Genomics Part II

Applications of Sequencing Technology

Biomedical Data Science: Mining and Modeling
CB&B 752 • MB&B 452
Matt Simon
January 17, 2020

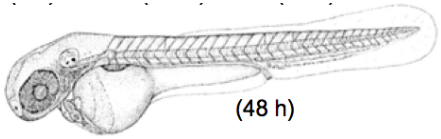
Overview

- Genomics I (Wednesday's lecture): Focus on sequencing technology and genomes.

- Genomics II: (Today's lecture): Focus on applications of sequencing technology.
    1. Annotation of the genome in chromatin
    2. Regulation of gene expression at the level of RNA

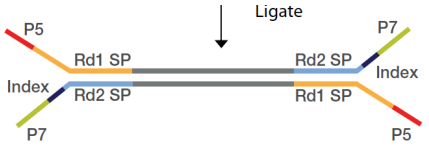
Workflow

1. Isolation of sample.



e.g., Isolate DNA and shear.

2. Library preparation



e.g., Clean up and ligate Y-adaptors.

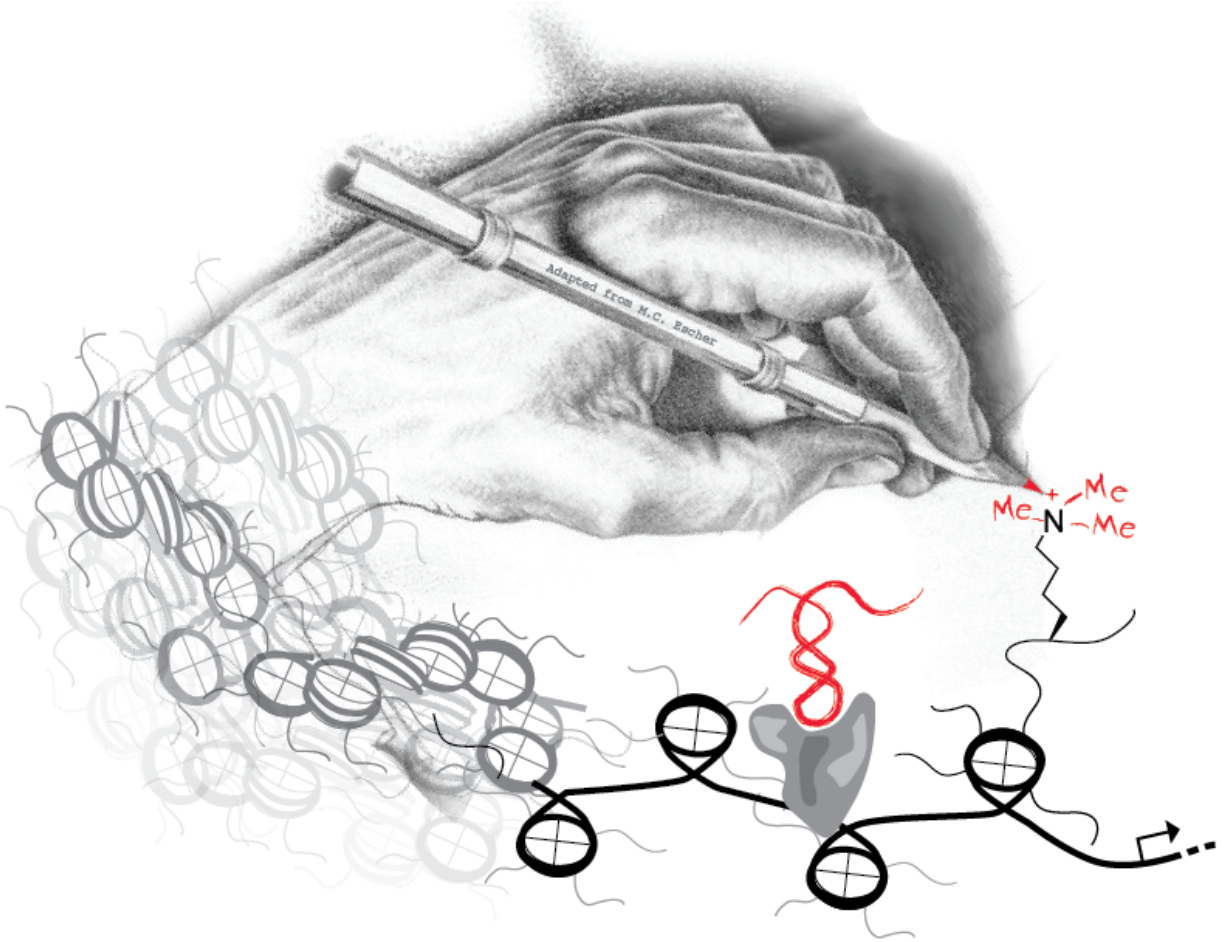
3. Sequencing

e.g., Illumina HiSeq

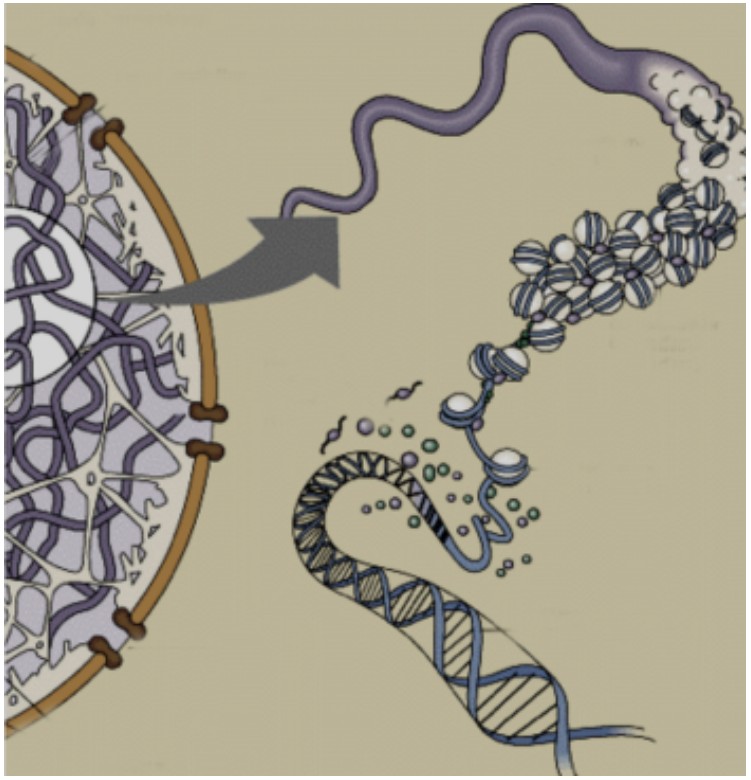
4. Analysis

e.g., Map to genome and interpret.

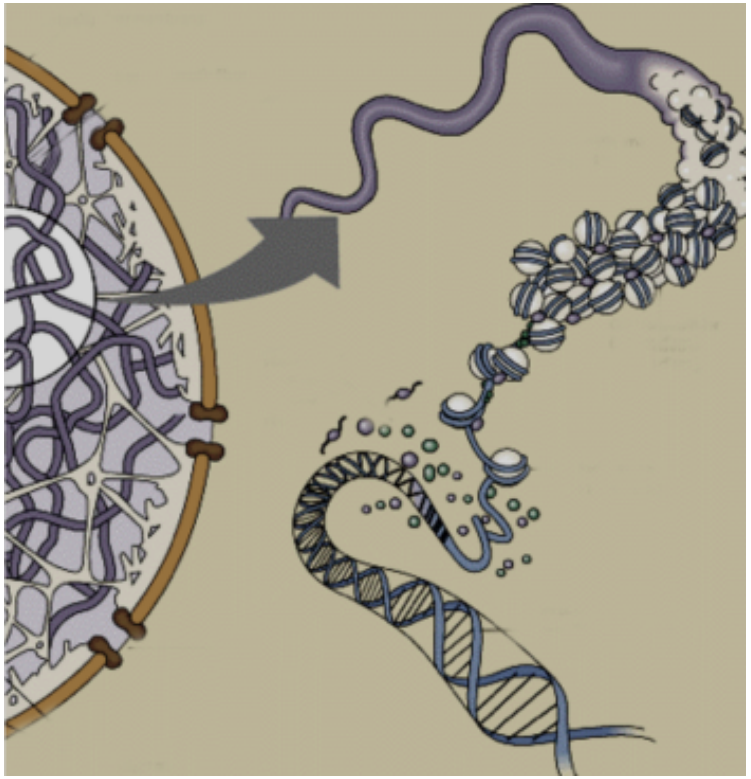
Part 1. How do cells annotate their genomes?

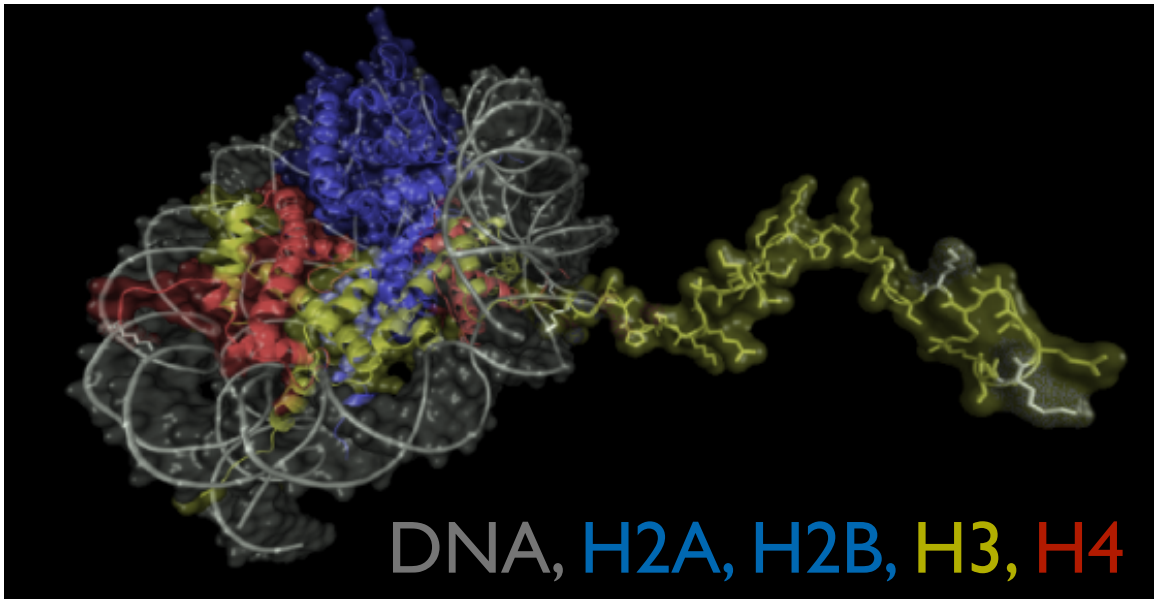


DNA in the cell is packaged into chromatin



DNA in the cell is packaged into chromatin

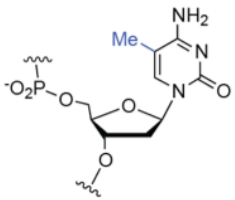




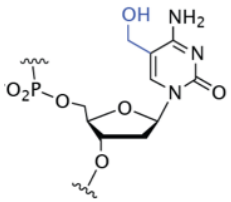
Modeled nucleosome based on Luger et al., *Nature* **1997** *389*, 251.

Summary and nomenclature of common covalent modifications.

Summary and nomenclature of common covalent modifications.

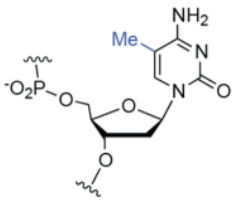


mC

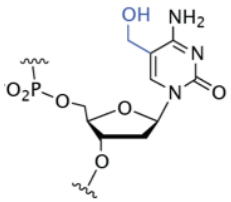


hmC

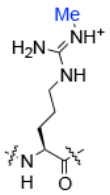
Summary and nomenclature of common covalent modifications.



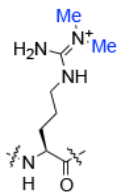
mC



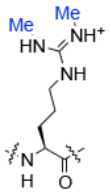
hmC



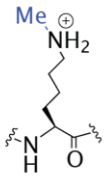
Rme1



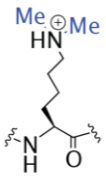
Rme2a



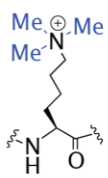
Rme2s



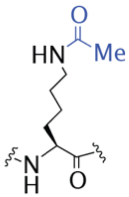
Kme1



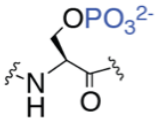
Kme2



Kme3

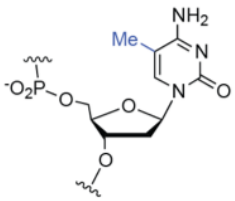


Kac

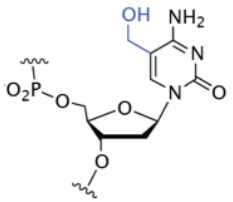


Sph

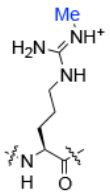
Summary and nomenclature of common covalent modifications.



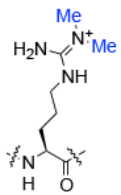
mC



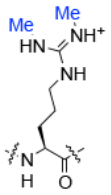
hmC



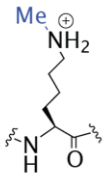
Rme1



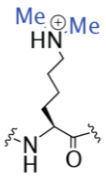
Rme2a



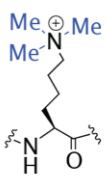
Rme2s



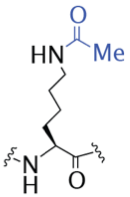
Kme1



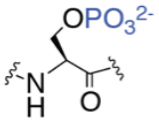
Kme2



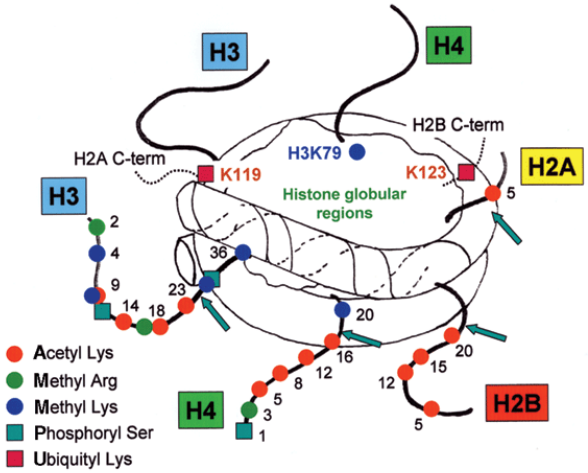
Kme3



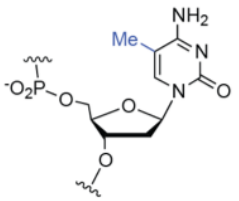
Kac



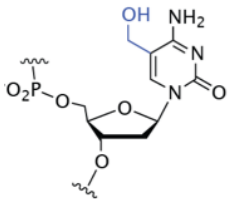
Sph



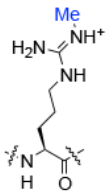
Summary and nomenclature of common covalent modifications.



