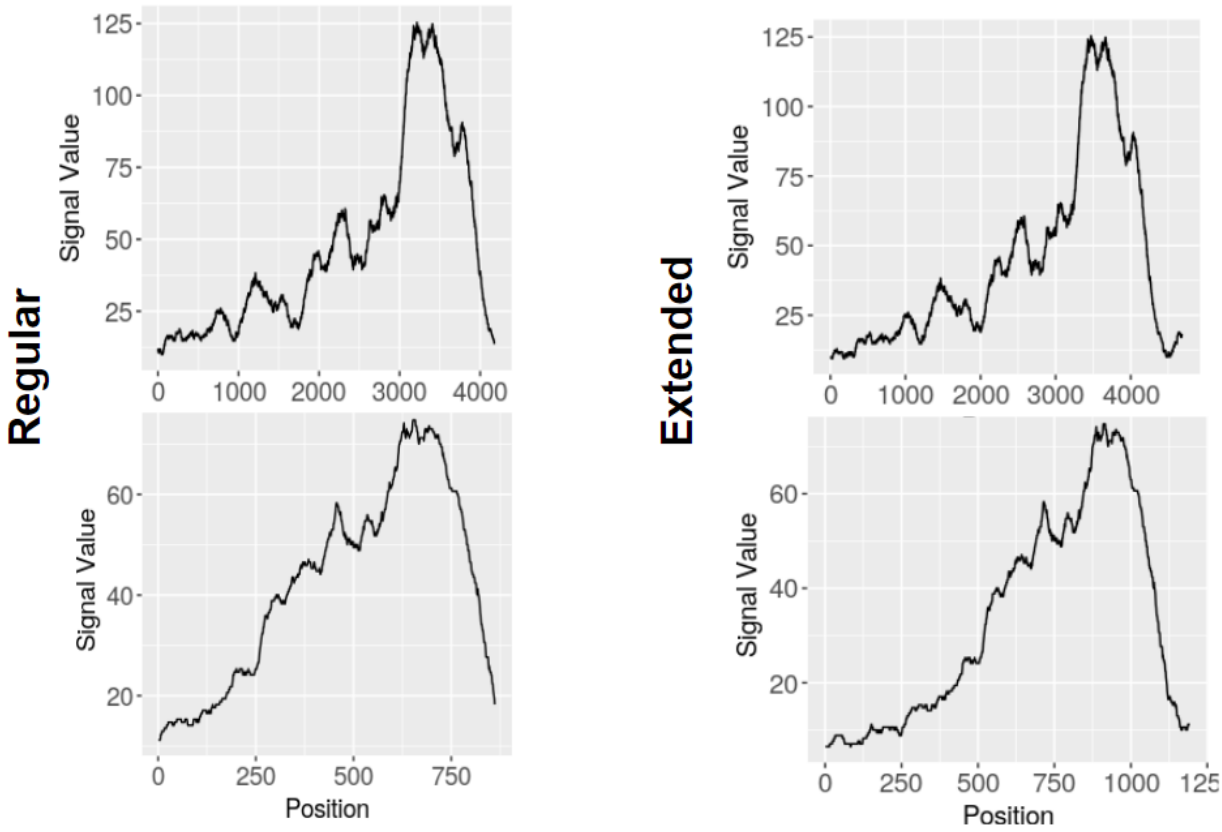

Figura 19. Ejemplos de señales anteriormente negativas clasificadas como negativas por los SVMs tanto lineales como de RBF, después de haber sido re-etiquetadas como positivas después del proceso de extensión.

valles significativos. Esto se debe en parte a que el intervalo tiene otras características pronunciadas, como la pendiente descendente que se muestra en el ejemplo extendido inferior derecho. Las áreas del valle deben permanecer sin cambios, y por lo tanto, si nuestro clasificador está basando fuertemente sus predicciones en la característica del área, no se observará ningún cambio general en la etiqueta.

Un ejemplo en el que clasificadores con diferente núcleo mostraban decisiones de etiquetado dispares se muestra en la Figura 20. Estas señales presentaban una extensión similar entre ellas; eran notablemente más cortas que los ejemplos anteriores. La extensión total fue mayor que el tamaño original de la señal, lo que cambió significativamente el carácter de ambas señales. Los clasificadores lineales determinaron que la señal anterior era negativa, mientras que los clasificadores con núcleo RBF la consideraron positiva y viceversa para el ejemplo inferior; los SVM lineales lo consideraron positivo, mientras que los de RBF lo consideraron negativo. Mientras que se requieren más ejemplos, el ejemplo superior derecho, que fue considerado positivo por la RBF

