Bioinformatics: Genomics Part II

Applications of Sequencing Technology

Matt Simon
Dept. of Molecular Biophysics & Biochemistry
Chemical Biology Institute

January 16, 2015

Overview

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

- Genomics II: (Today's lecture): Focus on applications of sequencing technology.

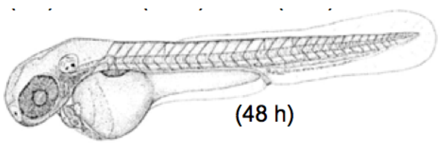
    1. Annotation of the genome in chromatin

    2. Regulation of gene expression at the level of RNA

Credit: Jim Noonan for many of the slides

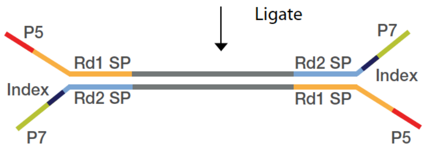
Review of how a sequencing experiment works

1. Isolation of sample.



e.g., Isolate DNA and shear.

2. Library preparation



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

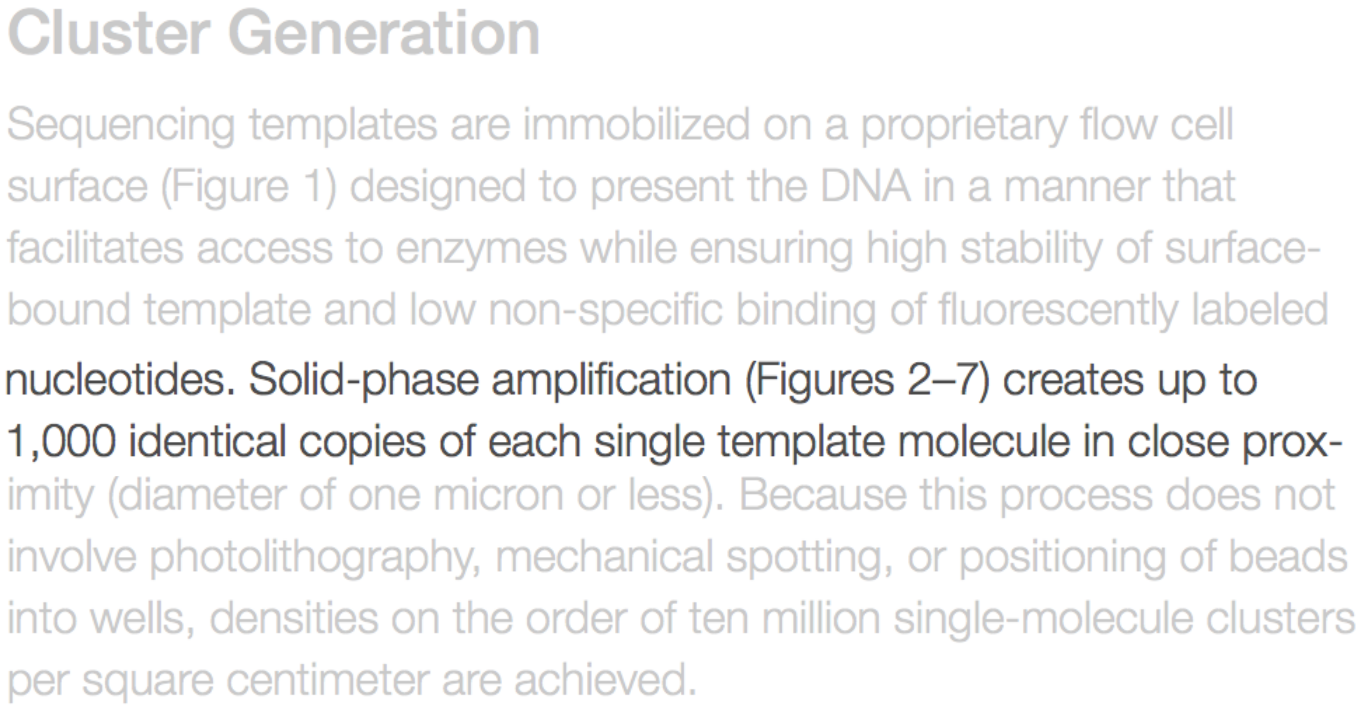
3. Sequencing

e.g., Illumina HiSeq

4. Analysis

e.g., Map to genome and interpret.

Q. How many cycles of PCR are used in flow cell generation?



Cluster Generation

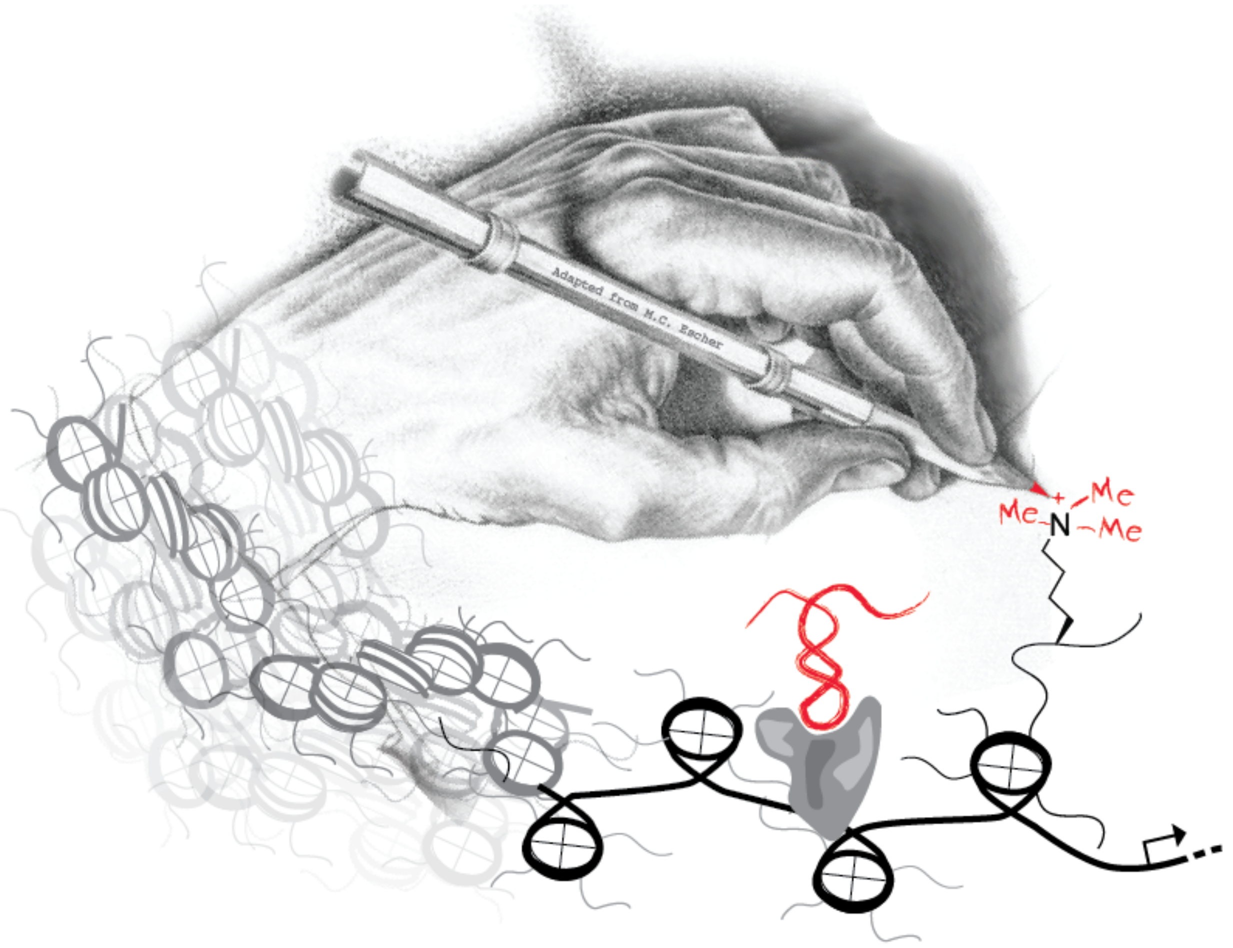
Sequencing templates are immobilized on a proprietary flow cell surface (Figure 1) designed to present the DNA in a manner that facilitates access to enzymes while ensuring high stability of surface-bound template and low non-specific binding of fluorescently labeled nucleotides. Solid-phase amplification (Figures 2–7) creates up to 1,000 identical copies of each single template molecule in close proximity (diameter of one micron or less). Because this process does not involve photolithography, mechanical spotting, or positioning of beads into wells, densities on the order of ten million single-molecule clusters per square centimeter are achieved.

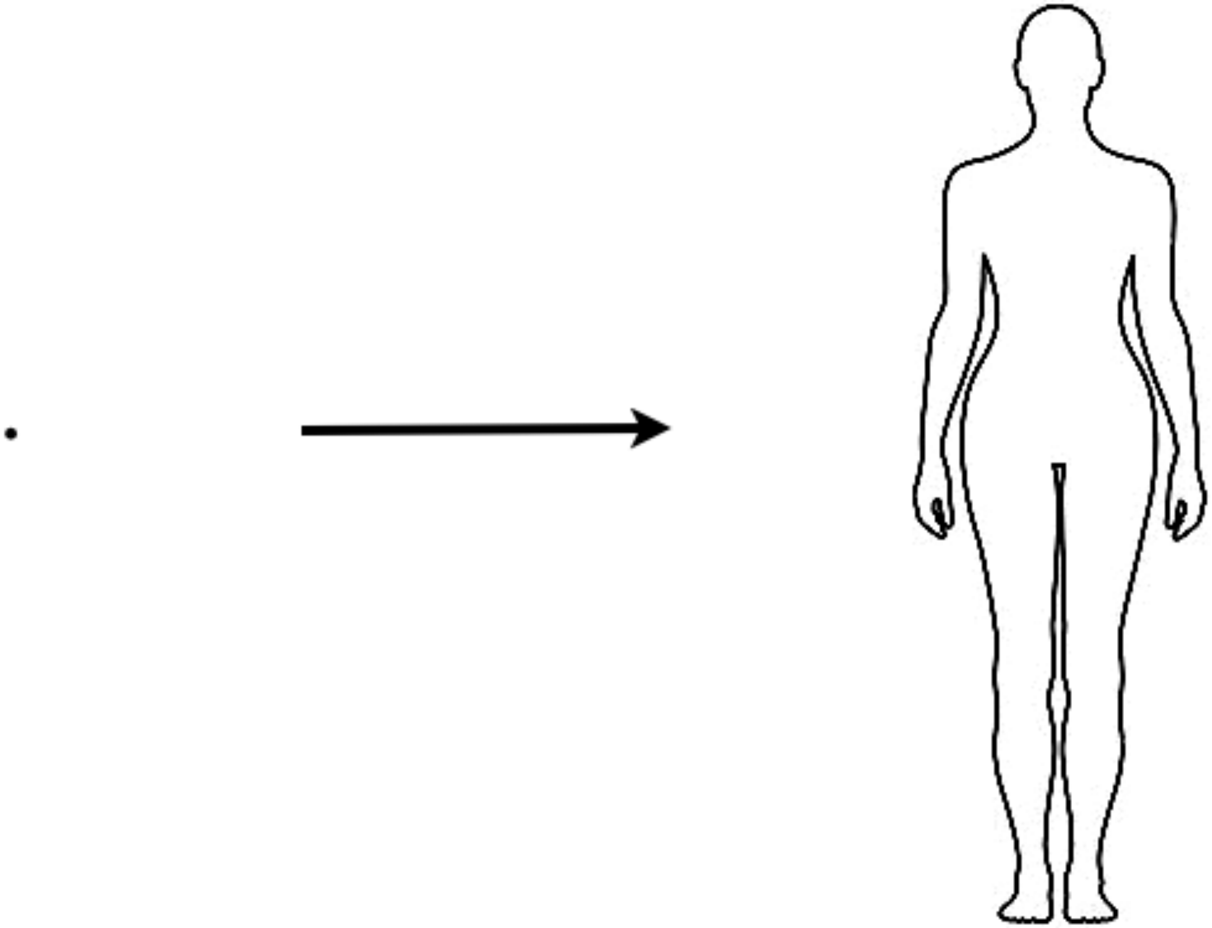
http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

$$1 \text{ molecule} \bullet 2^{n} = 1000 \text{ molecules}$$

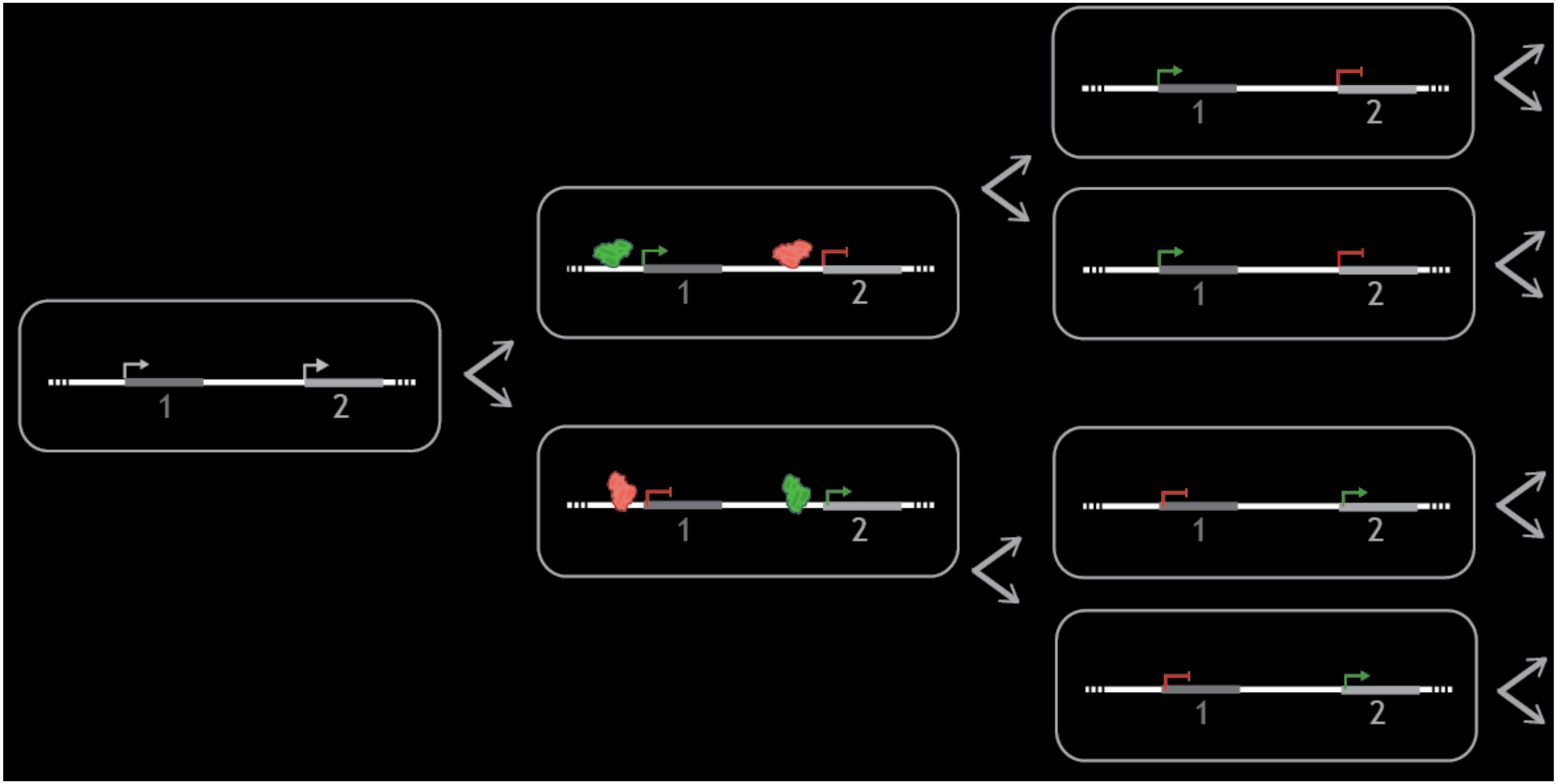
$$n \approx 10$$

Part 1. How do cells annotate their genomes?

