

Bioinformática y análisis de datos ómicos

TEMA 10: IDENTIFICACIÓN DE ALTERACIONES EPIGENÓMICAS



Ignacio Varela Egocheaga

DEPARTAMENTO DE BIOLOGÍA MOLECULAR

Este material se publica bajo la siguiente licencia:

[Creative Commons BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)

Tema 10: Identificación de alteraciones epigenómicas

- 10.1 Identificación de patrones de metilación
- 10.2 Identificación de sitios de unión de proteínas
- 10.3 Caracterización de alteraciones en la accesibilidad de la cromatina.
- 10.4 Caracterización de la estructura tridimensional de la cromatina
- 10.5 Formato de archivos de regiones genómicas

Tema 10: Identificación de alteraciones epigenómicas

10.1 Identificación de patrones de metilación

10.2 Identificación de sitios de unión de proteínas

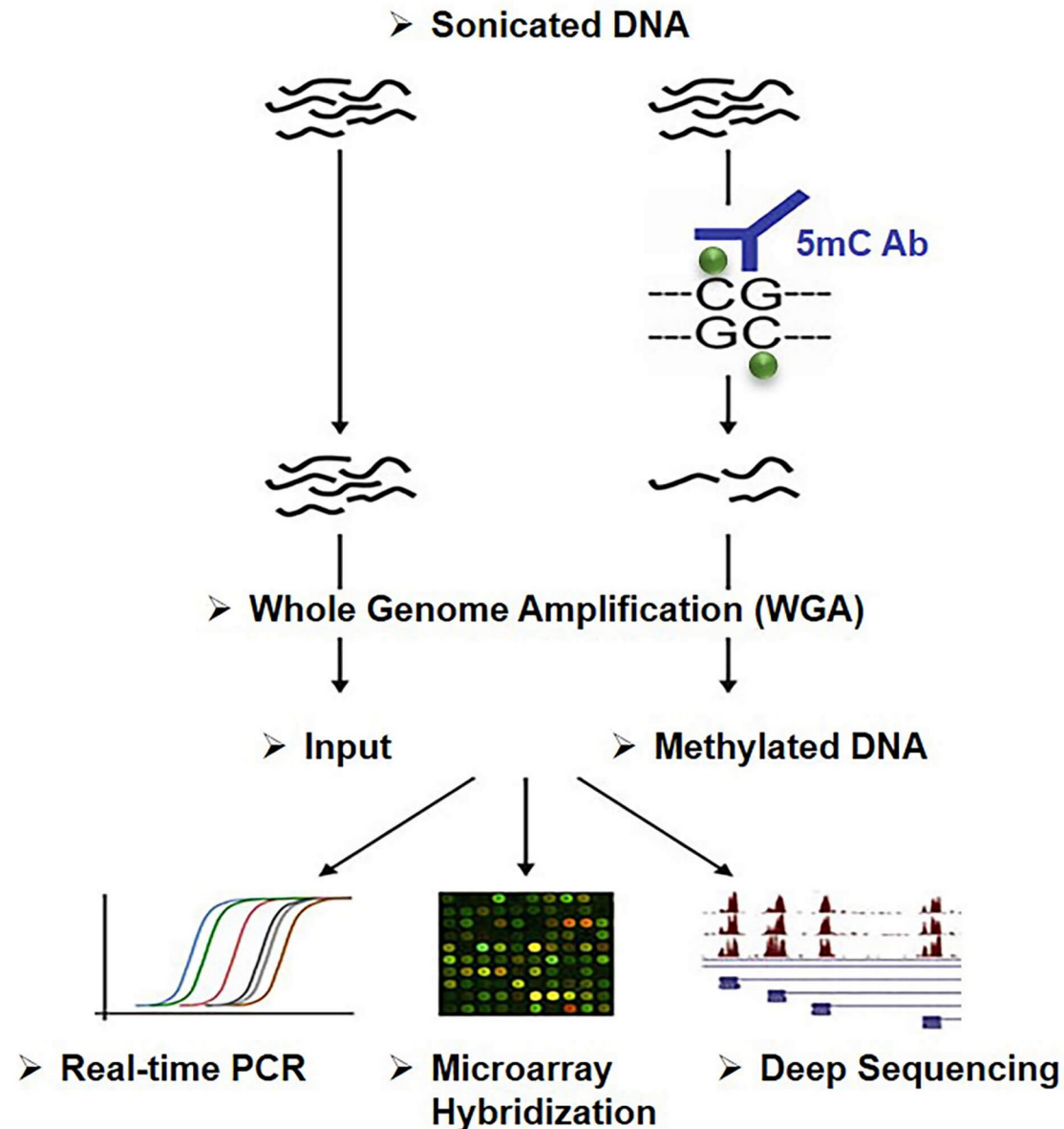
10.3 Caracterización de alteraciones en la accesibilidad de la cromatina.

10.4 Caracterización de la estructura tridimensional de la cromatina

10.5 Formato de archivos de regiones genómicas

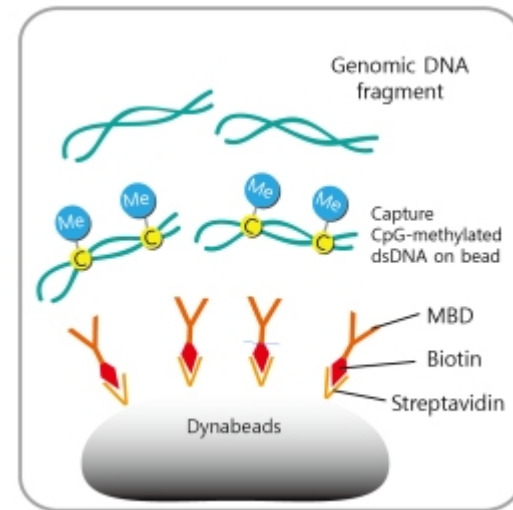
Enriquecimiento de regiones metiladas

Anticuerpos anti-5mC

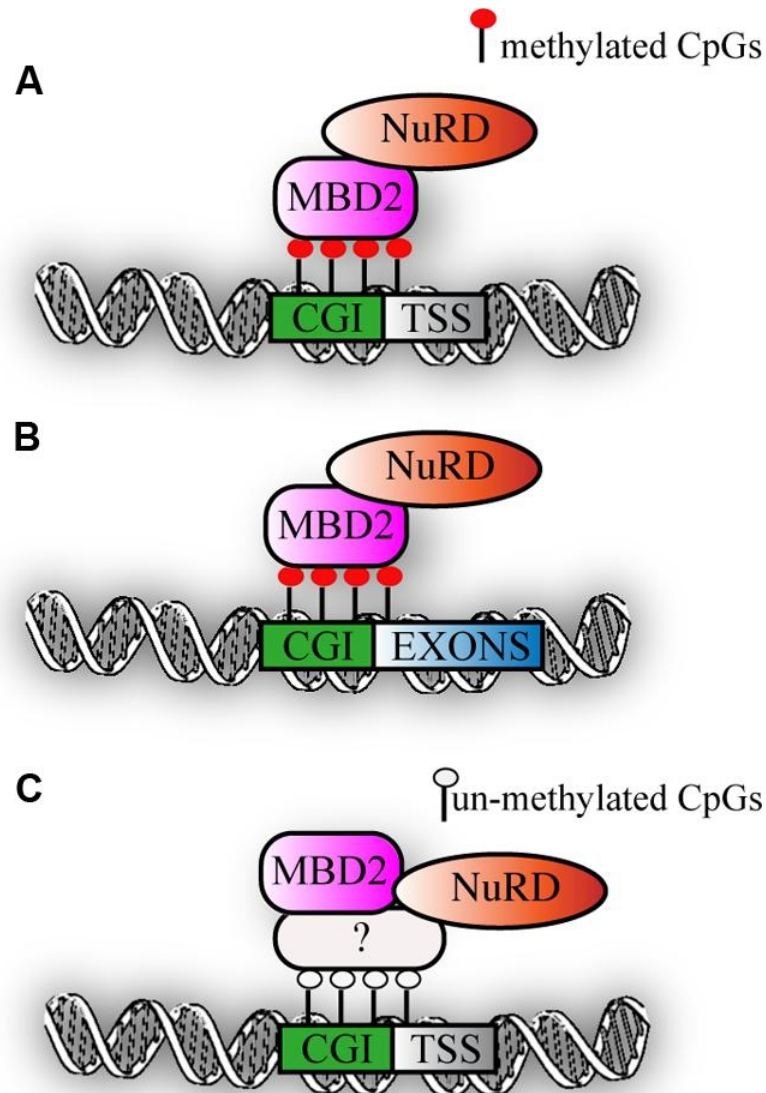


Enriquecimiento de regiones metiladas

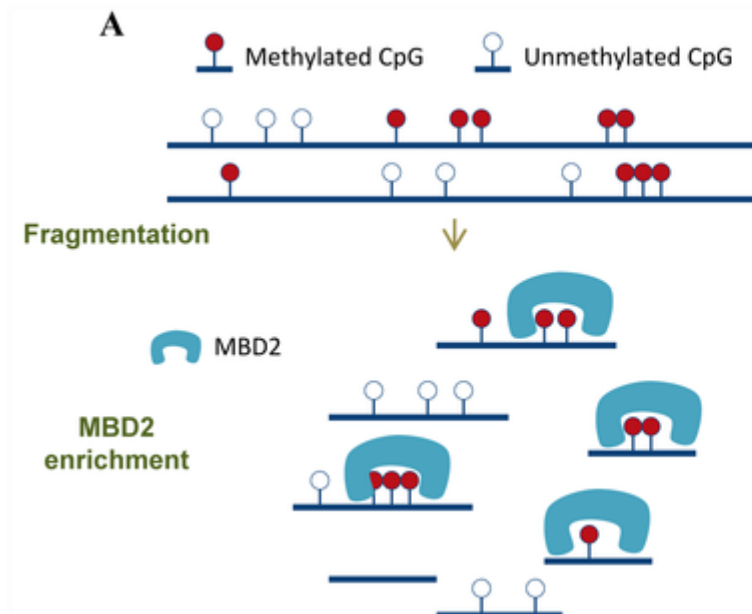
Dominio de unión a DNA metilado de MBD2



https://www.dnalink.com/english/service/capture_sequencing.html

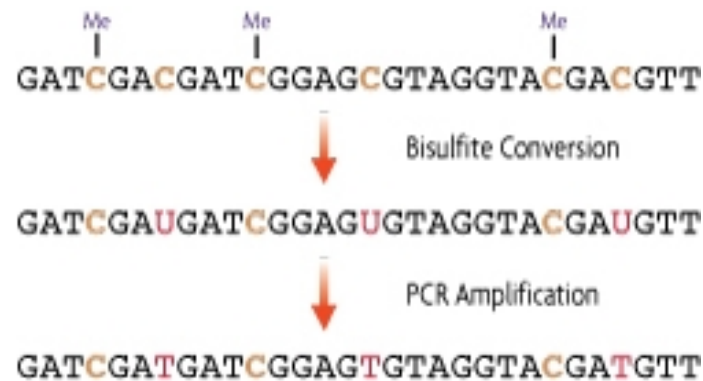
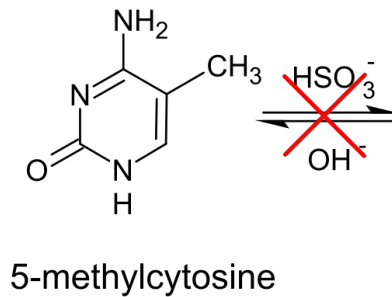
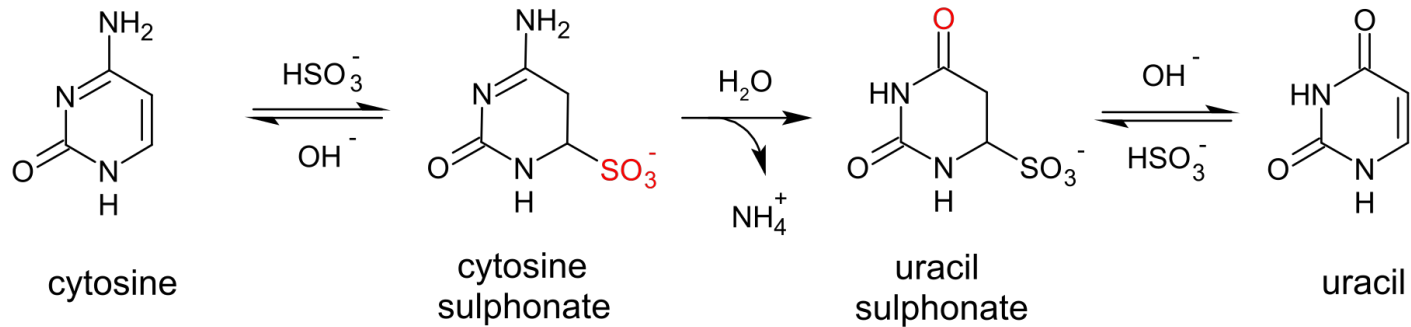


https://www.researchgate.net/figure/MBD2-binding-to-chromatin-A-Schematic-representation-of-MBD2-binding-to-highly_fig1_270003573



https://www.researchgate.net/figure/Overview-of-experimental-design-and-analysis-A-Major-steps-of-MBD-seq-experiment_fig3_51511829