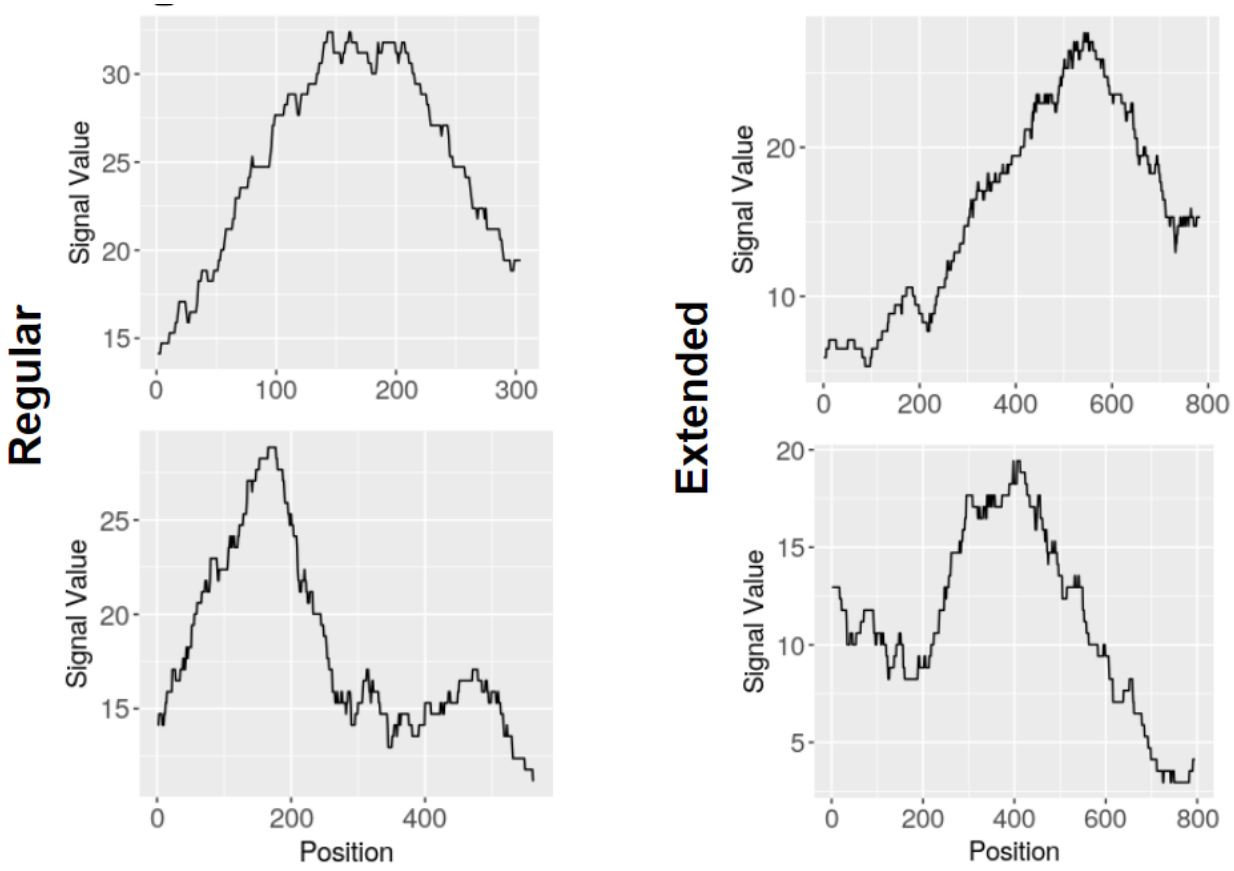

Figura 20. Ejemplos de señales anteriormente negativas con reetiquetado dispar por los SVM lineales y de RBF, después de haber sido re-etiquetadas positivas posteriores al proceso de extensión.

parece tener un valle de mayor longitud, donde el ejemplo inferior derecho, considerado positivo por nuestros núcleos lineales, parecía tener un valle más corto, pero más pronunciado. Esto puede ser indicativo de que los límites de las decisiones lineales y de RBF difieren significativamente en cuanto a la preferencia de área, incluso cuando los valores de extensión son relativamente similares, lo que refuerza la relevancia de nuestra abstracción inspirada en el área en contraste con utilizar solamente las características descriptivas.

Motivados por lo anterior, se implementó un conjunto adicional de experimentos, denominados predicciones cruzadas. El objetivo de estas predicciones fue determinar si un clasificador entrenado exclusivamente en señales extendidas tendría un mejor rendimiento en lugar de utilizar señales extendidas para la predicción en un modelo entrenado en señales regulares, a fin de verificar si el clasificador estaba aprendiendo información adicional relevante para la detección de nuevos intervalos ATAC-seq después de la extensión. Con la excepción de la especificidad, cada métrica vio una mejo-

ra después del entrenamiento exclusivamente en señales extendidas, para predicciones extendidas, siguiendo las tendencias observadas en clasificadores previamente discutidos. Esto puede sugerir que los clasificadores extendidos no están ignorando completamente la información proveniente de los verdaderos positivos recientemente encontrados. Esto aboga por un criterio negativo verdadero más robusto y mayor equilibrio de muestra, idealmente provisto mediante la inclusión de ejemplos polarizantes con una verdad básica “conocida“, tanto para los casos negativos como para los positivos.

Se podrían tomar medidas adicionales antes del análisis de la muestra para verificar la veracidad biológica del carácter de una señal. Revisar los controles utilizados en el flujo de trabajo antes de la determinación de picos para tomar en cuenta los niveles generales de enriquecimiento y el número de lecturas, el ajuste de los parámetros de la determinación de picos o la contemplación de medidas genómicas adicionales, como la estructura de cromatina tridimensional, entre otros criterios, podría dar lugar a señales con un carácter robusto, lo que haría que nuestras predicciones fueran menos propensas al ruido proveniente de artefactos experimentales, lo cual nos permitiría ajustar nuestros clasificadores de acuerdo con el fenómeno biológico del que se trate.

Capítulo 6. Conclusiones

6.1. Sumario

Esta tesis trató sobre la predicción de regiones de cromatina abierta a partir de datos específicos de ChIP-seq de H3k27ac. Se propuso un conjunto de seis descriptores para representar las señales de ChIP-seq con el objetivo de construir un modelo de aprendizaje. El modelo aquí elaborado consiste en una máquina de soporte vectorial que utiliza los descriptores propuestos, construidos a partir de datos genómicos de acuerdo con una abstracción biológicamente inspirada para determinar de forma binaria si una determinada señal proveniente de ChIP-seq es propensa a corresponder a una región de cromatina abierta como la determina señales de ATAC-seq. Para verificar la efectividad del modelo se contruyó un paquete en R denominado histoneSig, el cual está disponible para descarga y utilización en el repositorio GitHub.