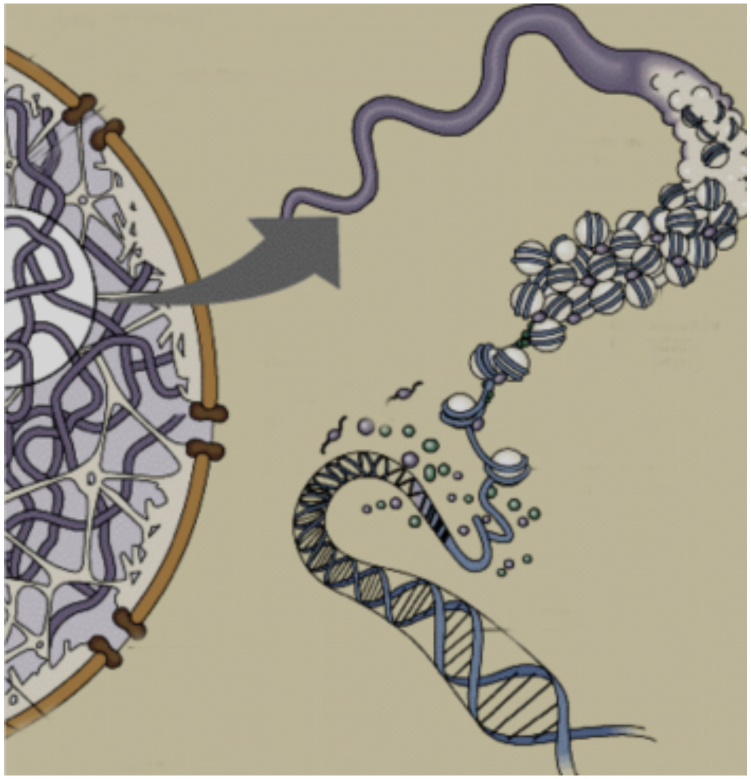


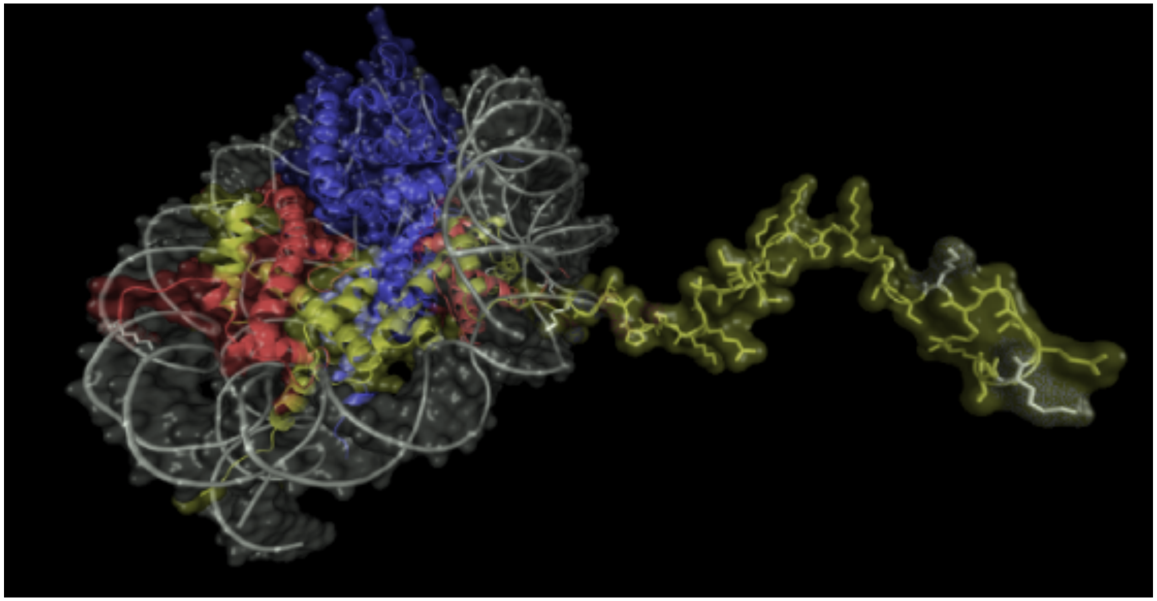


How is gene expression regulated and faithfully inherited?



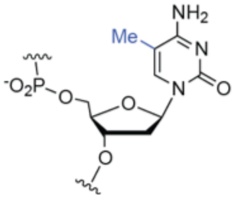
DNA in the cell is packaged into chromatin



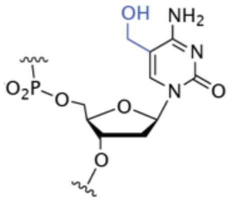


Modeled nucleosome based on Luger et al., *Nature* **1997** *389,* 251.

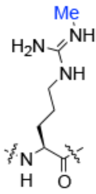
Summary and nomenclature of common covalent modifications.



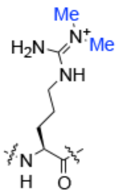
mC



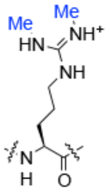
hmC



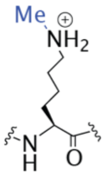
Rme1



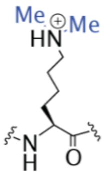
Rme2a



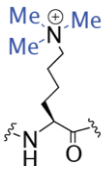
Rme2s



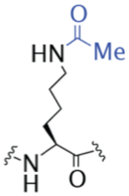
Kme1



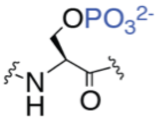
Kme2



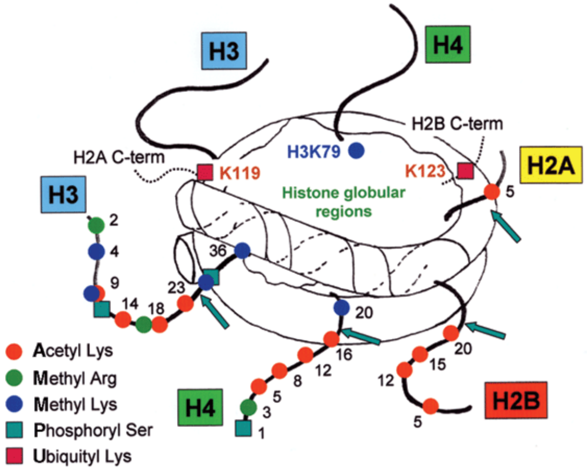
Kme3



Kac



Sph



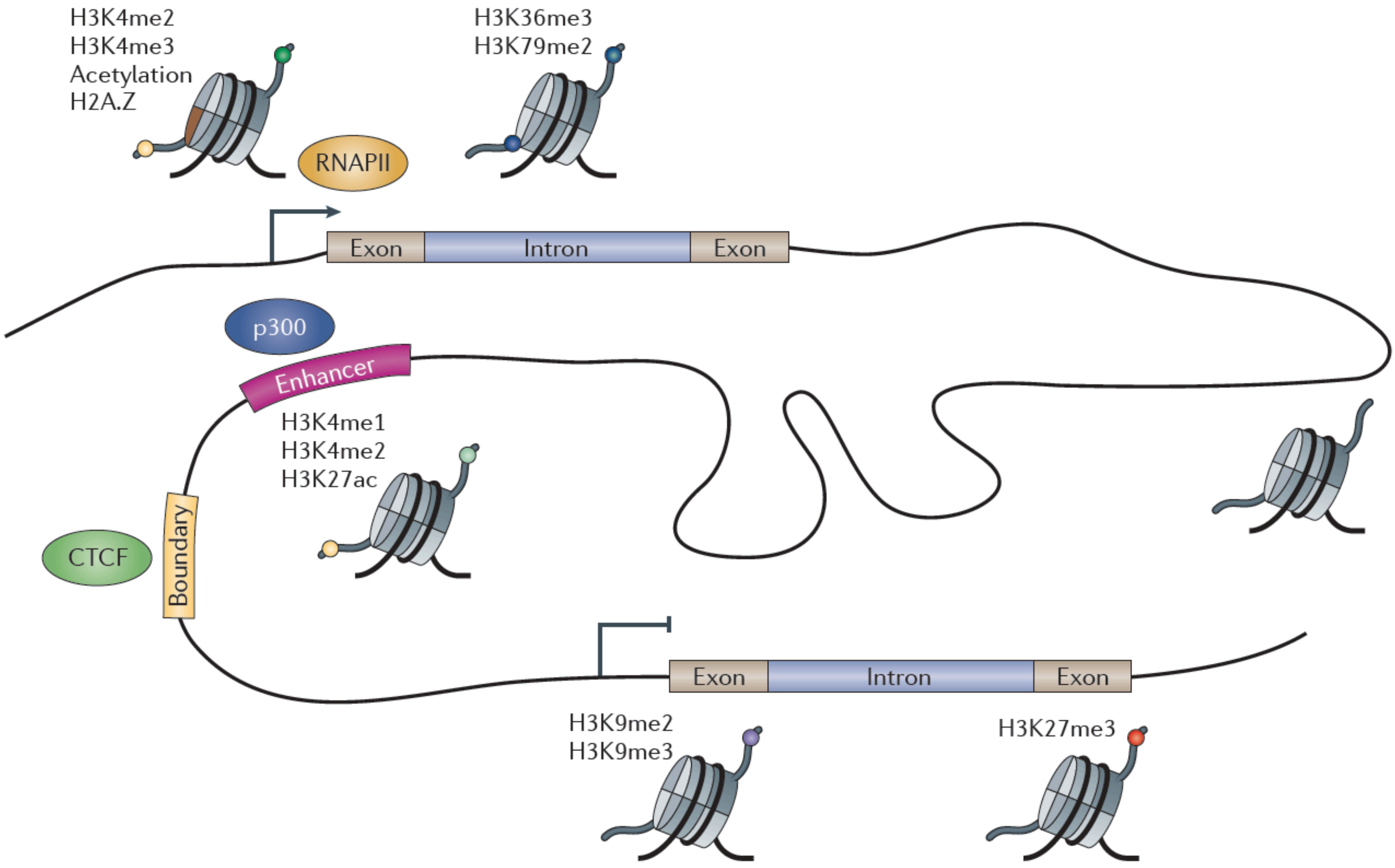
$$\underbrace{\text{H3}}_{\text{Histone}}\quad\underbrace{\text{K27}}_{\text{Residue}}\quad\underbrace{\text{ac}}_{\text{Modification}}$$

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).

Chromatin modifications correlate with different genomic functions.



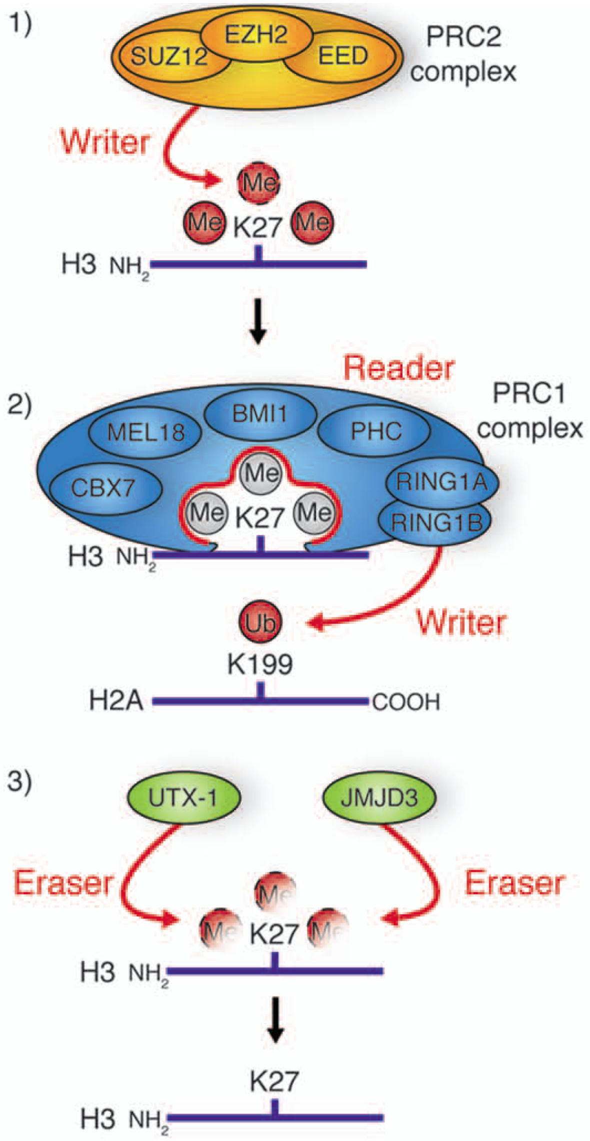
Zhou *et al*. *Nat Rev Genet* 12:7 (2011)

Installing, binding, and removing modifications

Writers
- HMTs
- HATs
- Kinases
- Ubiquitin-ligases
- Sumo-ligases
- ADP-ribosyltransferases

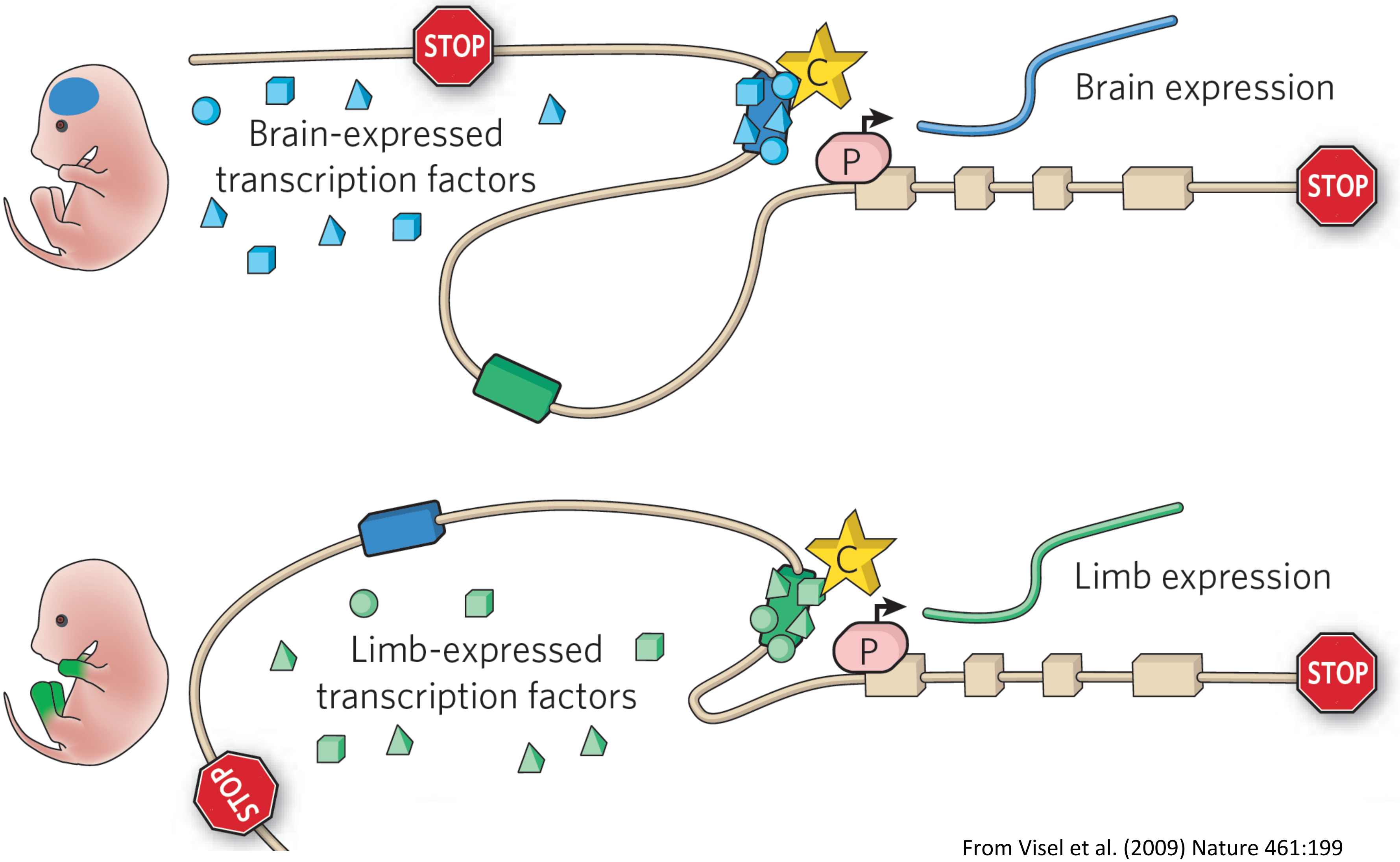
Readers
- Chromodomains
- Bromodomains
- SH2 domains
- FHA domains
- and many others...