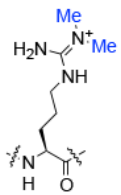
mC



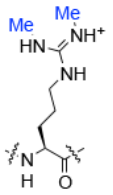
hmC



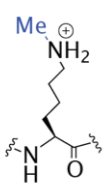
Rme1



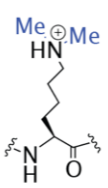
Rme2a



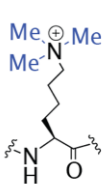
Rme2s



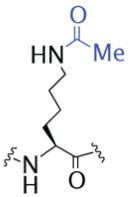
Kme1



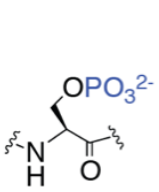
Kme2



Kme3



Kac



Sph

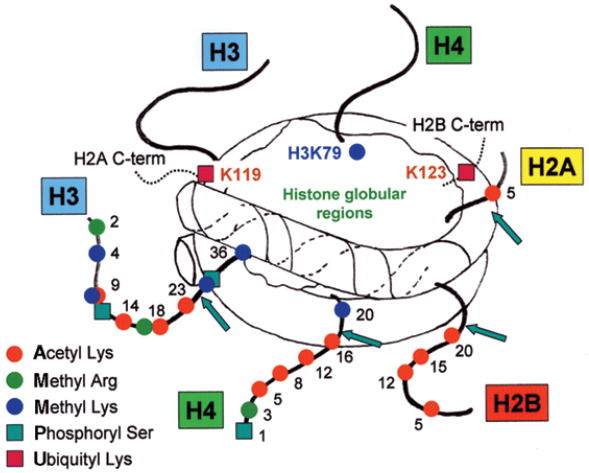
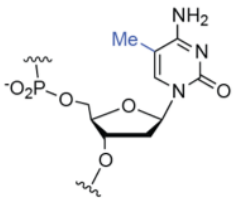


Table 1 The Brno nomenclature for histone modifications

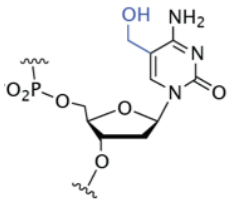
Modifying group	Amino acid(s) modified	Level of modification	Abbreviation for modification[a]	Examples of modified residues[b]
Acetyl-	Lysine	mono-	ac	H3K9ac
Methyl-	Arginine	mono-	me1	H3R17me1
	Arginine	di-, symmetrical	me2s	H3R2me2s
	Arginine	di-, asymmetrical	me2a	H3R17me2a
	Lysine	mono-	me1	H3K4me1
	Lysine	di-	me2	H3K4me2
	Lysine	tri-	me3	H3K4me3
Phosphoryl-	Serine or threonine	mono-	ph	H3S10ph
Ubiquityl-	Lysine	mono-[c]	ub1	H2BK123ub1
SUMOyl-	Lysine	mono-	su	H4K5su[d]
ADP ribosyl-	Glutamate	mono-	ar1	H2BE2ar1
	Glutamate	poly-	arn	H2BE2arn[d]

Turner, B. M. Reading signals on the nucleosome with a new nomenclature for modified histones. *Nat Struct Mol Biol* 12, 110–112 (2005).

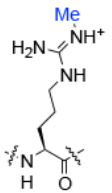
Summary and nomenclature of common covalent modifications.



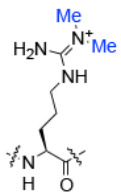
mC



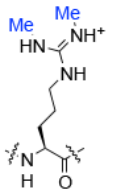
hmC



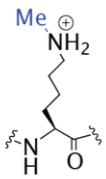
Rme1



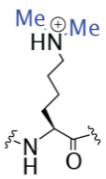
Rme2a



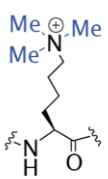
Rme2s



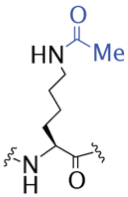
Kme1



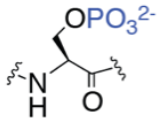
Kme2



Kme3



Kac



Sph

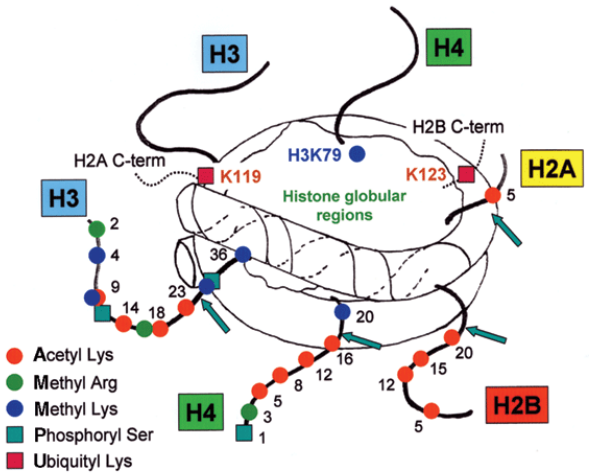


Table 1 The Brno nomenclature for histone modifications

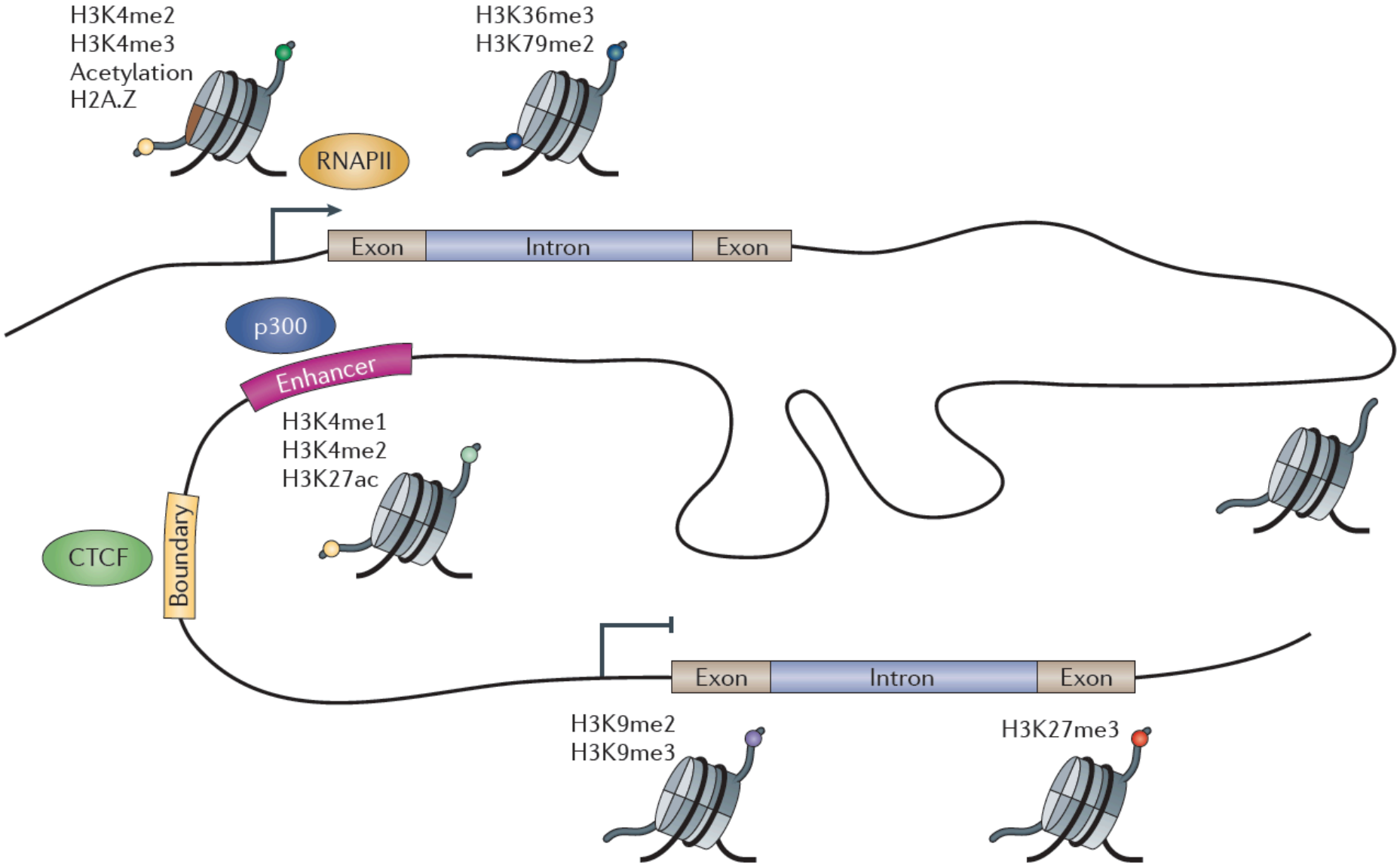
Modifying group	Amino acid(s) modified	Level of modification	Abbreviation for modification[a]	Examples of modified residues[b]
Acetyl-	Lysine	mono-	ac	H3K9ac
Methyl-	Arginine	mono-	me1	H3R17me1
	Arginine	di-, symmetrical	me2s	H3R2me2s
	Arginine	di-, asymmetrical	me2a	H3R17me2a
	Lysine	mono-	me1	H3K4me1
	Lysine	di-	me2	H3K4me2
	Lysine	tri-	me3	H3K4me3
Phosphoryl-	Serine or threonine	mono-	ph	H3S10ph
Ubiquityl-	Lysine	mono-[c]	ub1	H2BK123ub1
SUMOyl-	Lysine	mono-	su	H4K5su[d]
ADP ribosyl-	Glutamate	mono-	ar1	H2BE2ar1
	Glutamate	poly-	arn	H2BE2arn[d]

H3 **K27** **ac**

Histone Residue Modification

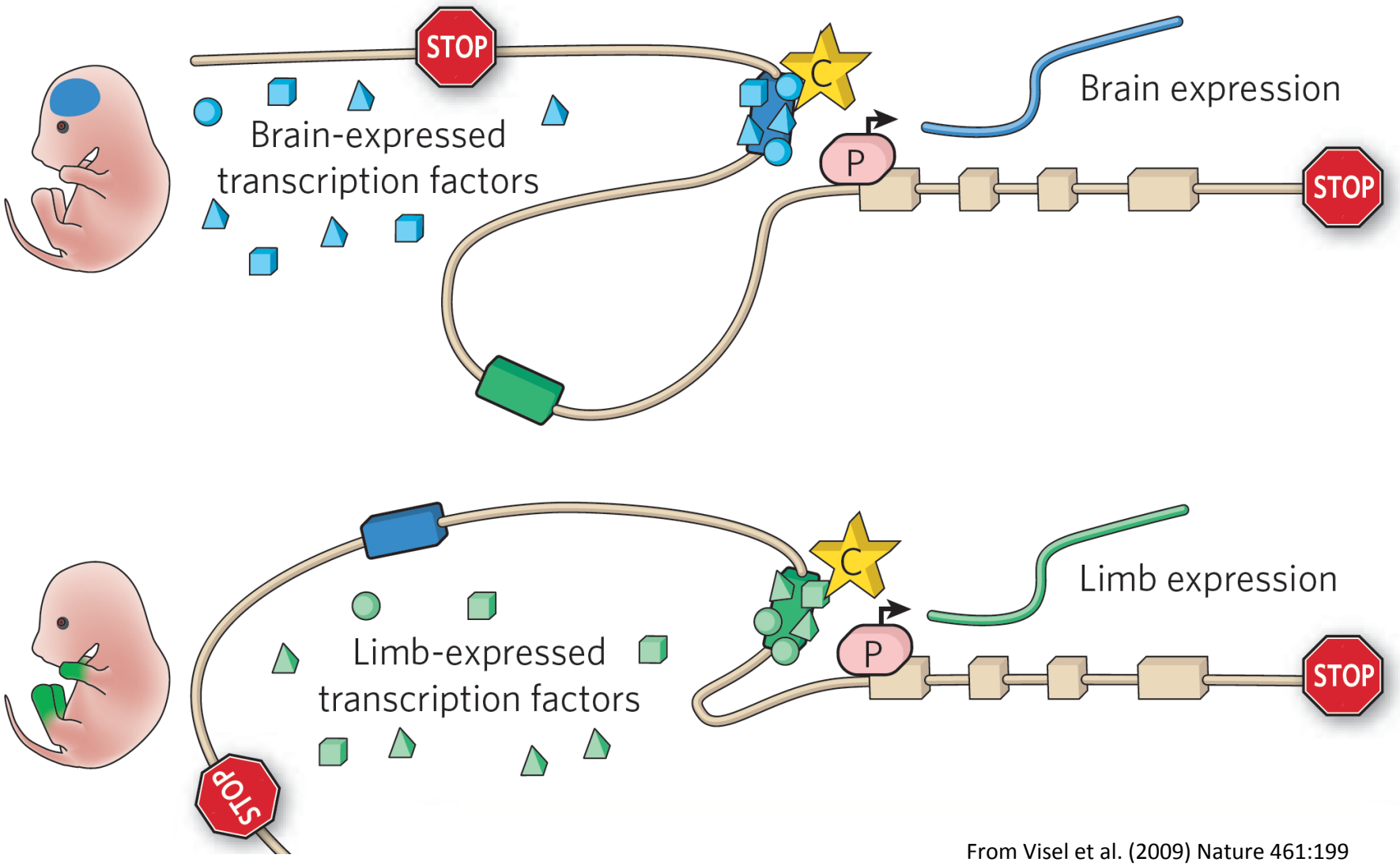
Turner, B. M. Reading signals on the nucleosome with a new nomenclature for modified histones. *Nat Struct Mol Biol* 12, 110–112 (2005).

Chromatin modifications correlate with different genomic functions.



Zhou *et al*. *Nat Rev Genet* 12:7 (2011)