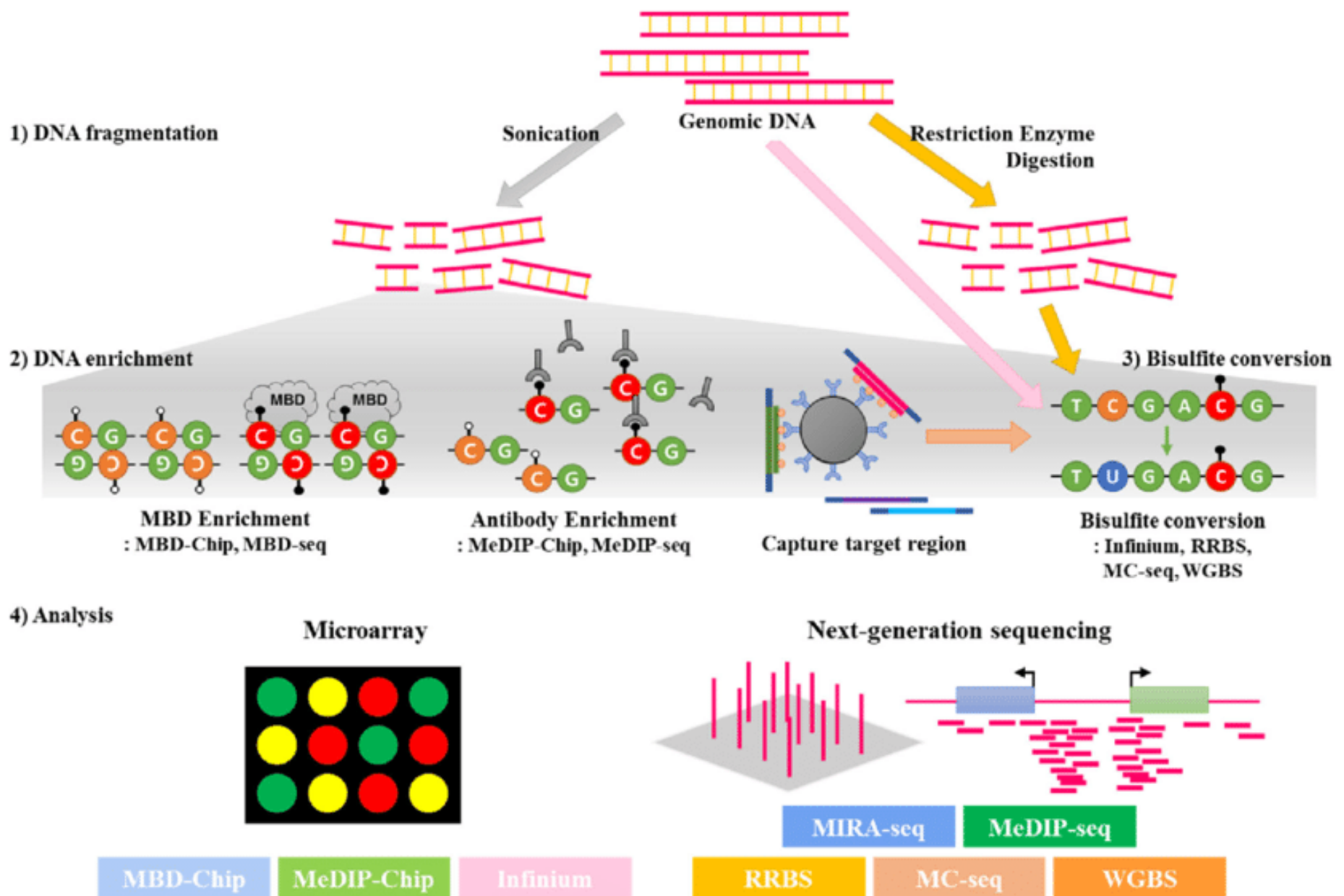
Conversión con bisulfito



https://en.wikipedia.org/wiki/Bisulfite_sequencing

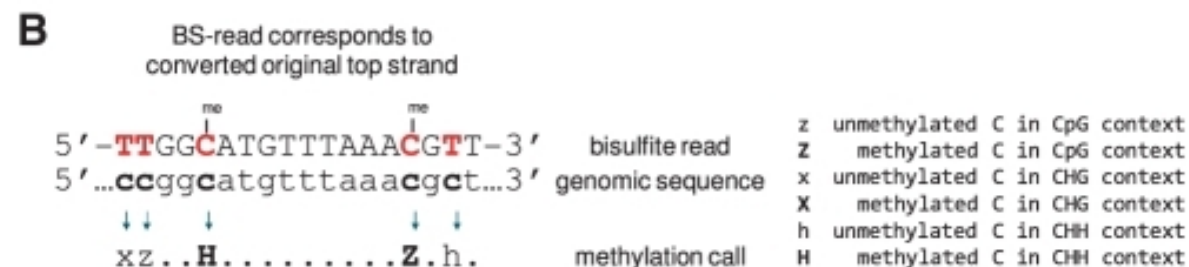
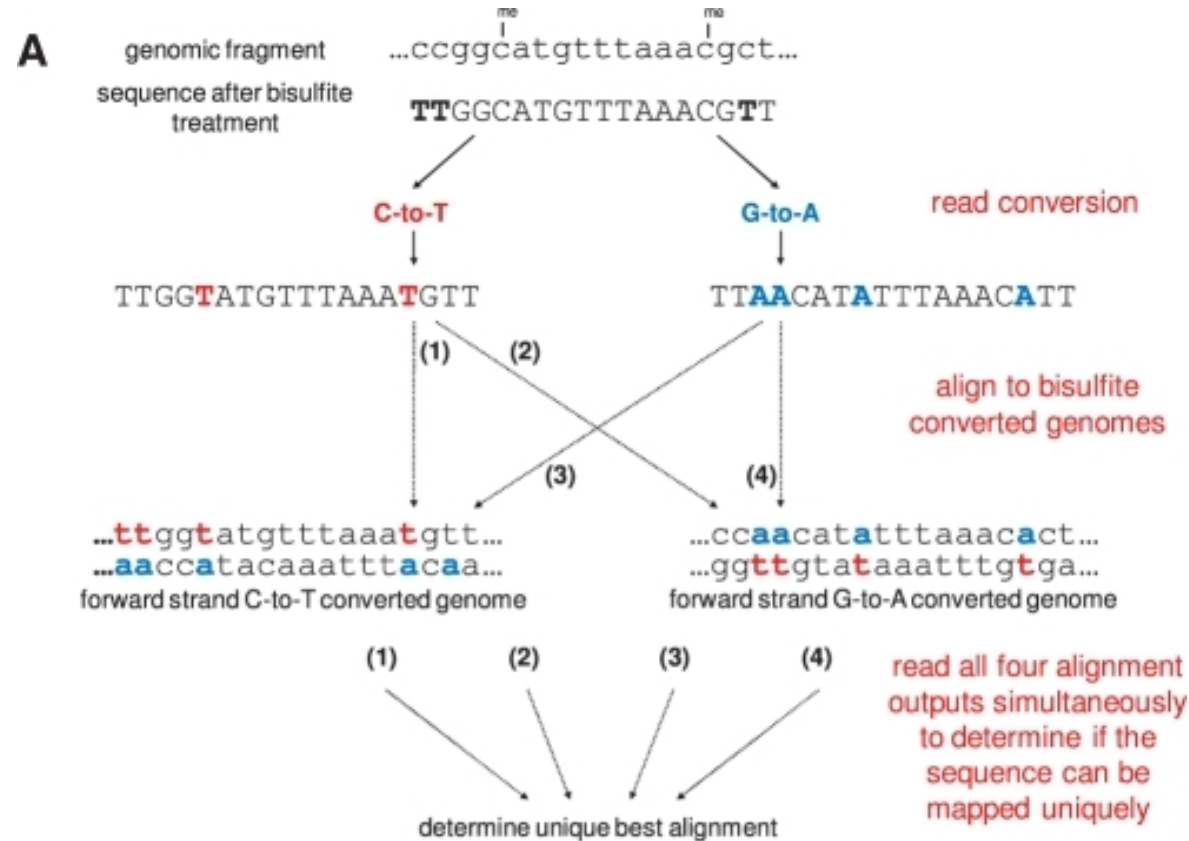
<https://www.activemotif.com/catalog/695/bisulfite-conversion>

