Bioinformatics Lab

Ivan Gesteira Costa & Zhijian Li
Institute for Computational Genomics

Institute for Computational Genomics
RWTH AACHEN UNIVERSITY

Objectives

- Hands on introduction to bioinformatics programing
- Review basic biological/computational aspects
    1. basics of molecular biology
    2. basics of sequencing
    3. basics bioinformatics problems
        - short sequences read alignment
        - differential peak calling
        - motif detection
        - footprint detection

Objectives

- **Introduction to Bioinformatics Frameworks/Tools**

  1. **biological sequence data formats/handling**
     - **Biopython, Pysam**
  2. **bioinformatics tools**
     - **BWA (aligner), MACS/ODIN(Peak caller), RGT/ bedtools (interval algebra), regulatory genomics toolbox (RGT)**

Grading/Online material

Evaluation:

- **20% prototypes**
- **60% final project**
- **20% presentation**

Extra-work (anyone from media informatics?):

- **research report**

References/Courses Online

http://costalab.org/teaching/bioinformatics-software-lab-2018/

Introduction to Molecular Biology

Institute for Computational Genomics
RWTH AACHEN UNIVERSITY

Understanding Live in a Molecular Level

How is genetic information inherited?

How the genetic information influence cellular processes?

How genes work together to promote particular molecular functions?