Erasers
- HDMTs
- HDACs
- Phosphatases
- Deubiquitinases



Tollervey and Lunyak (2012) Epigenetics 7:823
Ram *et al.*, *Cell* 147:1628 (2011)

Regulation is temporally and specially controlled

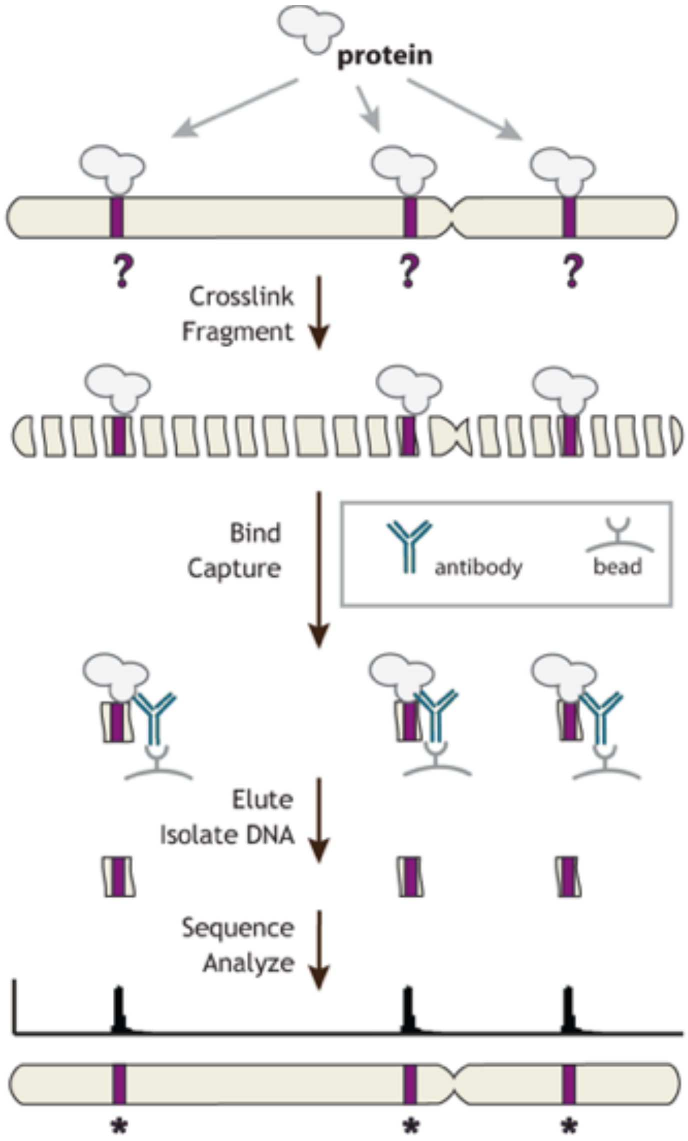

STOP
Brain-expressed transcription factors
C
P
Brain expression
STOP
C
P
Limb expression
Limb-expressed transcription factors
STOP
STOP

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.
2. Where do transcription factors bind?
   C. ChIP-seq of transcription factors (or in high res, ChIP-exo)
   D. Nucleosome mapping (MNase-Seq).
3. Where are different histone modifications found?
   E. ChIP-Seq of histone modifications.
   F. ChIP-Seq of chromatin writers, readers and erasers.
4. Where is RNA polymerase transcribing?
   G. ChIP-Seq of polymerase.
   H. GRO-Seq and NET-Seq to measure RNA in the polymerase active site..
5. How is the genome organized in 3D?
   I. 4C/5C/Hi-C to measure chromatin conformation.

Localization of 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.

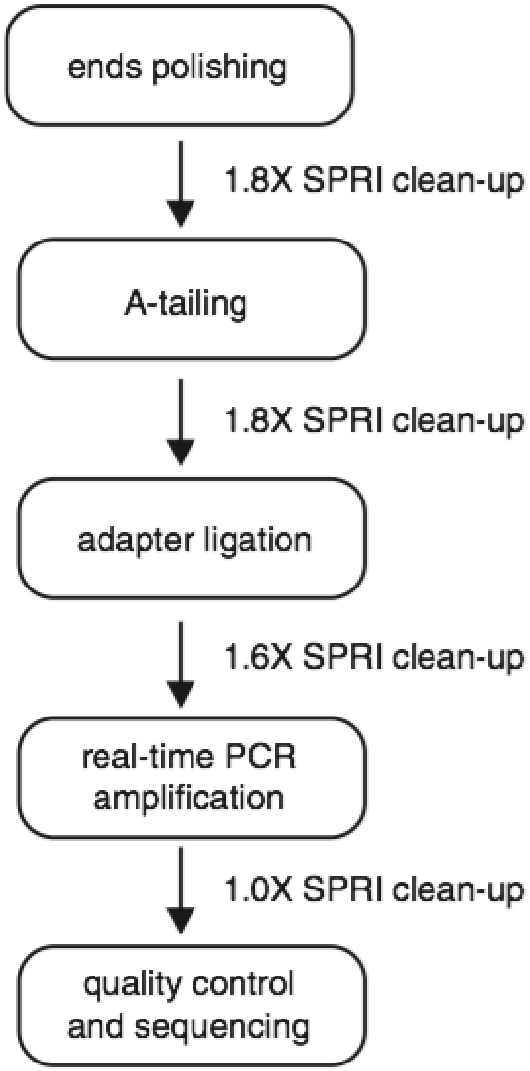
2. **Shear chromatin** to smaller pieces.

   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.

Preparing a Seq library using ChIP-enriched DNA.



ends polishing

1.8X SPRI clean-up

A-tailing

1.8X SPRI clean-up

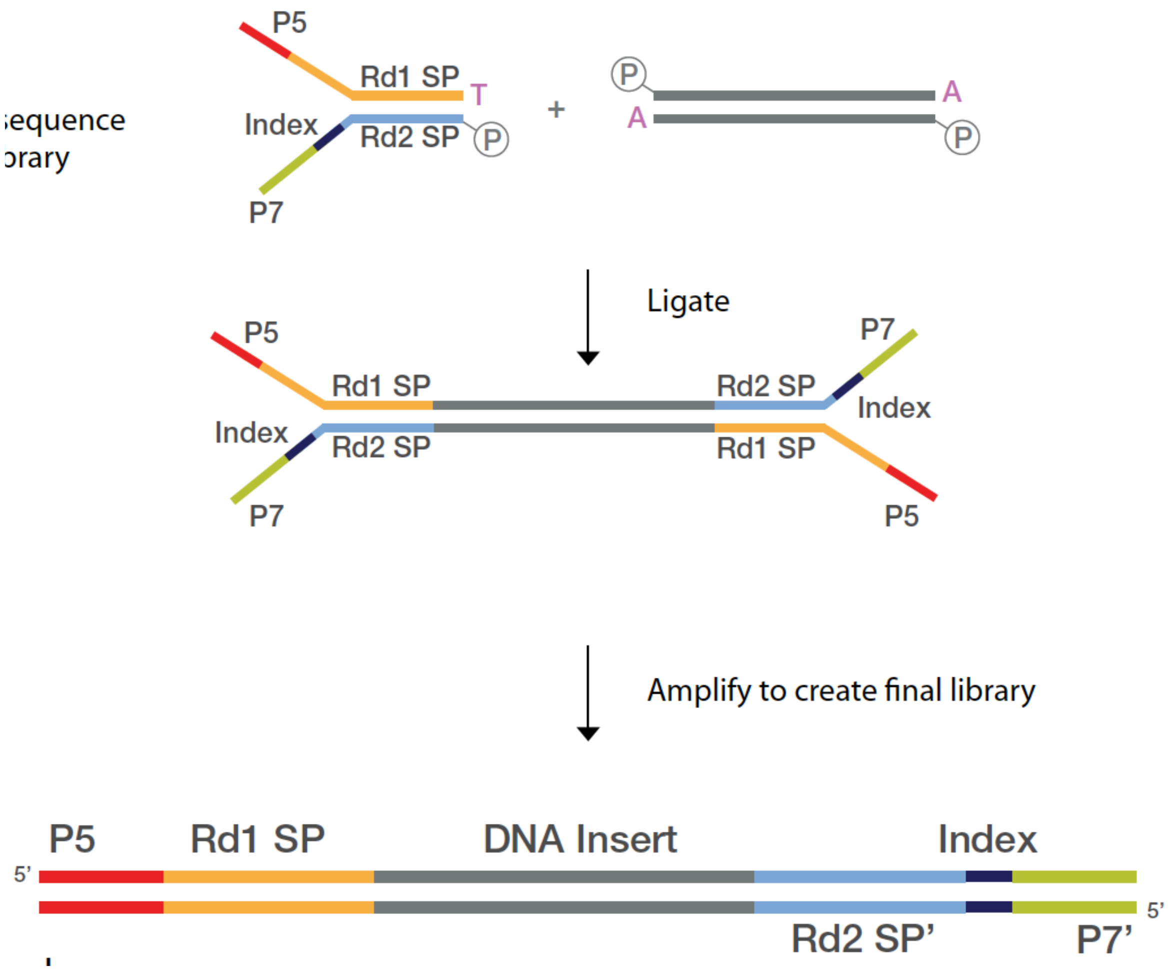
adapter ligation

1.6X SPRI clean-up

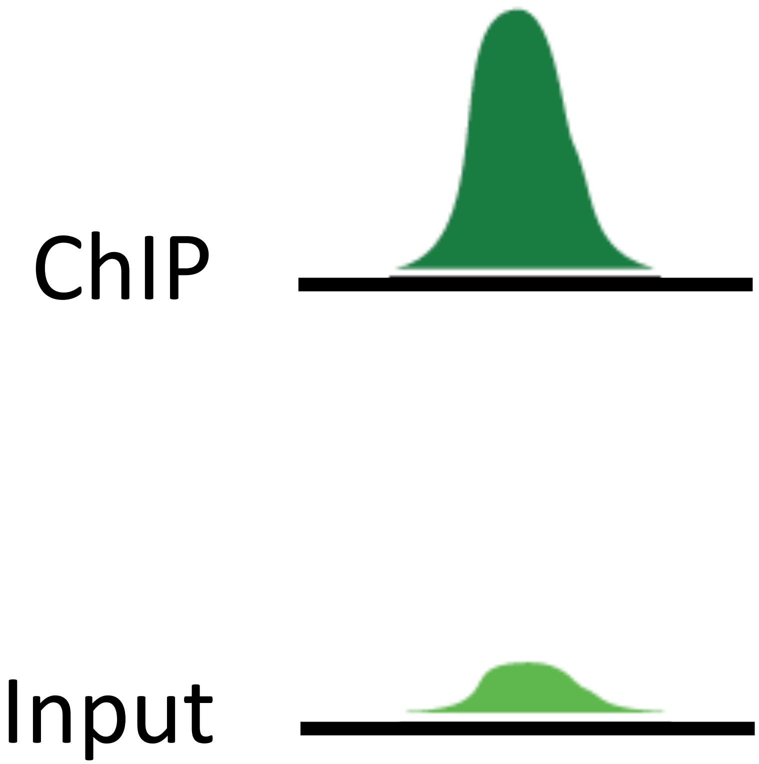
real-time PCR amplification

1.0X SPRI clean-up

quality control and sequencing



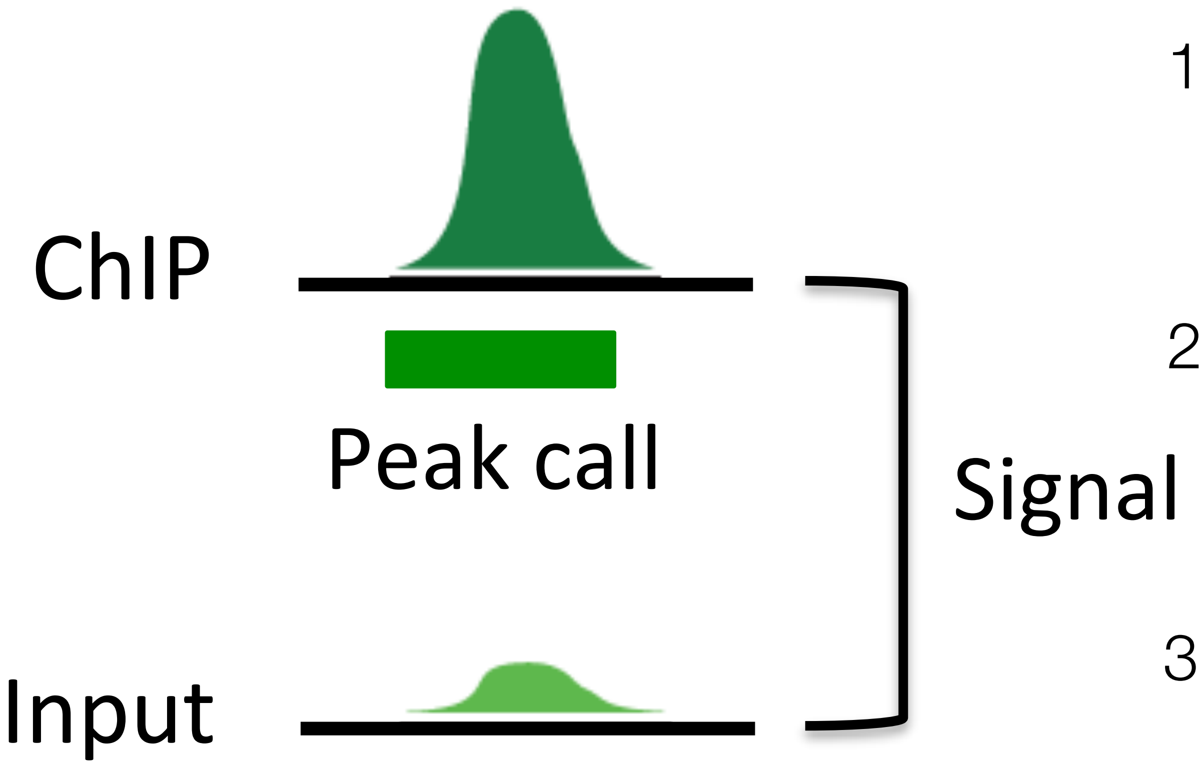
Determining sites of enrichment from ChIP-Seq