Distintas variantes de análisis de metilación

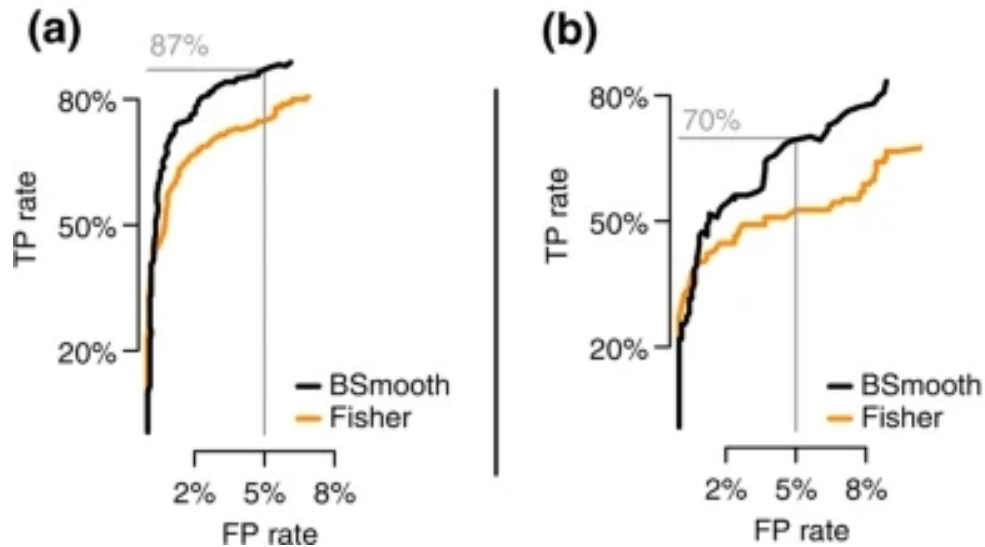


Alineamiento de lecturas convertidas

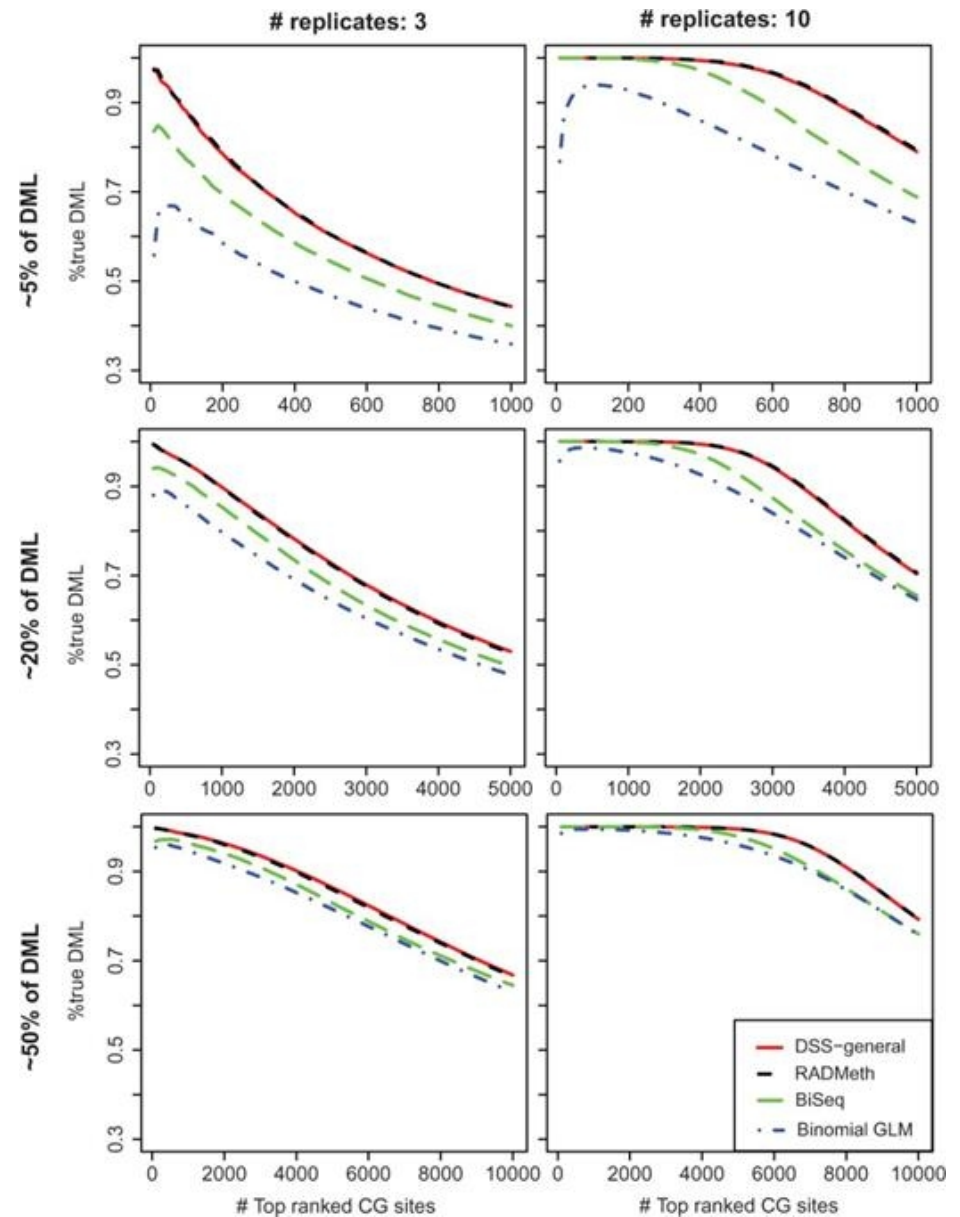
Alineadores específicos como **Bismark** o **BS-Seeker2** permiten convertir in vitro el genoma de referencia para el alineamiento y cuantificar las regiones metiladas



Identificación en el grado de metilación



Hansen K et al. Genome Biol 13:R83



Park Y et al. Bioinformatics 32:1446-1453

Tema 10: Identificación de alteraciones epigenómicas

10.1 Identificación de patrones de metilación

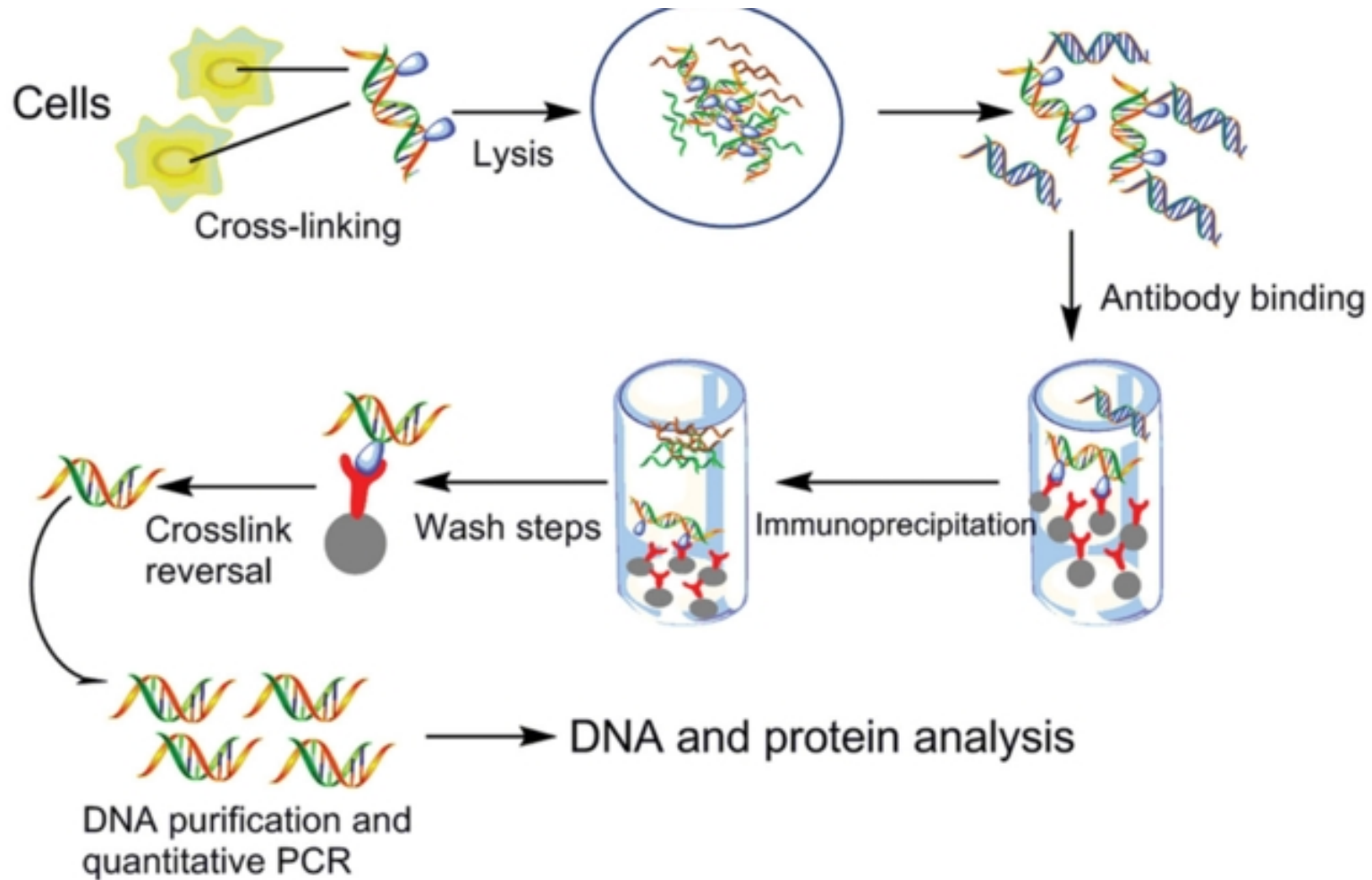
10.2 Identificación de sitios de unión de proteínas

10.3 Caracterización de alteraciones en la accesibilidad de la cromatina.

10.4 Caracterización de la estructura tridimensional de la cromatina

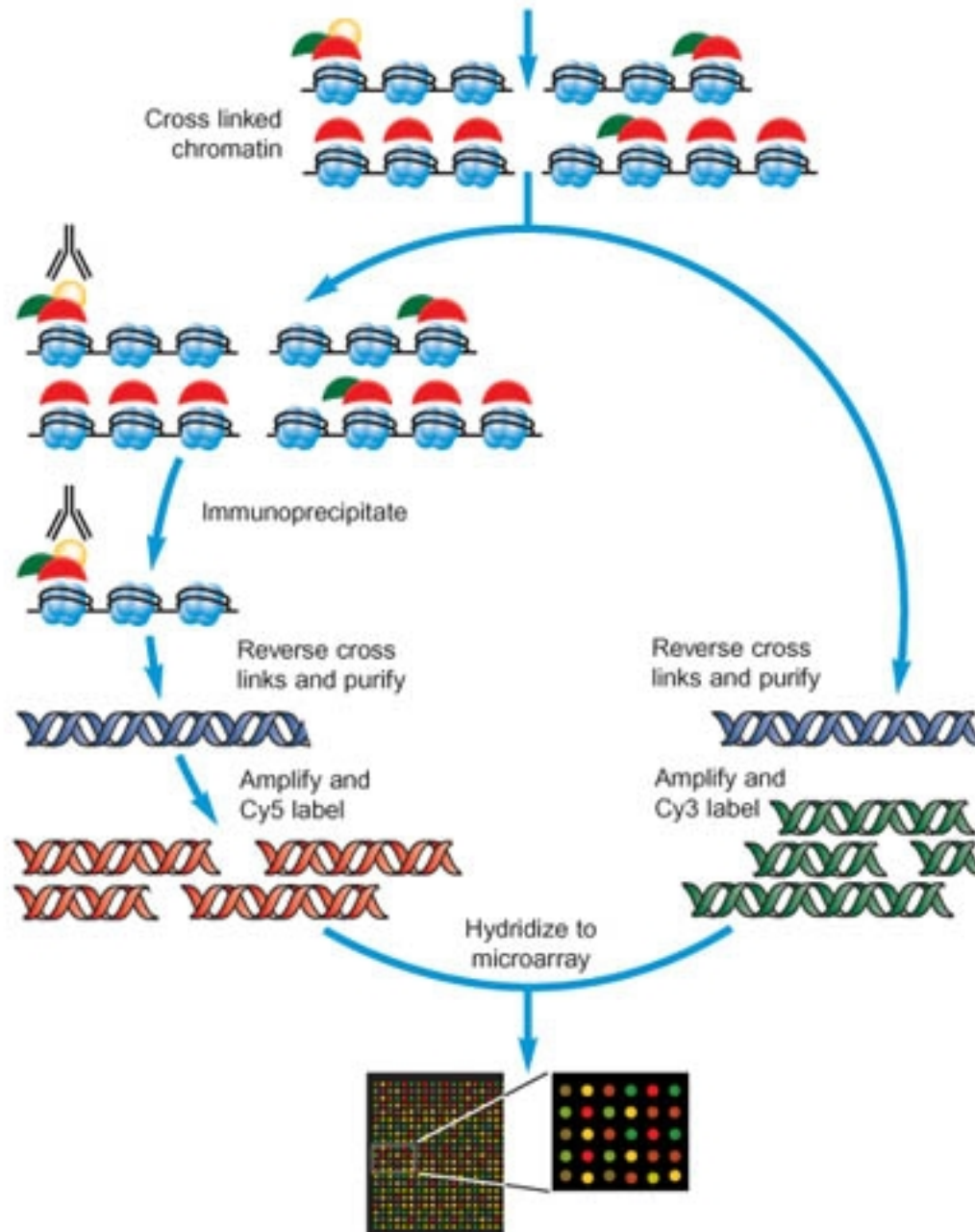
10.5 Formato de archivos de regiones genómicas

Inmunoprecipitación de cromatina



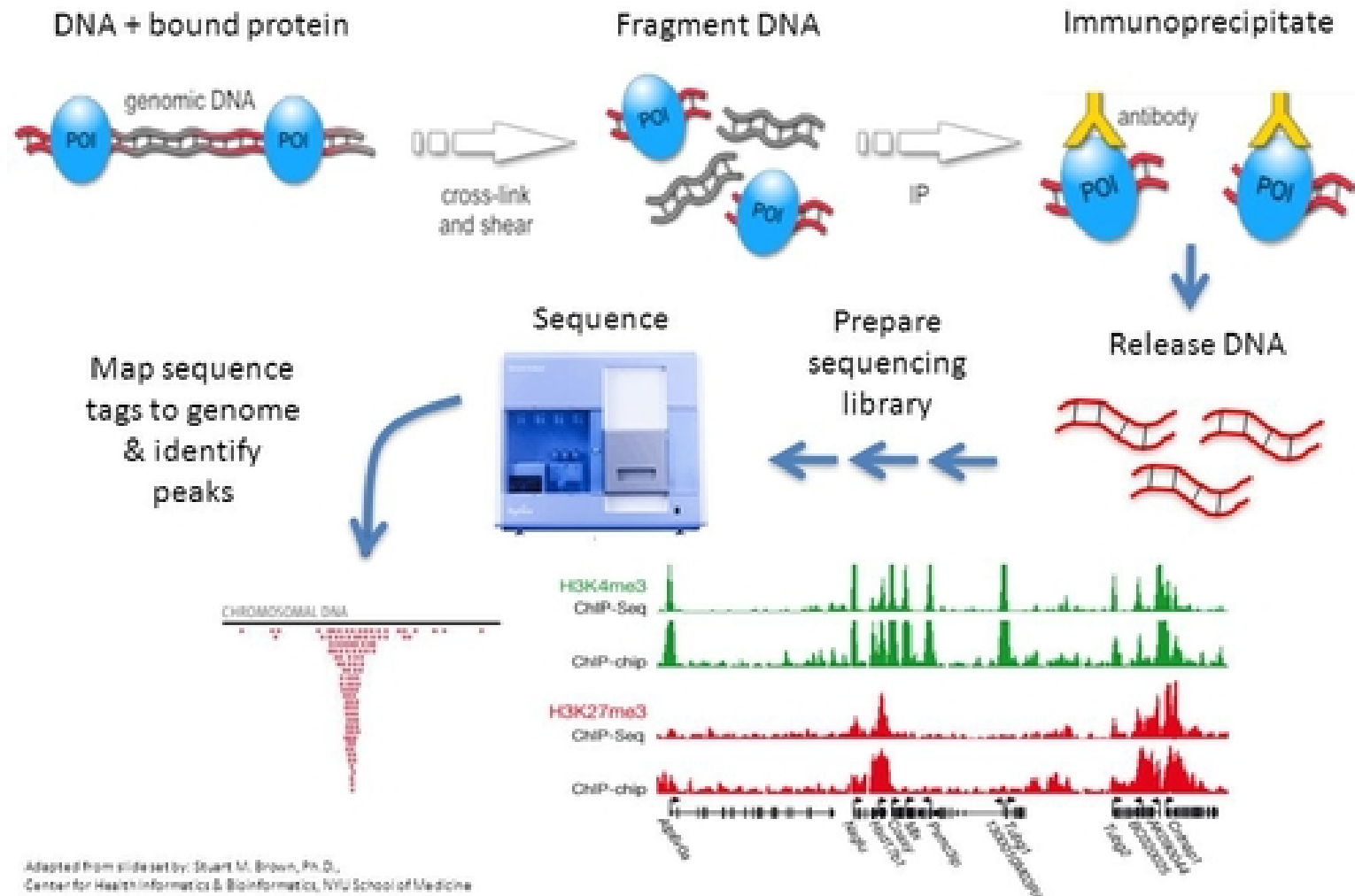
<https://www.sigmaaldrich.com/ES/es/technical-documents/technical-article/genomics/gene-expression-and-silencing/chromatin-immunoprecipitation-chip>