Regulation is temporally and specially controlled



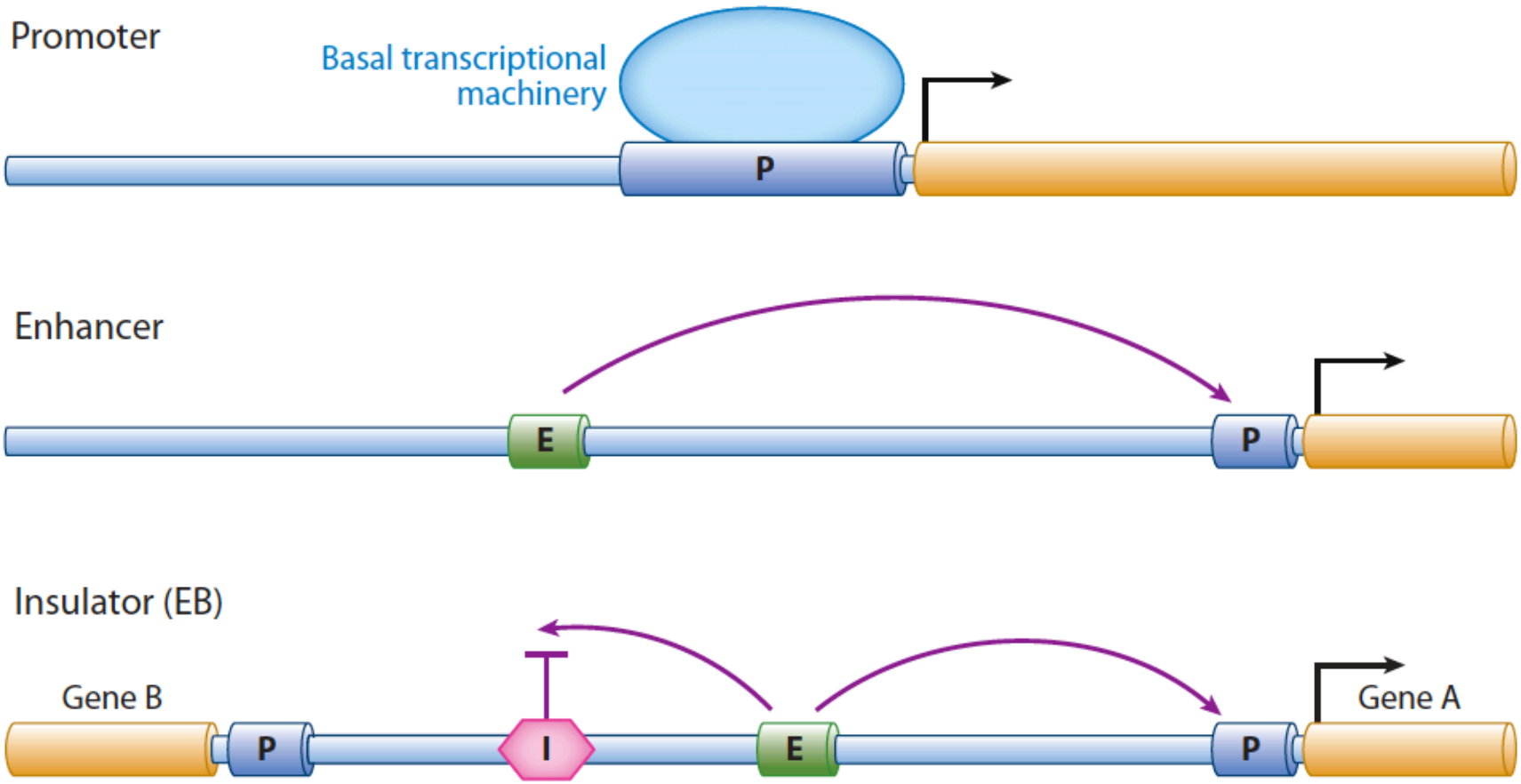
From Visel et al. (2009) Nature 461:199

Using sequencing to annotate the genome

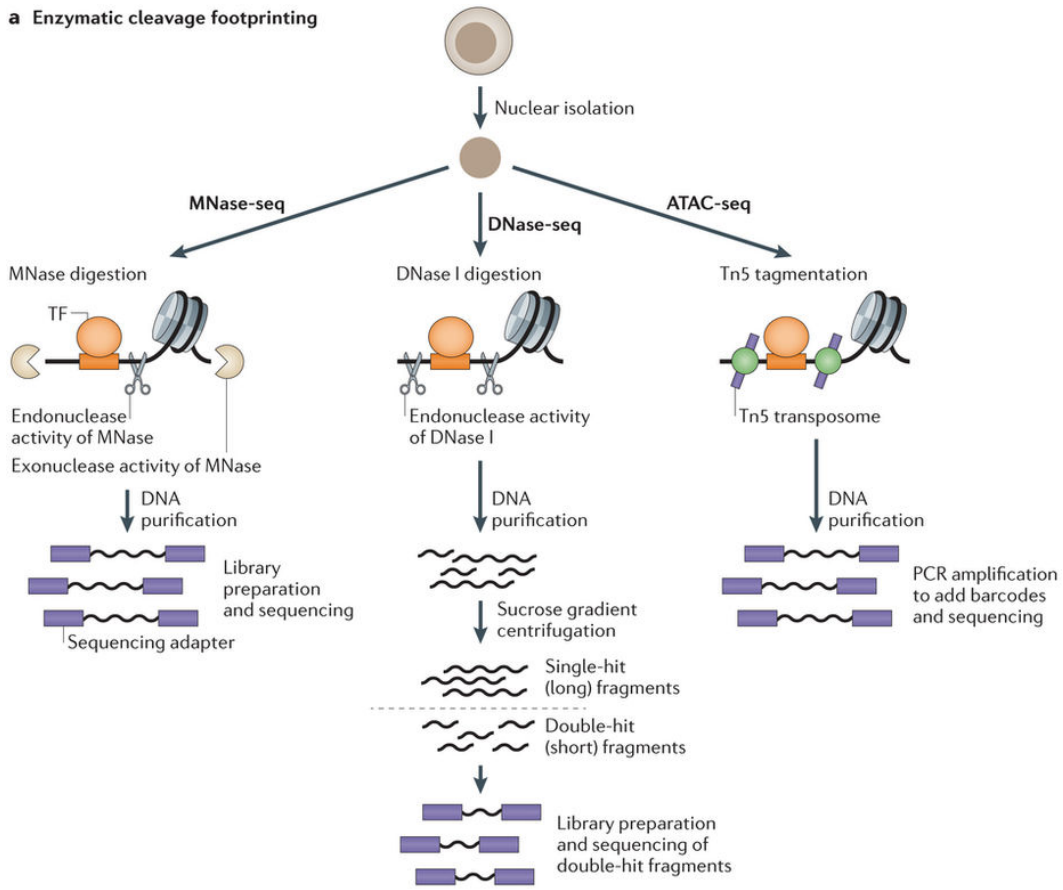
1. Where are the cis-acting regulatory elements in DNA?
   A. DNase I hyper-sensitivity mapping (**DNase-Seq**).
   B. **FAIRE** to map regulatory elements.
   C. **ATAC-Seq** to map regulatory elements.
2. Where do transcription factors bind?
   D. **ChIP-seq** of transcription factors (or in high res, ChIP-exo)
   E. **CUT&RUN** and **TAG&RUN** for small scale/single cell analysis.
3. Where are different histone modifications found?
   F. **ChIP-Seq** of histone modifications.
   G. **ChIP-Seq** of chromatin writers, readers and erasers.
4. Where is RNA polymerase transcribing?
   H. **ChIP-Seq** of polymerase.
   I. **GRO-Seq, PRO-Seq** and **NET-Seq** to measure RNA polymerase activity.
5. How is the genome organized in 3D?
   J. **4C/5C**/**Hi-C** to measure chromatin conformation.

Targeted approaches v Global approaches

How do we identify regulatory elements in the genome?



Using differences in biochemical properties of regulatory elements to identify them by Seq

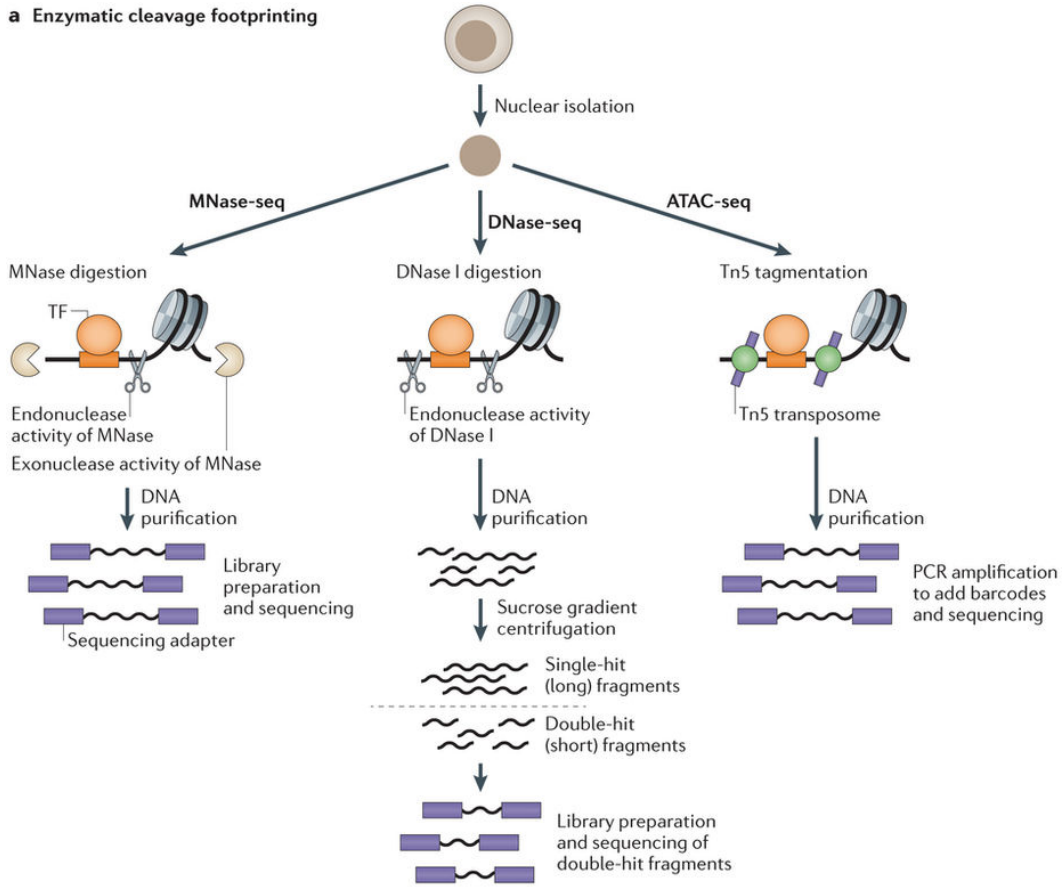


1. **Transcription factor binding** frequently deforms the B-form DNA, making it hypersensitive to DNase I and transposases.

.

Zentner GE, Henikoff S. High-resolution digital profiling of the epigenome. Nat Rev Genet. 2014;15: 814–827. doi:10.1038/nrg3798

Using differences in biochemical properties of regulatory elements to identify them by Seq



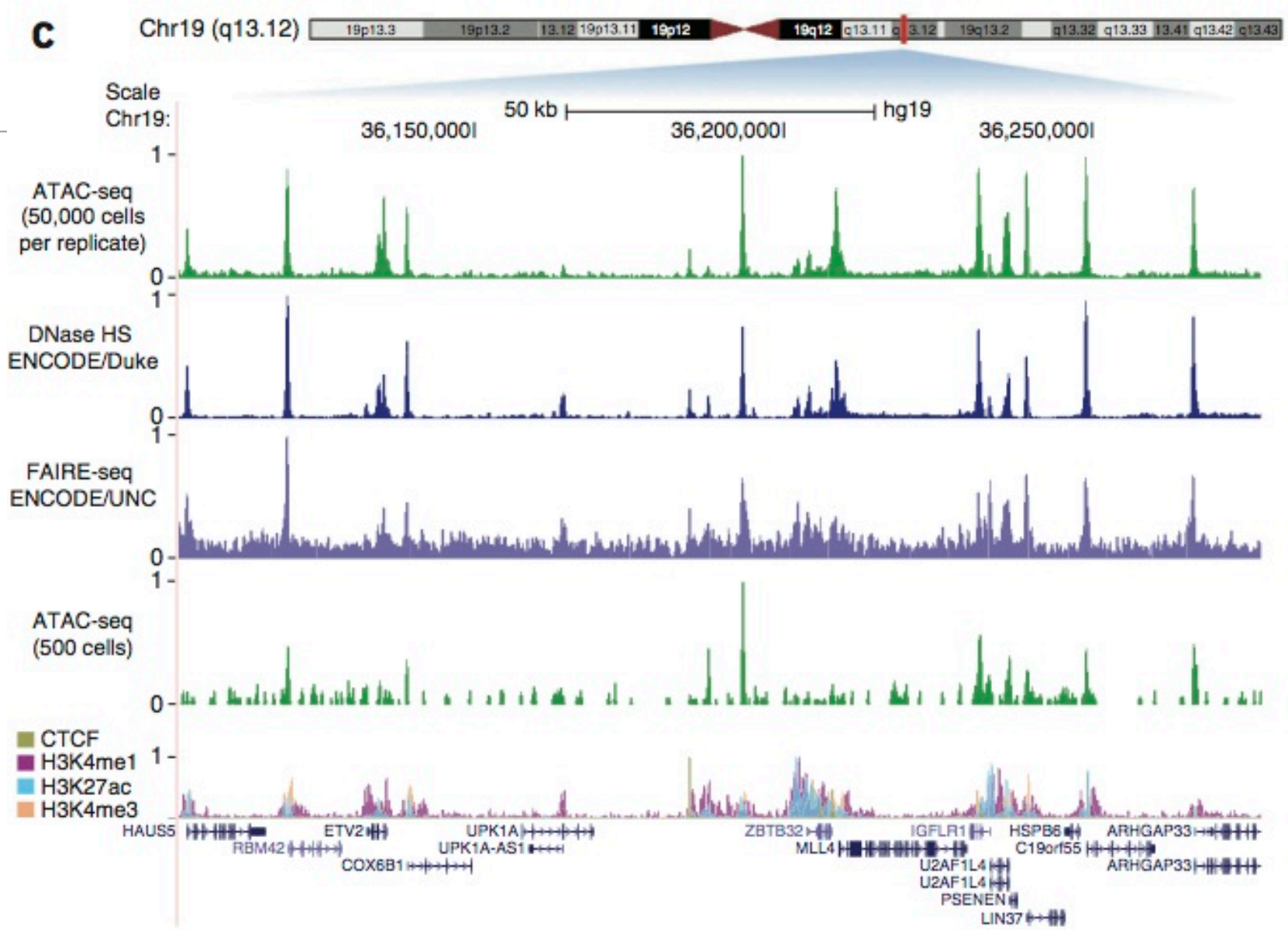
Zentner GE, Henikoff S. High-resolution digital profiling of the epigenome. Nat Rev Genet. 2014;15: 814–827. doi:10.1038/nrg3798

1. **Transcription factor binding** frequently deforms the B-form DNA, making it hypersensitive to DNase I and transposases.

.

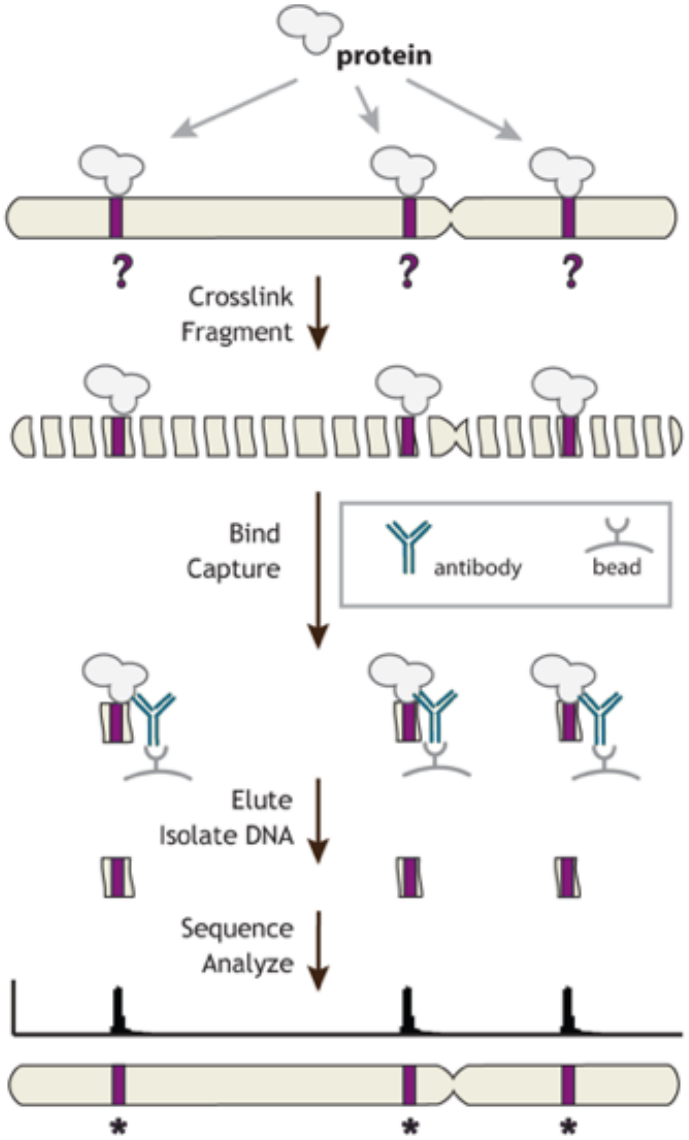
Changes in **accessibility of chromatin** can provide information about regulation

-ATAC-seq (shown)
-MNase-Seq (shown).
-DNase-Seq (shown).
-FAIRE-Seq (not shown).



Buenrostro JD, Giresi PG, Zaba LC, Chang HY, and Greenleaf WJ. (2013) "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." *Nature Methods*

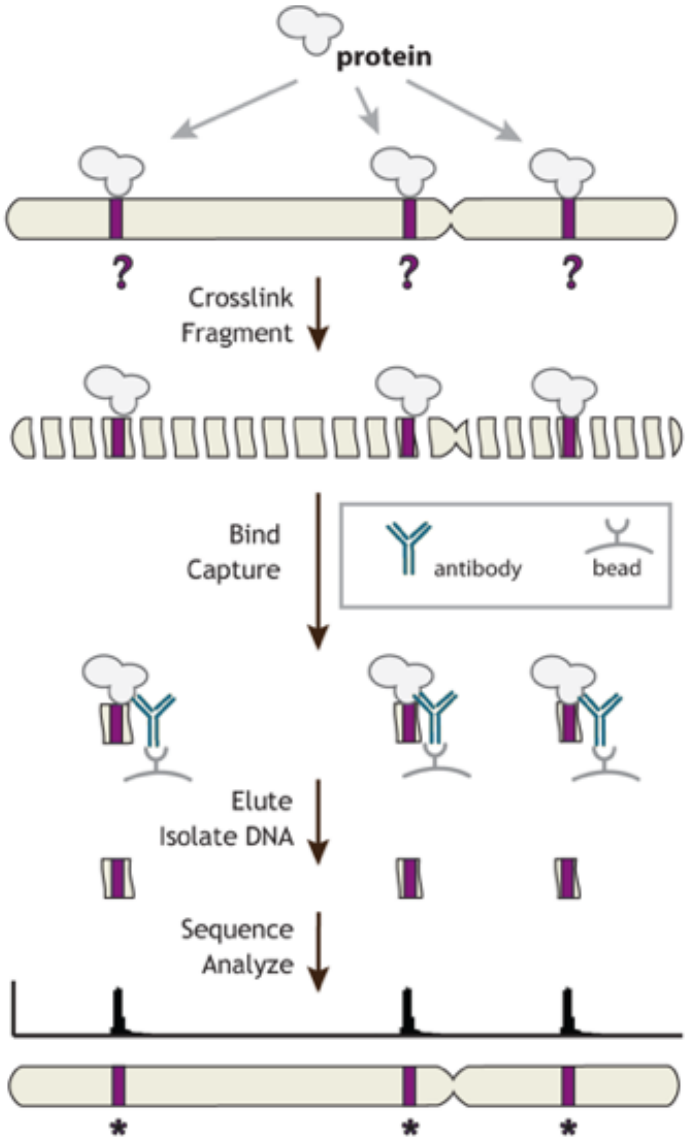
Localization of *specific proteins* in the genome with chromatin immunoprecipitation (ChIP-Seq)



1. **Crosslink** the cells with formaldehyde to “fix” factors in place.