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

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

Determining sites of enrichment from ChIP-Seq

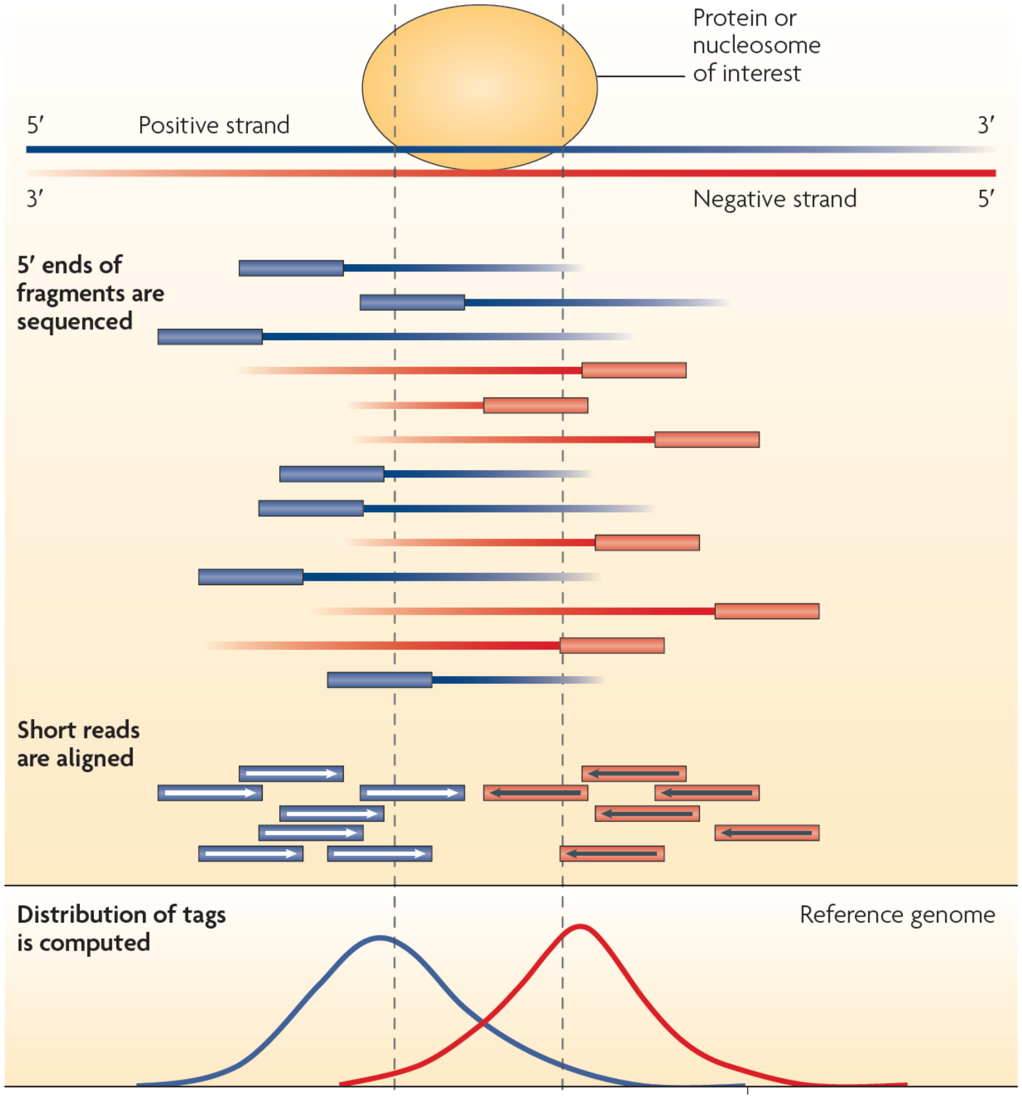


1. **Align** reads to the genome.

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

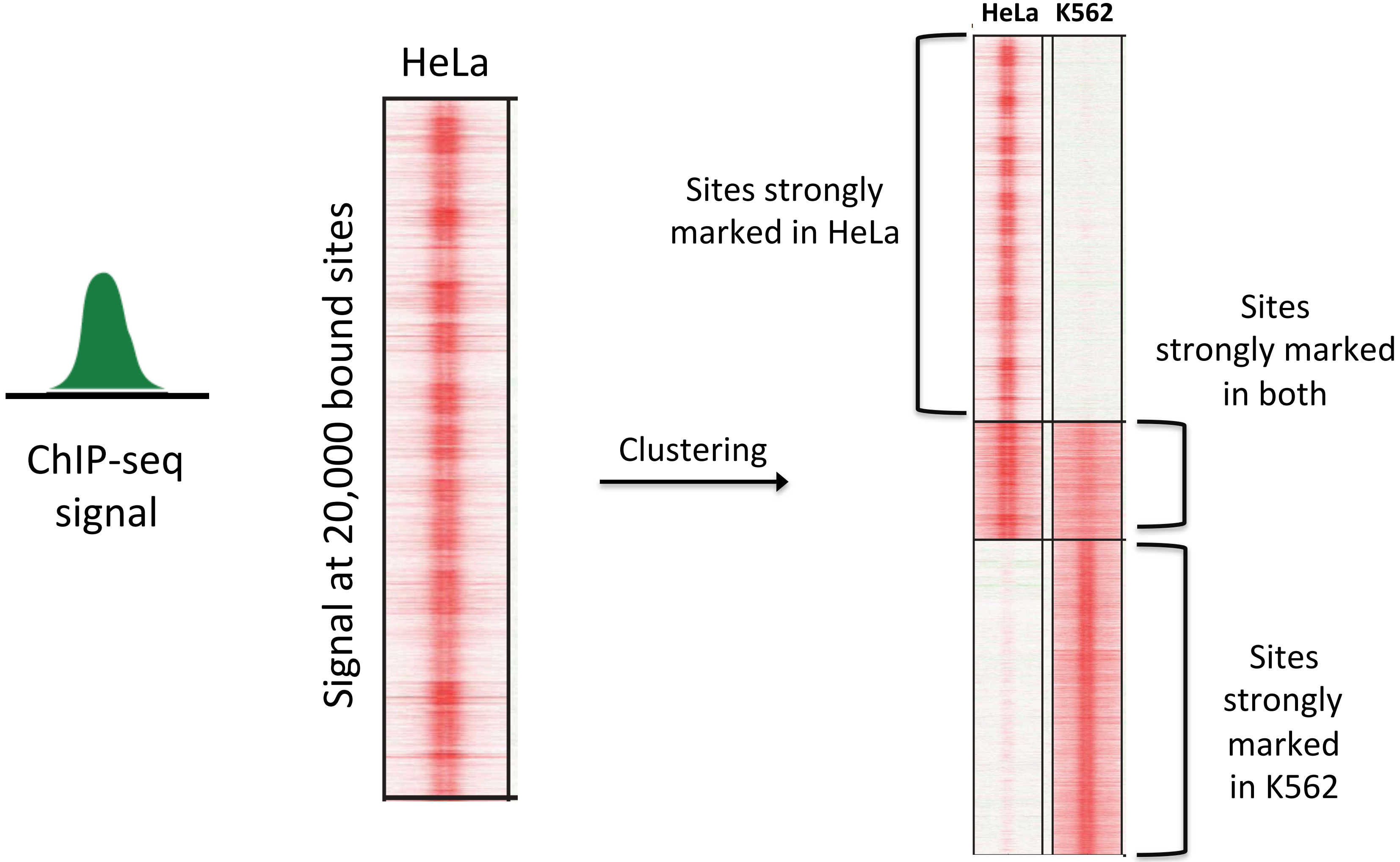
3. **Call peaks** to determine statistically significant sites of enrichment.

Avoiding artifacts using features in Seq data



From Park (2009) Nat Rev Genet 10:669

ChIP-Seq signals reveal difference between cells



ChIP-seq signal

HeLa

Signal at 20,000 bound sites

Clustering

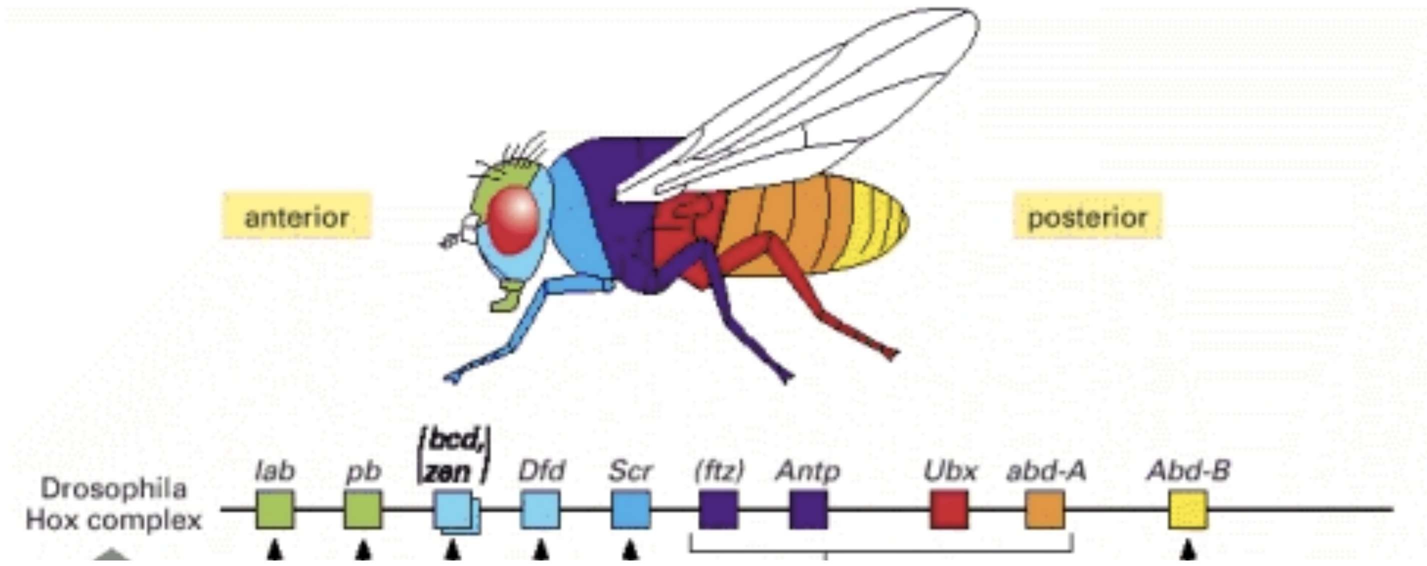
HeLa K562

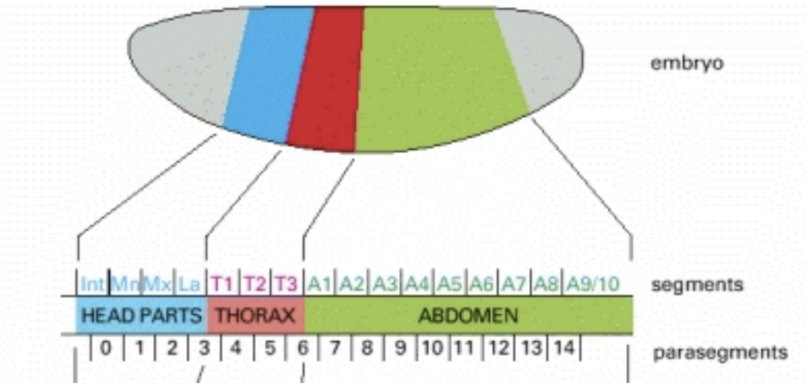
Sites strongly marked in HeLa

Sites strongly marked in both

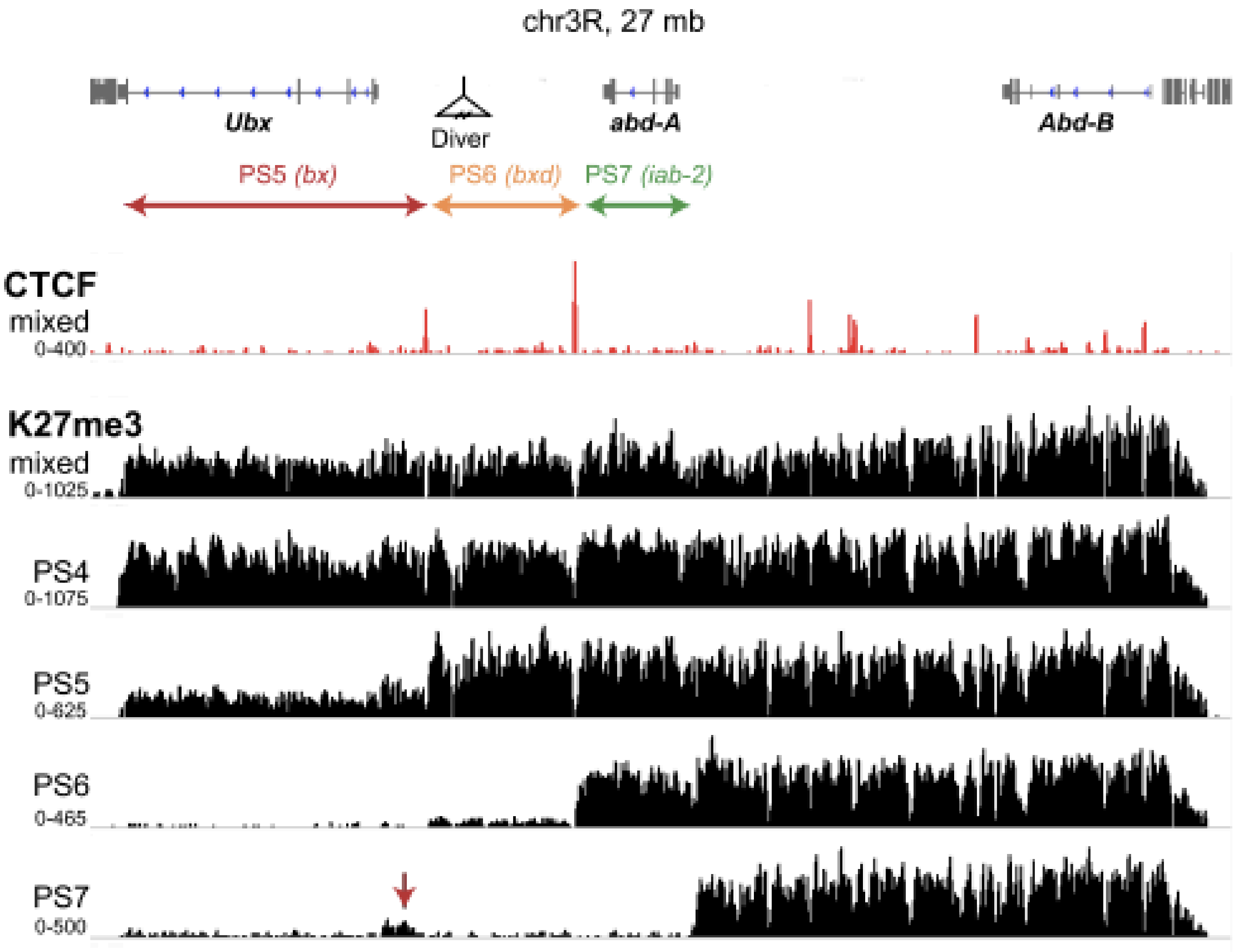
Sites strongly marked in K562

Example: Anterior-to-posterior body plan in flies

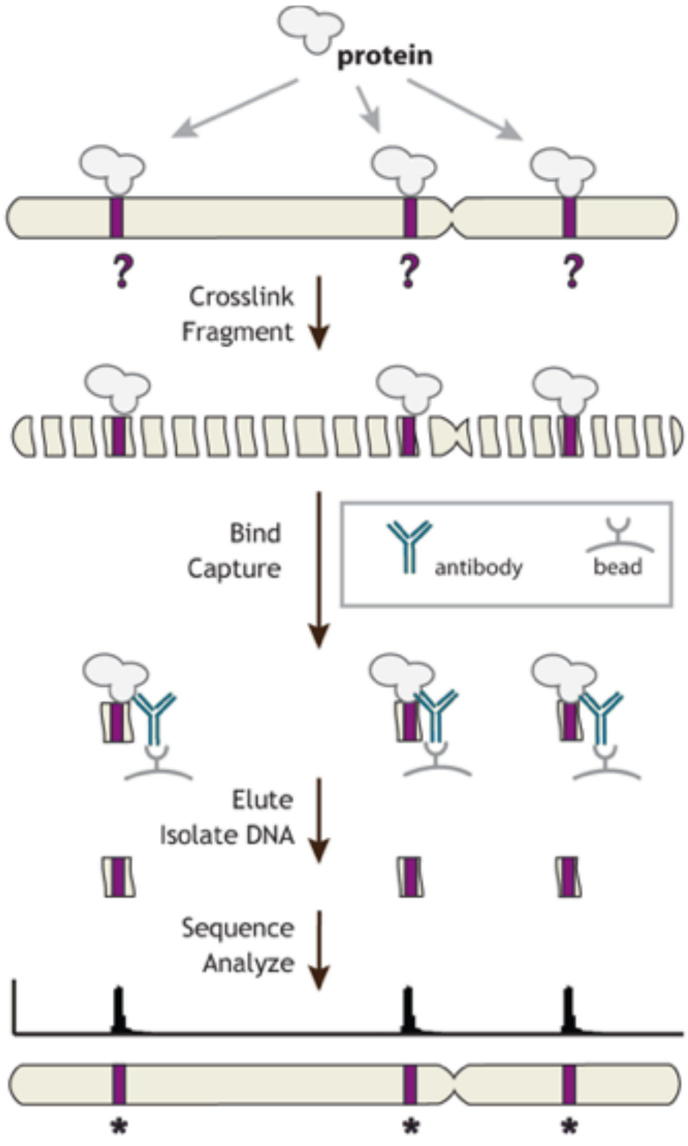




ChIP of CTCF and H3K27me3 in fly development



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.