ChIP on chip



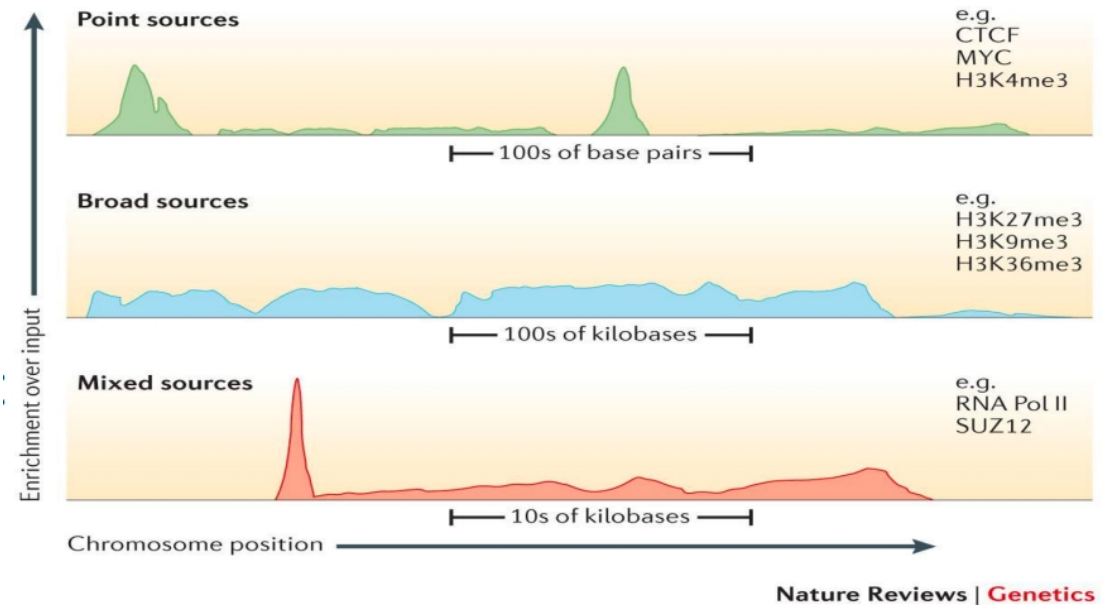
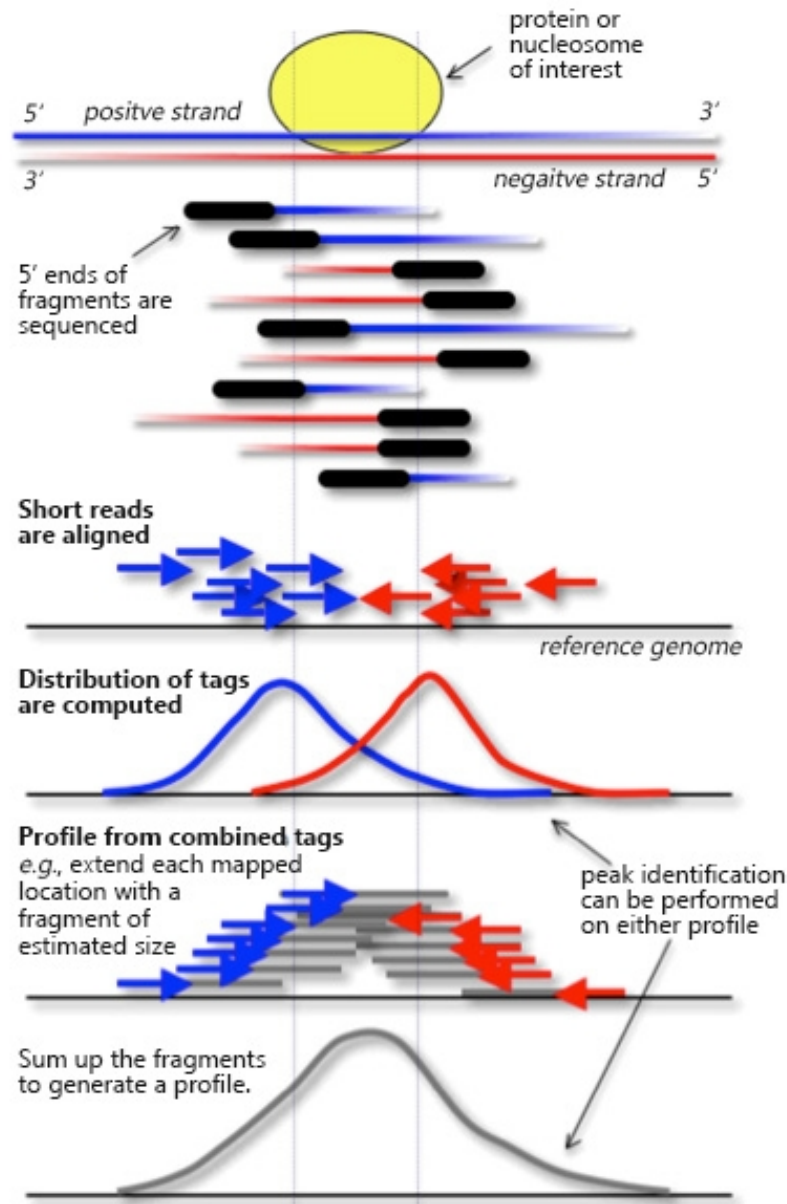
ChIP-seq

ChIP-seq overview



<https://www.indiamart.com/proddetail/chip-sequencing-service-15911548030.html>

Identificación de las regiones de unión



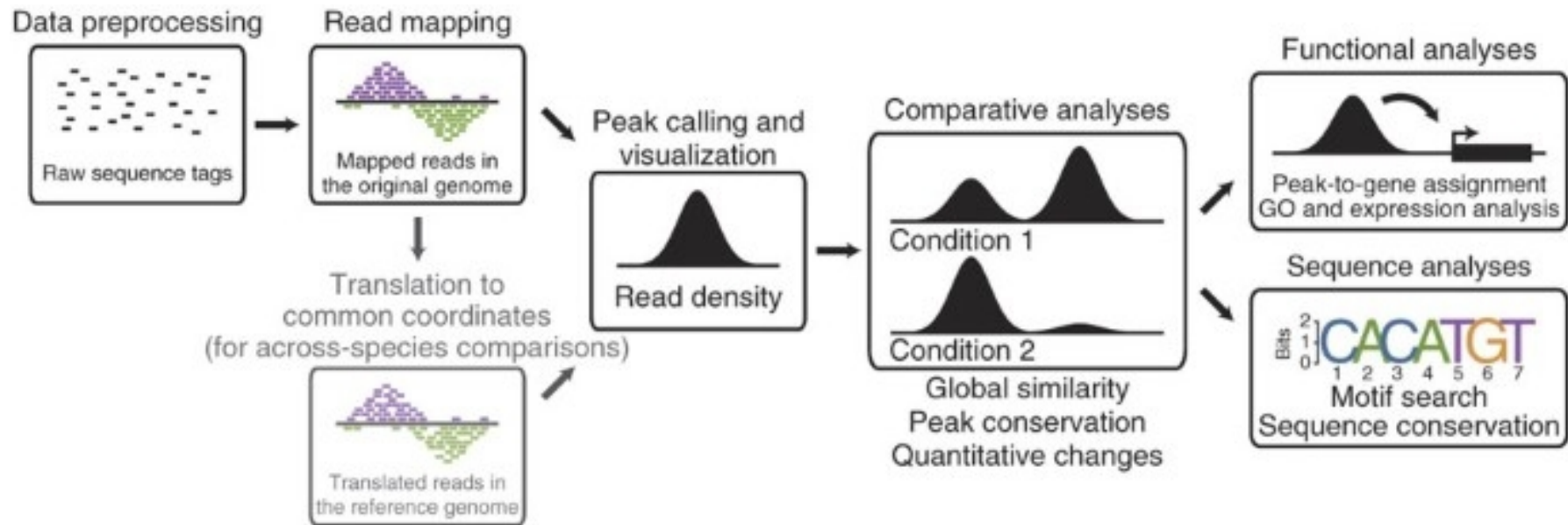
https://ressources.france-bioinformatique.fr/sites/default/files/EBAI2017_ChIPSeq_a_imprimer.pdf

Programas como **MACs** permiten la identificación de regiones o “picos” de enriquecimiento y tiene parámetros para buscar picos estrechos o más extendidos.

<https://www.profacgen.com/chromatin-immunoprecipitation-chip-assay-service.htm>

Identificación de diferencias de enriquecimiento

Una vez identificadas las regiones enriquecidas, la calidad cuantitativa de la NGS permite identificar diferencias en el factor de enriquecimiento con diversas herramientas como **DESeq2**.

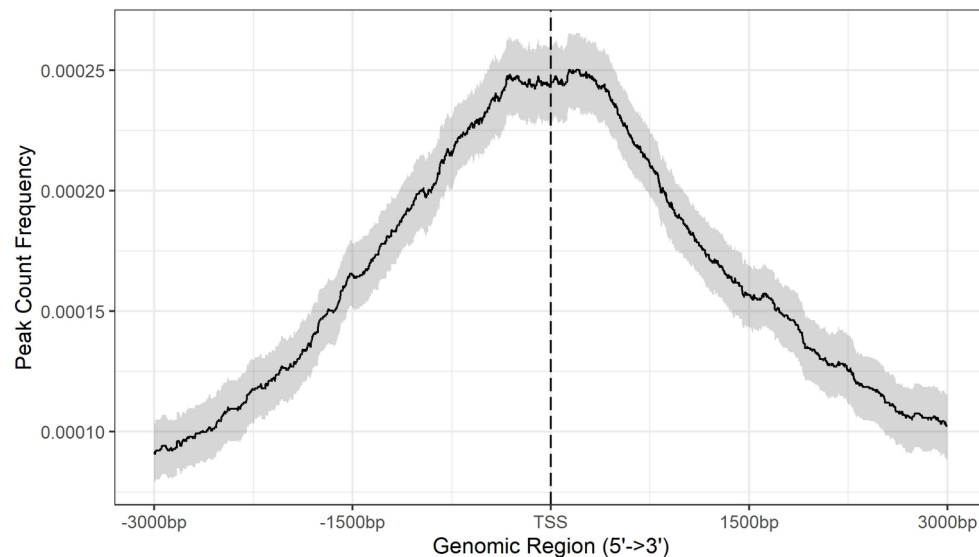
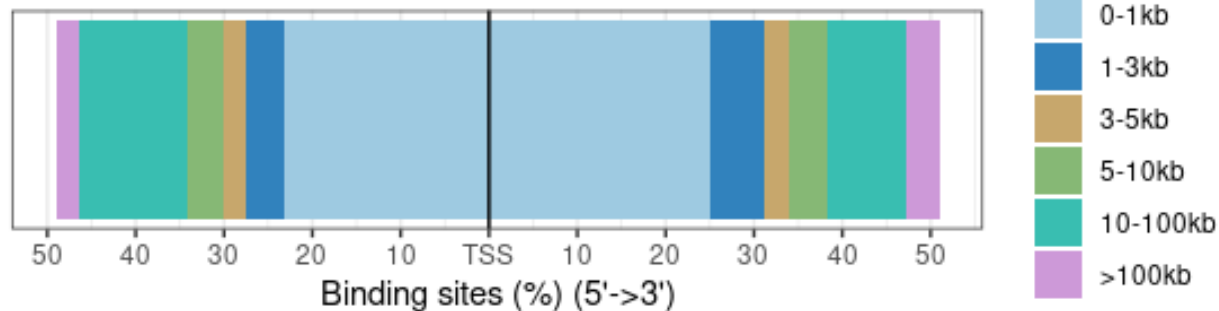