   Exception: Native ChIP with histone antibodies.

Localization of *specific proteins* in the genome with chromatin immunoprecipitation (ChIP-Seq)



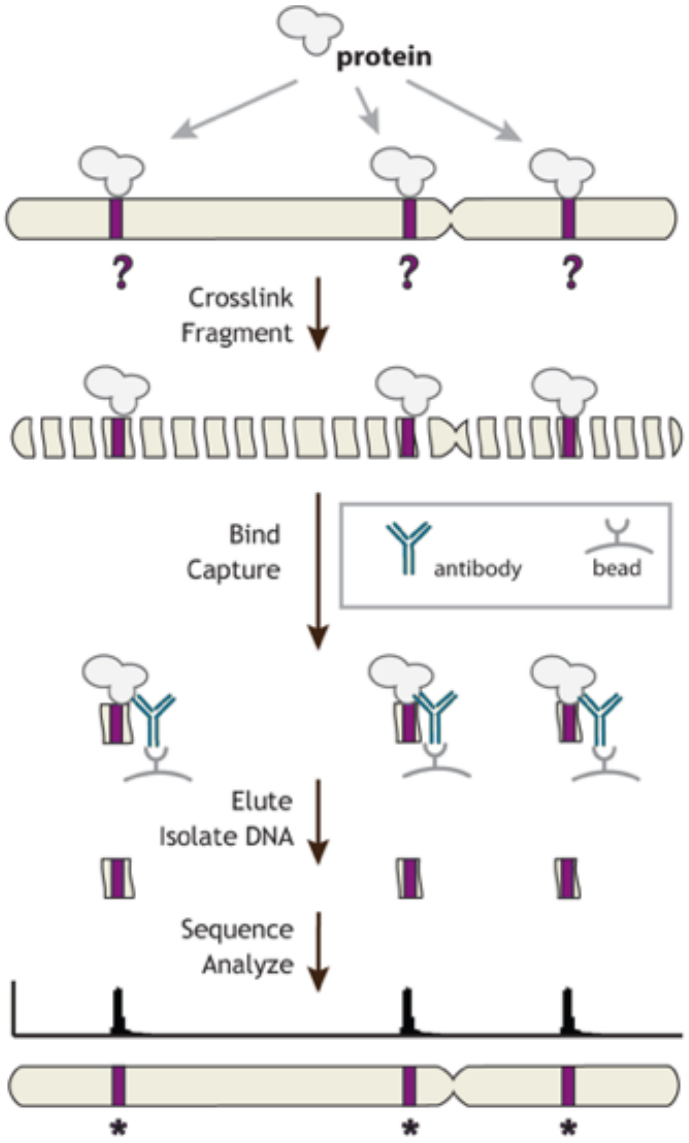
1. **Crosslink** the cells with formaldehyde to "fix" factors in place.

   Exception: Native ChIP with histone antibodies.

2. **Shear chromatin** to smaller pieces.

   Shear size determines resolution.
   Note: ChIP-exo uses an exonuclease at a later step to increase resolution.

Localization of *specific proteins* in the genome with chromatin immunoprecipitation (ChIP-Seq)



1. **Crosslink** the cells with formaldehyde to "fix" factors in place.

   Exception: Native ChIP with histone antibodies.

2. **Shear chromatin** to smaller pieces.

   Shear size determines resolution.
   Note: ChIP-exo uses an exonuclease at a later step to increase resolution.

3. **Enrich** target using an antibody.

   Enrichment is only as good as the antibody.

Determining sites of enrichment from ChIP-Seq

ChIP

1. **Align** reads to the genome.

Determining sites of enrichment from ChIP-Seq

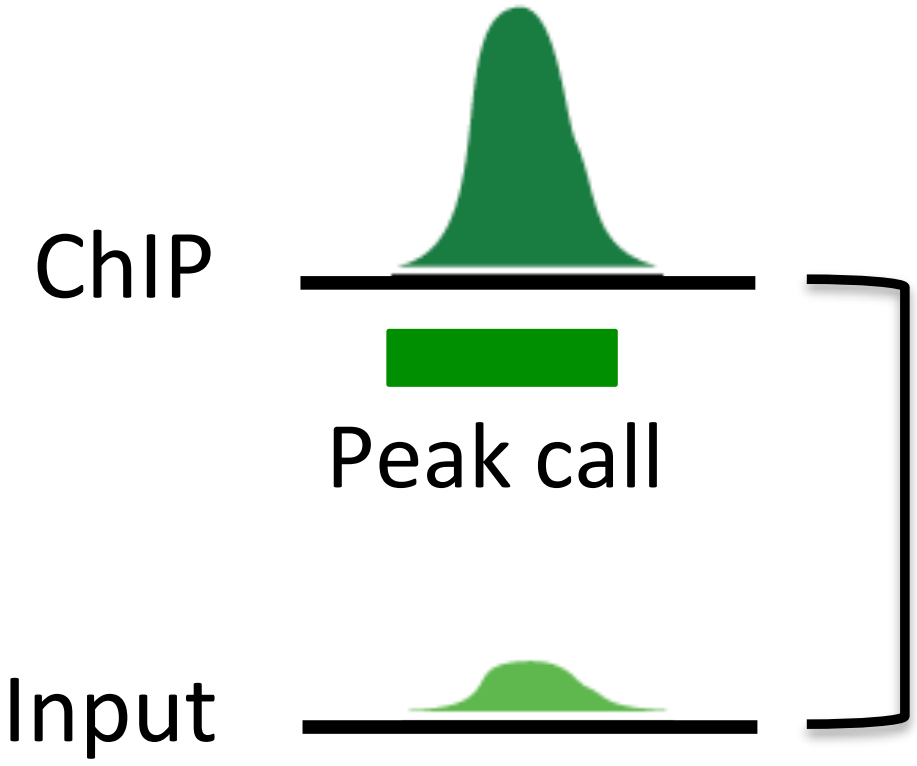
ChIP

Input

1. **Align** reads to the genome.

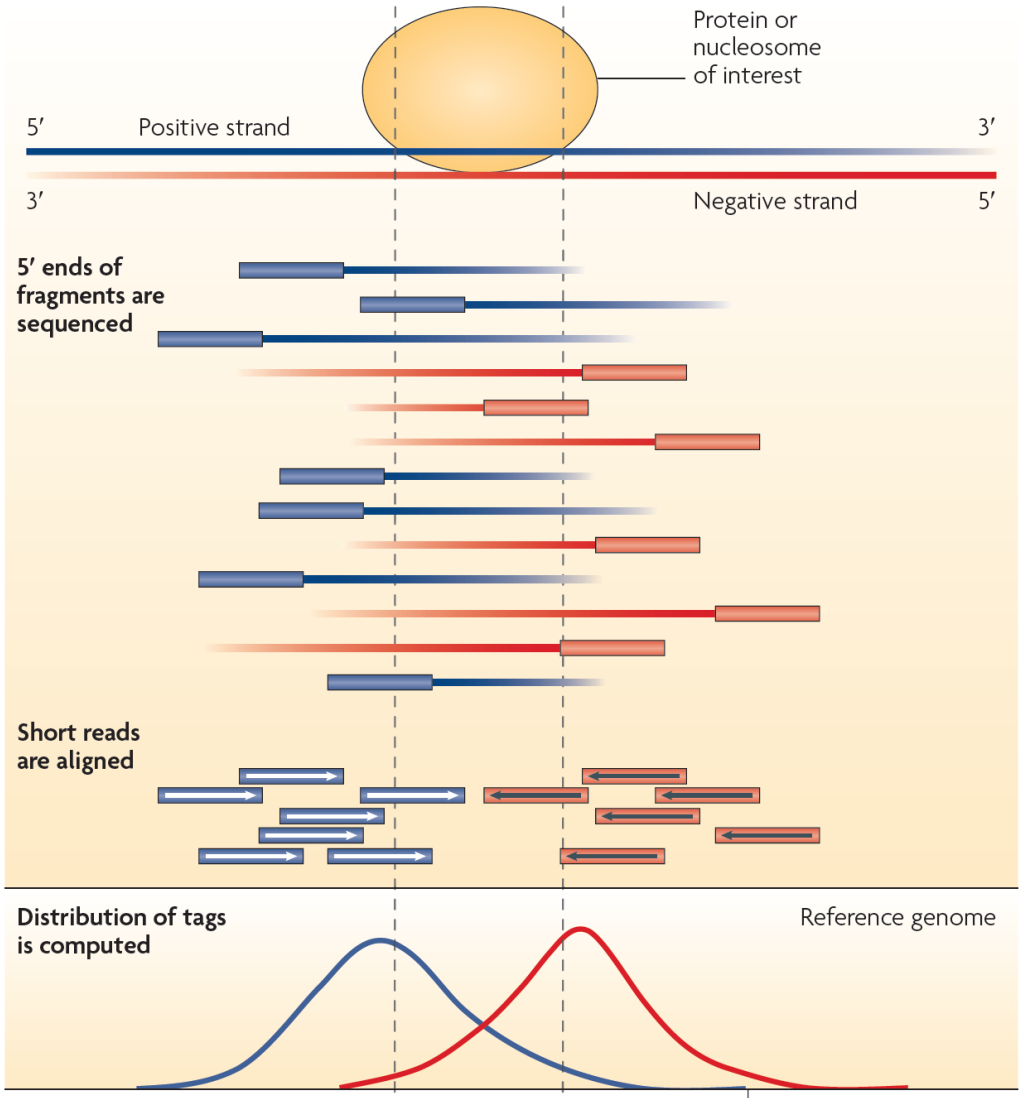
2. **Compare to input** to look for enrichment.
   Input coverage is not even.

Determining sites of enrichment from ChIP-Seq



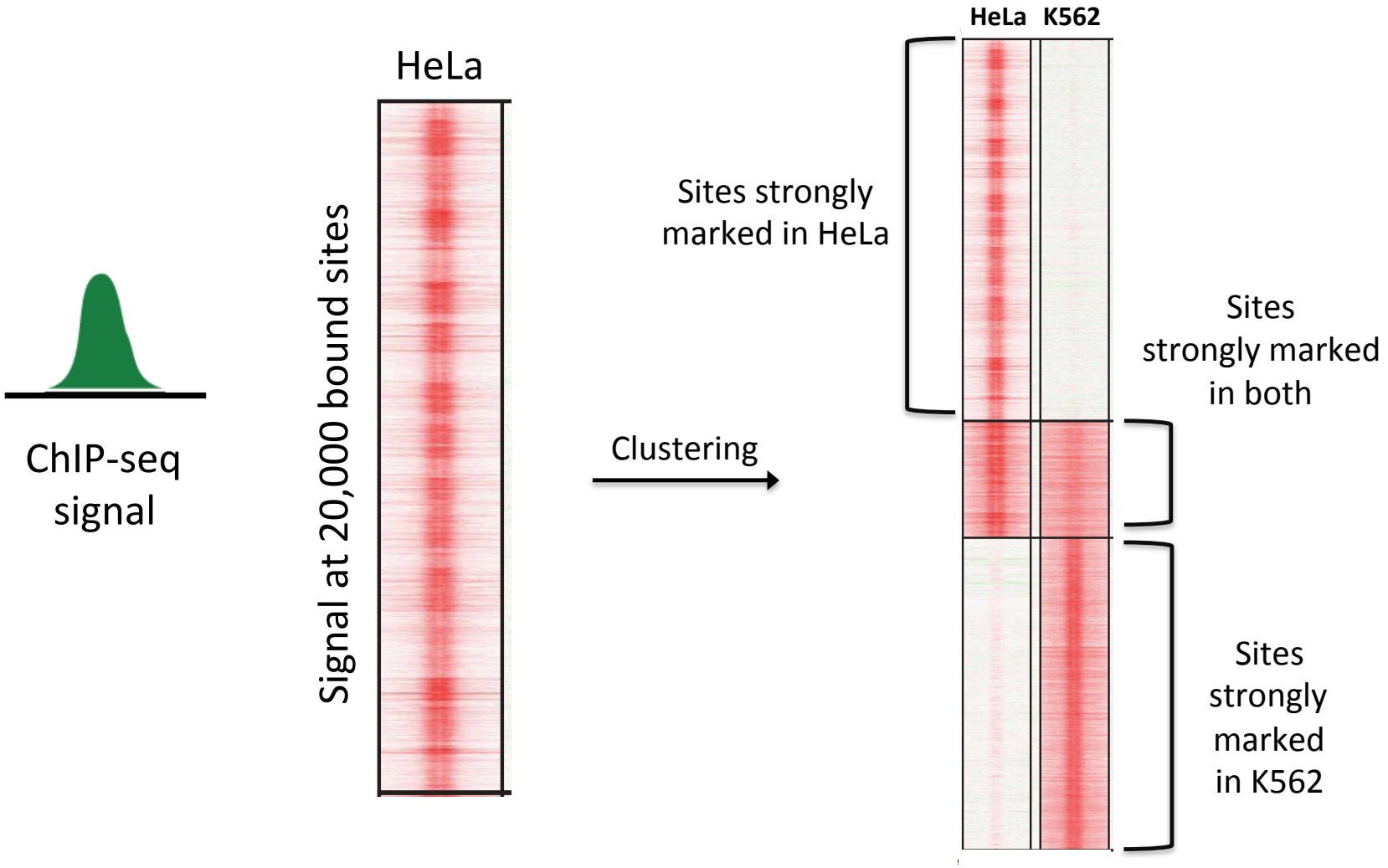
1. **Align** reads to the genome.
2. **Compare to input** to look for enrichment.
   Input coverage is not even.
3. **Call peaks** to determine statistically significant sites of enrichment.