6.2. Conclusiones

Utilizando esta abstracción y el modelo que la acompaña, el porcentaje máximo de intervalos ATAC recuperables por ChIP en el conjunto de datos A549 proporcionado para los conjuntos regulares y extendidos es de 62.48 % y 65.16 %, respectivamente. Nuestros SVMs de mayor rendimiento, que tienen un núcleo de función de base radial, pueden capturar hasta el 85.01 % y 82.84 % de los intervalos mencionados anteriormente para los casos de señales regulares y extendidas.

Generalmente, los núcleos no lineales muestran una sensibilidad mucho mayor, a expensas de poca especificidad frente a los núcleos lineales, sin una ganancia de precisión altamente perceptible. Al cambiar de núcleo lineal a RBF en el caso de las señales regulares, se observó una ganancia de sensibilidad del 9.3 % para una pérdida del 1.41 % de especificidad. Una ganancia de sensibilidad similar de 7.09 % a un costo de 0.31 % de especificidad fue exhibida para el caso de la señal extendida.

La implementación de la selección de características nos permitió mostrar que existe una clara distinción de rendimiento entre las características asociadas al enrique-

cimiento en las mediciones de muestra y de distancia genómica; todas las características asociadas a la intensidad máxima de la señal, una medida de enriquecimiento, presentaban un mayor poder predictivo sobre las características descriptivas de la distancia genómica.

Aunque la ampliación de la señal base hace que nuestras máquinas de entrenamiento tengan un rendimiento inferior al de las señales regulares en las métricas de rendimiento cuantitativas comúnmente usadas, esto permite a los clasificadores capturar regiones ATAC únicas que antes no habían sido detectadas por un modelo entrenado sólo con señales regulares. Las señales más largas parecen conservar el carácter de su etiqueta original después de la extensión.

Proporcionamos nuestro modelo en un formato listo para usar para pruebas adicionales y refinamiento futuro del modelo. *histoneSig*, el paquete utilizado a lo largo de esta tesis para cálculos de valle y determinación de características asociadas está disponible en github[1] para uso público.

6.3. Trabajo Futuro

- Hacer que el clasificador propuesto sea capaz de determinar cuántas regiones de cromatina abierta son capturadas, y hasta qué grado de accesibilidad, por cada señal asociada a un intervalo determinado.
- Hacer que nuestro clasificador sea sensible a los diferentes niveles de enriquecimiento presentes en los datos de origen del experimento que se analice.
- Encontrar, desarrollar y probar métodos adicionales y refinar aún más las abstracciones biológicas para obtener características de señal epecíficas altamente generalizables por familia de factor de transcripción.
- Analizar las regiones de cromatina abierta para determinar si existen características adicionales asociadas a estas regiones que permitan a nuestro clasificador tener un desempeño igual o mejor que el establecido por las métricas de desempeño actuales.

[1] https://github.com/semibah/histoneSig

- Explorar otros paradigmas de construcción de características, basados en otros paradigmas de aprendizaje de máquina. El uso de la Transformación Wavelet, como lo sugiere Hidalgo H., está siendo considerado como una perspectiva de investigación futura para ser utilizada en conjunto con métodos de aprendizaje profundo, específicamente, Redes Neuronales Convolucionales.

Literatura citada

Abu-Mostafa, Y. S., Magdon-Ismail, M., y Lin, H.-T. (2012). *Learning From Data*. AML-Book.

Akbani, R., Kwek, S., y Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. En: *Machine Learning: ECML 2004*. Springer Berlin Heidelberg, pp. 39–50.

Alan V. Oppenheim, Alan S. Willsky, I. T. Y. (1982). *Signals and Systems*. Prentice-Hall Signal Processing Series. Prentice-Hall, primera edición.

Alberts, B. (2002). *Molecular Biology of the Cell, Fourth Edition*. Garland Science.

Aleksic, J., Carl, S. H., y Frye, M. (2014). Beyond library size: a field guide to NGS normalization. *bioRxiv*.

Alipanahi, B., Delong, A., Weirauch, M. T., y Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, **33**(8): 831–838.

Allis, C. D. y Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. *Nature Reviews Genetics*, **17**(8): 487–500.

Angermueller, C., Pärnamaa, T., Parts, L., y Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, **12**(7): 878.

Arvey, A., Agius, P., Noble, W. S., y Leslie, C. (2012). Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Research*, **22**(9): 1723–1734.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., y Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Computational Biology*, **9**(11): e1003326.

Bannister, A. J. y Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, **21**(3): 381–395.

Bao, Y., Vinciotti, V., Wit, E., y Hoen, P. A. C. (2013). Joint modeling of ChIP-seq data via a markov random field model. *Biostatistics*, **15**(2): 296–310.

Baranello, L., Kouzine, F., Sanford, S., y Levens, D. (2015). ChIP bias as a function of cross-linking time. *Chromosome Research*, **24**(2): 175–181.

Bardet, A. F., He, Q., Zeitlinger, J., y Stark, A. (2011). A computational pipeline for comparative ChIP-seq analyses. *Nature Protocols*, **7**(1): 45–61.

Bártová, E., Krejčí, J., Harničarová, A., Galiová, G., y Kozubek, S. (2008). Histone modifications and nuclear architecture: A review. *Journal of Histochemistry & Cytochemistry*, **56**(8): 711–721.

Behjati, S. y Tarpey, P. S. (2013). What is next generation sequencing? *Archives of disease in childhood - Education & practice edition*, **98**(6): 236–238.

Benveniste, D., Sonntag, H.-J., Sanguinetti, G., y Sproul, D. (2014). Transcription factor binding predicts histone modifications in human cell lines. *Proceedings of the National Academy of Sciences*, **111**(37): 13367–13372.

Besser, J., Carleton, H., Gerner-Smidt, P., Lindsey, R., y Trees, E. (2018). Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clinical Microbiology and Infection*, **24**(4): 335–341.