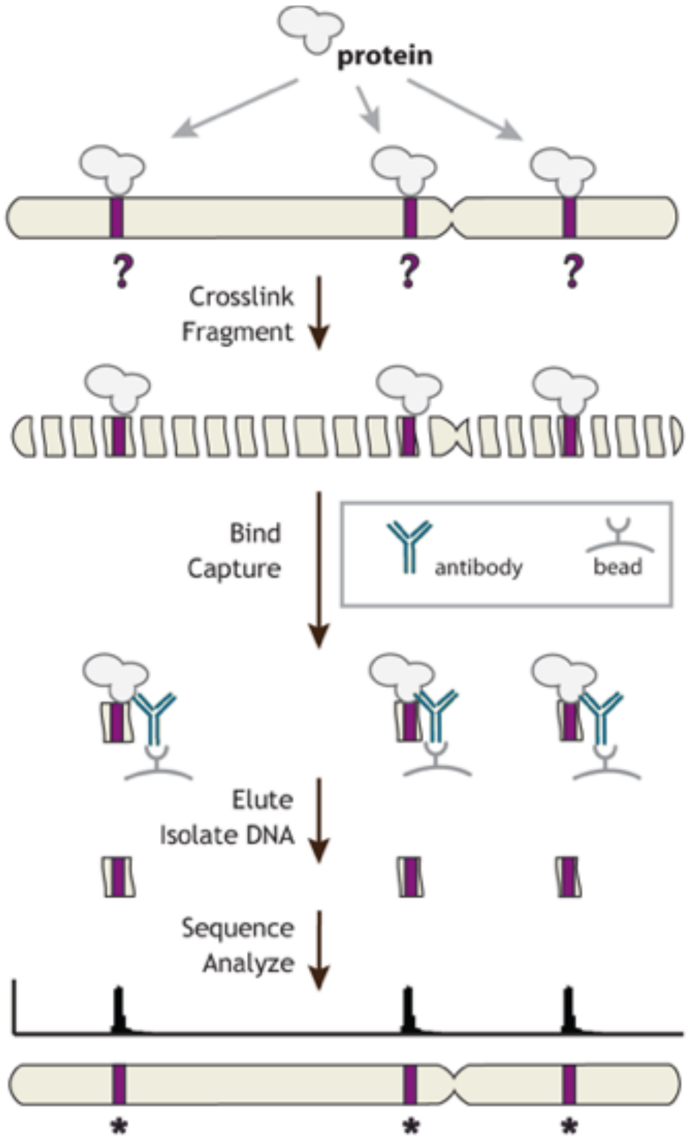
3. Enrichment is dependent on the **accessibility of the epitope.**

   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

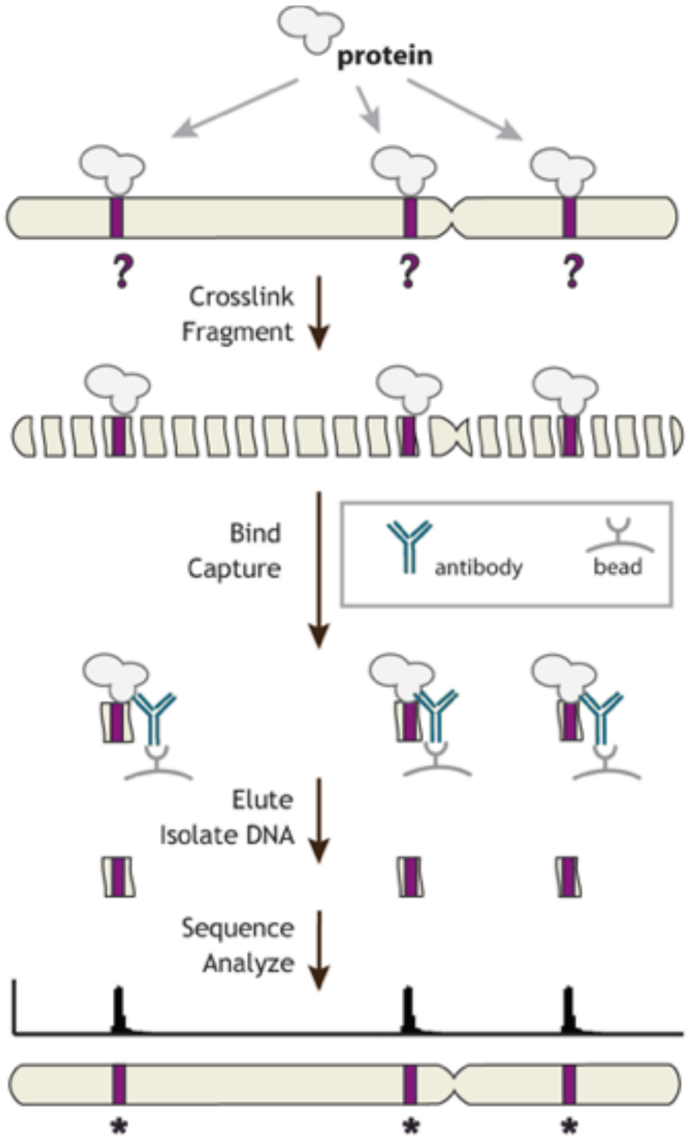


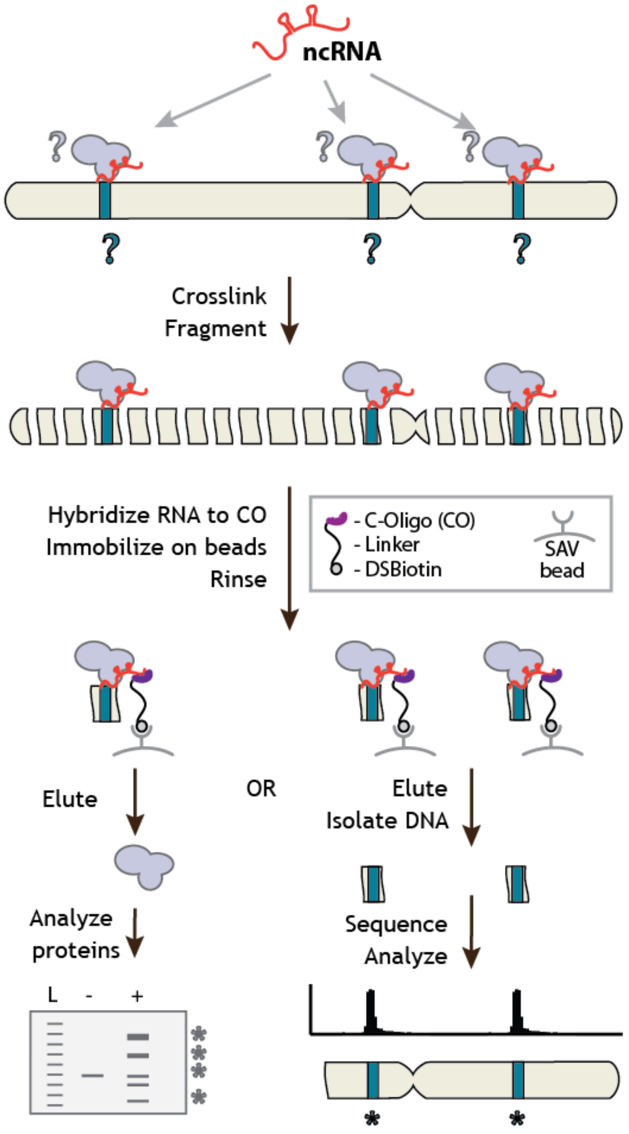
1. Using a nuclease to get very **higher resolution** (ChIP-exo).
.

2. Analysis of **nucleosome turnover** and exchange.

3. Extension to **RNA factors**.

Extension to RNA factors: CHART, ChIRP and RAP



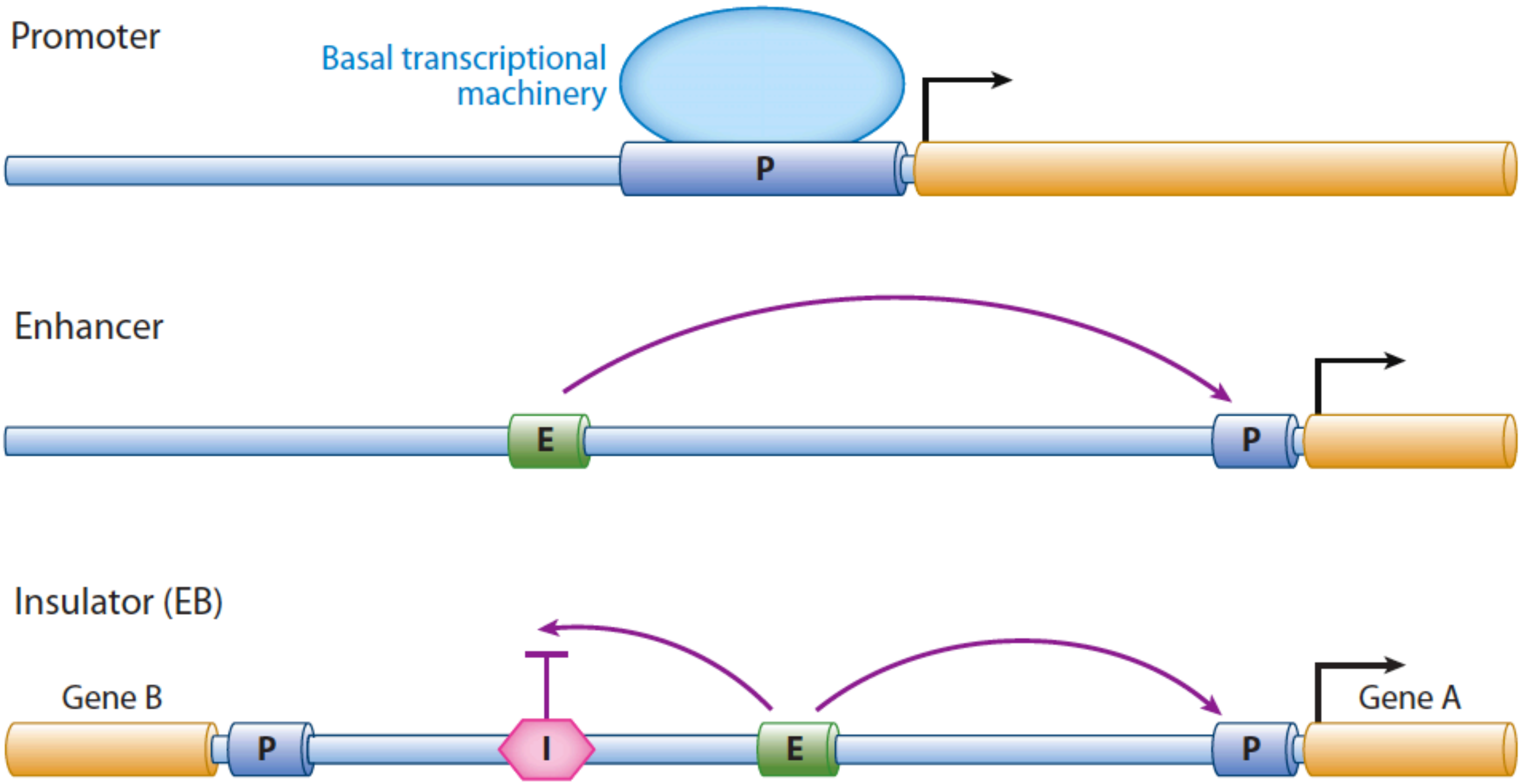


Using sequencing to annotate the genome

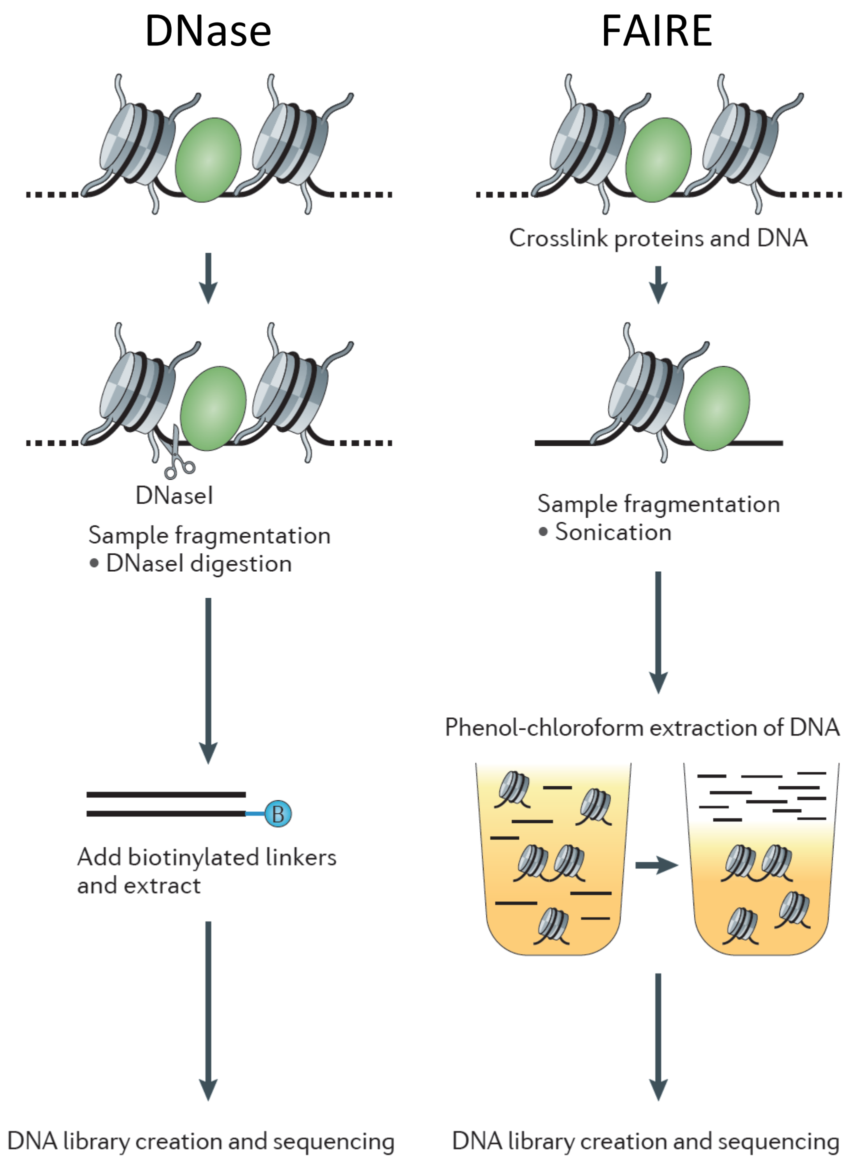
1. Where are the cis-acting regulatory elements in DNA?
   A. DNase I hyper-sensitivity mapping (**DNase-Seq**).
   B. **FAIRE** to map regulatory elements.
2. Where do transcription factors bind?
   C. **ChIP-seq** of transcription factors (or in high res, ChIP-exo)
   D. Nucleosome mapping (**MNase-Seq**).
3. Where are different histone modifications found?
   E. **ChIP-Seq** of histone modifications.
   F. **ChIP-Seq** of chromatin writers, readers and erasers.
4. Where is RNA polymerase transcribing?
   G. **ChIP-Seq** of polymerase.
   H. **GRO-Seq** and **NET-Seq** to measure RNA in the polymerase active site..
5. How is the genome organized in 3D?
   I. **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


