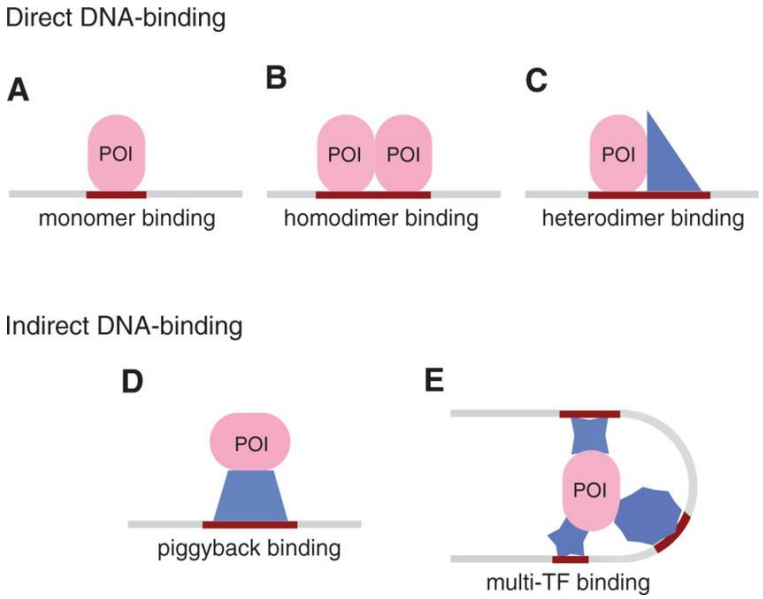
Análisis secuencias reguladoras

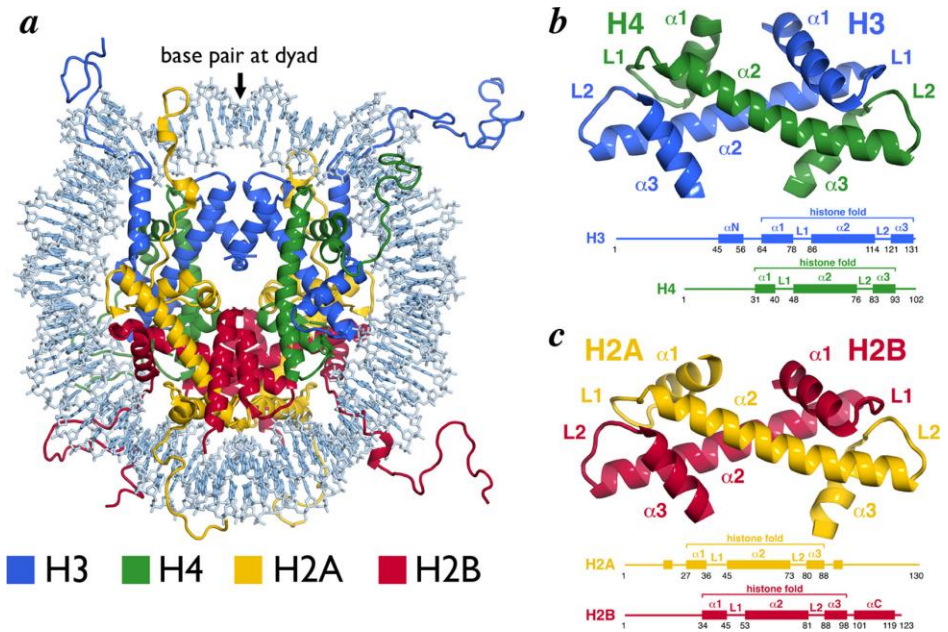
Guillermo Barturen Briñas
(gbarturen@ugr.es)

Proteínas de unión al ADN

- Las proteínas de unión al ADN desempeñan un papel fundamental en las células, regulando procesos como la transcripción, el barajamiento de exones, la replicación o la reparación el ADN entre otros.