Bishop, C. (2006). *Pattern recognition and machine learning*. Springer. New York.

Borboudakis, G. y Tsamardinos, I. (2019). Forward-backward selection with early dropping. *The Journal of Machine Learning Research*, **20**(1): 276–314.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., y Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**(2): 311–322.

Brahmachari, Shruti Jain (auth.), W. D. O. W. K.-H. C. H. Y. e. (2013). *Encyclopedia of Systems Biology*. Springer-Verlag New York, primera edición.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., y Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, **10**(12): 1213–1218.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., y Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**(7561): 486–490.

Busby, M., Xue, C., Li, C., Farjoun, Y., Gienger, E., Yofe, I., Gladden, A., Epstein, C. B., Cornett, E. M., Rothbart, S. B., Nusbaum, C., y Goren, A. (2016). Systematic comparison of monoclonal versus polyclonal antibodies for mapping histone modifications by ChIP-seq. *Epigenetics & Chromatin*, **9**(1).

C. David Allis, Thomas Jenuwein, D. R. M.-L. C. (2007). *Epigenetics*. Cold Spring Harbor Laboratory Press, primera edición.

Calviello, A. K., Hirsekorn, A., Wurmus, R., Yusuf, D., y Ohler, U. (2019). Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. *Genome Biology*, **20**(1).

Cazaly, E., Saad, J., Wang, W., Heckman, C., Ollikainen, M., y Tang, J. (2019). Making sense of the epigenome using data integration approaches. *Frontiers in Pharmacology*, **10**: 126.

Chandrashekar, G. y Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, **40**(1): 16–28.

Chang, H., Liu, Y., Xue, M., Liu, H., Du, S., Zhang, L., y Wang, P. (2016). Synergistic action of master transcription factors controls epithelial-to-mesenchymal transition. *Nucleic Acids Research*, **44**(6): 2514–2527.

Charlet, J., Duymich, C. E., Lay, F. D., Mundbjerg, K., Sørensen, K. D., Liang, G., y Jones, P. A. (2016). Bivalent regions of cytosine methylation and h3k27 acetylation suggest an active role for DNA methylation at enhancers. *Molecular Cell*, **62**(3): 422–431.

Chen, L. (2009). *Biomolecular networks : methods and applications in systems biology*. Wiley. Hoboken, N.J.

Chen, L., Wang, C., Qin, Z. S., y Wu, H. (2015). A novel statistical method for quantitative comparison of multiple ChIP-seq datasets. *Bioinformatics*, **31**(12): 1889–1896.

Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., Silva, T. C., Groeneveld, C., Wong, C. K., Cho, S. W., Satpathy, A. T., Mumbach, M. R., Hoadley, K. A., Robertson, A. G., Sheffield, N. C., Felau, I., Castro, M. A. A., Berman, B. P., Staudt, L. M., Zenklusen, J. C., Laird, P. W., Curtis, C., Greenleaf, W. J., y and, H. Y. C. (2018). The chromatin accessibility landscape of primary human cancers. *Science*, **362**(6413): eaav1898.

Cortes, C. y Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**(3): 273–297.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., y Jaenisch, R. (2010). Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, **107**(50): 21931–21936.

Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, **36**(5): 961–1005.

Diaz, A., Park, K., Lim, D. A., y Song, J. S. (2012). Normalization, bias correction, and peak calling for ChIP-seq. *Statistical Applications in Genetics and Molecular Biology*, **11**(3).

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Gall, C. L., Schaeffer, B., Crom, S. L., Guedj, M., y and, F. J. (2012). A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, **14**(6): 671–683.

Ding, S., Zhu, H., Jia, W., y Su, C. (2011). A survey on feature extraction for pattern recognition. *Artificial Intelligence Review*, **37**(3): 169–180.

Ernst, J. y Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, **9**(3): 215–216.

Fejes, A. P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M., y Jones, S. J. M. (2008). FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, **24**(15): 1729–1730.

Ferguson, J. P., Cho, J. H., y Zhao, H. (2012). A new approach for the joint analysis of multiple chip-seq libraries with application to histone modification. *Statistical Applications in Genetics and Molecular Biology*, **11**(3).

Fisher, Doug; Lenz, H.-J. (1996). *[Lecture Notes in Statistics] Learning from Data Volume 112 || A Comparative Evaluation of Sequential Feature Selection Algorithms*, Vol. 10.1007/978-1-4612-2404-4.

Fletcher, T. (2009). Support vector machines explained. *Tutorial paper*.

Fonseca, N. A., Marioni, J., y Brazma, A. (2014). RNA-seq gene profiling - a systematic empirical comparison. *PLoS ONE*, **9**(9): e107026.

Furey, T. S. (2012). ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, **13**(12): 840–852.

Fyodorov, D. V., Zhou, B.-R., Skoultchi, A. I., y Bai, Y. (2017). Emerging roles of linker histones in regulating chromatin structure and function. *Nature Reviews Molecular Cell Biology*, **19**(3): 192–206.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., y Zhang, J. (2004). *Genome Biology*, **5**(10): R80.

George, J., Uyar, A., Young, K., Kuffler, L., Waldron-Francis, K., Marquez, E., Ucar, D., y Trowbridge, J. J. (2016). Leukaemia cell of origin identified by chromatin landscape of bulk tumour cells. *Nature Communications*, **7**(1).

Gerland, U., Moroz, J. D., y Hwa, T. (2002). Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proceedings of the National Academy of Sciences*, **99**(19): 12015–12020.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., y Snyder, M. (2007). What is a gene, post-ENCODE? history and updated definition. *Genome Research*, **17**(6): 669–681.

Gligorijević, V., Barot, M., y Bonneau, R. (2017). deepNF: Deep network fusion for protein function prediction.