DNase
FAIRE
Crosslink proteins and DNA
DNaseI
Sample fragmentation
• DNaseI digestion
Sample fragmentation
• Sonication
Phenol-chloroform extraction of DNA
B
Add biotinylated linkers and extract
DNA library creation and sequencing
DNA library creation and sequencing

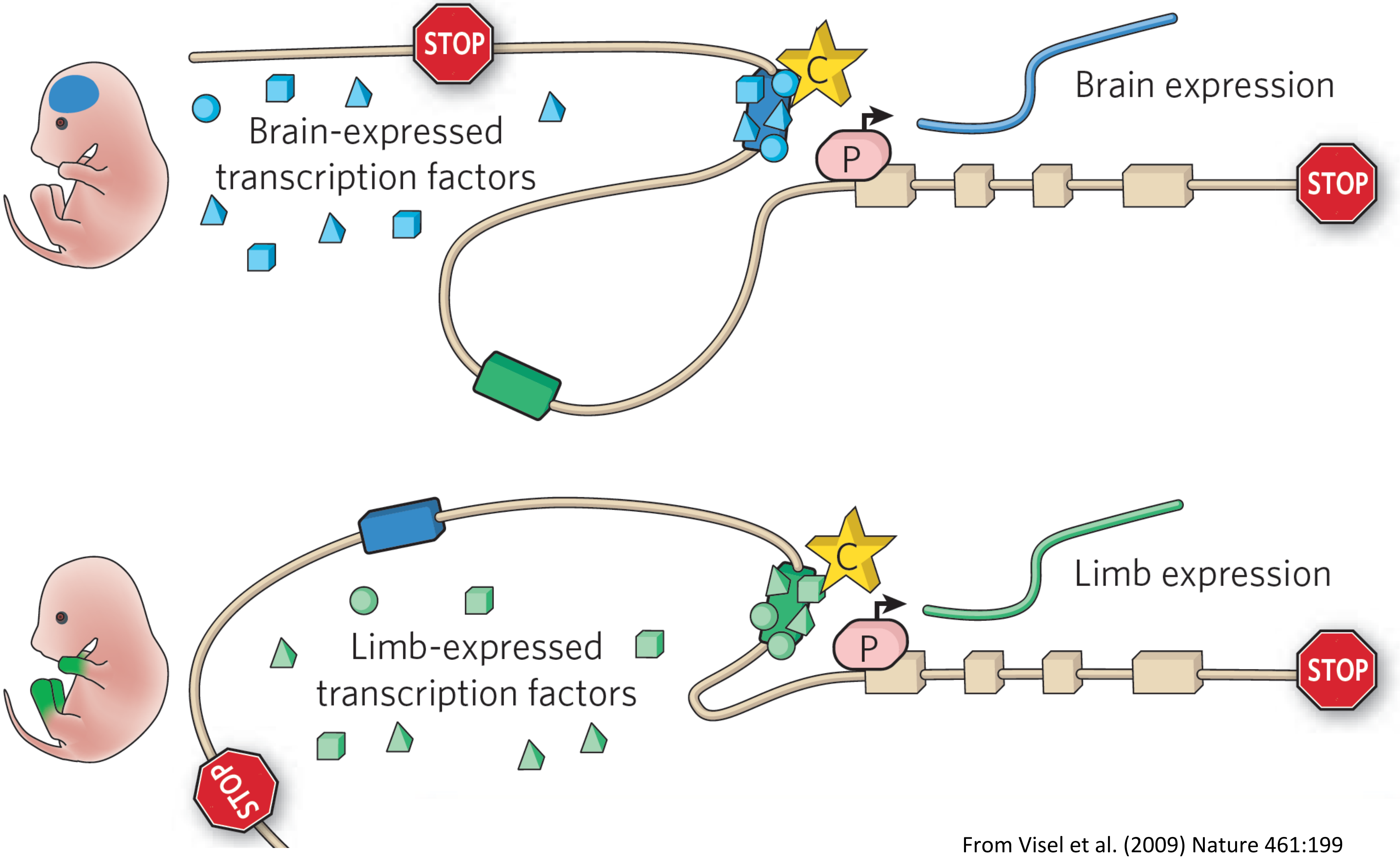
1. **Transcription factor binding** frequently deforms the B-form DNA, making it hypersensitive to DNase I.
.

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

    -FAIRE-seq (shown)

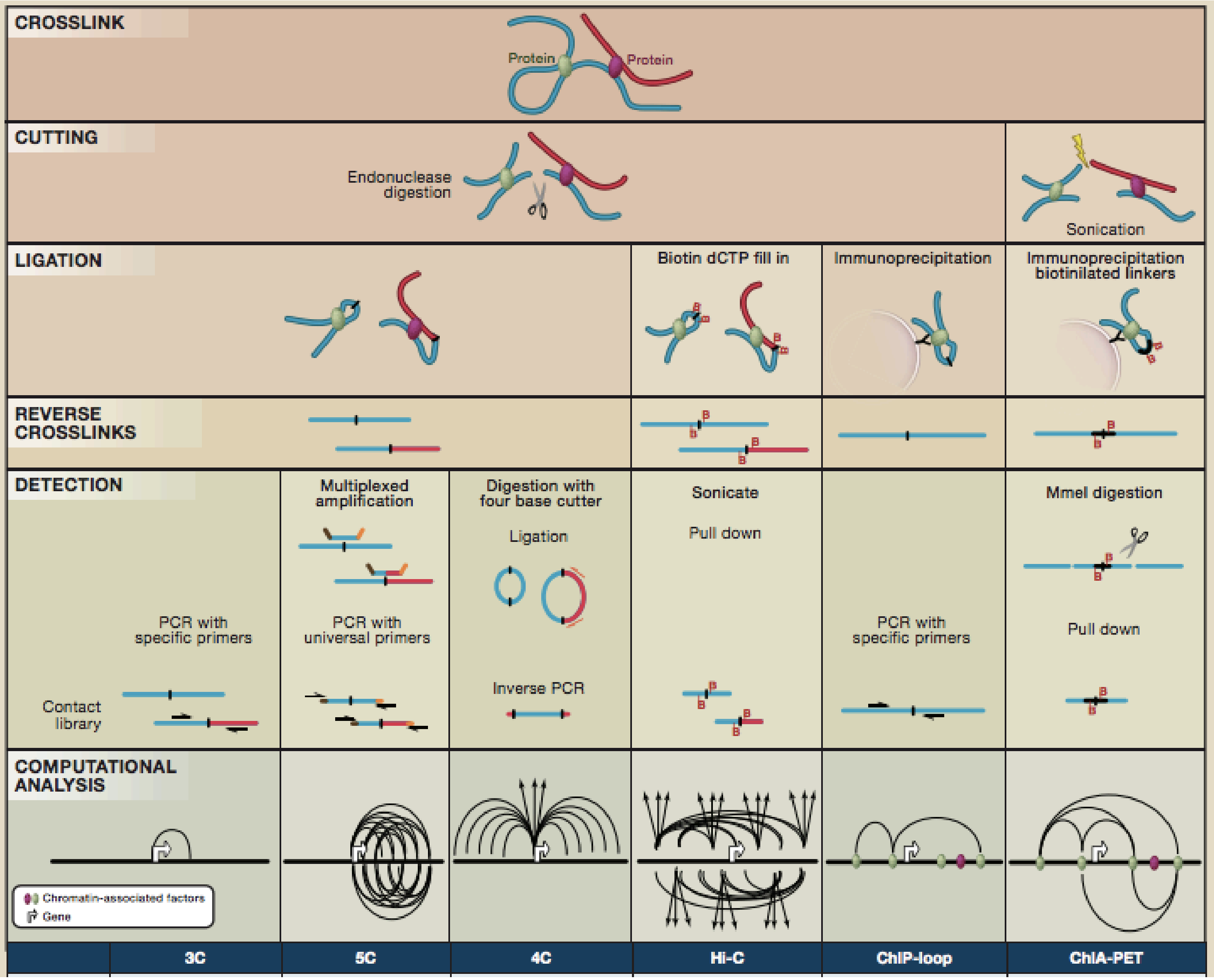
    -MNase-Seq (not shown).

The 3D organization of the genome is important



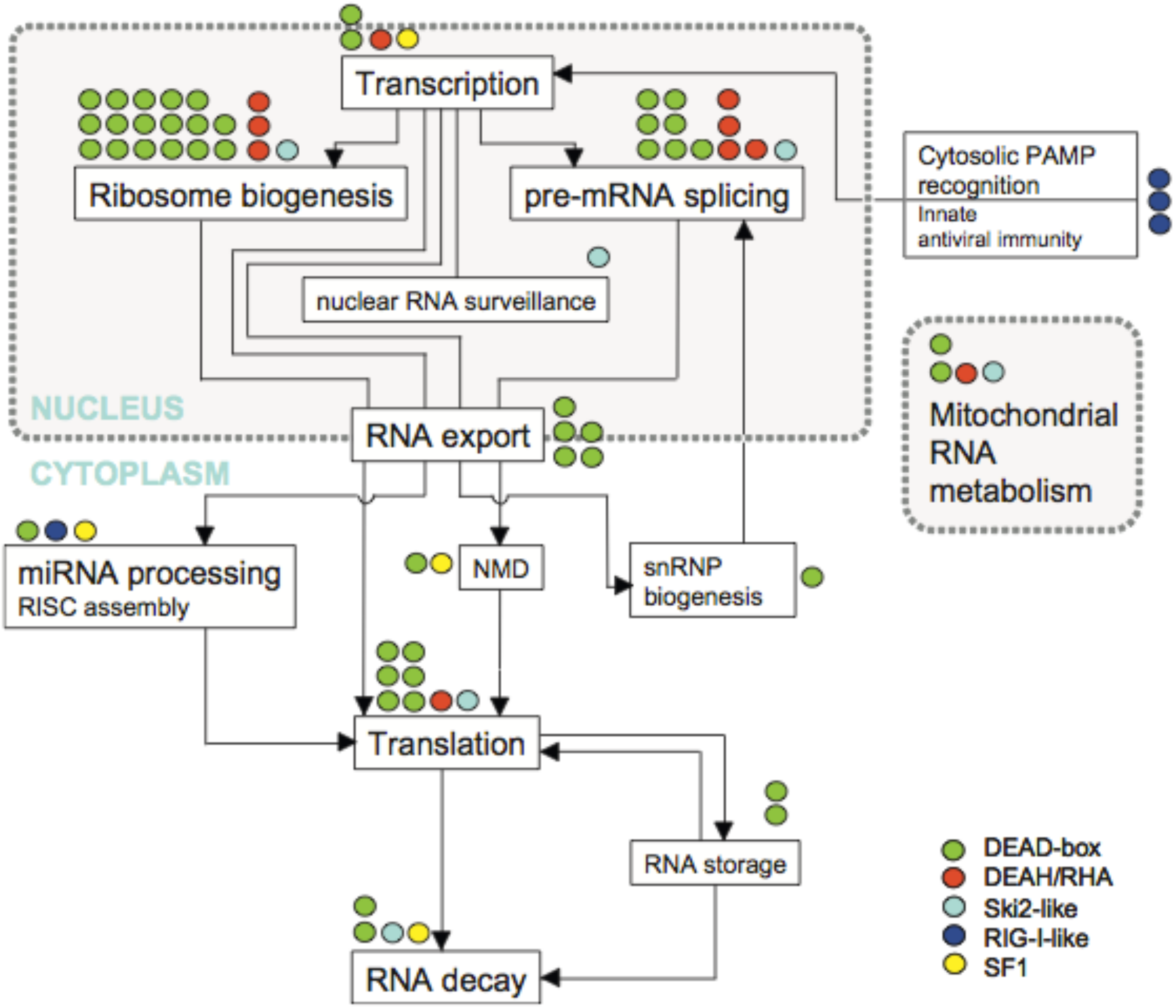
From Visel et al. (2009) Nature 461:199

Techniques to analyze chromatin conformation



Hakim & Misteli, Cell (2012)

Gene expression is also controlled at the level of RNA

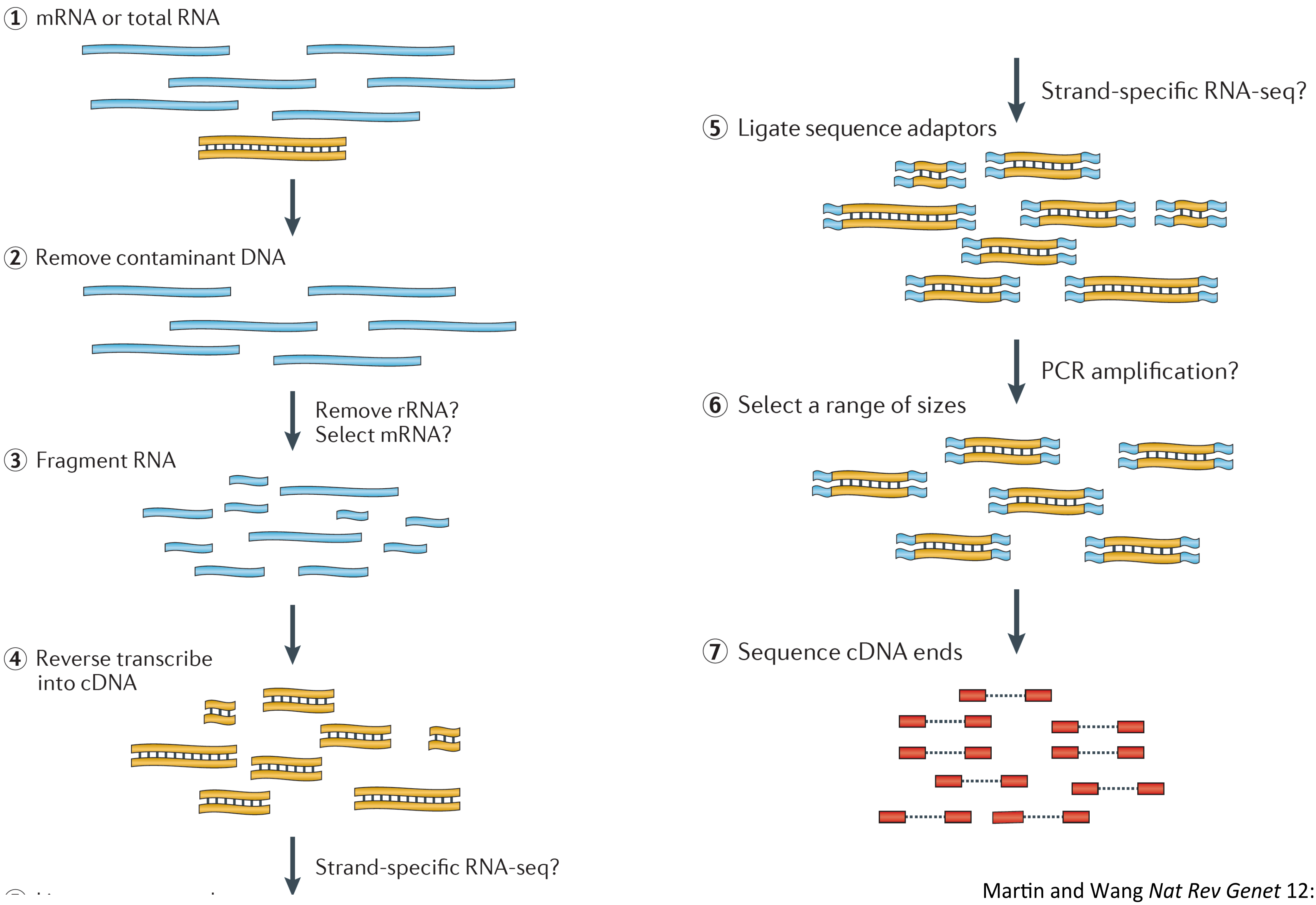


Part 2: RNA-Seq and applications of RNA-Seq

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



① mRNA or total RNA

② Remove contaminant DNA

Remove rRNA?
Select mRNA?