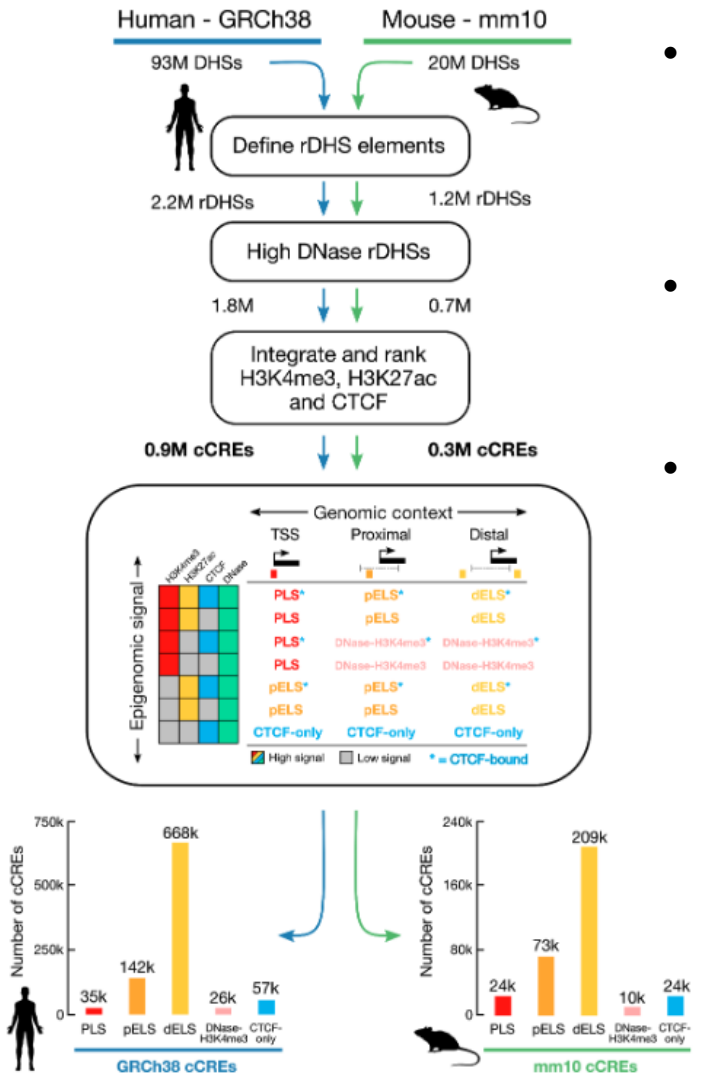
- Estas proteínas incluyen tanto **factores de transcripción (TFs)** que se unen e interacción con el ADN en regiones concretas determinadas por la secuencia de ADN, como **histonas**.





- Las histonas conforman el nucleosoma, estructura fundamental de la cromatina, mientras que los TFs pueden interaccionar de diferentes formas con el ADN: de manera directa solas o en dímeros, o de manera indirecta uniéndose con otras proteíans que tienen los dominios de unión al ADN.

The encyclopedia of DNA elements (ENCODE)



- ENCODE es una iniciativa que comienza tras completar la secuenciación del genoma humano (2003), con el objetivo de identificar y estudiar los elementos funcionales del genoma humano.
- Inicialmente se seleccionó un 1% del genoma humano para ser anotado funcionalmente en algunas líneas celulares.
- Actualmente contiene información para miles de tipos celulares y tejidos a genoma completo, siendo uno de los proyectos más longevos y con mayor número de publicaciones usando sus datos (que se encuentran disponibles).

https://www.encodeproject.org/

ENCODE

Registry V3 (Latest Version)

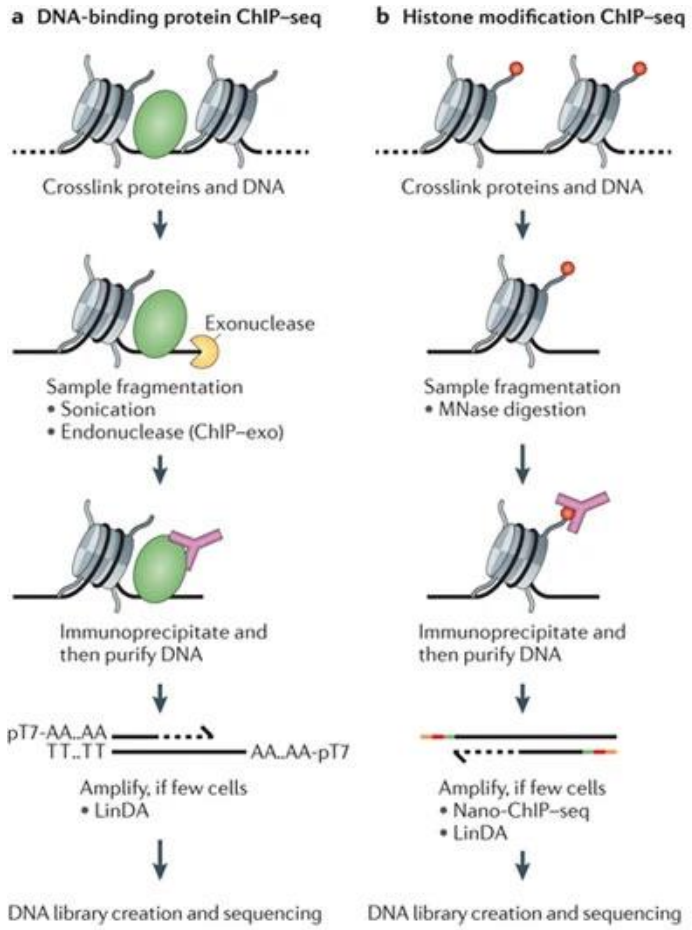
Release date: 2021
URL: screen.encodeproject.org
Human Genome assembly: hg38
Human cCRE count: 1,063,878
Human cell and tissue types covered: 1,518
Mouse Genome assembly: mm10
Mouse cCRE count: 313,838
Mouse cell and tissue types covered: 169