Genetic Information - DNA

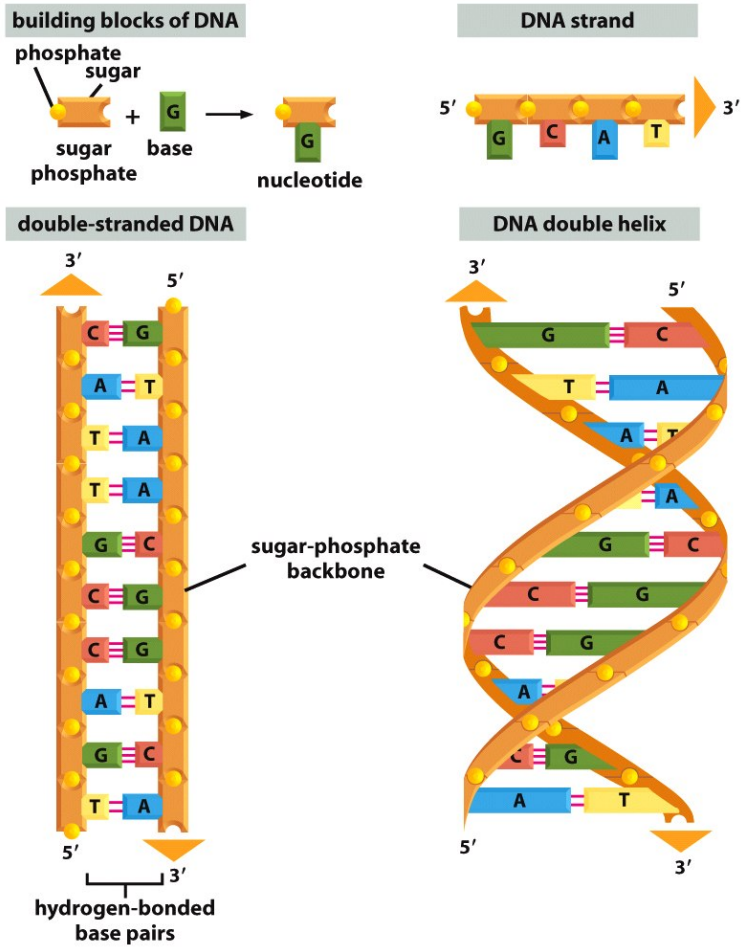
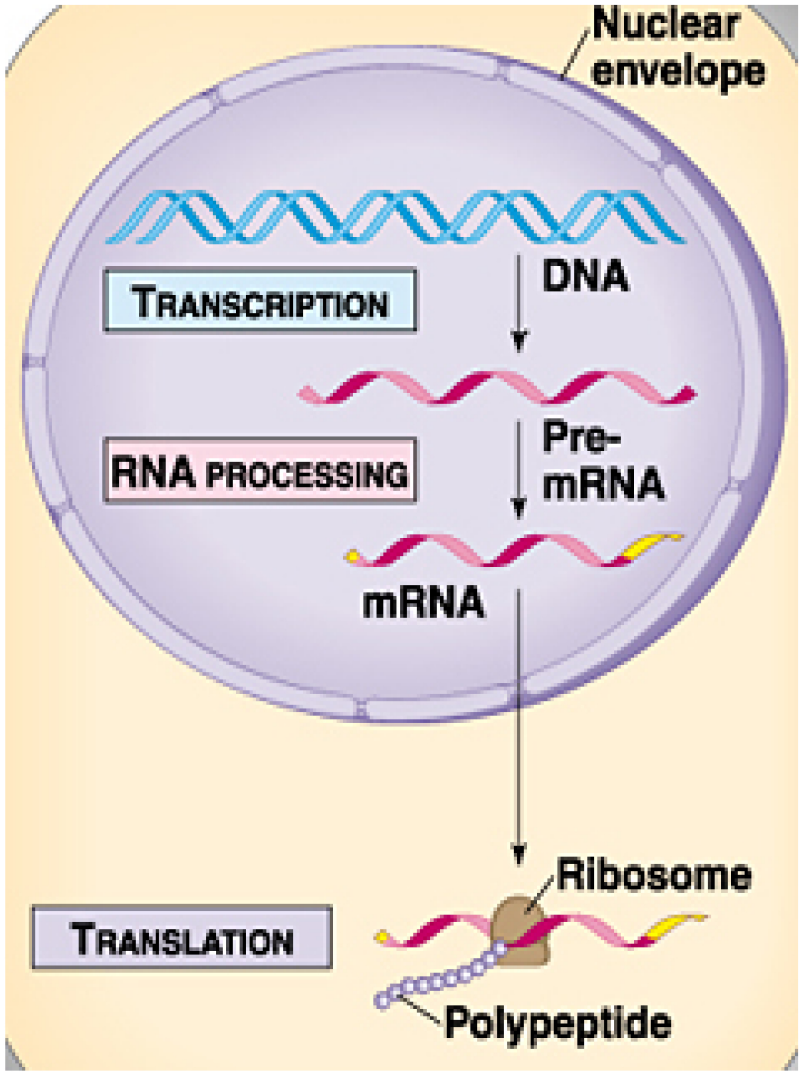


Figure 4-3 Molecular Biology of the Cell 5/e (© Garland Science 2008)

DNA *(Deoxyribonucleic)*

- chain of nucleic acids
- 4 bases: A;C;G;T
- forms DNA duplexes with paring A = T e C = G

Central Dogma - Transcription



Transcription

- ***DNA to RNA***

RNA (ribonucleic acid)

- **single stranded**
- **4 bases: A;C;G;U**
- **unstable**
- **transport of information from nucleus to cytoplasm**