Goodfellow, I. (2016). *Deep learning*. The MIT Press. Cambridge, Massachusetts.

Goodwin, S., McPherson, J. D., y McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**(6): 333–351.

Grossman, S. R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B. E., Mikkelsen, T. S., y Lander, E. S. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proceedings of the National Academy of Sciences*, **114**(7): E1291–E1300.

Guo, Y., Mahony, S., y Gifford, D. K. (2012). High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Computational Biology*, **8**(8): e1002638.

Gusmao, E. G., Dieterich, C., Zenke, M., y Costa, I. G. (2014). Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics*, **30**(22): 3143–3151.

Guyon, I. (2006). *Feature Extraction Foundations and Applications. Pattern Recognition*. Springer.

Guyon, I., Weston, J., Barnhill, S., y Vapnik, V. (2002). *Machine Learning*, **46**(1/3): 389–422.

Haar, A. (1910). Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, **69**(3): 331–371.

Hammond, C. M., Strømme, C. B., Huang, H., Patel, D. J., y Groth, A. (2017). Histone chaperone networks shaping chromatin function. *Nature Reviews Molecular Cell Biology*, **18**(3): 141–158.

Hansen, K. D., Wu, Z., Irizarry, R. A., y Leek, J. T. (2011). Sequencing technology does not eliminate biological variability. *Nature Biotechnology*, **29**(7): 572–573.

Harmanci, A., Rozowsky, J., y Gerstein, M. (2014). MUSIC: identification of enriched regions in ChIP-seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biology*, **15**(10).

Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York.

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Nieuwerburgh, F. V., Salomon, D. R., y Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, **56**(2).

Hild, K., Erdogmus, D., Torkkola, K., y Principe, J. (2006). Feature extraction using information-theoretic learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(9): 1385–1392.

Hira, Z. M. y Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, **2015**: 1–13.

Hocking, T. D., Goerner-Potvin, P., Morin, A., Shao, X., Pastinen, T., y Bourque, G. (2016). Optimizing ChIP-seq peak detectors using visual labels and supervised machine learning. *Bioinformatics*, p. btw672.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., y Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, **9**(5): 473–476.

Hornik, K. y Kuan, C.-M. (1992). Convergence analysis of local feature extraction algorithms. *Neural Networks*, **5**(2): 229–240.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., *et al.* (2003). A practical guide to support vector classification.

Huan Liu, Hiroshi Motoda (auth.), H. L. H. M. e. (1998). *Feature Extraction, Construction and Selection: A Data Mining Perspective*. The Springer International Series in Engineering and Computer Science 453. Springer US, primera edición.

Ian H. Witten, E. F. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, segunda edición.

Izzo, A. y Schneider, R. (2016). The role of linker histone h1 modifications in the regulation of gene expression and chromatin dynamics. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, **1859**(3): 486–495.

Jain, A. K., Murty, M. N., y Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, **31**(3): 264–323.

Jain, D., Baldi, S., Zabel, A., Straub, T., y Becker, P. B. (2015). Active promoters give rise to false positive 'phantom peaks' in ChIP-seq experiments. *Nucleic Acids Research*, **43**(14): 6959–6968.

Jankowski, A., Tiuryn, J., y Prabhakar, S. (2016). Romulus: robust multi-state identification of transcription factor binding sites from DNase-seq data. *Bioinformatics*, **32**(16): 2419–2426.

Jiang, S. y Mortazavi, A. (2018a). Integrating ChIP-seq with other functional genomics data. *Briefings in Functional Genomics*, **17**(2): 104–115.

Jiang, S. y Mortazavi, A. (2018b). Integrating ChIP-seq with other functional genomics data. *Briefings in Functional Genomics*, **17**(2): 104–115.

John, S., Sabo, P. J., Thurman, R. E., Sung, M.-H., Biddie, S. C., Johnson, T. A., Hager, G. L., y Stamatoyannopoulos, J. A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics*, **43**(3): 264–268.

Jung, S., Angarica, V. E., Andrade-Navarro, M. A., Buckley, N. J., y del Sol, A. (2017). Prediction of chromatin accessibility in gene-regulatory regions from transcriptomics data. *Scientific Reports*, **7**(1).

Kagohara, L. T., Stein-O'Brien, G. L., Kelley, D., Flam, E., Wick, H. C., Danilova, L. V., Easwaran, H., Favorov, A. V., Qian, J., Gaykalova, D. A., y Fertig, E. J. (2017). Epigenetic regulation of gene expression in cancer: techniques, resources and analysis. *Briefings in Functional Genomics*, **17**(1): 49–63.

Kallio, A. y Elo, L. L. (2013). Optimizing detection of transcription factor-binding sites in ChIP-seq experiments. En: *Methods in Molecular Biology*. Humana Press, pp. 181–191.

Karimzadeh, M. y Hoffman, M. M. (2018). Virtual chip-seq: Predicting transcription factor binding by learning from the transcriptome. *bioRxiv*.

Keilwagen, J., Posch, S., y Grau, J. (2019). Accurate prediction of cell type-specific transcription factor binding. *Genome Biology*, **20**(1).

Kharchenko, P. V., Tolstorukov, M. Y., y Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology*, **26**(12): 1351–1359.

Kim, S. y Shendure, J. (2019). Mechanisms of interplay between transcription factors and the 3d genome. *Molecular Cell*.

Klemm, S. L., Shipony, Z., y Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, **20**(4): 207–220.

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, **128**(4): 693–705.

Krizhevsky, A., Sutskever, I., y Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. En: *Advances in neural information processing systems*. pp. 1097–1105.

Kuhn, M. *et al.* (2008). Building predictive models in r using the caret package. *Journal of statistical software*, **28**(5): 1–26.