Técnica ChIP-seq

- La **immunoprecipitación de cromatina seguida de secuenciación masiva** es actualmente la técnica por excelencia para identificar la localización y las modificaciones químicas de las proteínas que interaccionan con el ADN a genoma completo.

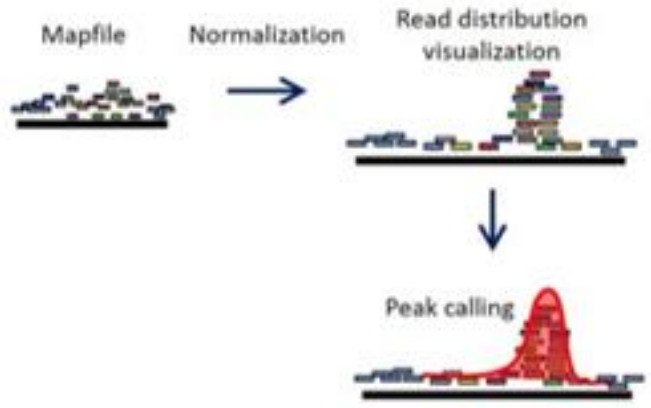
- La metodología se basa en generar una **unión covalente** entre las proteínas y el ADN, lo que interfiere en la dinámica de la regulación fijando estas en su localización (*cross-linking*). Una vez fijadas, se fragmenta la muestra y las proteínas o modificaciones de interés se capturan mediante inmunoprecipitación, arrastrando con ellas la secuencia de ADN ligada.

- Una vez recuperados los anticuerpos, se purifica el ADN y se secuencia. Obteniendo por lo tanto **lecturas sólo de los fragmentos capturados por los anticuerpos**, que se alinean frente al genoma de referencia y se cuantifica su presencia por nucleótido en el genoma dando una idea de dónde y qué actividad presenta dicha proteína en la muestra analizada.



https://doi.org/10.1038/nrg3306

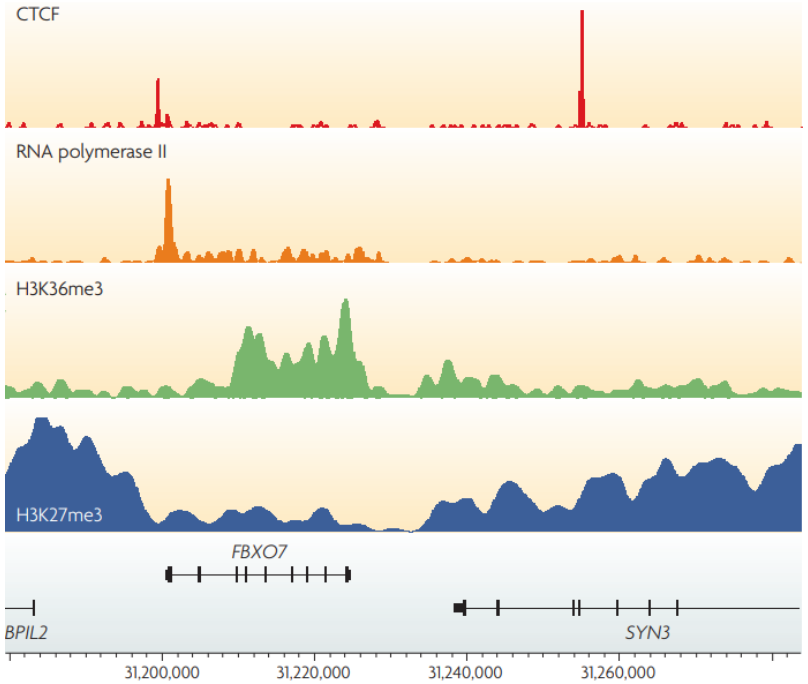
Análisis ChIP-seq



- Una vez obtenidos los conteos por posición, esta información debe ser interpretada para identificar las regiones donde ocurren las interacciones y diferenciarlas del ruido (debido a un mal alineamiento, baja especificidad del anticuerpo...), a este proceso se le denomina ***peak calling***.