Central Dogma - Transcription

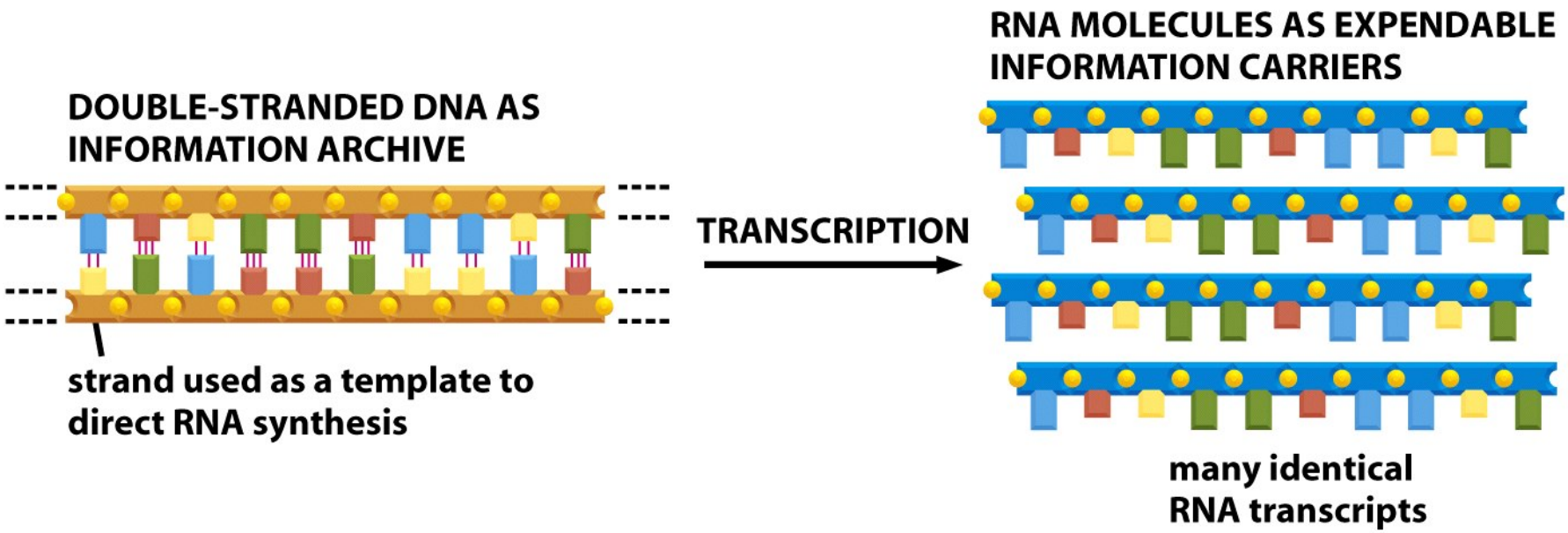
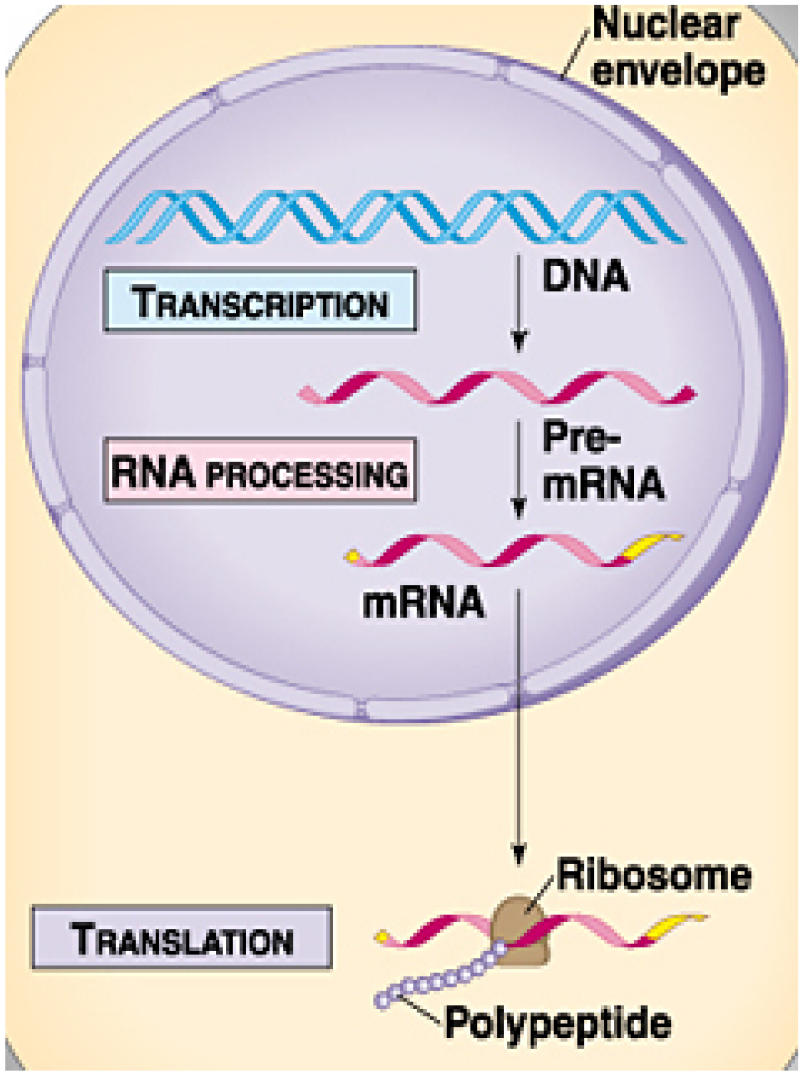