Avoiding artifacts using features in Seq data

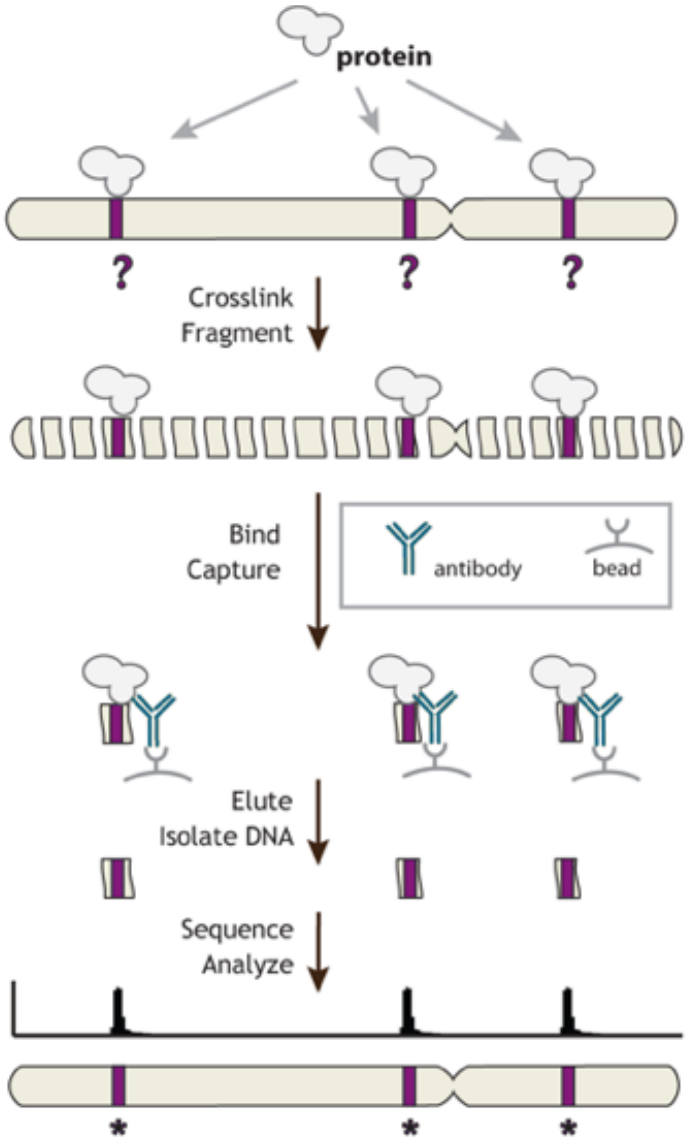


From Park (2009) Nat Rev Genet 10:669

ChIP-Seq signals reveal difference between cells

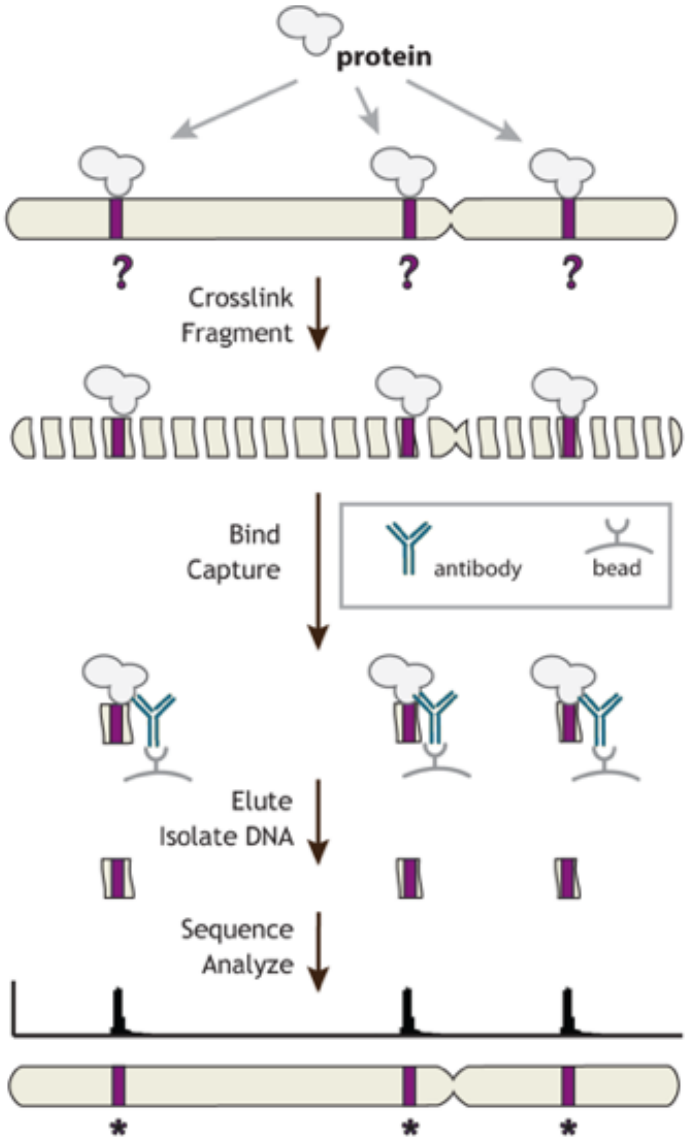


Limitations of ChIP-Seq



1. **Cross linking** efficiency is not necessarily uniform.

Limitations of ChIP-Seq

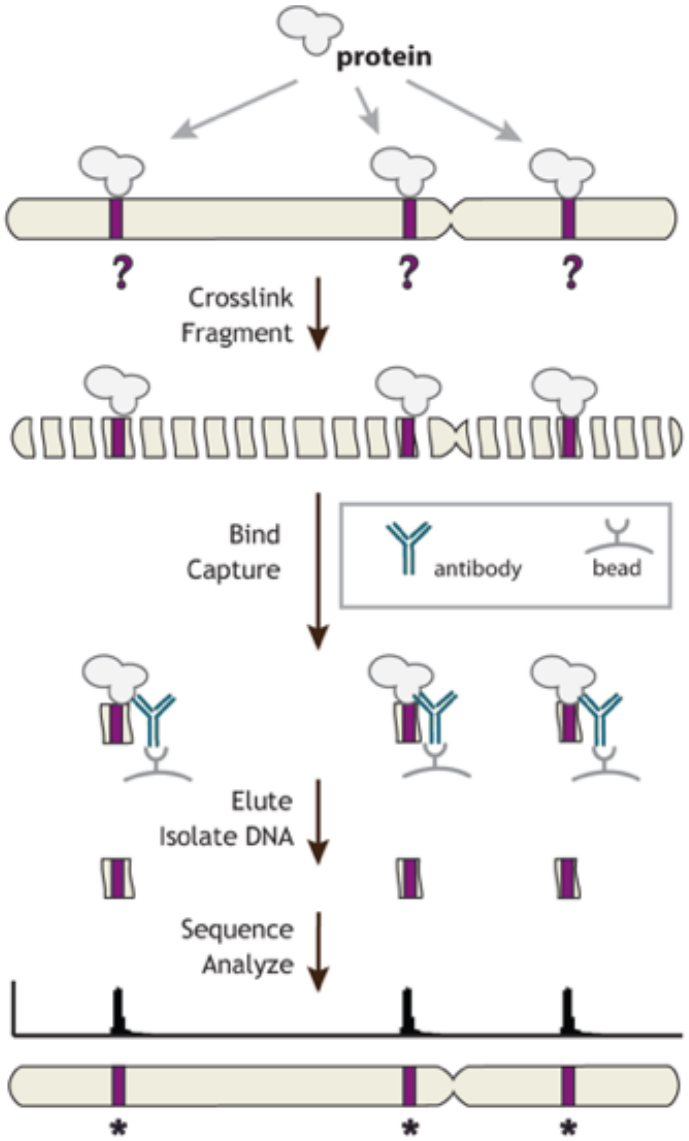


1. **Cross linking** efficiency is not necessarily uniform.

.

2. Enrichment is dependent on the **quality of antibody.**
   e.g., Site and degree of histone modifications.

Limitations of ChIP-Seq



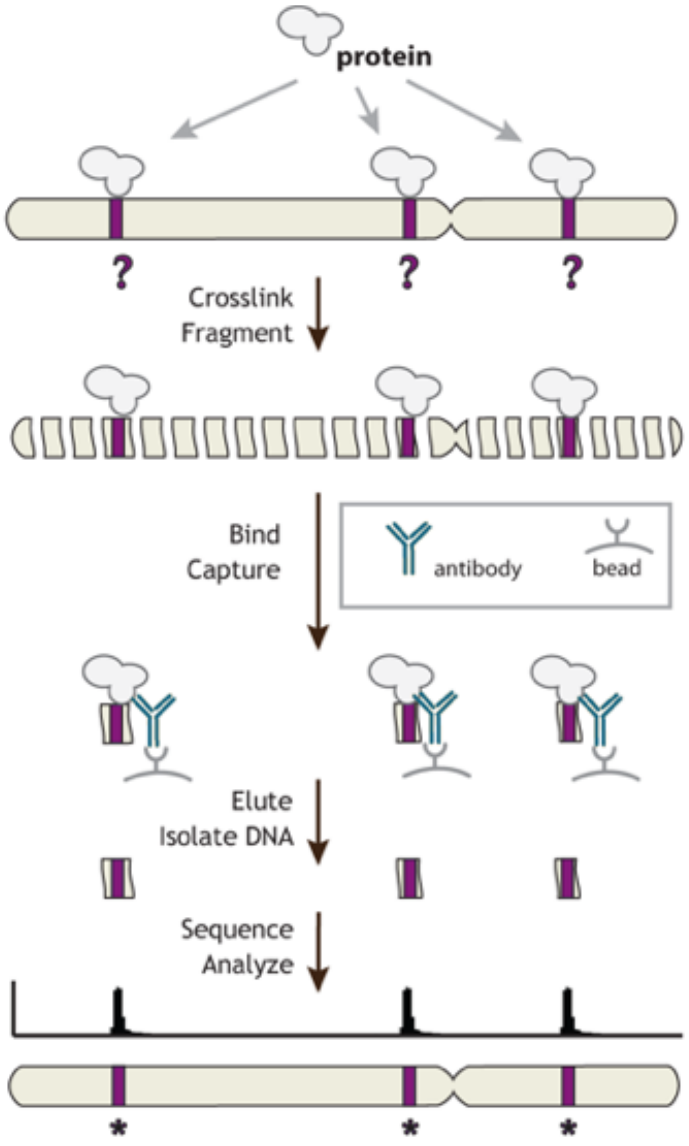
1. **Cross linking** efficiency is not necessarily uniform.

.

2. Enrichment is dependent on the **quality of antibody.**
   e.g., Site and degree of histone modifications.

3. Enrichment is dependent on the **accessibility of the epitope.**
   Comparing different sites to each other in the genome can be problematic.

Limitations of ChIP-Seq



1. **Cross linking** efficiency is not necessarily uniform.

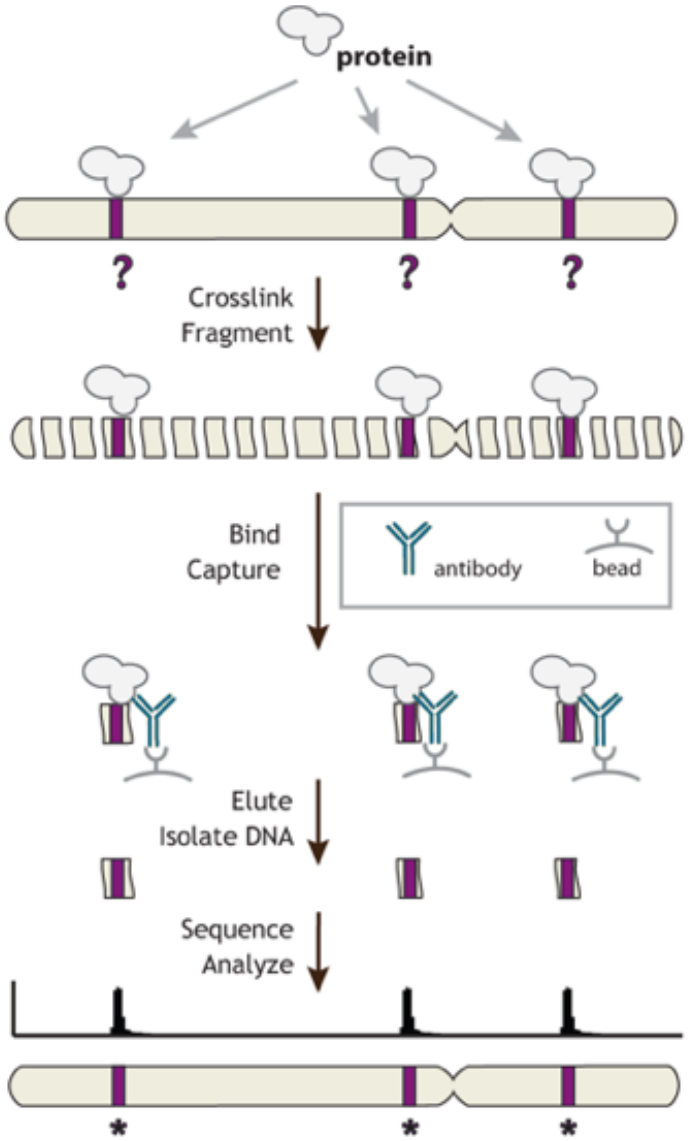
.

2. Enrichment is dependent on the **quality of antibody.**
   e.g., Site and degree of histone modifications.

3. Enrichment is dependent on the **accessibility of the epitope.**
   Comparing different sites to each other in the genome can be problematic.

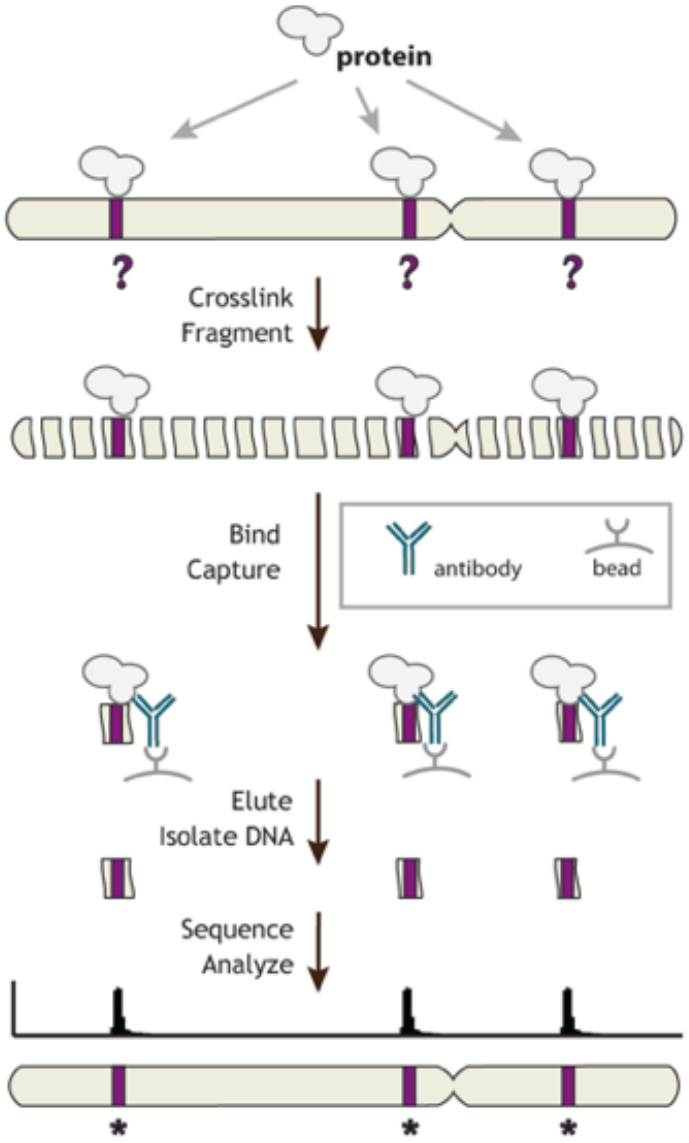
4. Output is **descriptive**.
   Hard to infer function without more experimentation.