Bardet AF et al. Nat Protocols 7:45-61

Estudio de la localización de las regiones de unión

Distintas herramientas como **ChIPseeker** o **GREAT** permiten analizar los sitios de unión con respecto a elementos funcionales del genoma

Distribution of transcription factor-binding loci relative to TSS



Identificación de motivos de unión

Herramientas como **HOMMER** permite identificar la presencia de motivos de secuencia en las regiones enriquecidas más frecuentes de lo esperado por azar.



$$SimilarityScore = \frac{1}{MotifLength} \sum_i^{MotifLength} - \frac{(Observed_i - Expect_i)}{Expect_i}$$

$$Observed_i = \sum_j^{A,C,G,T} - (freq1_{ij} - freq2_{ij})^2$$

$$Expect_i = \sum_j^{A,C,G,T} \sum_k^{A,C,G,T} \frac{-(freq1_{ij} - freq2_{ik})^2}{4}$$

Rank	Motif	P-value	log P-value	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		1e-1835	-4.228e+03	28.11%	5.16%	37.7bp (63.1bp)	NFkB-p65(RHD)/GM12787-p65-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)
2		1e-1716	-3.953e+03	34.50%	8.65%	47.8bp (62.6bp)	PB0058.1_Sfpi1_1 More Information Similar Motifs Found	motif file (matrix)
3		1e-1585	-3.651e+03	28.85%	6.39%	41.8bp (62.8bp)	MA0102.1_Cebpa More Information Similar Motifs Found	motif file (matrix)
4		1e-1004	-2.314e+03	25.07%	7.22%	49.2bp (61.0bp)	NF-E2(bZIP)/K562-NFE2-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)
5		1e-262	-6.039e+02	5.52%	1.27%	47.8bp (59.7bp)	Oct2(POU/Homeobox)/Bcell-Oct2-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)
6		1e-198	-4.565e+02	6.42%	2.09%	49.9bp (53.2bp)	c-Jun-CRE(bZIP)/K562-cJun-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)
7		1e-146	-3.370e+02	7.66%	3.30%	51.7bp (58.9bp)	RUNX1(Runt)/Jurkat-RUNX1-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)

Tema 10: Identificación de alteraciones epigenómicas

10.1 Identificación de patrones de metilación

10.2 Identificación de sitios de unión de proteínas

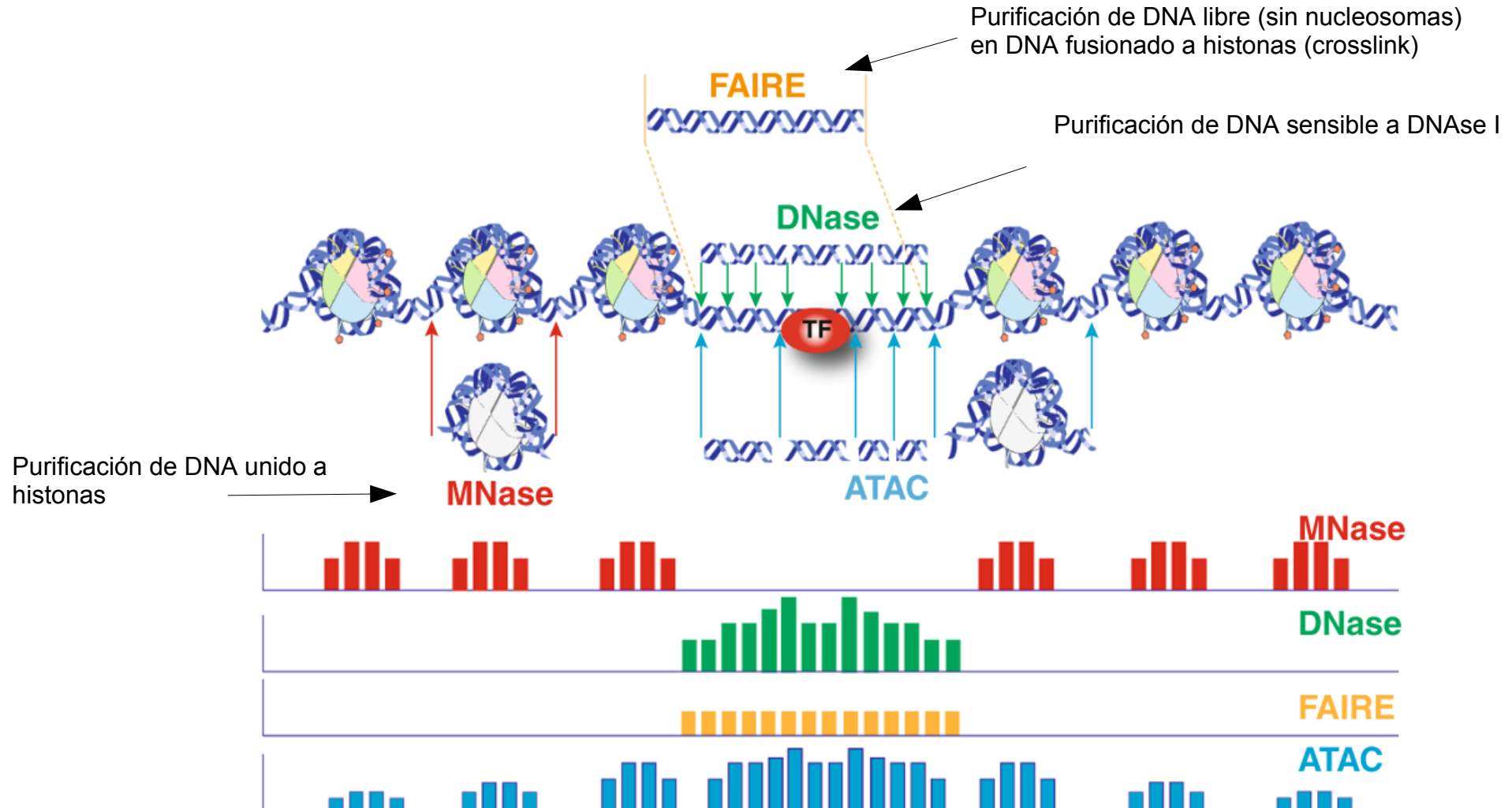
10.3 Caracterización de alteraciones en la accesibilidad de la cromatina.

10.4 Caracterización de la estructura tridimensional de la cromatina

10.5 Formato de archivos de regiones genómicas

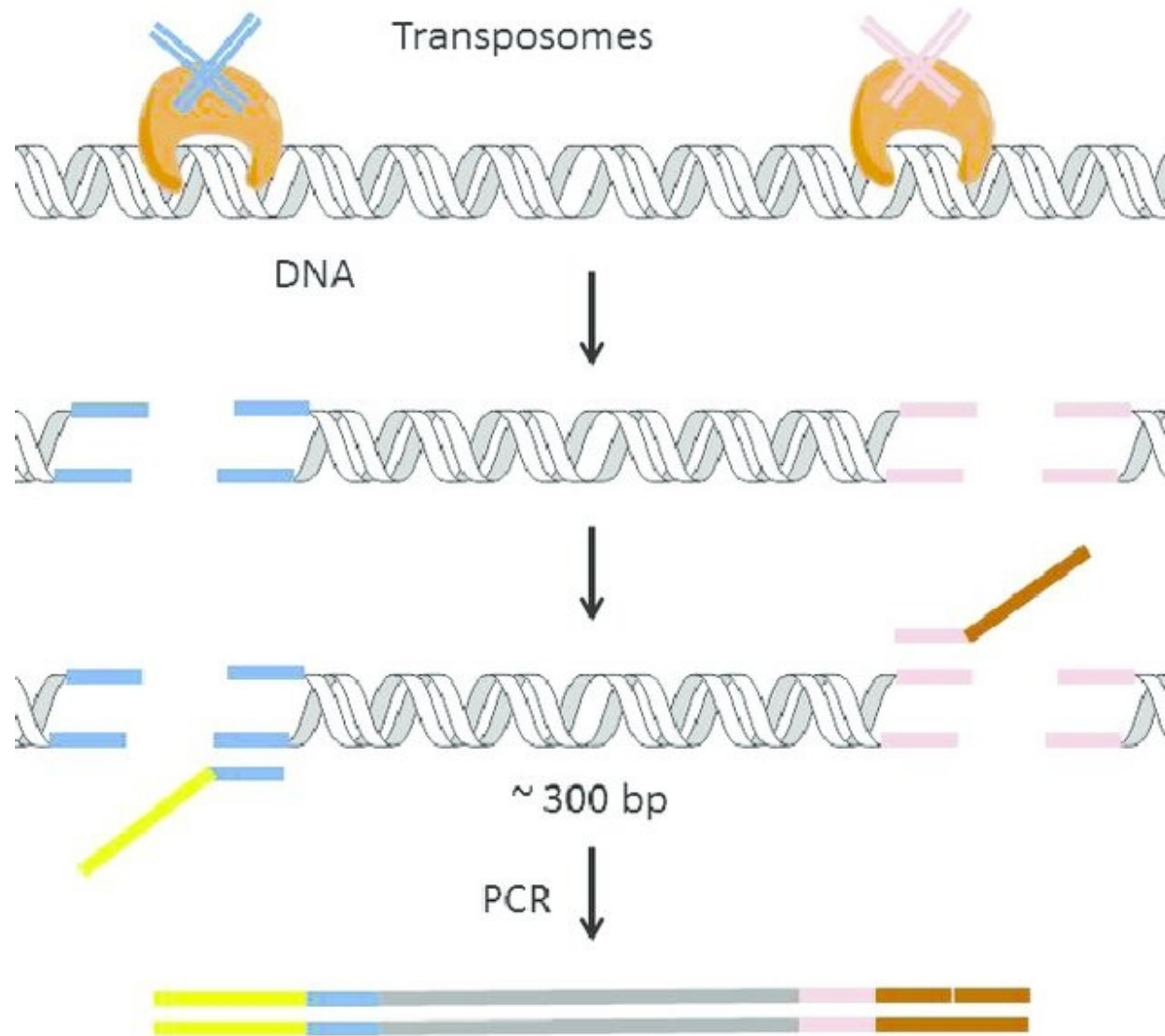
Generación de librerías en regiones abiertas de la cromatina

Existen distintos protocolos de generación de librerías para identificar las regiones de cromatina abierta.