Figure 1-5 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Transcription - copy of DNA information to RNA (T to U)

Central Dogma - Translation



Translation

- *RNA to Protein*
- performed by the ribosome
- follows the genetic code

Proteins

- single stranded chain
- 20 amino acids
- assumes 3D structure
- main functional entities in the cell

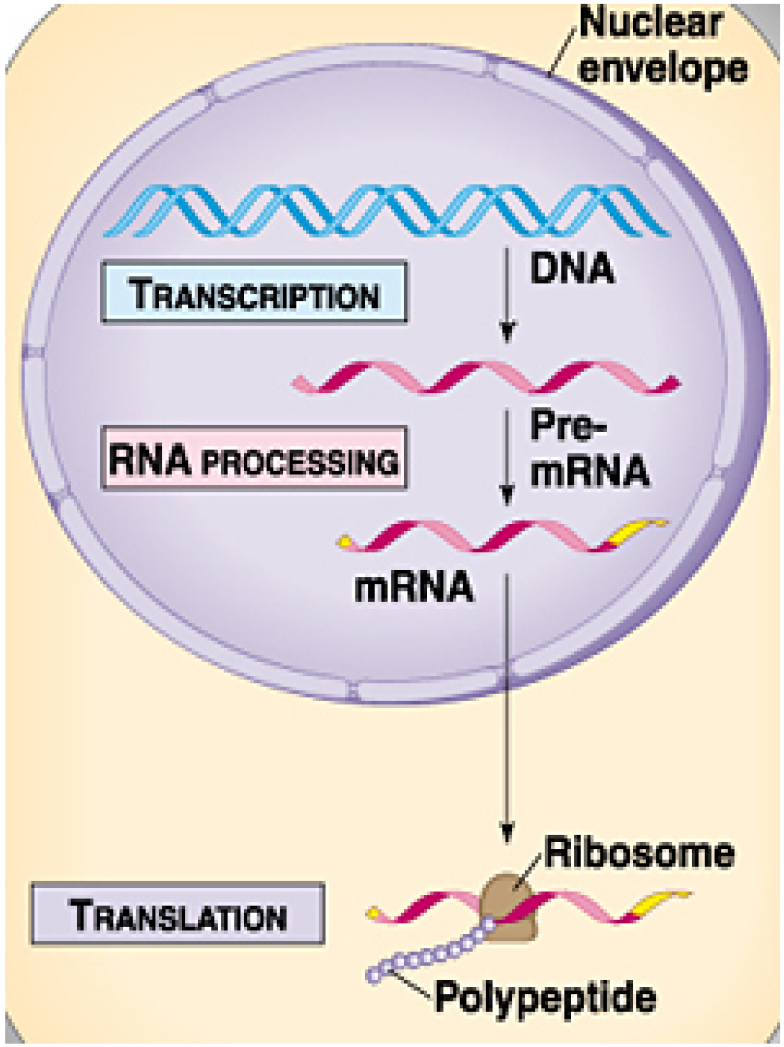
Genetic Code - Translation

GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

Figure 6-50 Molecular Biology of the Cell 5/e (© Garland Science 2008)

triples of RNA bases encodes a amino acid

Central Dogma



- Dogma: information flux DNA -> mRNA -> Proteins
- Gene: DNA segment coding a protein.
- Transcript: RNA segment associated to a gene.
- Genes is associated to one proteins and one function*

* Genes might be associated to many proteins

Control of Gene Expression

How is the expression of genes controlled?