Extensions of ChIP



1. Using a nuclease to achieve **higher resolution** (ChIP-exo).

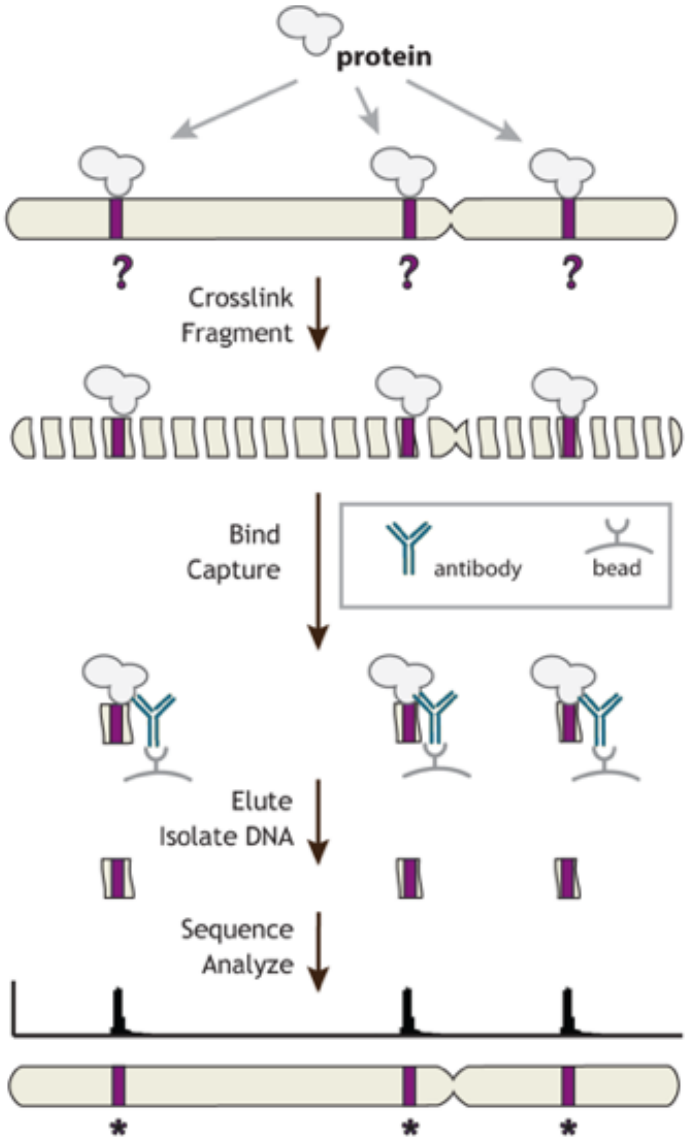
Extensions of ChIP



1. Using a nuclease to achieve **higher resolution** (ChIP-exo).

2. Analysis of **small samples or single cells** (CUT&RUN or CUT&Tag).

Extensions of ChIP

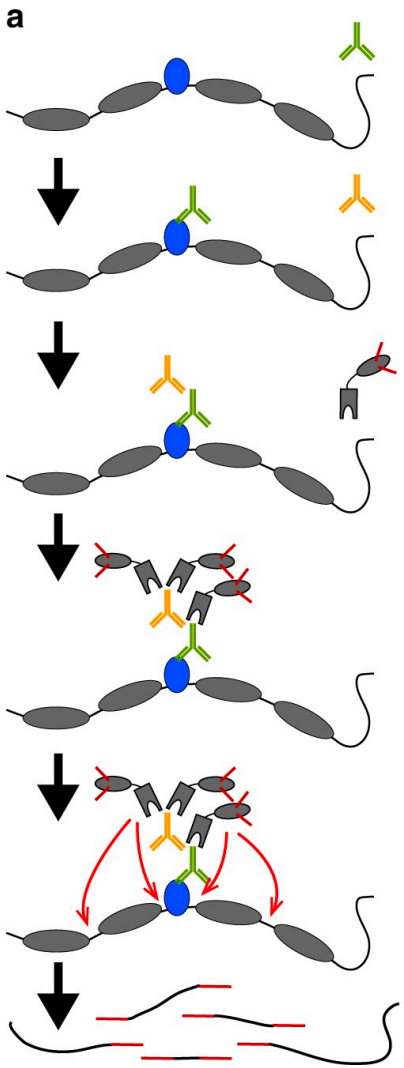


1. Using a nuclease to achieve **higher resolution** (ChIP-exo).

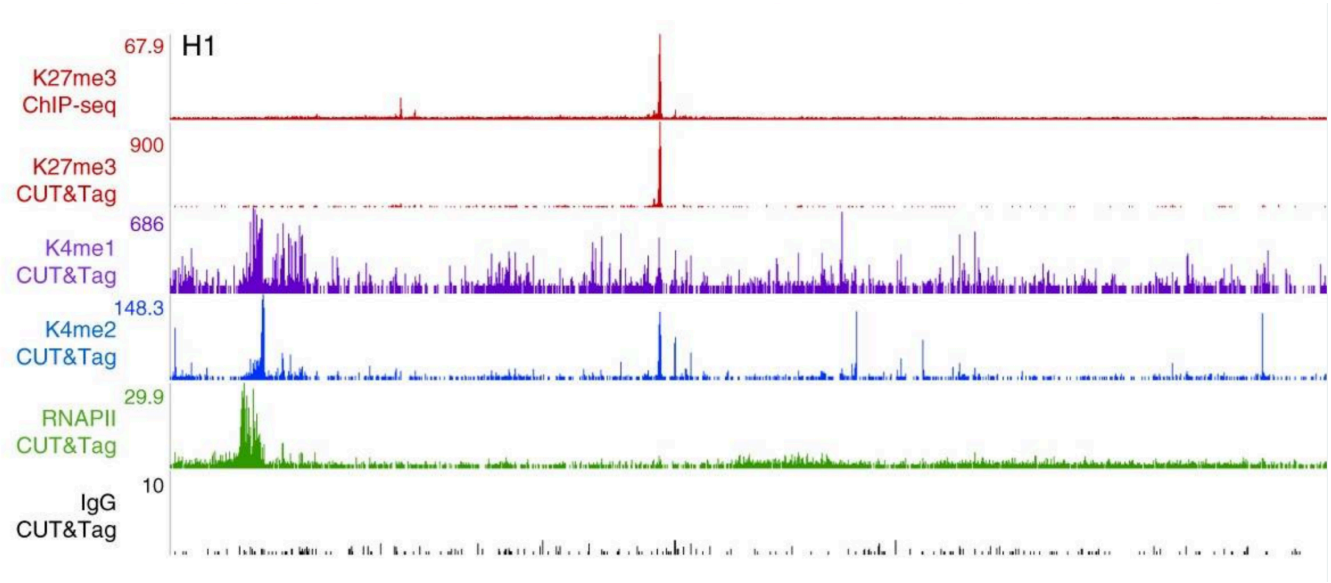
2. Analysis of **small samples or single cells** (CUT&RUN or CUT&Tag).

3. Extension to **RNA factors**.

Extensions of ChIP: CUT&Tag

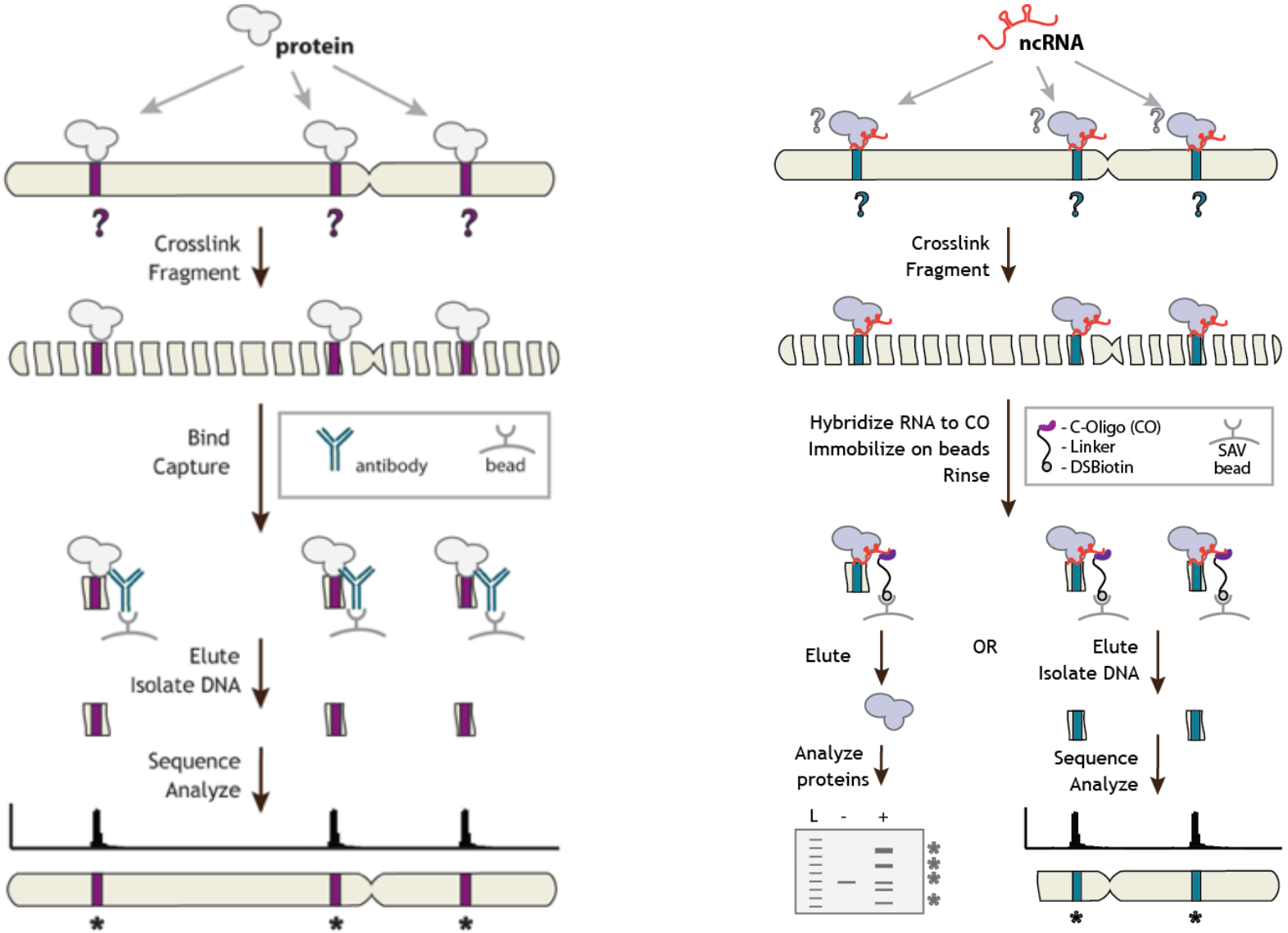


Concept: Use factor-specific antibodies to target a transposes to direct the addition of DNA tags.

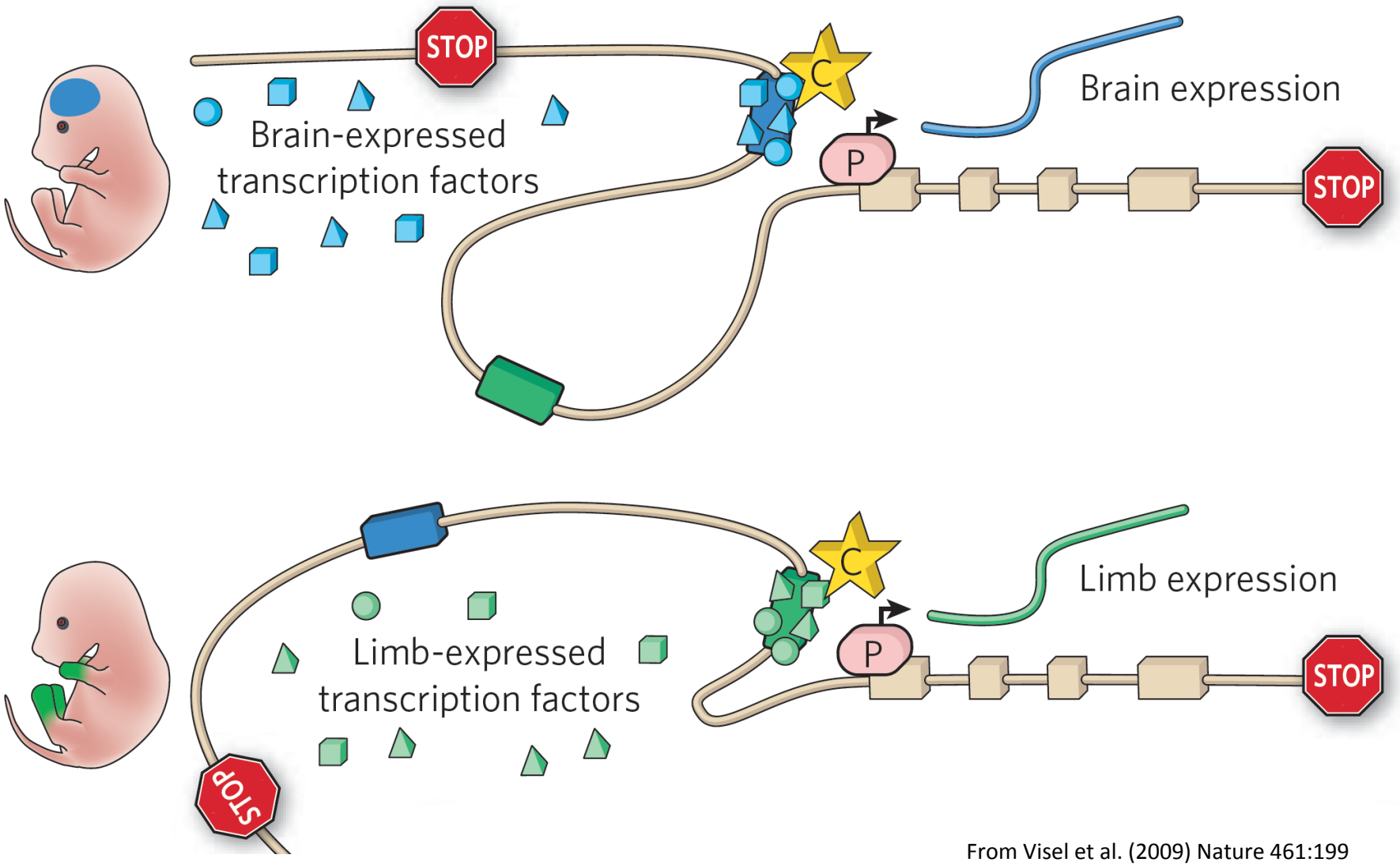


Kaya-Okur…& Henikoff (2019) CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun*

Extension to RNA factors: CHART, ChIRP and RAP



The 3D organization of the genome is important



From Visel et al. (2009) Nature 461:199

Ligation-based methods to study 3D conformation

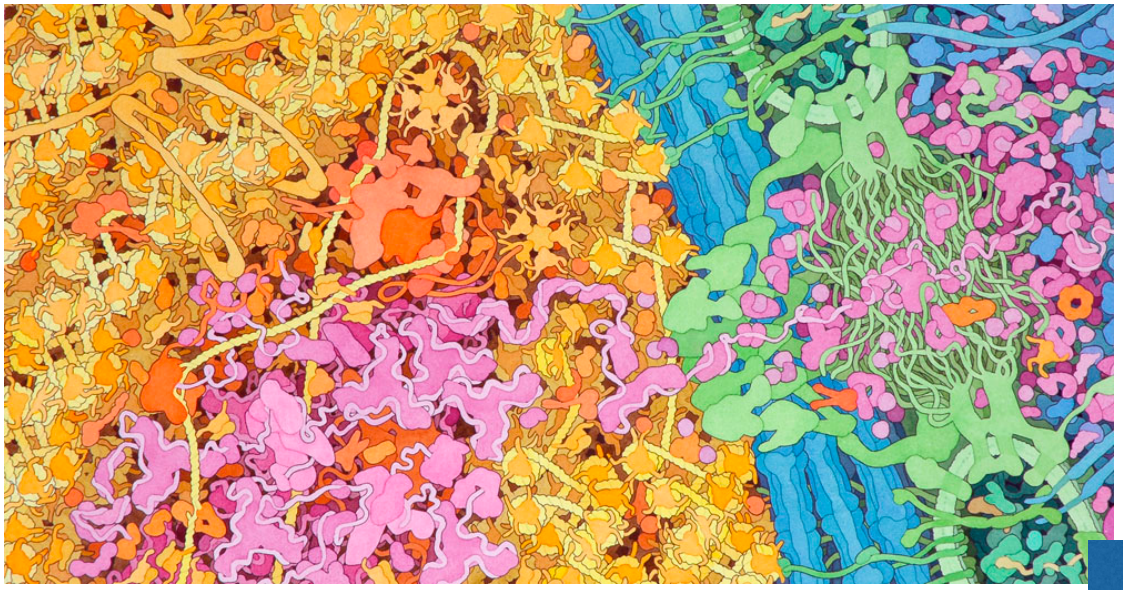
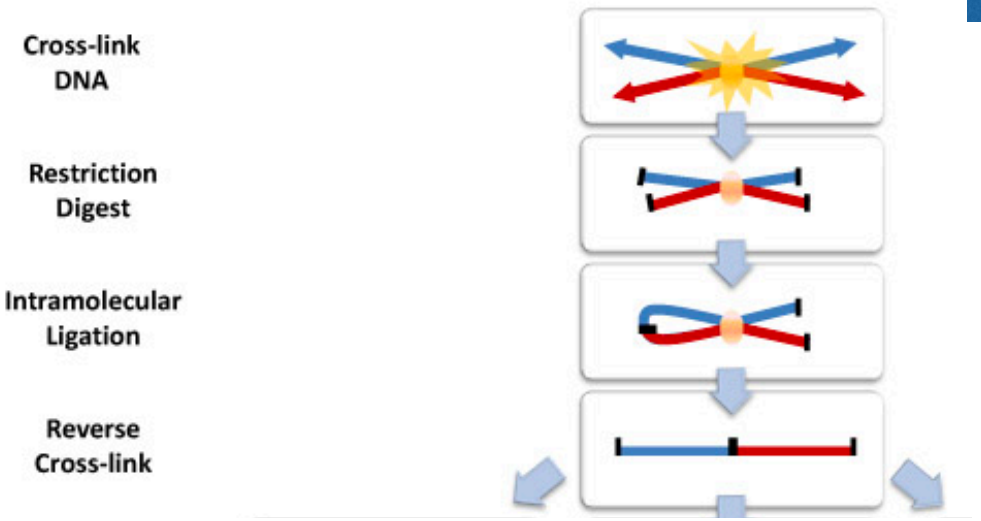
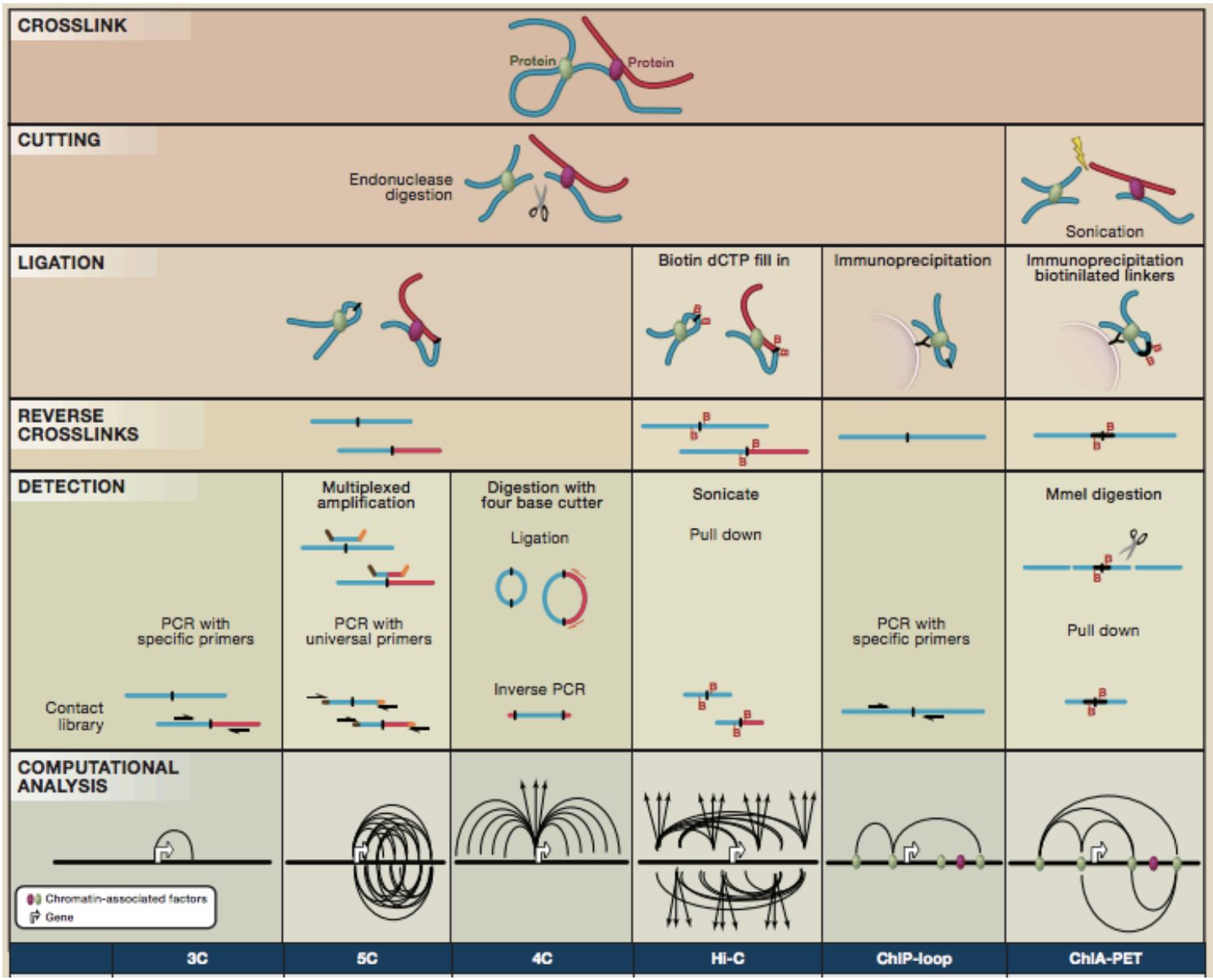