- Existen numerosos algoritmos para resolver este problema ruido-señal, generalmente basados en la idea de que el **ruido debe distribuirse de manera uniforme** a lo largo del genoma e independientemente de hebra.

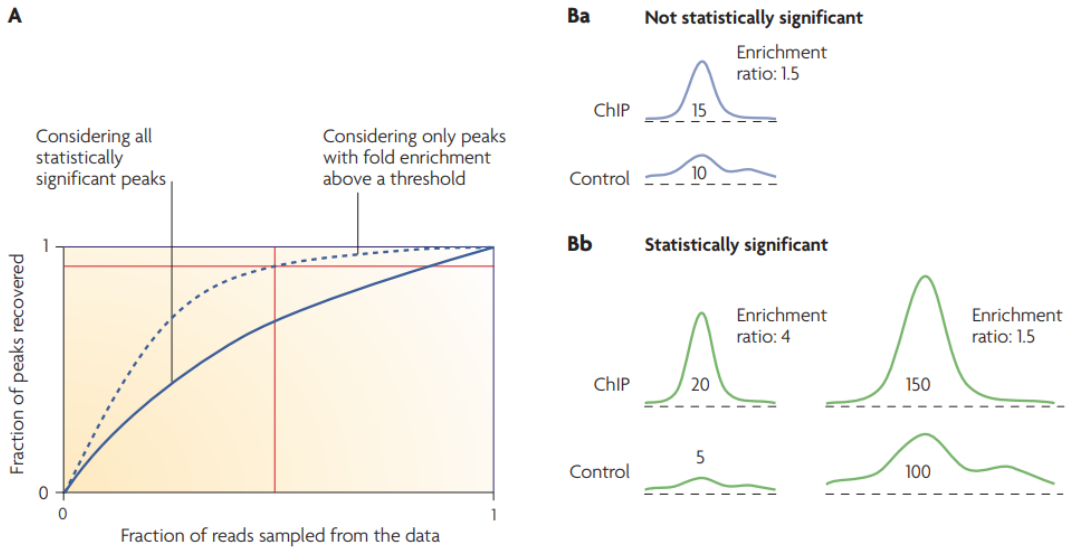
- Estos algoritmos pueden estar diseñados para identificar picos más estrechos o más anchos. Generalmente, los **factores de transcripción presentan picos más estrechos que las modificaciones de histonas**, debido a su funcionalidad. Y entre las modificaciones de histonas, las marcas represivas suelen presentar un perfil más extenso que las relacionadas con la activación de la transcripción.



https://doi.org/10.1038/nrg3306

Análisis ChIP-seq

- Un problema añadido a la hora de diseñar y analizar un experimento de ChIP-Seq es **definir la cobertura óptima** a la que secuenciar. En el caso de la secuenciación de ADN, vamos a secuenciar el genoma completo y sobre su tamaño calculamos la cobertura media que necesitamos. Sin embargo, en este caso no sabemos exactamente cuántas regiones ni que tamaño tendrán, por lo que la estimación del número de lecturas necesarias es complicada.
- Para determinar la cobertura necesaria para un experimento concreto podemos recurrir a **métodos empíricos como el re-muestreo** para estudiar a qué cobertura satura la identificación de picos. Esta metodología suele funcionar relativamente bien en el caso de los factores de transcripción, no tanto con las marcas de histonas donde es difícil llegar a la saturación.



- Esta decisión suele ser un compromiso entre lo que se pretende detectar como significativo y lo realmente funcional. Cuanto más se secuencie más picos significativos se encontrarán, pero su diferencia con un control será menor.