Certain proteins (transcription factors) bind to DNA and initiate transcription

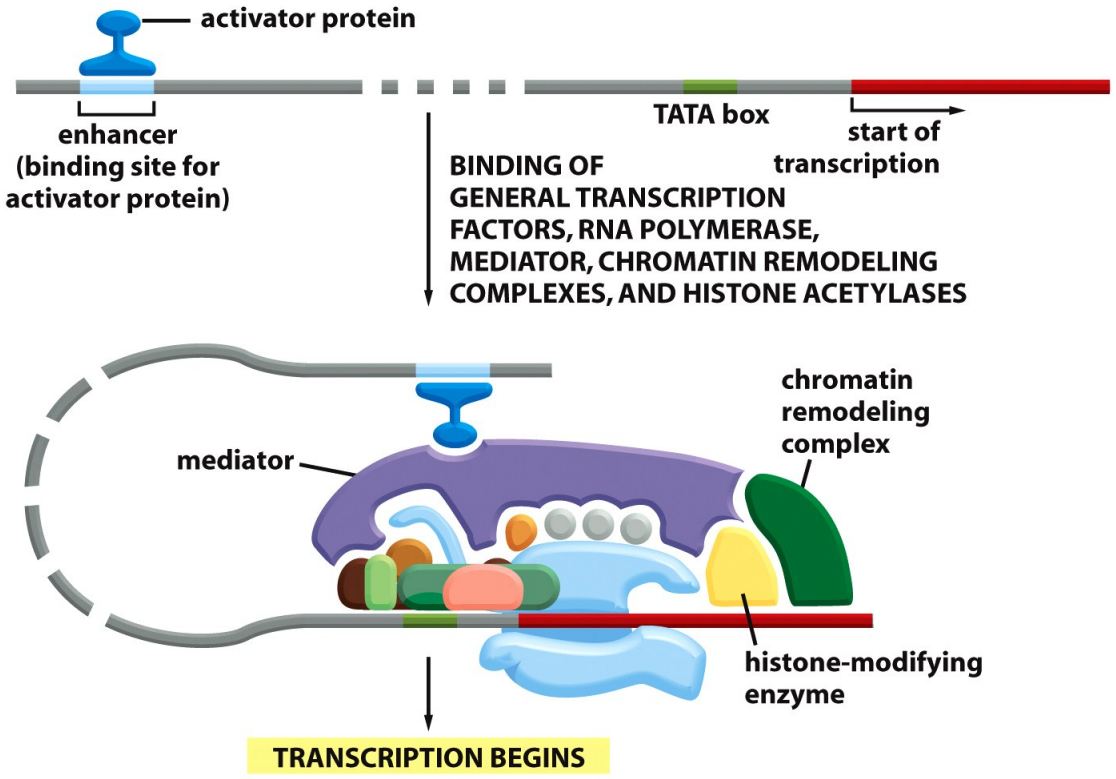