Kumar, V., Muratani, M., Rayan, N. A., Kraus, P., Lufkin, T., Ng, H. H., y Prabhakar, S. (2013a). Uniform, optimal signal processing of mapped deep-sequencing data. *Nature Biotechnology*, **31**(7): 615–622.

Kumar, V., Muratani, M., Rayan, N. A., Kraus, P., Lufkin, T., Ng, H. H., y Prabhakar, S. (2013b). Uniform, optimal signal processing of mapped deep-sequencing data. *Nature Biotechnology*, **31**(7): 615–622.

Laajala, T. D., Raghav, S., Tuomela, S., Lahesmaa, R., Aittokallio, T., y Elo, L. L. (2009). A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics*, **10**(1): 618.

Lal, T. N., Chapelle, O., Weston, J., y Elisseeff, A. (2006). Embedded methods. En: *Feature Extraction*. Springer Berlin Heidelberg, pp. 137–165.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., y Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, **22**(9): 1813–1831.

Lange, U. C. y Schneider, R. (2010). What an epigenome remembers. *BioEssays*, **32**(8): 659–668.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., y Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, **9**(8): e1003118.

Le, Q. V. *et al.* (2015). A tutorial on deep learning part 2: Autoencoders, convolutional neural networks and recurrent neural networks. *Google Brain*, pp. 1–20.

LeCun, Y., Bengio, Y., y Hinton, G. (2015). Deep learning. *Nature*, **521**(7553): 436–444.

Lee, H., Grosse, R., Ranganath, R., y Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. En: *Proceedings of the 26th annual international conference on machine learning*. ACM, pp. 609–616.

Li, F. y Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, **105**(491): 1202–1214.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., y Liu, H. (2017). Feature selection. *ACM Computing Surveys*, **50**(6): 1–45.

Li, Q., Brown, J. B., Huang, H., y Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, **5**(3): 1752–1779.

Linnarsson, S. (2010). Recent advances in DNA sequencing methods – general principles of sample preparation. *Experimental Cell Research*, **316**(8): 1339–1343.

Liu, H. y Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, **17**(4): 491–502.

Lyons, R. G. (2004). *Understanding Digital Signal Processing*. Pearson Education, second edition edición.

Mack, S. C., Singh, I., Wang, X., Hirsch, R., Wu, Q., Villagomez, R., Bernatchez, J. A., Zhu, Z., Gimple, R. C., Kim, L. J., Morton, A., Lai, S., Qiu, Z., Prager, B. C., Bertrand, K. C., Mah, C., Zhou, W., Lee, C., Barnett, G. H., Vogelbaum, M. A., Sloan, A. E., Chavez, L., Bao, S., Scacheri, P. C., Siqueira-Neto, J. L., Lin, C. Y., y Rich, J. N. (2019). Chromatin landscapes reveal developmentally encoded transcriptional states that define human glioblastoma. *The Journal of Experimental Medicine*, **216**(5): 1071–1090.

Mahony, S., Edwards, M. D., Mazzoni, E. O., Sherwood, R. I., Kakumanu, A., Morrison, C. A., Wichterle, H., y Gifford, D. K. (2014). An integrated model of multiple-condition ChIP-seq data reveals predeterminants of cdx2 binding. *PLoS Computational Biology*, **10**(3): e1003501.

Makova, K. D. y Hardison, R. C. (2015). The effects of chromatin organization on variation in mutation rates in the genome. *Nature Reviews Genetics*, **16**(4): 213–223.

Mathelier, A. y Wasserman, W. W. (2013). The next generation of transcription factor binding site prediction. *PLoS Computational Biology*, **9**(9): e1003214.

Mathelier, A., Fornes, O., Arenillas, D. J., yu Chen, C., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., Zhang, A. W., Parcy, F., Lenhard, B., Sandelin, A., y Wasserman, W. W. (2015). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, **44**(D1): D110–D115.

Metzker, M. L. (2009). Sequencing technologies — the next generation. *Nature Reviews Genetics*, **11**(1): 31–46.

Meyer, C. A. y Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*, **15**(11): 709–721.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., Lin, C.-C., y Meyer, M. D. (2019). Package 'e1071'. *The R Journal*.

Miyamoto, K., Nguyen, K. T., Allen, G. E., Jullien, J., Kumar, D., Otani, T., Bradshaw, C. R., Livesey, F. J., Kellis, M., y Gurdon, J. B. (2018). Chromatin accessibility impacts transcriptional reprogramming in oocytes. *Cell Reports*, **24**(2): 304–311.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. The MIT Press, primera edición.

Nakato, R. y Shirahige, K. (2016). Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Briefings in Bioinformatics*, p. bbw023.

Navarro, F. C. P., Mohsen, H., Yan, C., Li, S., Gu, M., Meyerson, W., y Gerstein, M. (2019). Genomics and data science: an application within an umbrella. *Genome Biology*, **20**(1).

Nordström, K. J., Schmidt, F., Gasparoni, N., Salhab, A., Gasparoni, G., Kattler, K., Müller, F., Ebert, P., Costa, I. G., Pfeifer, N., Lengauer, T., Schulz, M. H., y and, J. W. (2019). Unique and assay specific features of NOMe-, ATAC- and DNase i-seq data.

Orenstein, Y. y Shamir, R. (2014). A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Research*, **42**(8): e63–e63.

Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, **10**(10): 669–680.

Park, Y. y Kellis, M. (2015). Deep learning for regulatory genomics. *Nature Biotechnology*, **33**(8): 825–826.

Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., y Pritchard, J. K. (2010). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, **21**(3): 447–455.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*.

Pudil, P., Novovičová, J., y Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, **15**(11): 1119–1125.