Image: David Goodsell



https://en.wikipedia.org/wiki/Chromosome_conformation_capture#/media/File:Chromosome_Conformation_Capture_Technology.jpg

Many techniques to analyze chromatin conformation



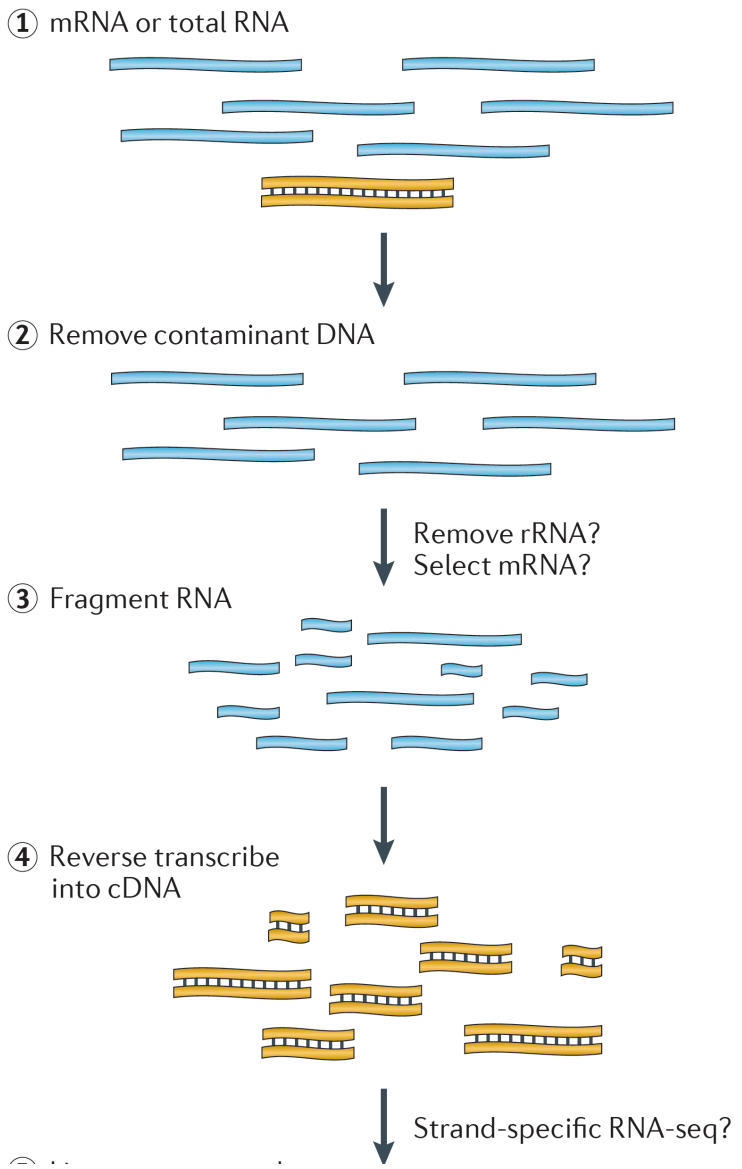
Hakim & Misteli, Cell (2012)

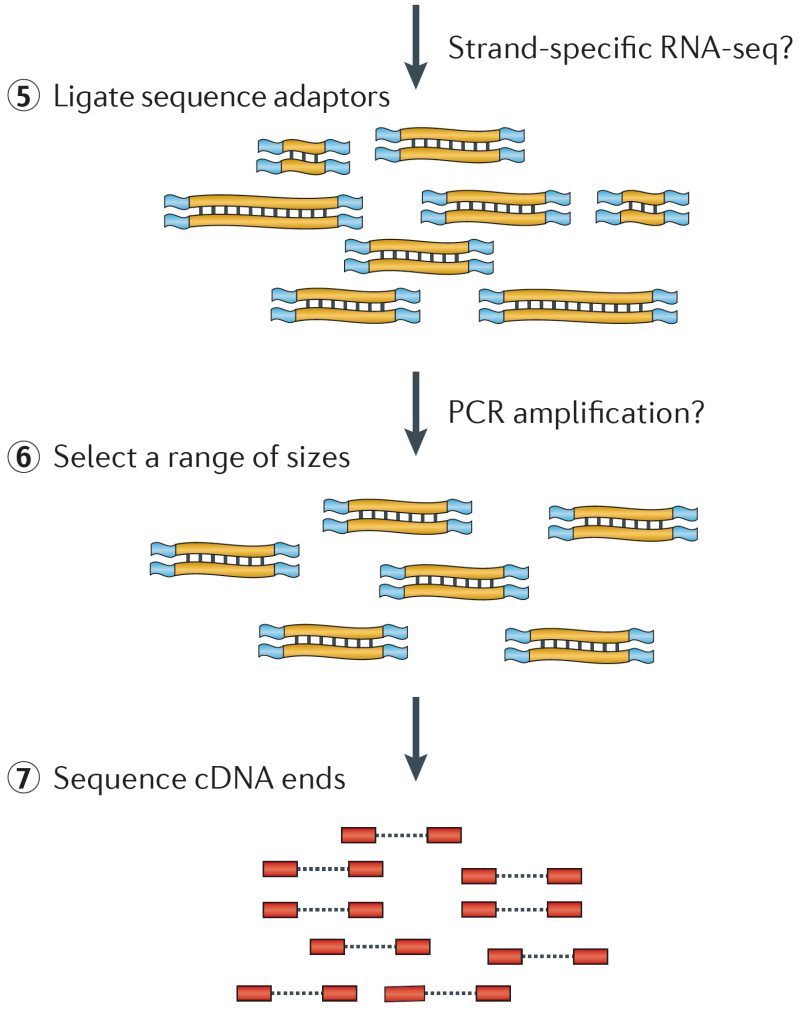
Part 2: RNA-Seq and applications of RNA-Seq

Using RNA-Seq to examine RNA

- Technical methodology
- Read mapping and normalization
- Estimating isoform-level gene expression
- De novo transcript reconstruction
- Sensitivity and sequencing depth
- Differential expression analysis

RNA-Seq workflow



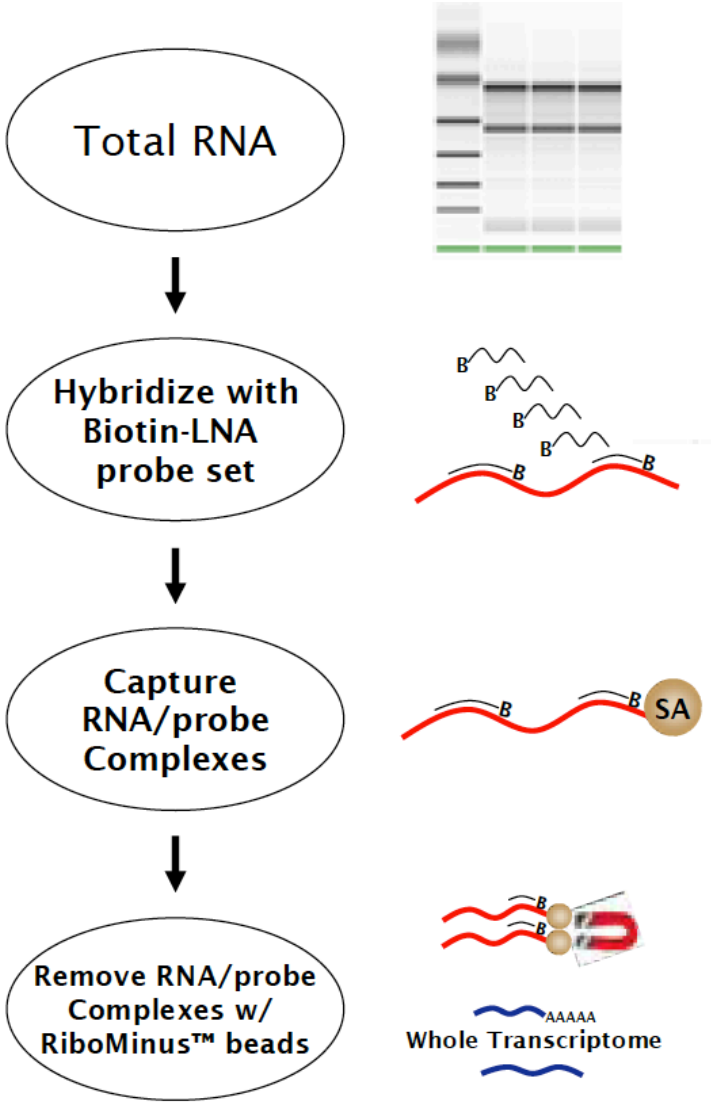


Martin and Wang *Nat Rev Genet* 12:671 (2011)

Some technical details specific to RNA-Seq

- Wide dynamic range of RNA concentrations.
- RNA is strand specific (unlike dsDNA)
- RNA degrades easily (RNase and spontaneous)
- RNA is processed (e.g., spliced)
- RNA has secondary structure (possible blocks to reverse transcriptase).

Ribosomal RNA will dominate the sequenced reads unless removed

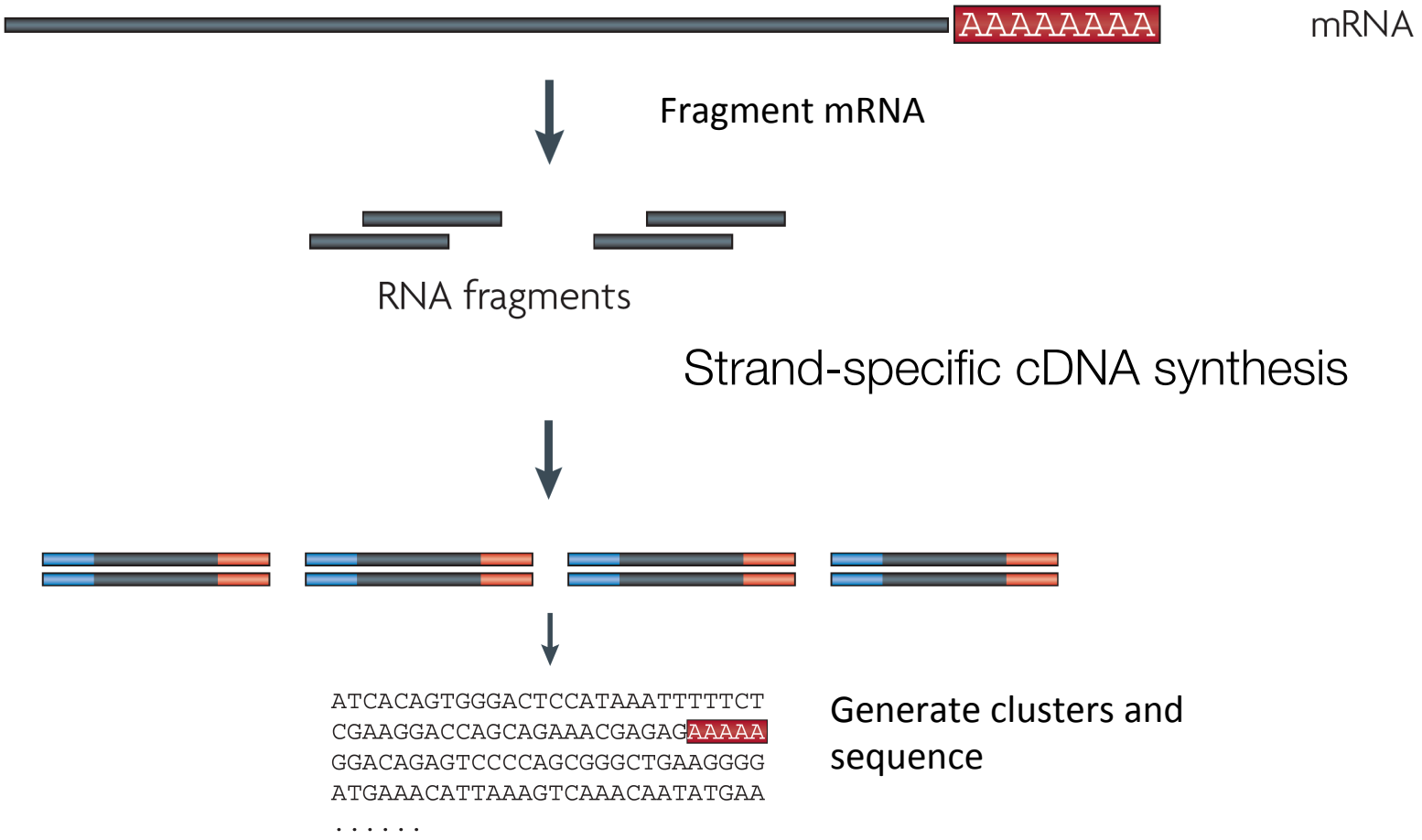


RiboMinus

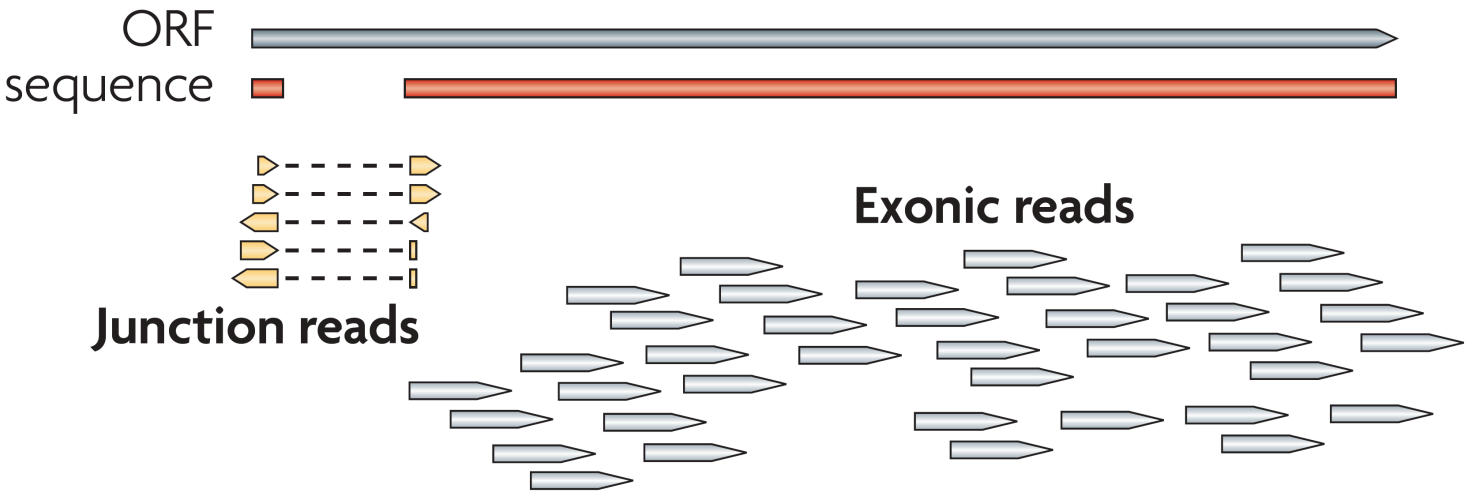
Illumina RNA-seq workflow

Capture poly-A RNA with poly-T oligo attached beads (100 ng total) (2x)