<https://yiweiniu.github.io/blog/2019/03/ATAC-seq-data-analysis-from-FASTQ-to-peaks/>

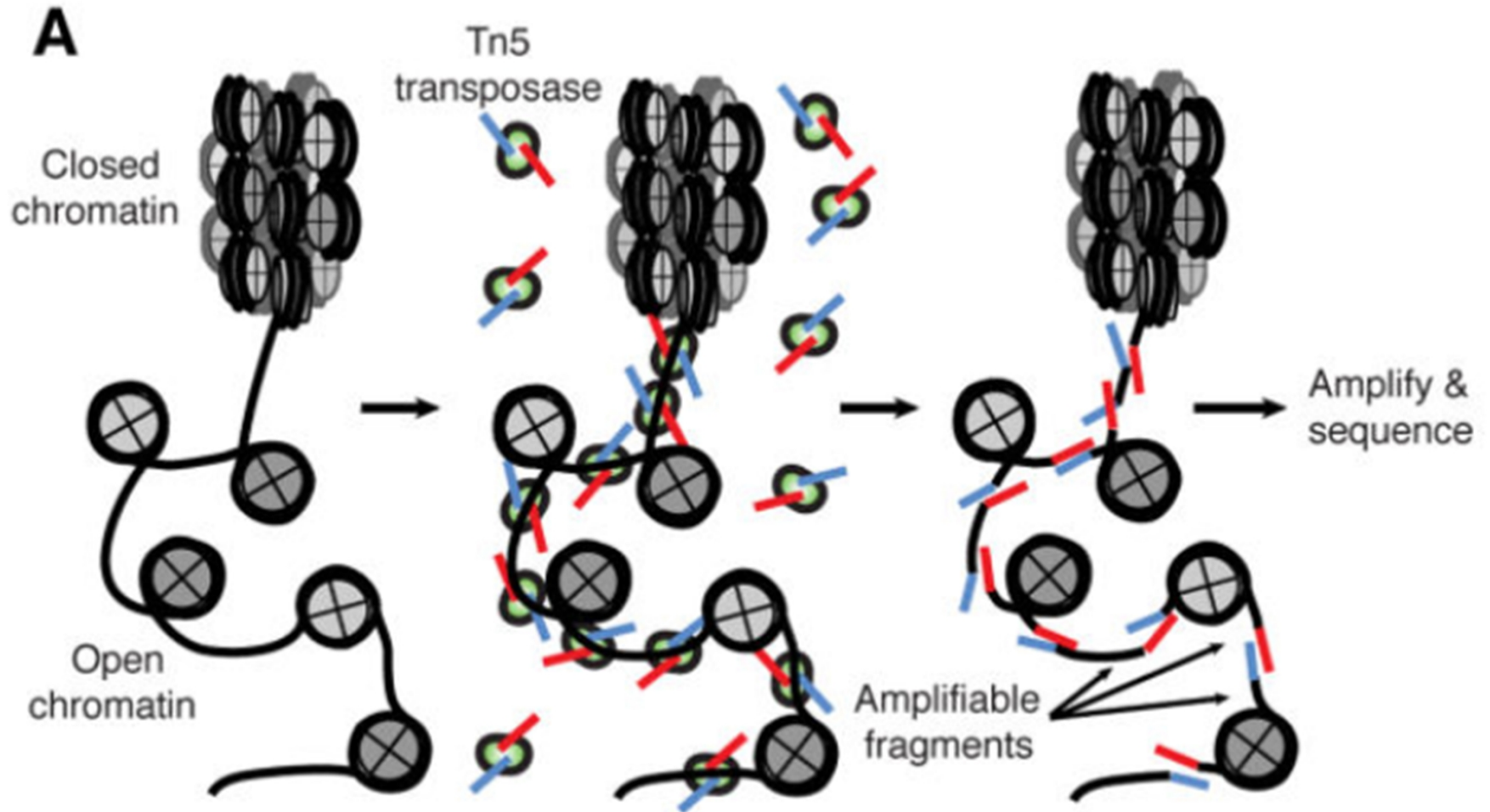
Generación de librerías utilizando trasposasa



Head SR *et al.* Biotechniques 56:61

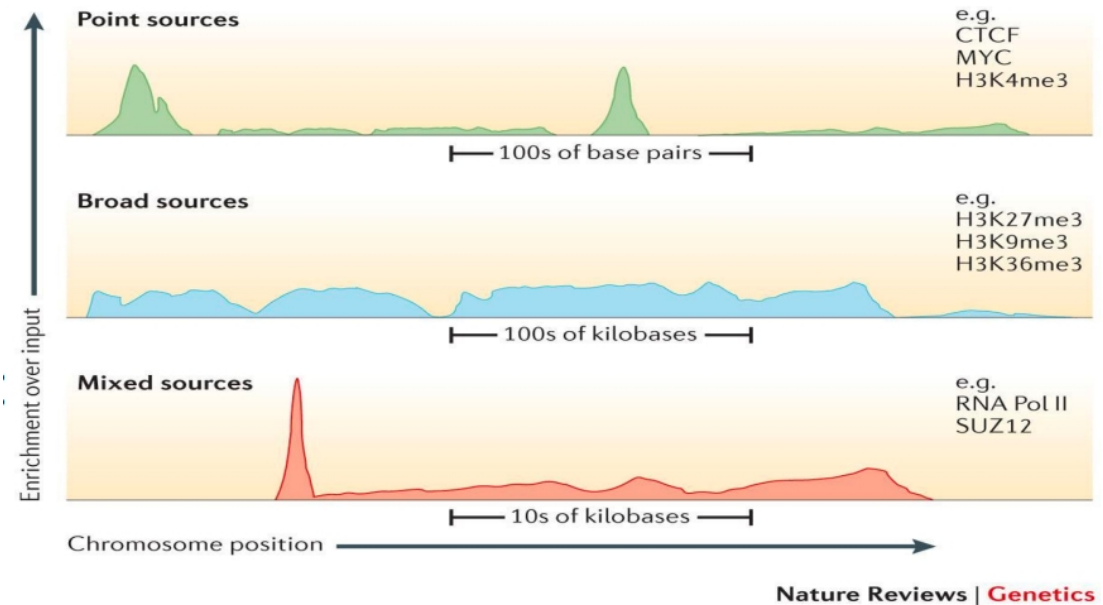
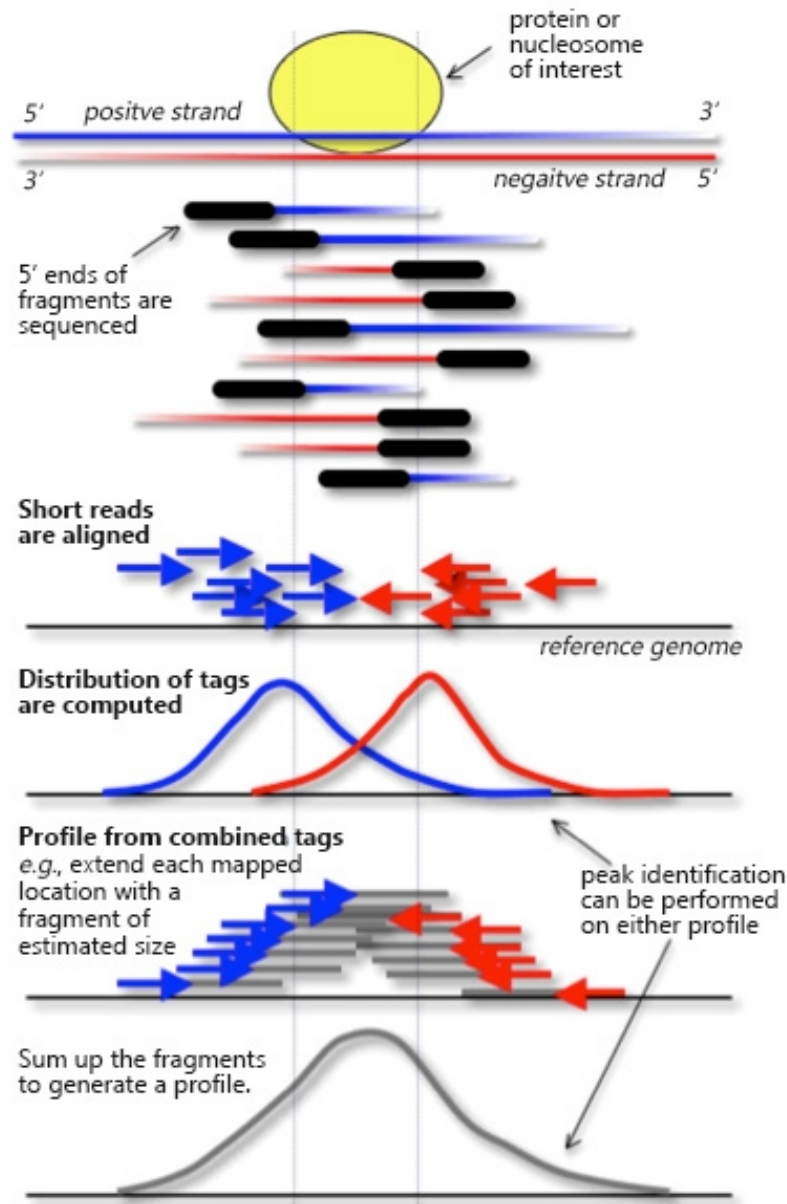
Identificación de regiones accesibles

El “ataque” de la trasposasa Tn5 sobre DNA no desnaturalizado permite identificar las regiones de cromatina abierta porque están más expuestas.



<https://informatics.fas.harvard.edu/atac-seq-guidelines.html>

Identificación de las regiones de unión



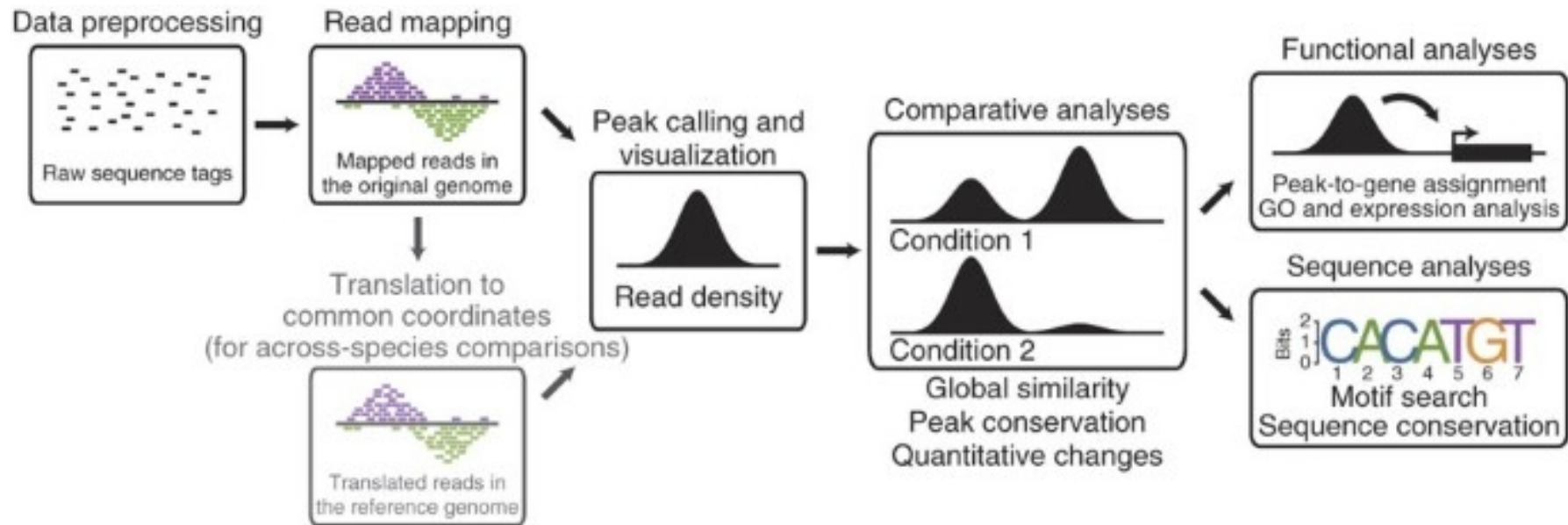
https://ressources.france-bioinformatique.fr/sites/default/files/EBAI2017_ChIPSeq_a_imprimer.pdf

Programas como **MACs** permiten la identificación de regiones o “picos” de enriquecimiento y tiene parámetros para buscar picos estrechos o más extendidos.