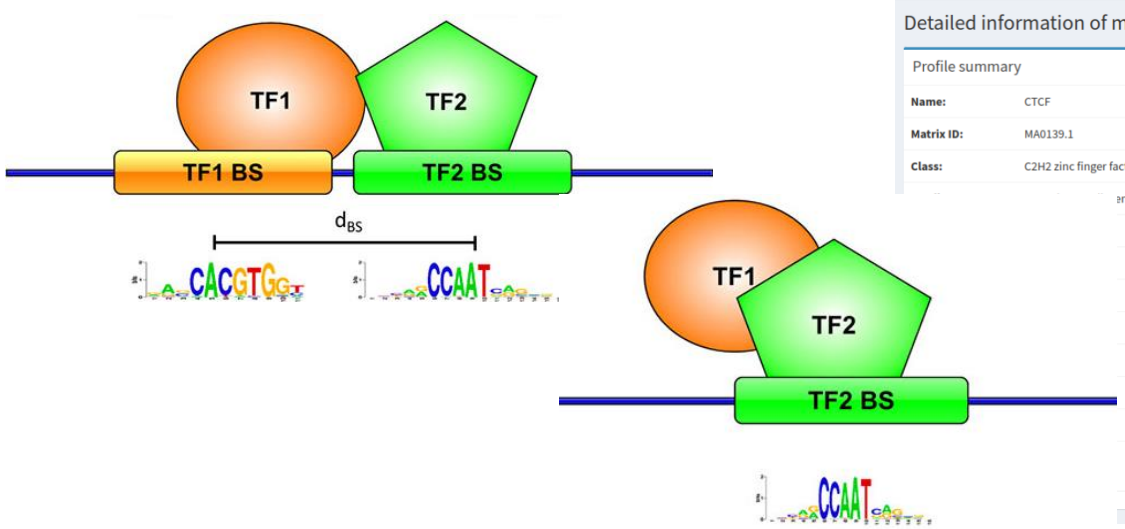
Identificación de motivos de unión

- Un motivo de unión de un TF al ADN es una **secuencia probabilística** compuesta por una matriz de frecuencias donde cada posición del motivo de unión contiene la frecuencia de cada nucleótido observada en todas las secuencias donde se ha observado interacción entre el TF y el ADN.

- La identificación de nuevos TFBS a partir de experimentos de ChIP-Seq nos permite identificar motivos probabilísticos que pueden usarse posteriormente para predecir otros posibles sitios de unión para dicho factor de transcripción.

- Esta aproximación tiene sus dificultades como proteínas que a veces interacciona directamente con el ADN y otras veces de manera indirecta a través de otras proteínas