Pundhir, S., Bagger, F. O., Lauridsen, F. B., Rapin, N., y Porse, B. T. (2016). Peak-valley-peak pattern of histone modifications delineates active regulatory elements and their directionality. *Nucleic Acids Research*, **44**(9): 4037–4051.

Qin, Q. y Feng, J. (2017). Imputation for transcription factor binding predictions based on deep learning. *PLOS Computational Biology*, **13**(2): e1005403.

Qin, Z. S., Yu, J., Shen, J., Maher, C. A., Hu, M., Kalyana-Sundaram, S., Yu, J., y Chinnaiyan, A. M. (2010). HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-seq data. *BMC Bioinformatics*, **11**(1): 369.

Quang, D. y Xie, X. (2017). FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data.

Rawat, W. y Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, **29**(9): 2352–2449.

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, **3**(Mar): 1371–1382.

Ruan, S. y Stormo, G. D. (2017). Inherent limitations of probabilistic models for protein-DNA binding specificity. *PLOS Computational Biology*, **13**(7): e1005638.

Sarda, S. y Hannenhalli, S. (2014). Next-generation sequencing and epigenomics research: A hammer in search of nails. *Genomics & Informatics*, **12**(1): 2.

Schadt, E. E., Turner, S., y Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, **19**(R2): R227–R240.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, **61**: 85–117.

Shalev-Shwartz, S. y Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. New York, NY, USA.

Shao, Z., Zhang, Y., Yuan, G.-C., Orkin, S. H., y Waxman, D. J. (2012). MAnorm: a robust model for quantitative comparison of ChIP-seq data sets. *Genome Biology*, **13**(3): R16.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., y Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, **15**(2): 121–132.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H. J., y Mann, R. S. (2011). Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. *Cell*, **147**(6): 1270–1282.

Slattery, M., Ma, L., Spokony, R. F., Arthur, R. K., Kheradpour, P., Kundaje, A., Negre, N., Crofts, A., Ptashkin, R., Zieba, J., Ostapenko, A., Suchy, S., Victorsen, A., Jameel, N., Grundstad, A. J., Gao, W., Moran, J. R., Rehm, E. J., Grossman, R. L., Kellis, M., y White, K. P. (2014). Diverse patterns of genomic targeting by transcriptional regulators in drosophila melanogaster. *Genome Research*, **24**(7): 1224–1235.

Stanković, R. S. y Falkowski, B. J. (2003). The haar wavelet transform: its status and achievements. *Computers & Electrical Engineering*, **29**(1): 25–44.

Starmer, J. y Magnuson, T. (2016). Detecting broad domains and narrow peaks in ChIP-seq data with hiddenDomains. *BMC Bioinformatics*, **17**(1).

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., y Robinson, G. E. (2015). Big data: Astronomical or genomical? *PLOS Biology*, **13**(7): 1–11.

Szalkowski, A. M. y Schmid, C. D. (2010). Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. *Briefings in Bioinformatics*, **12**(6): 626–633.

Sze, V., Chen, Y.-H., Yang, T.-J., y Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, **105**(12): 2295–2329.

Tamaru, H. (2010). Confining euchromatin/heterochromatin territory: jumonji crosses the line. *Genes & Development*, **24**(14): 1465–1478.

Tamil, E. M., Noor, M. H., Razak, Z., Noor, N. M., y Tamil, A. M. (2008). A review on feature extraction & classification techniques for biosignal processing (part v: Gait signal). En: *IFMBE Proceedings*. Springer Berlin Heidelberg, pp. 122–124.

Tang, J., Alelyani, S., y Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, p. 37.

Tarca, A. L., Carey, V. J., wen Chen, X., Romero, R., y Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS Computational Biology*, **3**(6): e116.

Taslim, C., Wu, J., Yan, P., Singer, G., Parvin, J., Huang, T., Lin, S., y Huang, K. (2009). Comparative study on ChIP-seq data: normalization and binding pattern characterization. *Bioinformatics*, **25**(18): 2334–2340.

Thomas, R., Thomas, S., Holloway, A. K., y Pollard, K. S. (2016). Features that define the best ChIP-seq peak calling algorithms. *Briefings in Bioinformatics*, p. bbw035.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., y Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, **489**(7414): 75–82.

Tie, F., Banerjee, R., Stratton, C. A., Prasad-Sinha, J., Stepanik, V., Zlobin, A., Diaz, M. O., Scacheri, P. C., y Harte, P. J. (2009). CBP-mediated acetylation of histone h3 lysine 27 antagonizes drosophila polycomb silencing. *Development*, **136**(18): 3131–3141.

Tsompana, M. y Buck, M. J. (2014). Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin*, **7**(1).

Tugrul, M., Paixao, T., Barton, N. H., y Tkacik, G. (2015). Dynamics of transcription factor binding site evolution. *PLOS Genetics*, **11**(11): 1–28.

Vaissiere, T., SAWAN, C., y HERCEG, Z. (2008). Epigenetic interplay between histone modifications and DNA methylation in gene silencing. *Mutation Research/Reviews in Mutation Research*, **659**(1-2): 40–48.

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M., y Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-seq data. *Nature Methods*, **5**(9): 829–834.

Voss, T. C. y Hager, G. L. (2013). Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Reviews Genetics*, **15**(2): 69–81.

Walker, J. S. (2008). *Primer on Wavelets and Their Scientific Applications*. Studies in Advanced Mathematics. Taylor, segunda edición.

Wang, J., Jia, S. T., y Jia, S. (2016). New insights into the regulation of heterochromatin. *Trends in Genetics*, **32**(5): 284–294.

Weirauch, M. T., , Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T. R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., Bussemaker, H. J., Morris, Q. D., Bulyk, M. L., Stolovitzky, G., y Hughes, T. R. (2013). Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, **31**(2): 126–134.

Wilbanks, E. G. y Facciotti, M. T. (2010). Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE*, **5**(7): e11471.