③ Fragment RNA

④ Reverse transcribe
into cDNA

Strand-specific RNA-seq?

Strand-specific RNA-seq?

⑤ Ligate sequence adaptors

PCR amplification?

⑥ Select a range of sizes

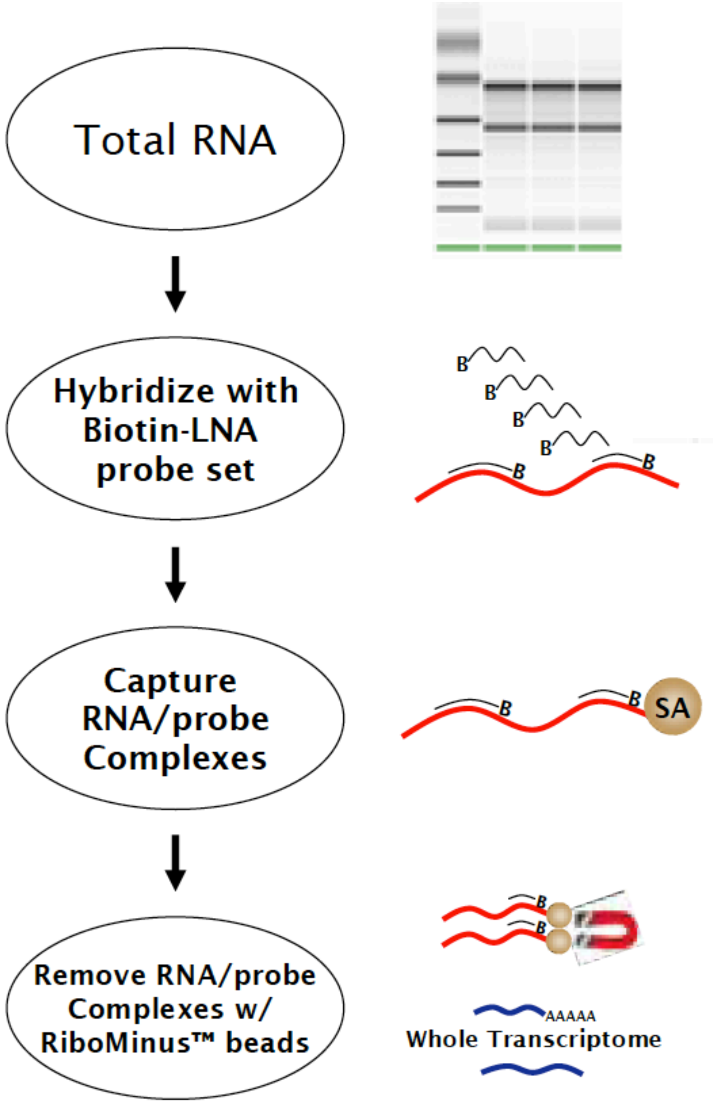
⑦ Sequence cDNA ends

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



Total RNA

Hybridize with Biotin-LNA probe set

Capture RNA/probe Complexes

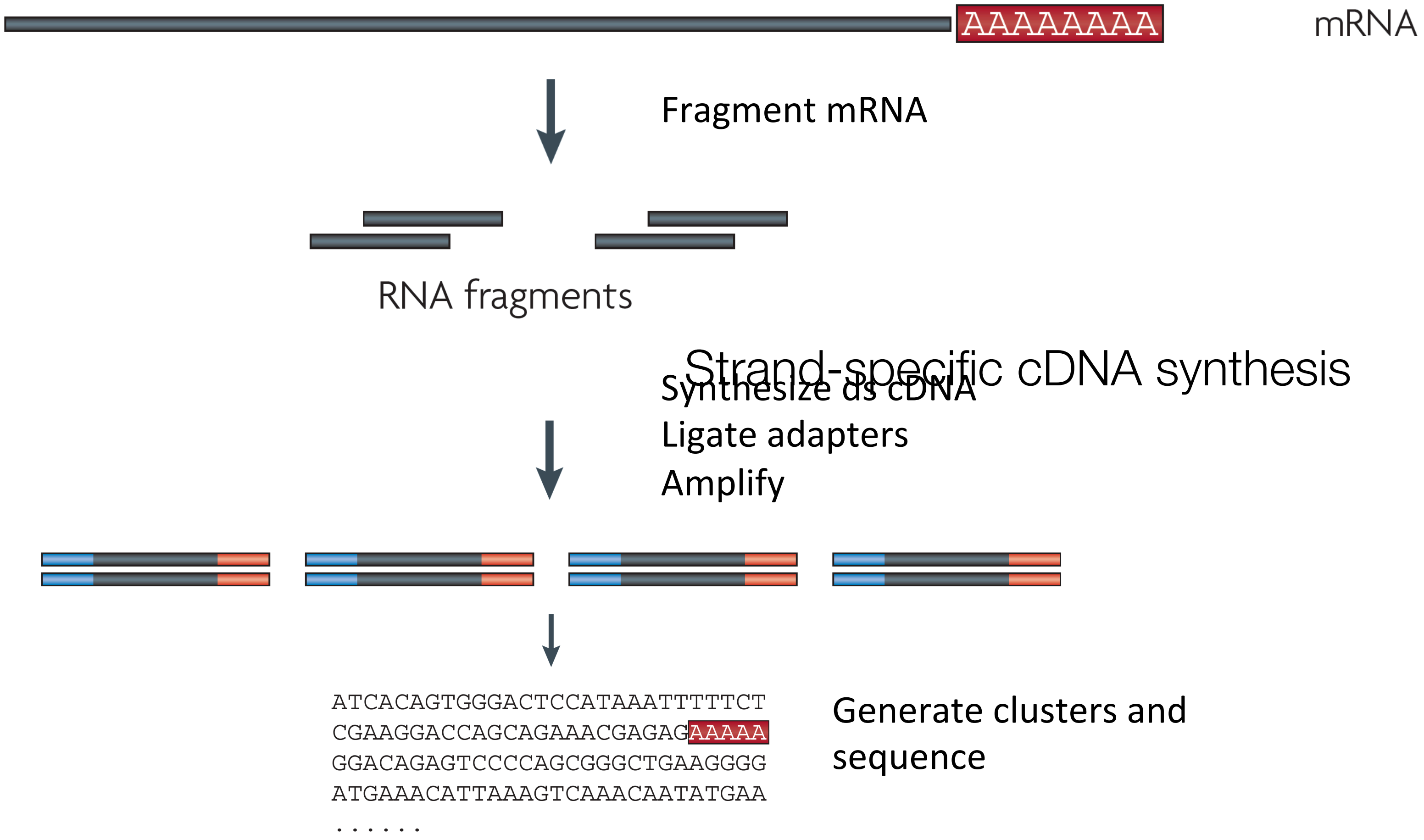
Remove RNA/probe Complexes w/ RiboMinus™ beads

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



Fragment mRNA

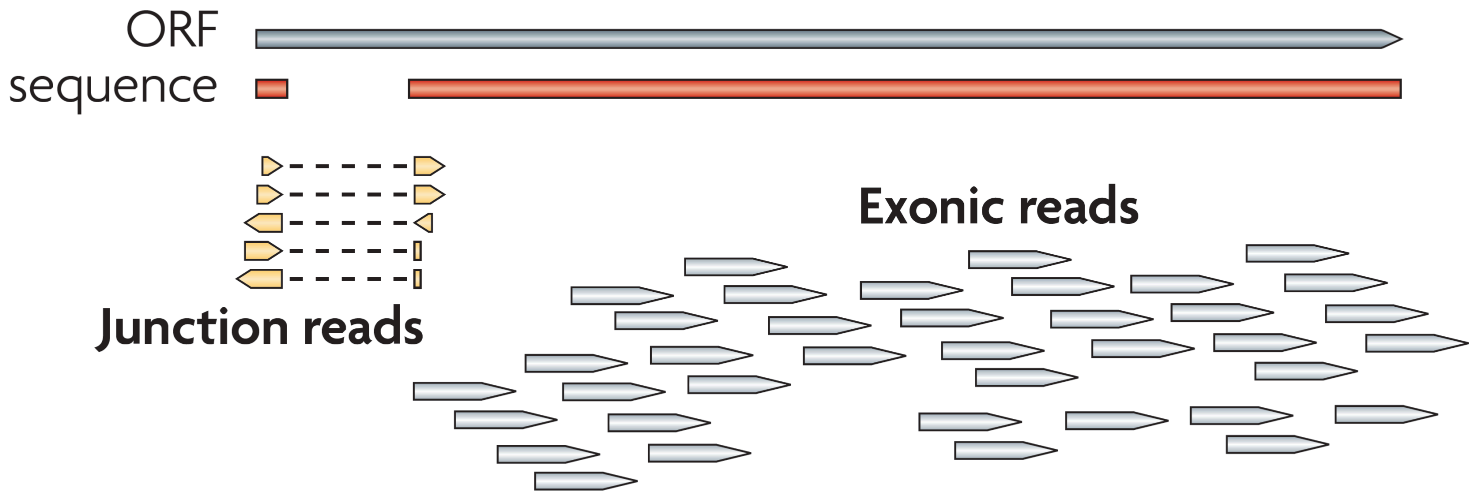
Strand-specific cDNA synthesis

Synthesize ds cDNA
Ligate adapters
Amplify

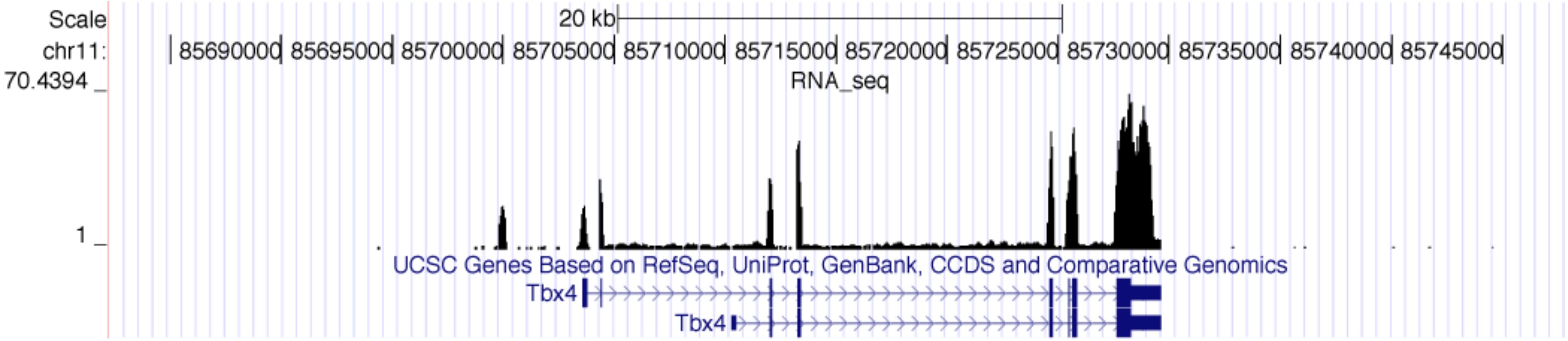
```
ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA
......
```

Generate clusters and sequence

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:

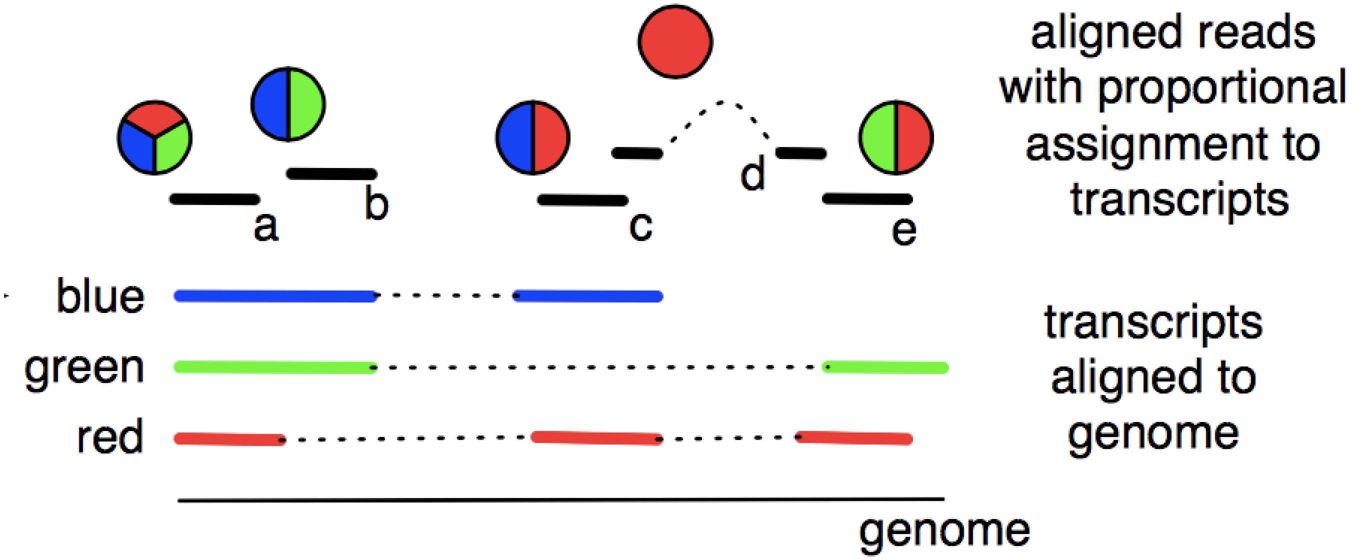
Assemble transcripts directly from reads
Allows transcriptome analyses of species without reference genomes

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

Normalization :

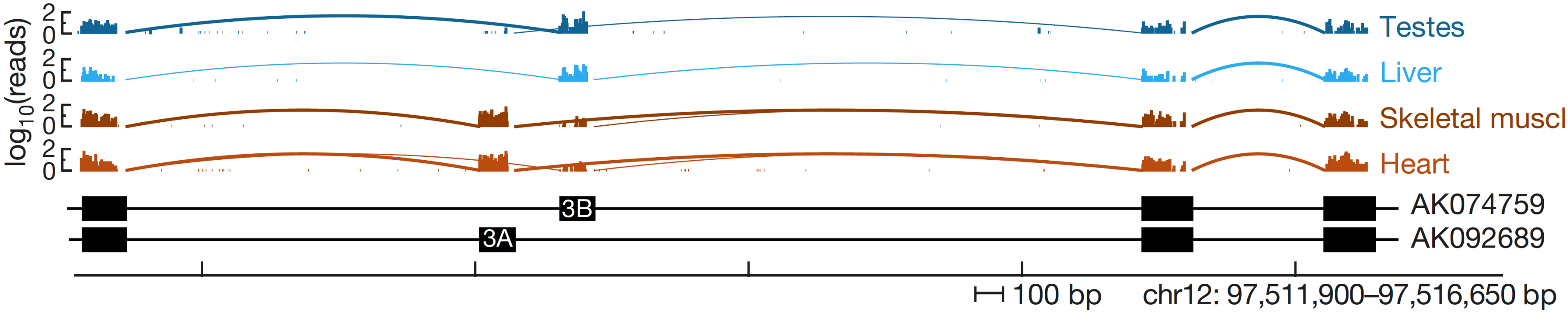
Internal: *Reads or Fragments* per kilobase of feature length per million mapped reads (RPKM or FPKM)