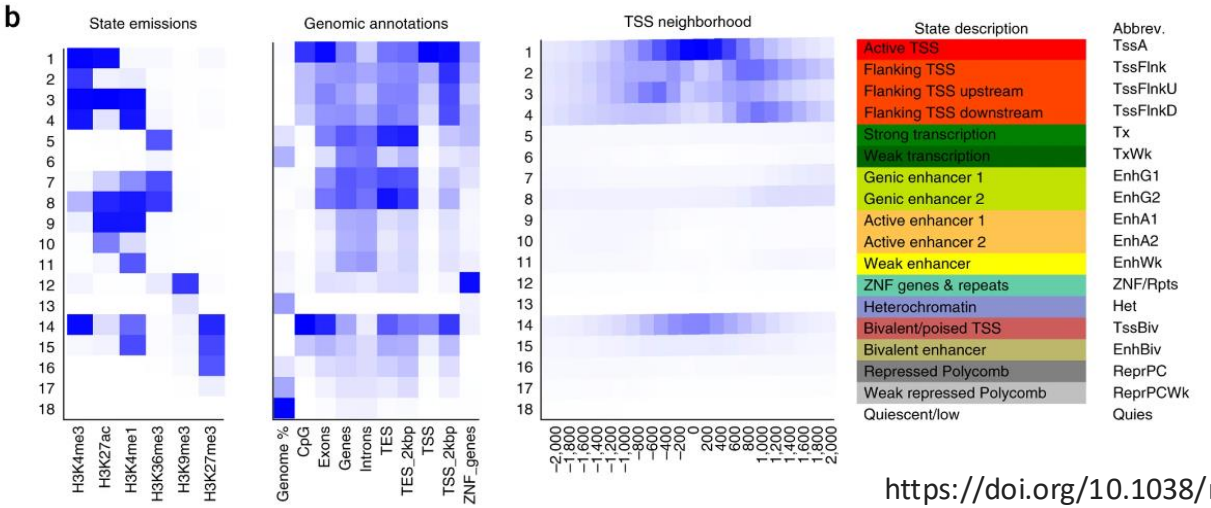
https://jaspar.genereg.net/



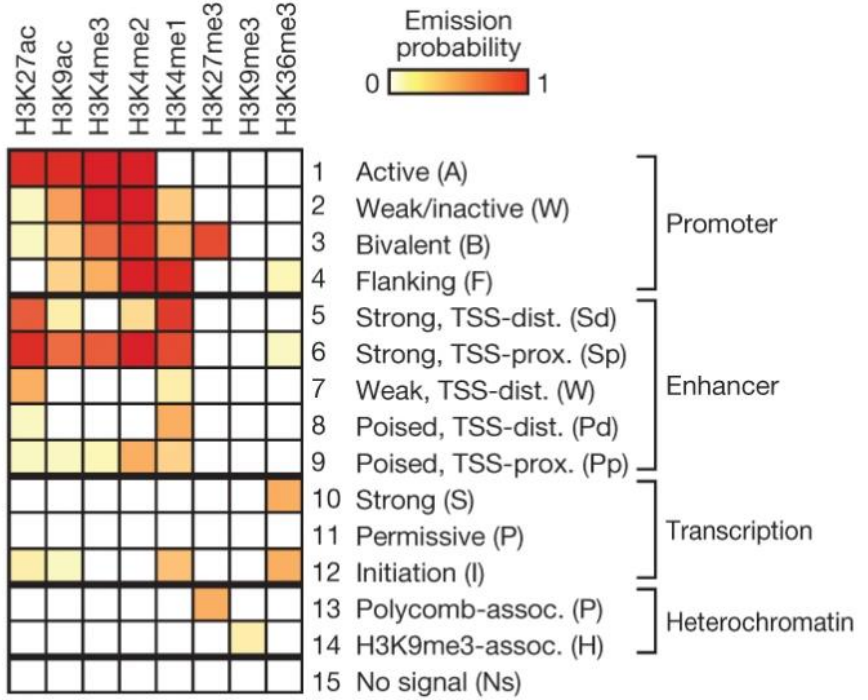
Estado de la cromatina

- Las modificaciones de las colas de las histonas pueden ser estudiadas de manera separada dando información importante de la regulación y la funcionalidad de ciertas regiones. Sin embargo, el **estudio combinado** de diferentes marcas **y su co-localización** con otros elementos reguladores amplían esta información, facilitando caracterizar el estado funcional de todas las regiones del genoma y descubrir nuevos elementos genómicos.

- La caracterización mediante múltiples modificaciones de las histonas se denomina **estado de la cromatina** y suele abordarse mediante algoritmos de segmentación de los picos identificados para todas las modificaciones, siendo los más utilizados algoritmos basados en **modelos ocultos de markov** (HMM).



https://doi.org/10.1038/nprot.2017.124



https://doi.org/10.1038/s41586-020-2093-3

Retos ChIP-seq

- Este método es capaz de **determinar la interacción** de una proteína a lo largo de un genoma o la presencia de ciertas **modificaciones en las histonas**.

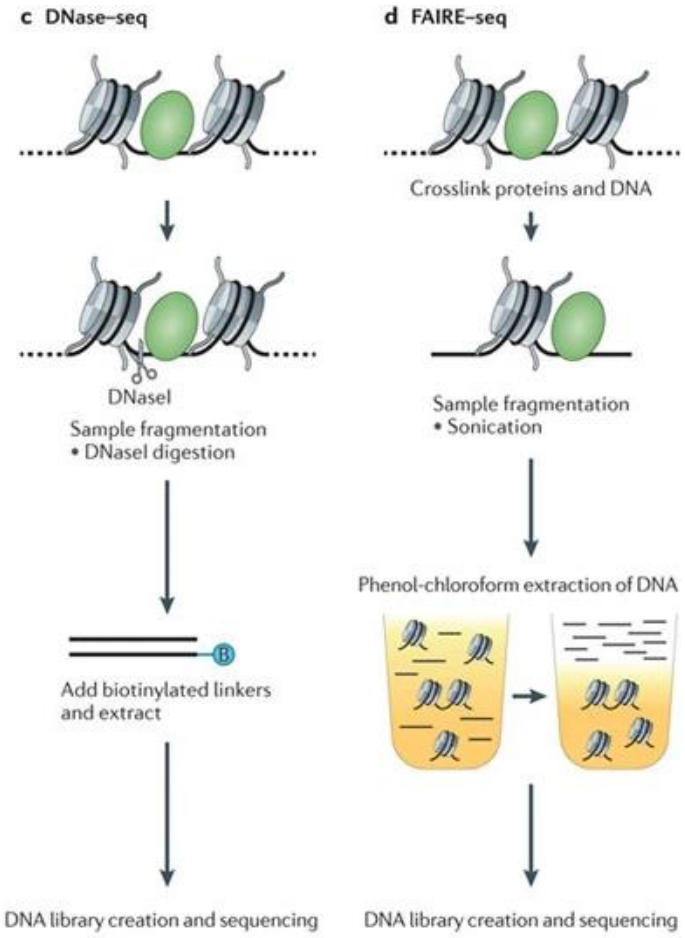
- Sin embargo, el éxito de la metodología depende del **desarrollo y la validación de anticuerpos muy específicos** para la molécula que se quiera estudiar. La especificidad de un anticuerpo incluso puede variar entre lotes de producción lo que complica poder replicar ciertos resultados.