Wong, K.-C., Li, Y., Peng, C., y Zhang, Z. (2014). SignalSpider: probabilistic pattern discovery on multiple normalized ChIP-seq signal profiles. *Bioinformatics*, **31**(1): 17–24.

Xin, B. y Rohs, R. (2018a). Relationship between histone modifications and transcription factor binding is protein family specific. *Genome Research*, **28**(3): 321–333.

Xin, B. y Rohs, R. (2018b). Relationship between histone modifications and transcription factor binding is protein family specific. *Genome Research*, **28**(3): 321–333.

Xu, H., Handoko, L., Wei, X., Ye, C., Sheng, J., Wei, C.-L., Lin, F., y Sung, W.-K. (2010). A signal–noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics*, **26**(9): 1199–1204.

Xu, T., Zheng, X., Li, B., Jin, P., Qin, Z., y Wu, H. (2018). A comprehensive review of computational prediction of genome-wide features. *Briefings in Bioinformatics*.

Yang, L., Orenstein, Y., Jolma, A., Yin, Y., Taipale, J., Shamir, R., y Rohs, R. (2017). Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Molecular Systems Biology*, **13**(2): 910.

Young, M. D., Willson, T. A., Wakefield, M. J., Trounson, E., Hilton, D. J., Blewitt, M. E., Oshlack, A., y Majewski, I. J. (2011). ChIP-seq analysis reveals distinct h3k27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Research*, **39**(17): 7415–7427.

Yuan, H., Kshirsagar, M., Zamparo, L., Lu, Y., y Leslie, C. S. (2019). BindSpace decodes transcription factor binding signals by large-scale sequence embedding. *Nature Methods*, **16**(9): 858–861.

Zeng, X., Sanalkumar, R., Bresnick, E. H., Li, H., Chang, Q., y Keleş, S. (2013). jMOSAiCS: joint analysis of multiple ChIP-seq datasets. *Genome Biology*, **14**(4): R38.

Zentner, G. E. y Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. *Nature Structural & Molecular Biology*, **20**(3): 259–266.

Zhang, L.-Q. y Li, Q.-Z. (2017). Estimating the effects of transcription factors binding and histone modifications on gene expression levels in human cells. *Oncotarget*, **8**(25).

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., y Liu, X. S. (2008). Model-based analysis of ChIP-seq (MACS). *Genome Biology*, **9**(9): R137.

Zhao, Y. y Stormo, G. D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature Biotechnology*, **29**(6): 480–483.

Zhou, J. y Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, **12**(10): 931–934.

Zhu, Y., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J. W., Ding, B., Li, N., Zheng, L., y Wang, W. (2016). Constructing 3d interaction maps from 1d epigenomes. *Nature Communications*, **7**(1).

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., y Telenti, A. (2018). A primer on deep learning in genomics. *Nature Genetics*, **51**(1): 12–18.

Zou, N., Zhu, Y., Zhu, J., Baydogan, M., Wang, W., y Li, J. (2015). A transfer learning approach for predictive modeling of degenerate biological systems. *Technometrics*, **57**(3): 362–373.

Anexo

Tabla 13. Orden de eliminación de variable e importancia de la característica de permutación (PIR) obtenida después de realizar RFE iterativo en señales regulares del cromosoma 8.

	Orden de eliminación de variable	PIR General
Máximo de Señal	6	499.97
Altura	5	357.42
Área	2	325.14
Extensión	1	216.88
Bps al siguiente	2	322.33
Bps al previo	3	354.95

Tabla 14. Orden de eliminación de variable e importancia de la característica de permutación (PIR) obtenida después de realizar RFE iterativo en señales extendidas del cromosoma 8.

	Orden de eliminación de variable	PIR General
Máximo de Señal	6	561.22
Altura	3	343.1377
Área	2	320.13
Extensión	1	226.89
Bps al siguiente	5	393.57
Bps al previo	4	365.91

Tabla 15. AuROC obtenida utilizando el método lineal RFE-SVM en señales regulares y extendidas del cromosoma 8.

Número de variables	Regular	Extendido
1 (MS)	0.83	0.80
2 (MS, AL)	0.83	0.80
3 (MS, AL, AR)	0.83	0.80
4 (MS, AL, AR, BpsS)	0.83	0.80
5 (MS, AL, AR, BpsS, BpsN)	0.83	0.80
6 (MS, AL, AR, BpsS, BpsN, EX)	0.83	0.80

Tabla 16. Orden de eliminación de variable e importancia de la característica de permutación (PIR) obtenida después de realizar RFE iterativo en señales regulares del cromosoma 21.

	Orden de eliminación de variable	PIR General
Máximo de Señal	6	92.16
Altura	4	71.54
Área	5	89.14
Extensión	1	48.14
Bps al siguiente	2	51.44
Bps al previo	3	54.94

Tabla 17. Orden de eliminación de variable e importancia de la característica de permutación (PIR) obtenida después de realizar RFE iterativo en señales extendidas del cromosoma 21.

	Orden de eliminación de variable	PIR General
Máximo de Señal	6	117.88
Altura	4	71.31
Área	5	73.95
Extensión	1	42.53
Bps al siguiente	2	61.54
Bps al previo	3	63.86

Tabla 18. AuROC obtenida utilizando el método lineal RFE-SVM en señales regulares y extendidas del cromosoma 21.

Número de variables	Regular	Extendido
1 (MS)	0.84	0.82
2 (MS, AL)	0.84	0.82
3 (MS, AL, AR)	0.84	0.82
4 (MS, AL, AR, BpsS)	0.84	0.82
5 (MS, AL, AR, BpsS, BpsN)	0.84	0.82
6 (MS, AL, AR, BpsS, BpsN, EX)	0.84	0.82