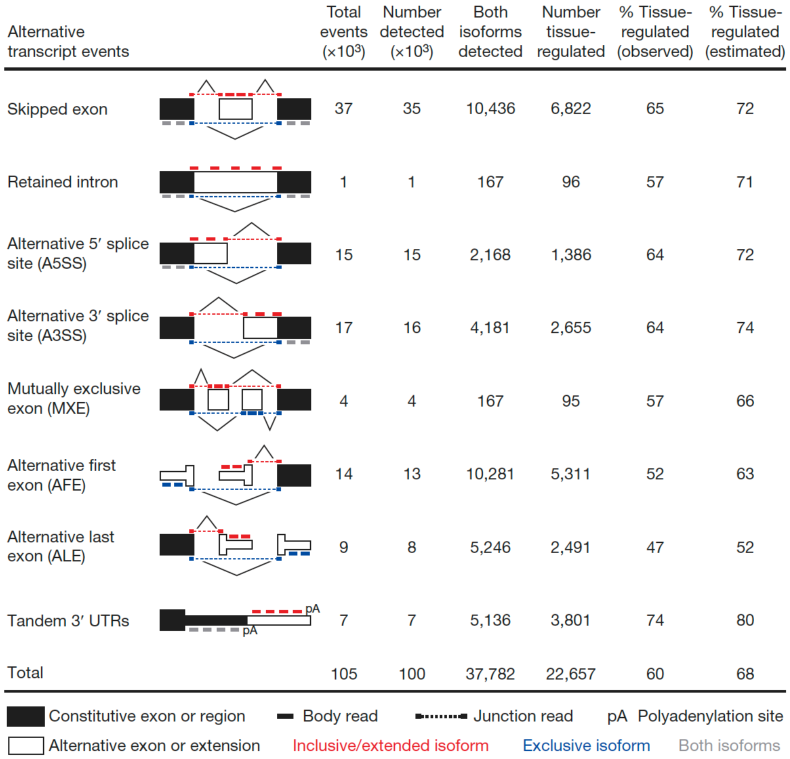
External: Reads relative to a standard “spike”



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

There is a lot of functional diversity in transcript isoforms





Alternative transcript events	Total events (×10^3)	Number detected (×10^3)	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

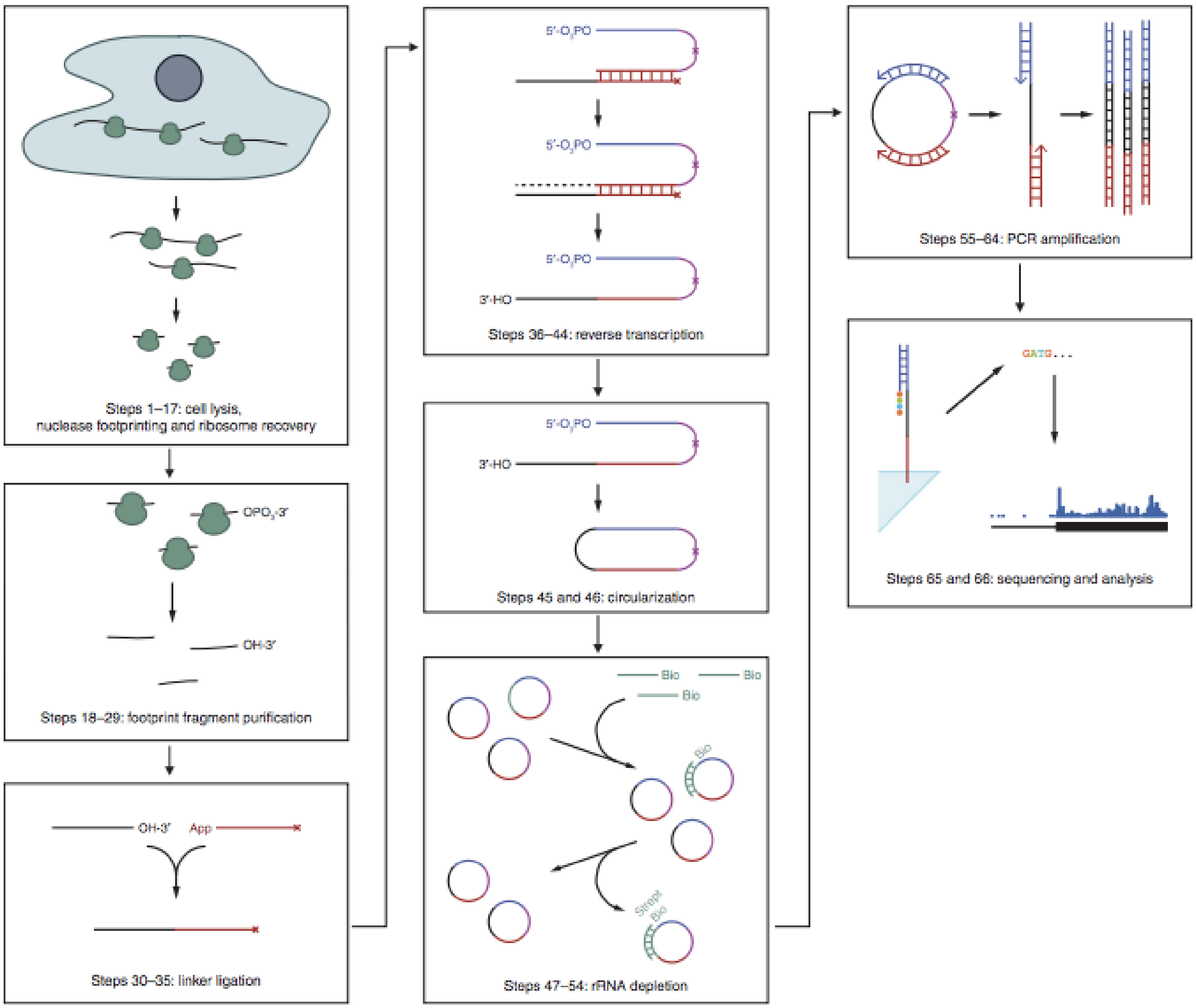
■ Constitutive exon or region — Body read ▪┄┄┄▪ Junction read pA Polyadenylation site

□ 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

Ribosome profiling to reveal translation



Steps 1–17: cell lysis, nuclease footprinting and ribosome recovery

Steps 18–29: footprint fragment purification

Steps 30–35: linker ligation

Steps 36–44: reverse transcription

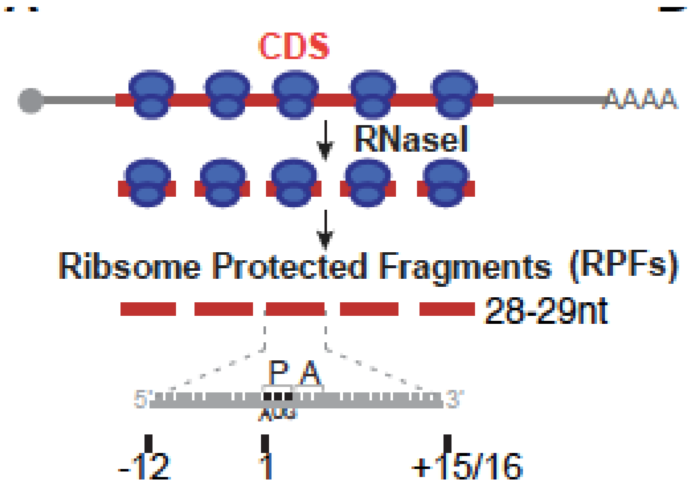
Steps 45 and 46: circularization

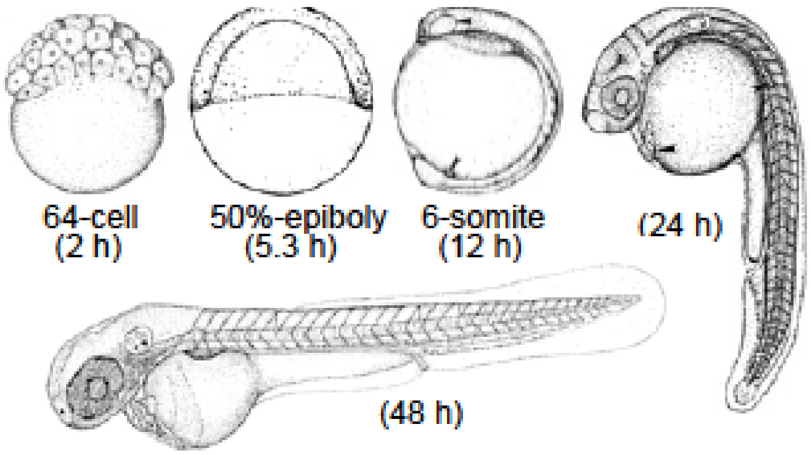
Steps 47–54: rRNA depletion

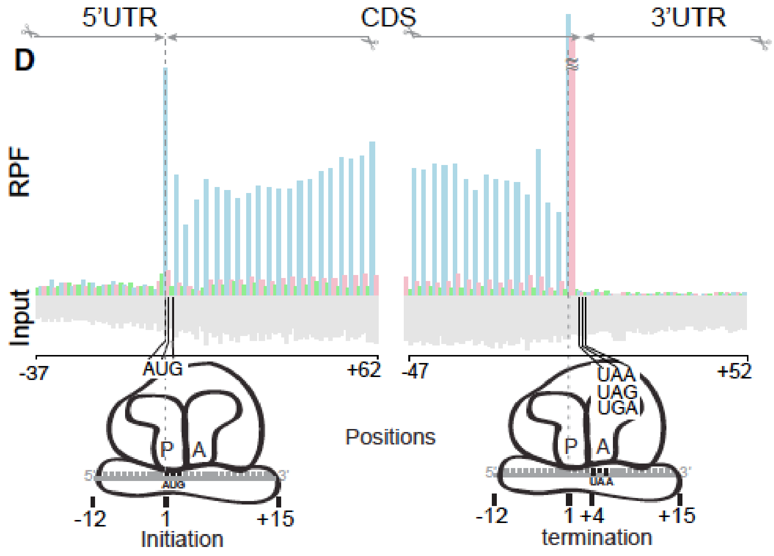
Steps 55–64: PCR amplification

Steps 65 and 66: sequencing and analysis

Ribosome foot printing can reveal which reading frame is translated.





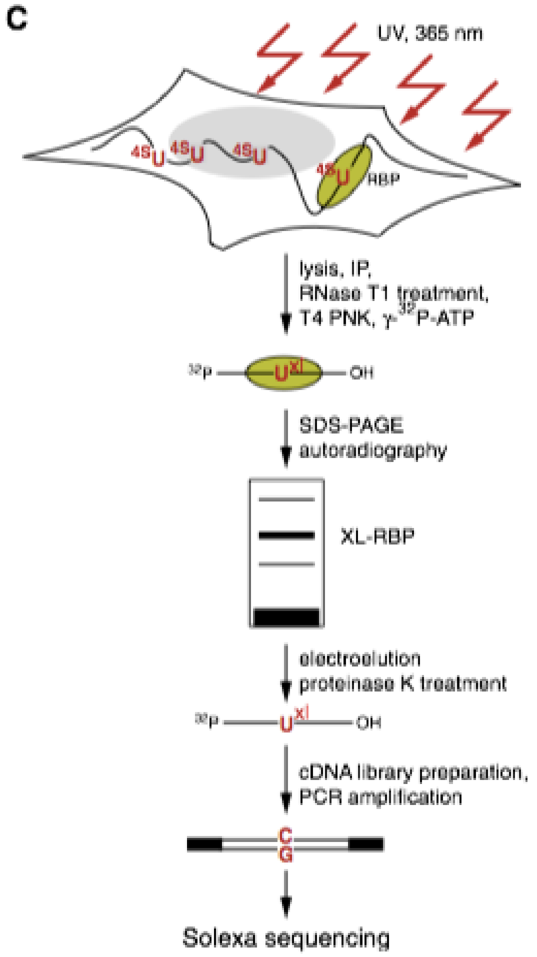


1. Bazzini, A. A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33,** 981–993 (2014).

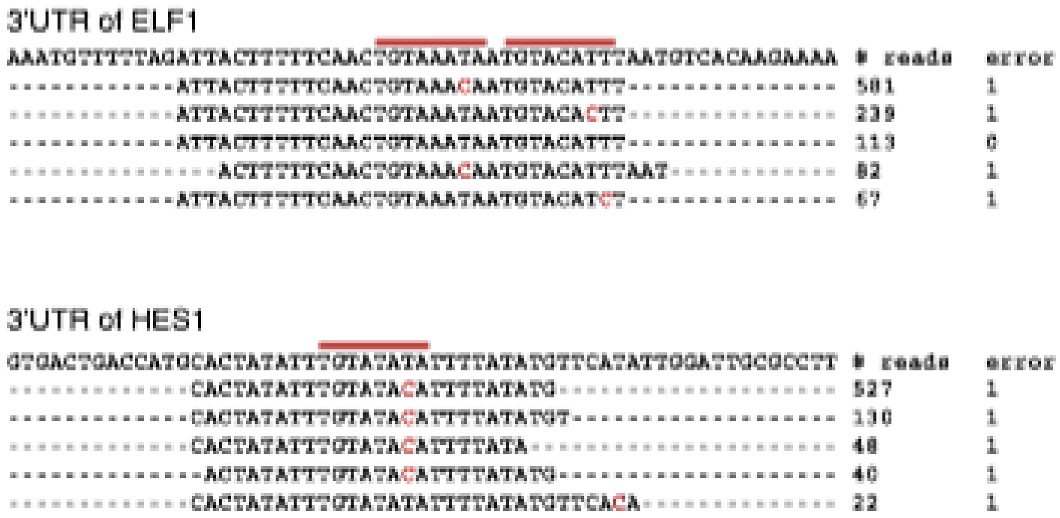
RNA-seq to examine protein-RNA interactions

PAR-CLIP

Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation



There are several methods to look at protein-RNA interactions using RNA-Seq such as RIP, CLIP and similar protocols.



3'UTR of ELF1

```
AAATGTTTTTAGATTACTTTTTCAACTGTAAATAATGTACATTTAATGTCACAAGAAAA  # reads  error
------------ATTACTTTTTCAACTGTAAACAATGTACATTT---------------  501      1
------------ATTACTTTTTCAACTGTAAATAATGTACACTT---------------  239      1
------------ATTACTTTTTCAACTGTAAATAATGTACATTT---------------  113      0
---------------ACTTTTTCAACTGTAAACAATGTACATTTAAT------------  82       1
------------ATTACTTTTTCAACTGTAAATAATGTACATCT---------------  67       1
```

3'UTR of HES1

```
GTGACTGACCATGCACTATATTTGTATATATTTTTATATGTTCATATTGGATTGCGCCTT  # reads  error
-------------CACTATATTTGTATACATTTTATATG---------------------  527      1
-------------CACTATATTTGTATACATTTTATATGT--------------------  130      1
-------------CACTATATTTGTATACATTTTATA-----------------------  48       1
--------------ACTATATTTGTATACATTTTATATG---------------------  40       1
-------------CACTATATTTGTATATATTTTTATATGTTCACA--------------  22       1
```

1. Hafner, M. *et al.* Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *CELL* **141,** 129–141 (2010).

Using sequencing to study _______________.

(noun)