<https://www.profacgen.com/chromatin-immunoprecipitation-chip-assay-service.htm>

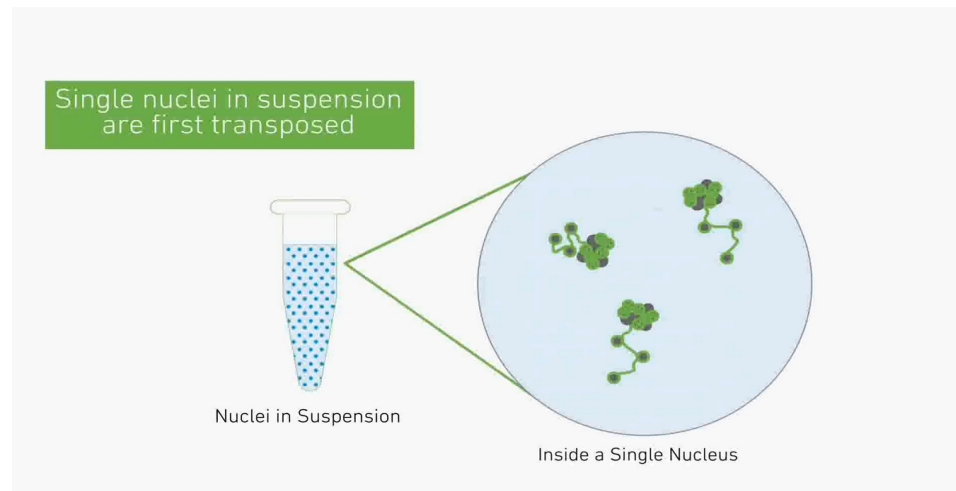
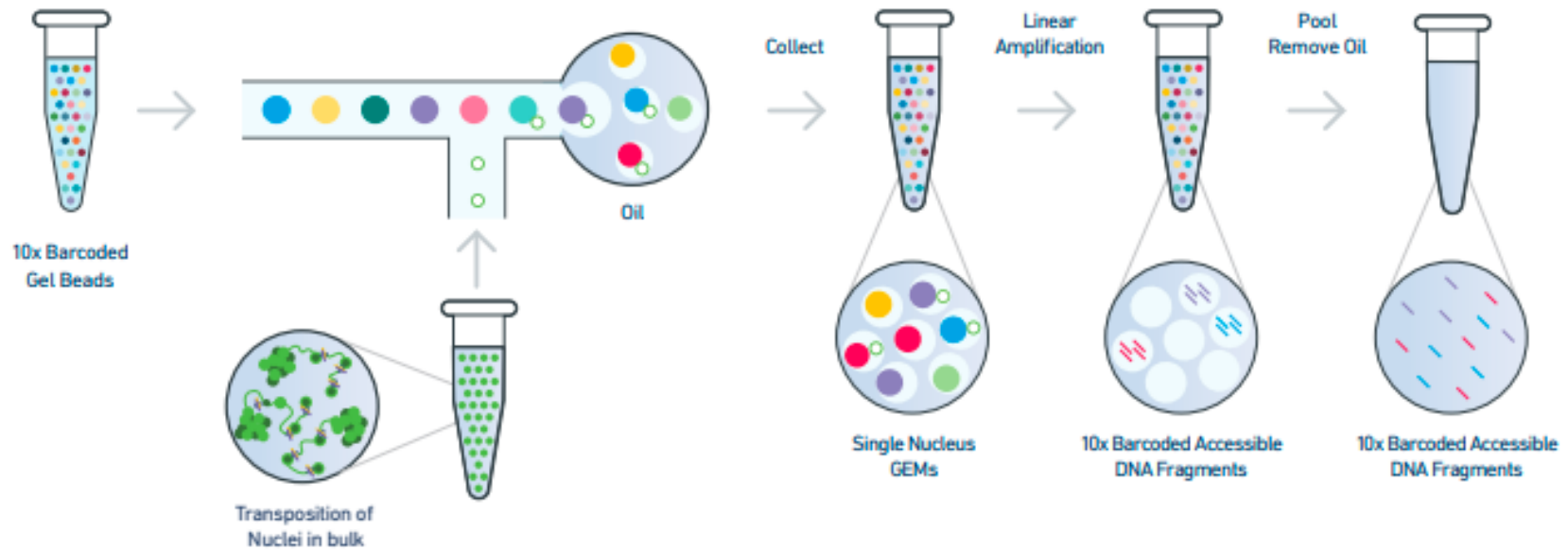
Identificación de diferencias de enriquecimiento

Una vez identificadas las regiones enriquecidas, la calidad cuantitativa de la NGS permite identificar diferencias en el factor de enriquecimiento con diversas herramientas como **DESeq2**.



Bardet AF et al. Nat Protocols 7:45-61

Single-cell ATAC-Seq



Tema 10: Identificación de alteraciones epigenómicas

10.1 Identificación de patrones de metilación

10.2 Identificación de sitios de unión de proteínas

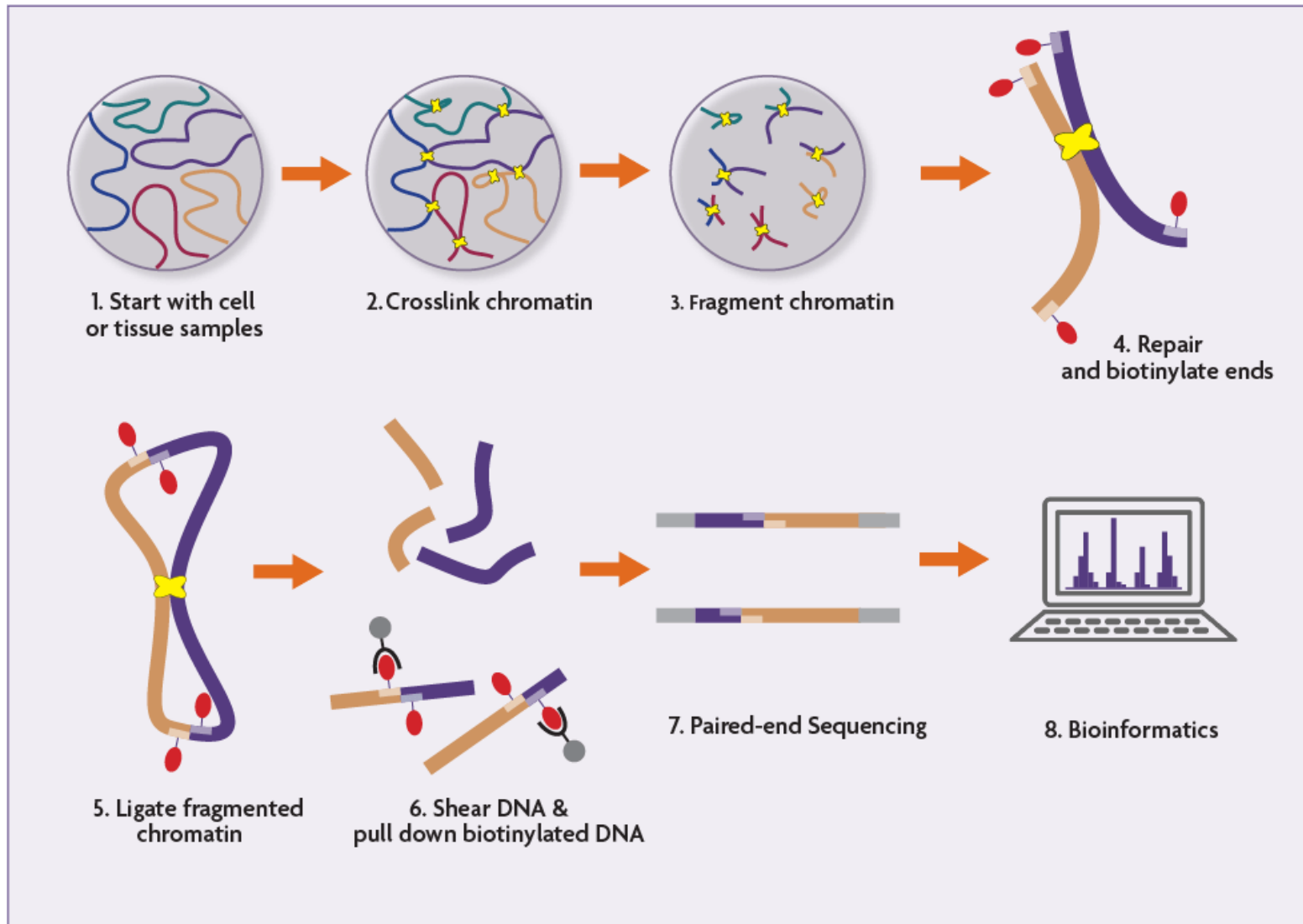
10.3 Caracterización de alteraciones en la accesibilidad de la cromatina.

10.4 Caracterización de la estructura tridimensional de la cromatina

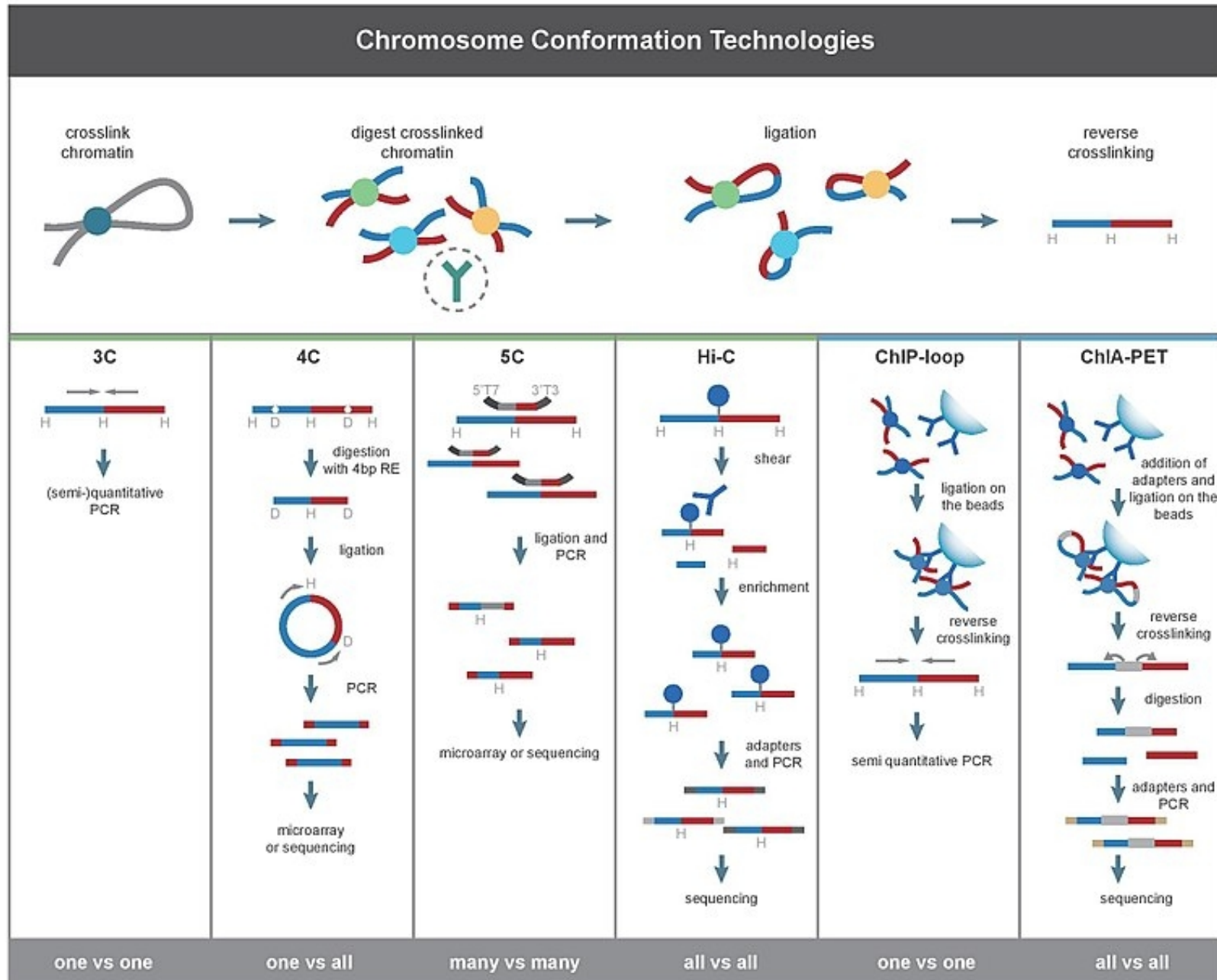
10.5 Formato de archivos de regiones genómicas

Identificación de regiones del DNA cercanas en su plegamiento 3D

En la generación de librerías, se generan moléculas híbridas (quimeras) pertenecientes an regiones que pueden estar distantes en la secuencia lineal del genoma pero se encuentran cercanas en el DNA plegado.