Figure 6-19 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Gene Expression

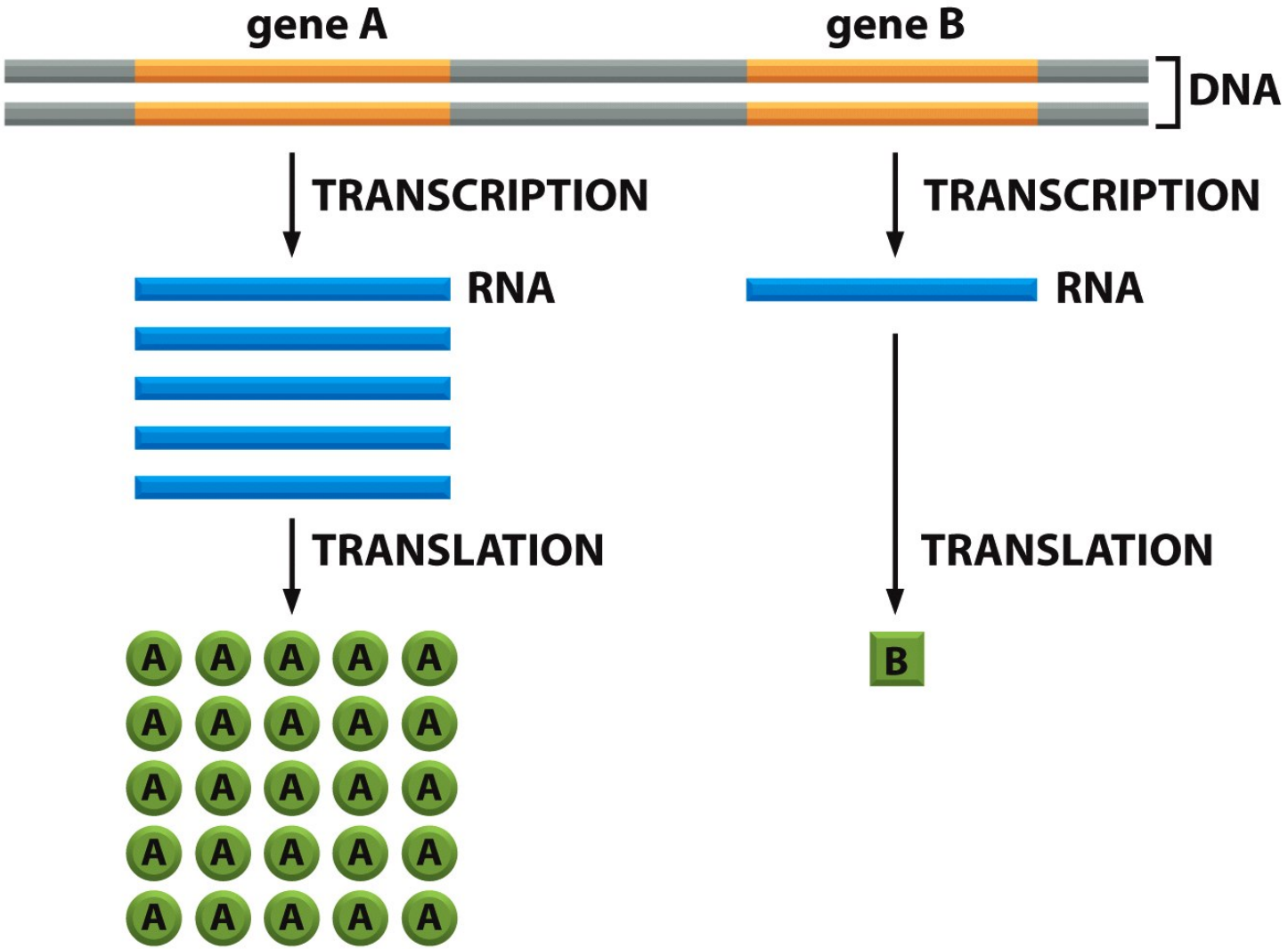
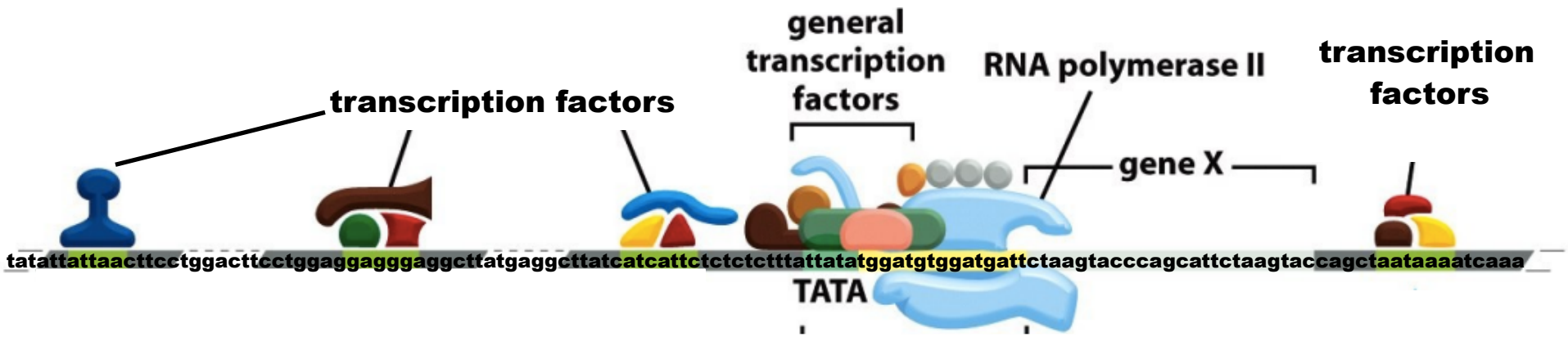