- Por otro lado, la técnica requiere de **conocimiento previos** de las proteínas/modificaciones a estudiar impidiendo un análisis global de los factores de regulación del genoma y es necesario un **gran número de células de partida** (~10 millones), lo que complica su uso en muestras clínicas.

- Durante los últimos años se han ido desarrollando adaptaciones a esta metodología que mejoran considerablemente la usabilidad y fiabilidad de la técnica: Nano-ChIP-Seq, uso de concentraciones variables de anticuerpos, ChiP-Exo, re-ChIP...

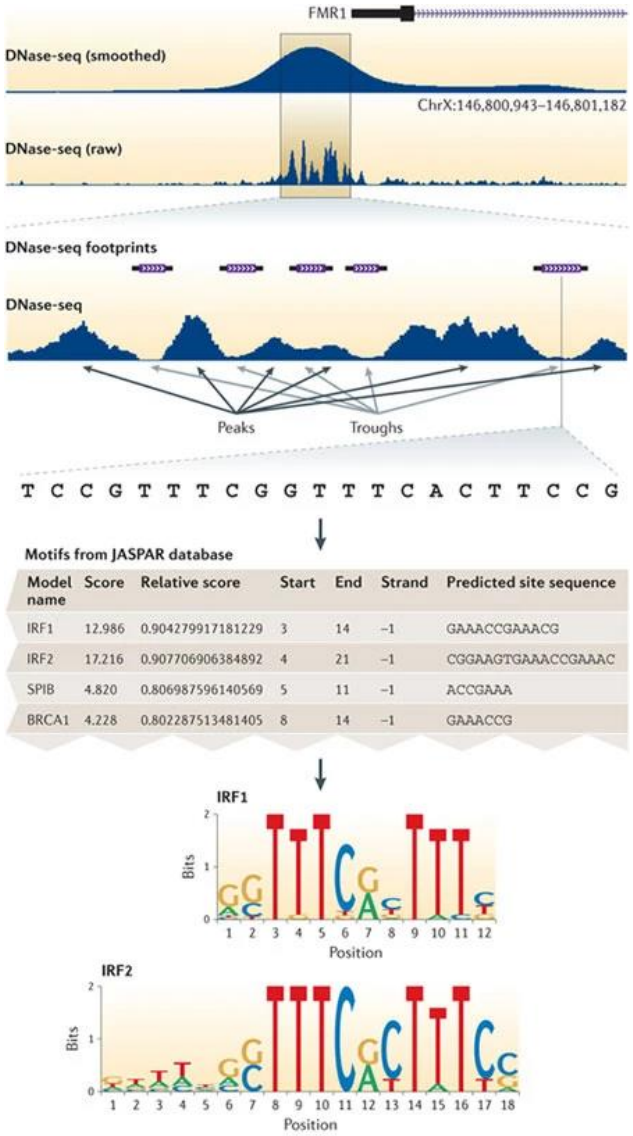
Identificar sitios abiertos de la cromatina

- El protocolo ChIP-Seq es muy complejo y requiere de información previa, además cada experimento debe centrarse en una proteína/modificación lo que implica múltiples experimentos para estudiar la regulación génica en un contexto amplio.
- La mayoría de los factores de transcripción necesitan una región libre de nucleosomas (región abierta de la cromatina, eucromatina) para interaccionar con el ADN. El desarrollo de tecnologías para localizar estas regiones y el amplio conocimiento de motivos de unión al ADN, permiten mediante un experimento (más simple que el ChIP-Seq) identificar regiones abiertas y predecir las los TFs que puedan estar interaccionando.
- Las metodologías más empleadas originalmente para identificar regiones abiertas de la cromatina a lo largo de todo el genoma fueron DNase-seq y FAIRE-seq. El primero se basa en la digestión no específica del ADN accesible por parte de la endonucleasa DNasa I, los fragmentos digeridos son capturados y secuenciados dando como resultado picos accesibles por la endonucleasa. Mientras que el FAIRE-Seq se basa en la unión covalente de las proteínas con el ADN, sonicación y extracción del ADN mediante fenol-cloroformo, quedando la fracción del ADN sin nucleosomas en la fase acuosa.