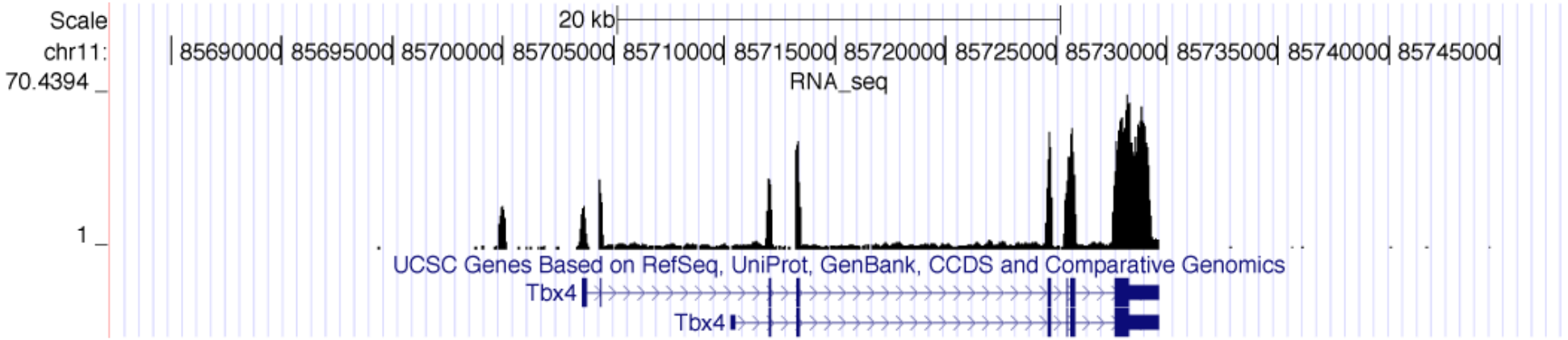
- RNA quality must be high – degradation produces 3' bias
- Non-poly-A RNAs are not recovered



RNA-Seq reads map mostly to exons



Martin and Wang *Nat Rev Genet* 12:671 (2011)



How does one analyze RNA levels from RNA-Seq?

Use existing gene annotation:

Align to genome plus annotated splices
Depends on high-quality gene annotation
Which annotation to use: RefSeq, GENCODE, UCSC?
Isoform quantification?
Identifying novel transcripts?

Reference-guided alignments:

Align to genome sequence
Infer splice events from reads
Allows transcriptome analyses of genomes with poor gene annotation

De novo transcript assembly:

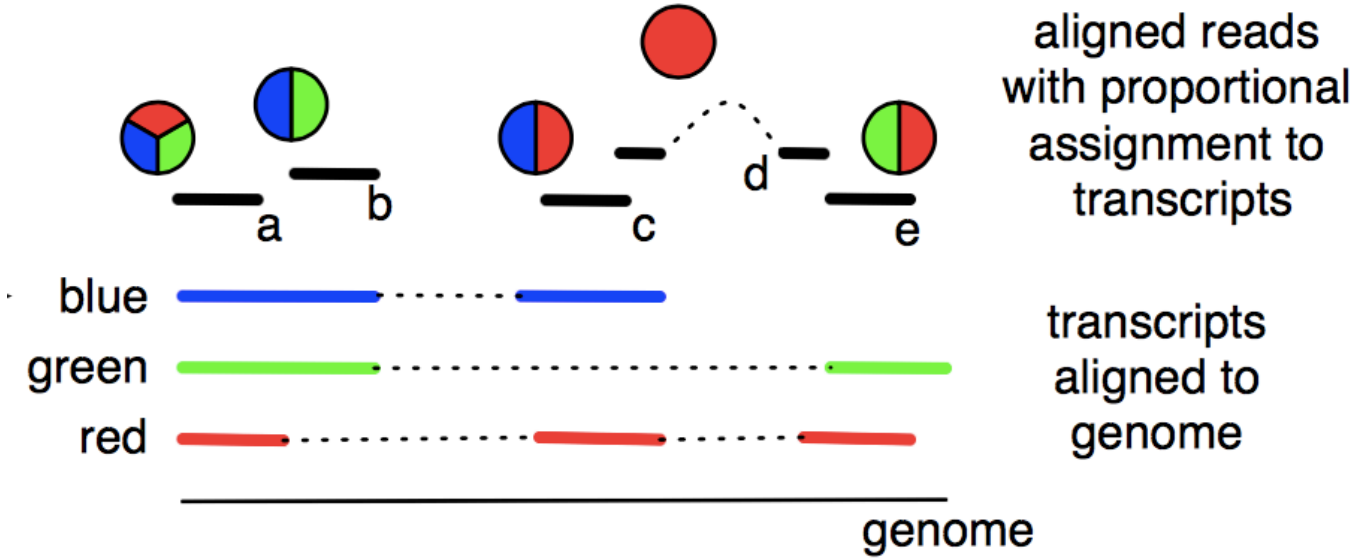
Assemble transcripts directly from reads
Allows transcriptome analyses of species without reference genomes

RNA-seq reads contain information about the abundance of different transcript isoforms

Normalization :

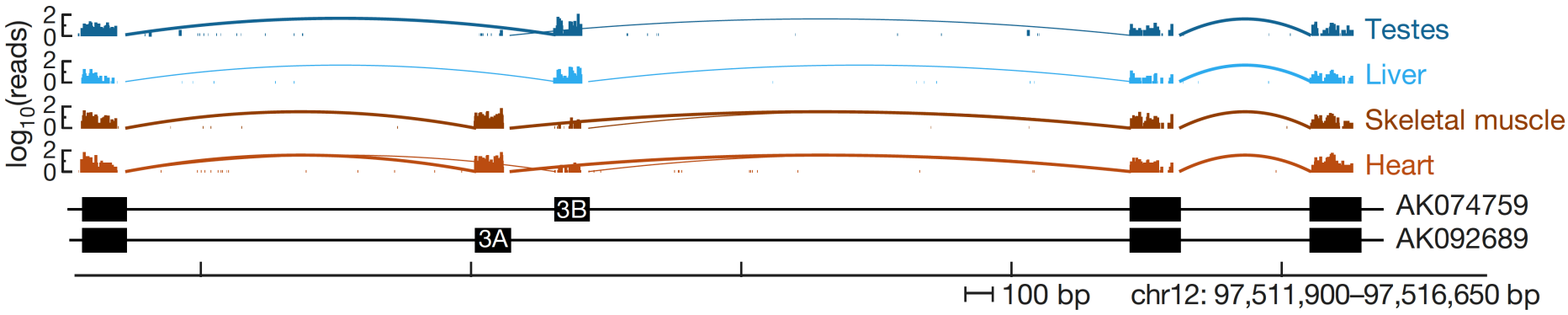
Internal: *Reads or Fragments* per kilobase of feature length per million mapped reads (RPKM or FPKM)

External: Reads relative to a standard “spike”



http://arxiv.org/pdf/1104.3889v2.pd

Functional diversity in transcript isoforms



Alternative transcript events	Total events (×10³)	Number detected (×10³)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon	37	35	10,436	6,822	65	72
Retained intron	1	1	167	96	57	71
Alternative 5′ splice site (A5SS)	15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)	17	16	4,181	2,655	64	74
Mutually exclusive exon (MXE)	4	4	167	95	57	66
Alternative first exon (AFE)	14	13	10,281	5,311	52	63
Alternative last exon (ALE)	9	8	5,246	2,491	47	52
Tandem 3′ UTRs	7	7	5,136	3,801	74	80
Total	105	100	37,782	22,657	60	68

Constitutive exon or region — Body read — Junction read — pA Polyadenylation site
Alternative exon or extension — Inclusive/extended isoform — Exclusive isoform — Both isoforms

Examples of applications of RNA-seq

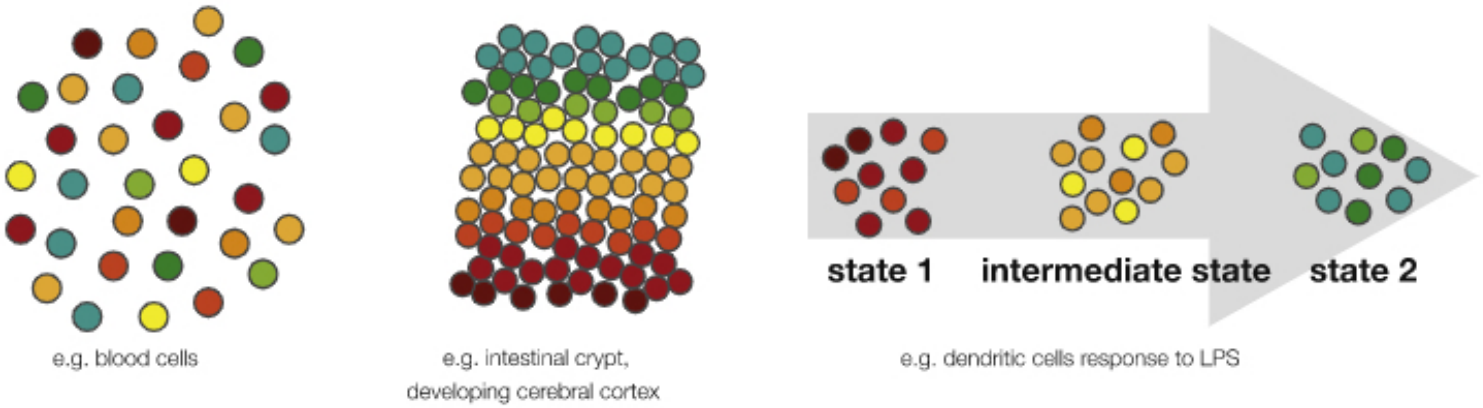
- Characterizing transcriptome complexity
  - Alternative splicing
- Differential expression analysis
  - Gene- and isoform-level expression comparisons
- Novel RNA species
  - lncRNAs and eRNAs
  - Pervasive transcription
- Translation
  - Ribosome profiling
- Allele-specific expression
- Measuring RNA half-lives and decay
- Examining protein-RNA interactions (CLIP, RIP, &c.)
- Effect of genetic variation on gene expression
  - Imprinting
  - RNA editing
  - Novel events

Examining cell heterogeneity with scRNA-seq

Bulk RNA-seq averages over the RNA content of many cells masking differences.

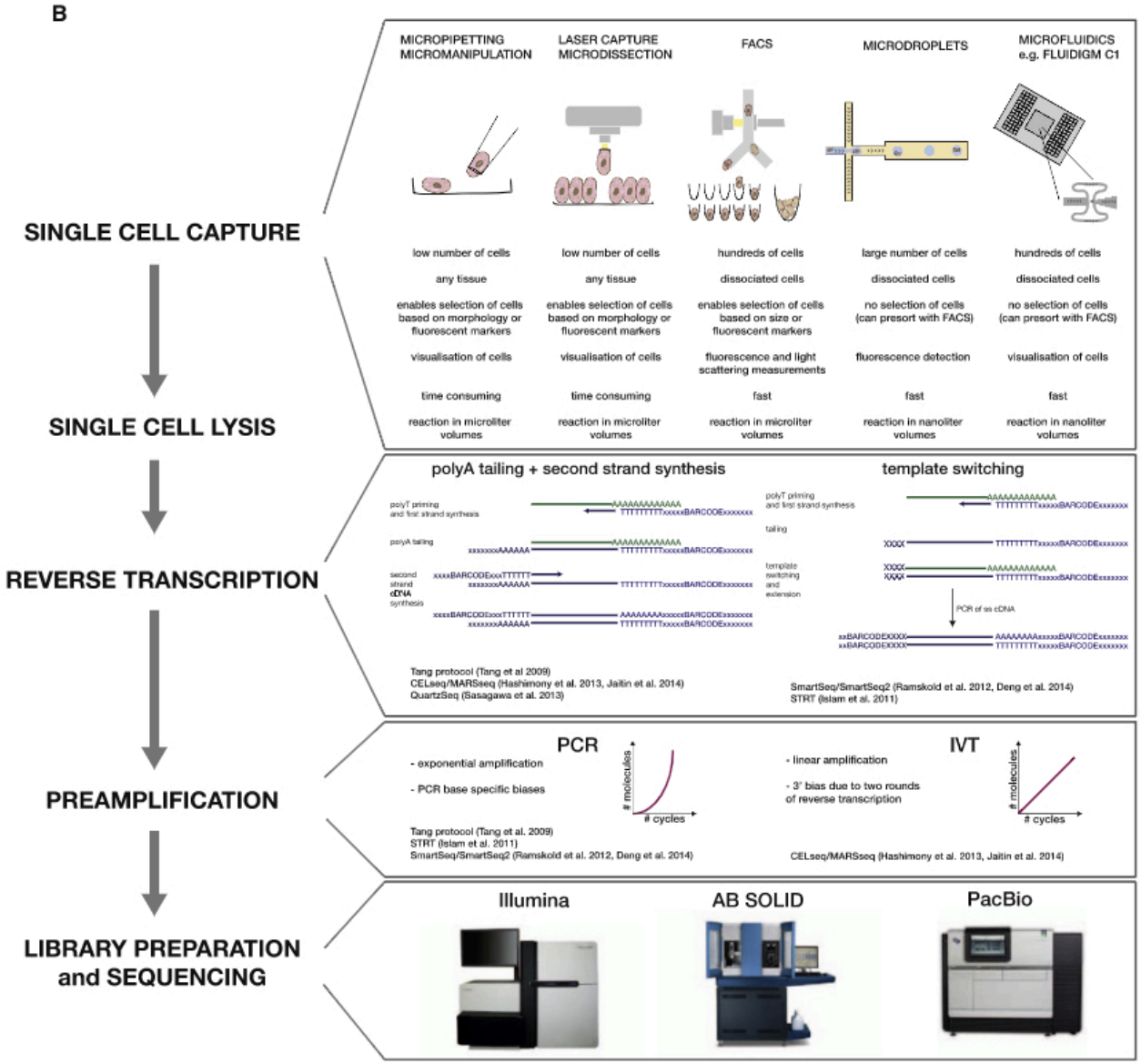
These differences can be revealed by sequencing the RNA from individual cells using single cell RNA-seq (scRNA-seq)

Analysis of RNA transcripts in individual cells can reveal rare cell populations and lineage trajectories.



Kolodziejczyk … & Teichmann (2015). The technology and biology of single-cell RNA sequencing. Mol Cell

Examining cell heterogeneity with scRNA-seq



Kolodziejczyk … & Teichmann (2015). The technology and biology of single-cell RNA sequencing. Mol Cell

Summary

- Genomics I: Deep sequencing gives us access to information on a genomic level.

- Genomics II: These approaches provide a diverse set of tools to study life at a genomic scale.

✴Sophisticated use of data from genomics requires an integrated understanding of the biological experiment, sample preparation and down stream computational analyses of the data.