Figure 6-3 Molecular Biology of the Cell 5/e (© Garland Science 2008)

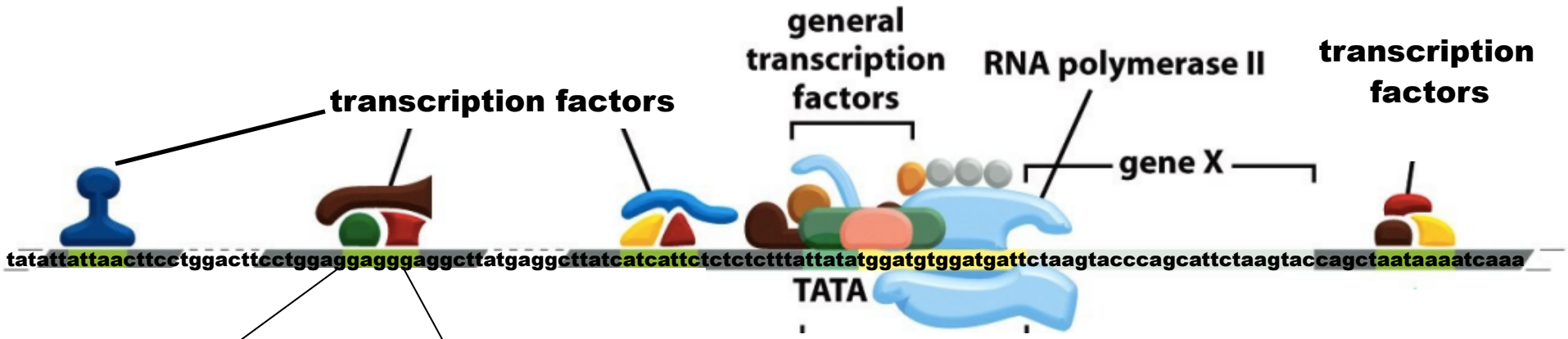
Gene Regulation / Motif Search

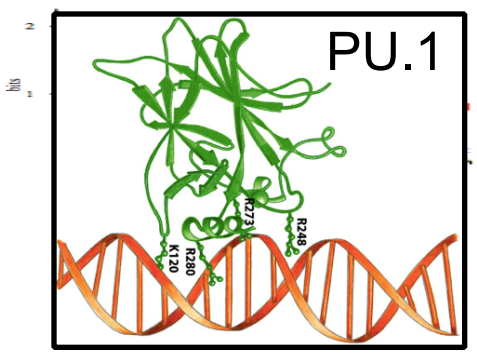
Regulatory Control – Protein-DNA interaction



Source: Alberts, B. et al. (2008) Garland Science, 5th ed.