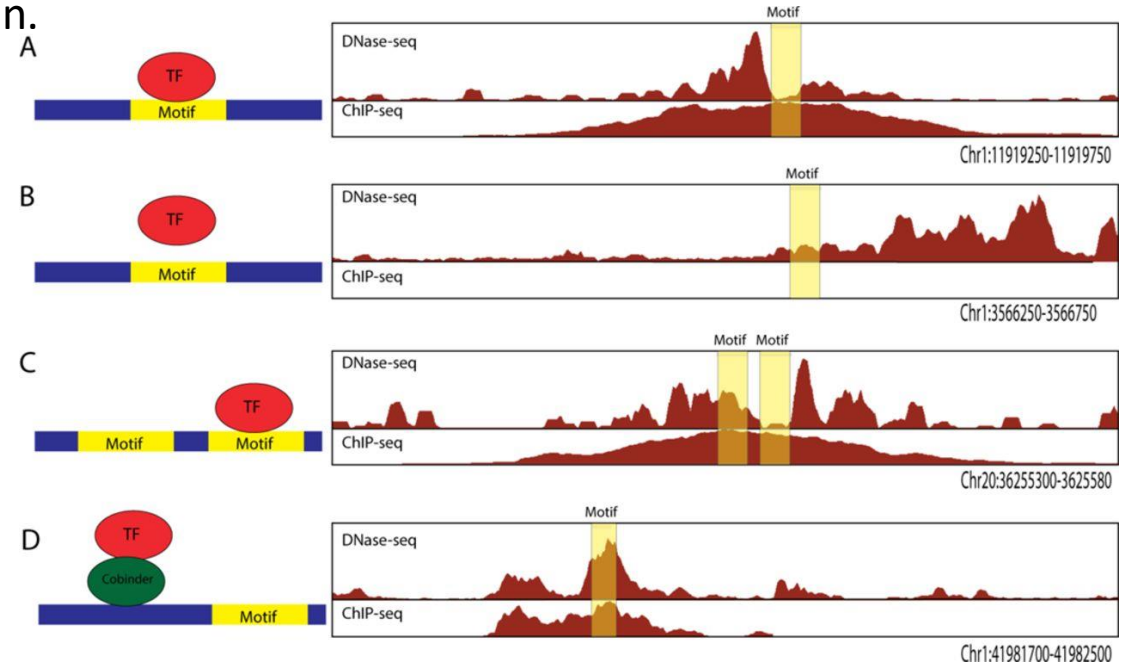
https://doi.org/10.1038/nrg3306

Identificar sitios abiertos de la cromatina



https://doi.org/10.1038/nrg3306

- Las regiones donde se concentran las **lecturas secuenciadas suelen ser regiones libres de nucleosomas**, indicando zonas abiertas de la cromatina. Sin embargo, para identificar posibles motivos de unión debemos fijarnos en los **valles dentro de las zonas accesibles (*footprints*)**, ya que estas son las regiones donde probablemente están los motivos de unión y por lo tanto el TF protegiendo de la digestión.



- Junto con experimentos de ChIP-Seq podemos identificar múltiples combinaciones de situaciones: interacciones, motivos sin interacción, múltiples motivos algunos con y otros sin interacción e interacciones indirectas.

Identificar sitios abiertos de la cromatina

- Actualmente el protocolo más utilizado es el ATAC-Seq. Este se basa una vez más en la accesibilidad que permiten los sitios no cubiertos por nucleosomas, pero en este caso se aísla el núcleo de las células conservando intacta la cromatina, para posteriormente **exponer el ADN a una transposasa (Tn5) que a su vez liga los adaptadores**. Sólo aquellos fragmentos en los que se ha producido la inserción de la Tn5 son secuenciados, identificados mediante un adaptador diferente en cada extremo.

- Este protocolo es mucho más sencillo, rápido y requiere de menos calibración que los dos anteriores, lo que también lo hace más replicable. Además, se puede realizar con muchas menos células. Esto lo ha convertido en el protocolo por excelencia actualmente, aunque por el momento sus sesgos son menos conocidos que para el DNase-seq.