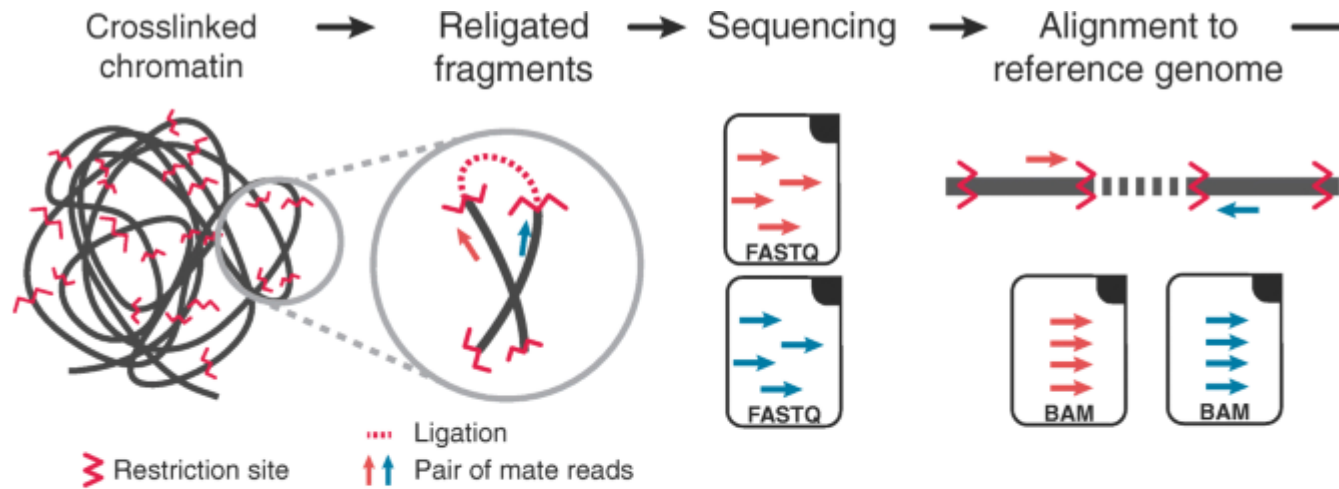
Distintas modalidades de análisis



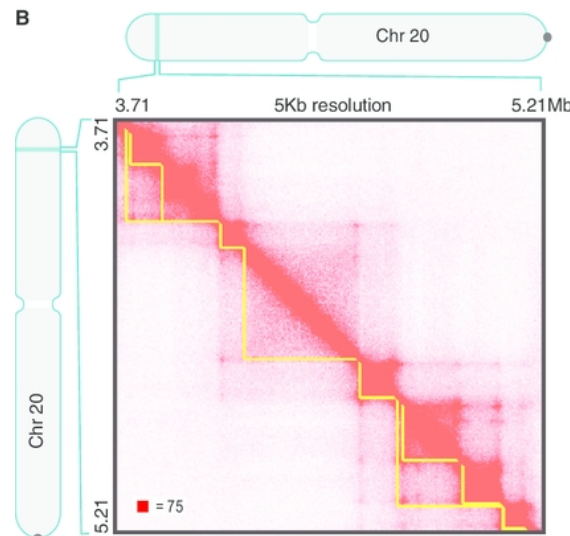
https://en.wikipedia.org/wiki/Chromosome_conformation_capture

Identificación de las regiones de contacto

Al llevar a cabo secuenciación pareada, se pueden identificar parejas de lecturas correspondientes a regiones distintas en el genoma

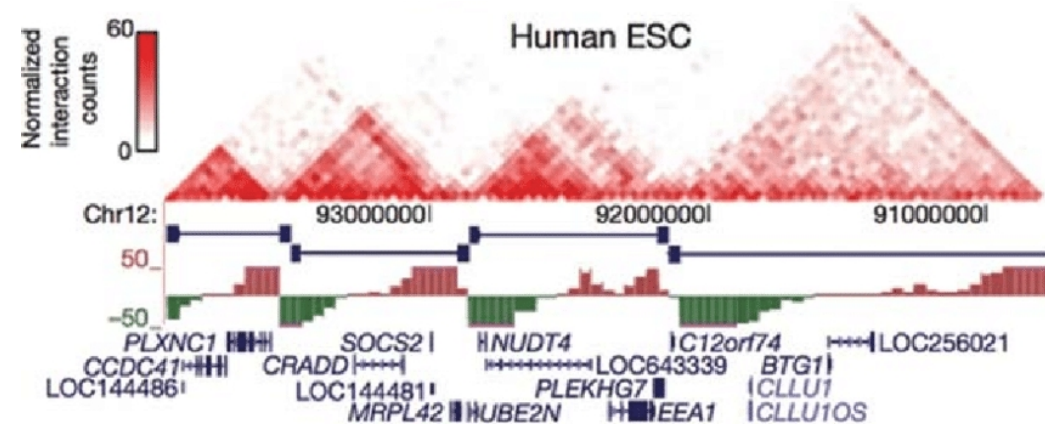
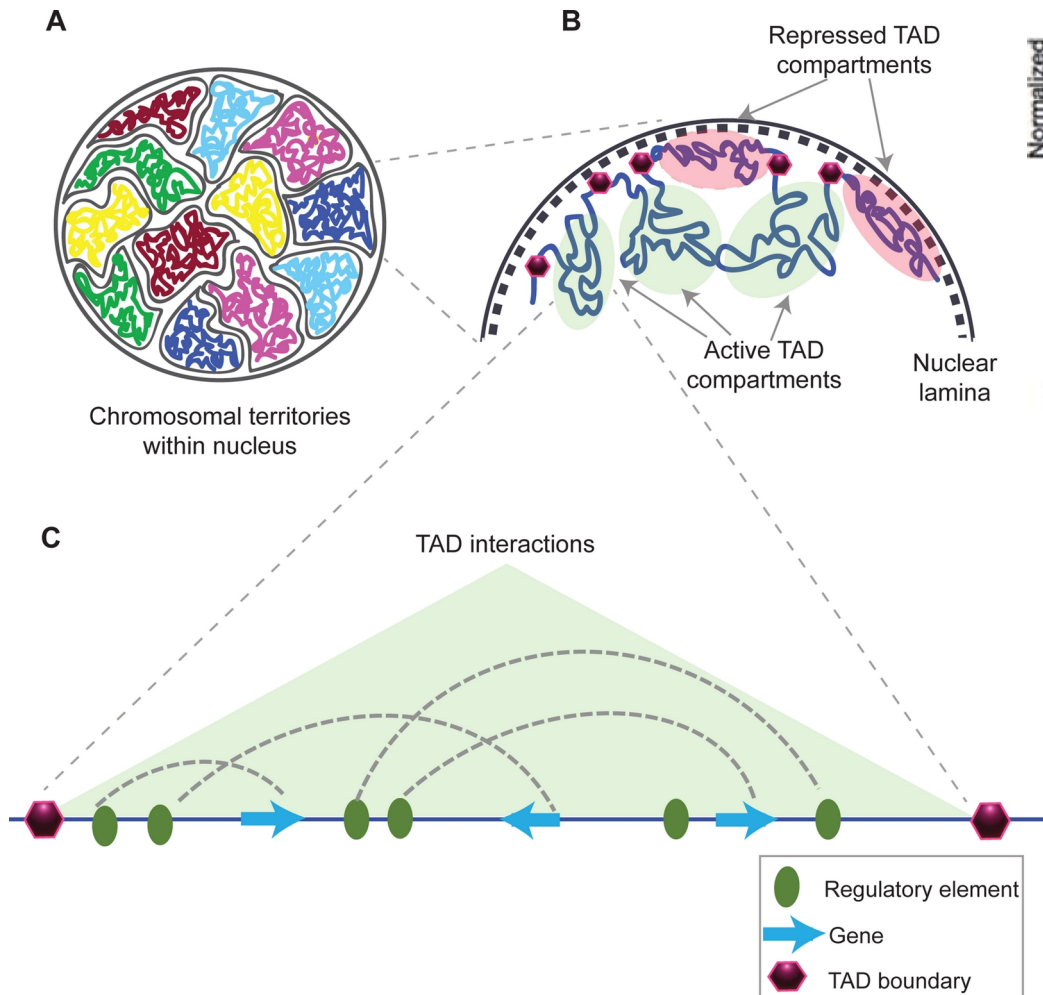


El número de lecturas de una determinada combinación de regiones determina la frecuencia con la que estas dos regiones se encuentran cercanas



Identificación de TADs

Un TAD (Topologically Associated Domain) está formado por una región del genoma aislada del resto en donde sus elementos interactúan entre ellos con alta frecuencia.



<http://symposium.cshlp.org/content/80/17>

Existen herramientas como **r3C** para identificar los puntos de contacto en datos de 4C y 5C así como para identificar TADs.

https://es.wikipedia.org/wiki/Dominios_asociados_topol%C3%B3gicamente

Tema 10: Identificación de alteraciones epigenómicas

10.1 Identificación de patrones de metilación

10.2 Identificación de sitios de unión de proteínas

10.3 Caracterización de alteraciones en la accesibilidad de la cromatina.

10.4 Caracterización de la estructura tridimensional de la cromatina

10.5 Formato de archivos de regiones genómicas

Archivos de regiones

Formato BED

The first three fields in each feature line are required:

1. **chrom** - name of the chromosome or scaffold. Any valid `seq_region_name` can be used, and chromosome names can be given with or without the 'chr' prefix.
2. **chromStart** - Start position of the feature in standard chromosomal coordinates (i.e. first base is 0).
3. **chromEnd** - End position of the feature in standard chromosomal coordinates

```
chr1 213941196 213942363
chr1 213942363 213943530
chr1 213943530 213944697
chr2 158364697 158365864
chr2 158365864 158367031
chr3 127477031 127478198
chr3 127478198 127479365
chr3 127479365 127480532
chr3 127480532 127481699
```

Optional fields

Nine additional fields are optional. Note that columns cannot be empty - lower-numbered fields must always be populated if higher-numbered ones are used.

4. **name** - Label to be displayed under the feature, if turned on in "Configure this page".
5. **score** - A score between 0 and 1000. See [track lines](#), below, for ways to configure the display style of scored data.
6. **strand** - defined as + (forward) or - (reverse).
7. **thickStart** - coordinate at which to start drawing the feature as a solid rectangle
8. **thickEnd** - coordinate at which to stop drawing the feature as a solid rectangle
9. **ItemRgb** - an RGB colour value (e.g. 0,0,255). Only used if there is a track line with the value of `itemRgb` set to "on" (case-insensitive).
10. **blockCount** - the number of sub-elements (e.g. exons) within the feature
11. **blockSizes** - the size of these sub-elements
12. **blockStarts** - the start coordinate of each sub-element

<https://www.ensembl.org/info/website/upload/bed.html>

Archivos de regiones

Formatos Wig y BigWig

Aportan un valor cuantitativo para cada posición del genoma. BigWig es el formato binario comprimido

variableStep format

This format is used for data with irregular intervals between new data values and data values:

```
variableStep chrom=chrN
[span=windowSize]
  chromStartA  dataValueA
  chromStartB  dataValueB
  ... etc ...  ... etc ...
```

The declaration line starts with the word **variableStep**
The span begins at each chromosome position

```
variableStep chrom=chr2
300701 12.5
300702 12.5
300703 12.5
300704 12.5
300705 12.5
```

is equivalent to:

```
variableStep chrom=chr2 span=5
300701 12.5
```

fixedStep format

This format is used for data with regular intervals between new data values and data values:

```
fixedStep chrom=chrN
start=position step=stepInterval
[span=windowSize]
  dataValue1
  dataValue2
  ... etc ...
```

The declaration line starts with the word **fixedStep** and includes specification of start position and step interval

```
fixedStep chrom=chr3 start=400601 step=100
11
22
33
```

displays the values 11, 22, and 33 as single-base regions on chromosome 3

```
fixedStep chrom=chr3 start=400601 step=100 span=5
11
22
33
```

<http://genome.ucsc.edu/goldenPath/help/wiggle.html>

Archivos de interacciones

Formato .pairs

Aporta información de cada pareja de lecturas y las regiones de contacto

index	name	description
1	read_id	the ID of the read as defined in fastq files
2	chrom1	the chromosome of the alignment on side 1
3	pos1	the 1-based genomic position of the outer-most (5') mapped bp on side 1
4	chrom2	the chromosome of the alignment on side 2
5	pos2	the 1-based genomic position of the outer-most (5') mapped bp on side 2
6	strand1	the strand of the alignment on side 1
7	strand2	the strand of the alignment on side 2

<https://pairtools.readthedocs.io/en/latest/formats.html>

Archivos de interacciones

Formato .cool

Añade información cuantitativa sobre el número de lecturas que apoyan cada interacción

chrom1	start1	end1	chrom2	start2	end2	count	balanced
chr18	10500000	10600000	chr18	10500000	10600000	1163	0.0482848
chr18	10500000	10600000	chr18	10600000	10700000	574	0.0397891
chr18	10500000	10600000	chr18	10700000	10800000	588	0.0312273
chr18	10500000	10600000	chr18	10800000	10900000	602	0.0255296
chr18	10500000	10600000	chr18	10900000	11000000	446	0.0186205
chr18	10500000	10600000	chr18	11000000	11100000	413	0.0167251
chr18	10500000	10600000	chr18	11100000	11200000	182	0.00995996
chr18	10500000	10600000	chr18	11200000	11300000	192	0.0109431
chr18	10500000	10600000	chr18	11300000	11400000	147	0.00932337

<https://www.biostars.org/p/327304/>

Bioinformática y análisis de datos ómicos