Motif Search – Computational Approach





Problems

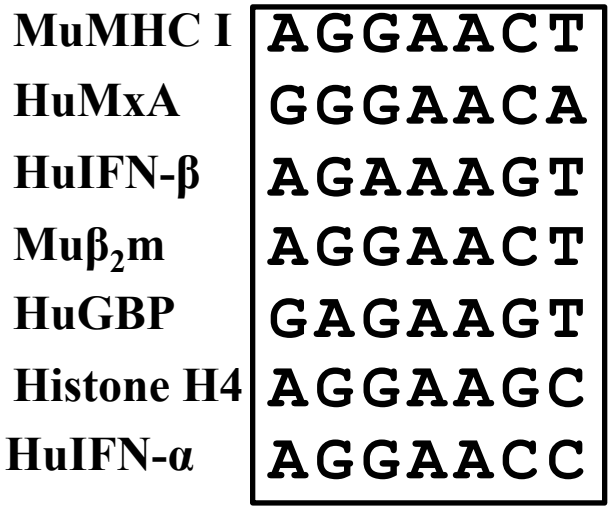
- motifs are small and degenerate (6-8 conserved positions)
- distal binding sites (>10^6 from genes)
- only part of motifs are known

Too many false positive predictions!

Source: Alberts, B. et al. (2008) Garland Science, 5th ed.

Model for DNA-protein binding

PU.1 binding sites ⟶

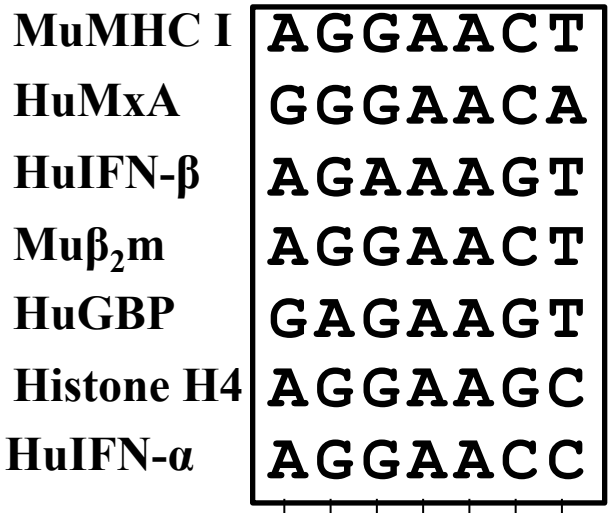


Kanno, Y. et al. (2005) Immune Cell-Specific Amplification of Interferon Signaling by the IRF-4/8-PU.1 Complex.

Model for DNA-protein binding

PU.1 binding sites →

Kanno, Y. et al. (2005) Immune Cell-Specific Amplification of Interferon Signaling by the IRF-4/8-PU.1 Complex.



PU.1 Position Weight Matrix (PWM) →

A	5	1	1	7	7	3	1
C	0	0	0	0	0	0	2
G	2	6	6	0	0	4	0
T	0	0	0	0	0	0	4

Model for DNA-protein binding

PU.1 binding sites →

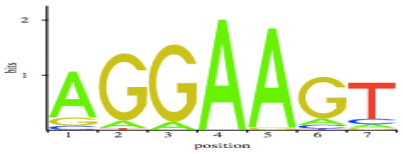
Kanno, Y. et al. (2005) Immune Cell-Specific Amplification of Interferon Signaling by the IRF-4/8-PU.1 Complex.

MuMHC I	AGGAACT
HuMxA	GGGAACA
HuIFN-β	AGAAAGT
Muβ_2m	AGGAACT
HuGBP	GAGAAGT
Histone H4	AGGAAGC
HuIFN-α	AGGAACC

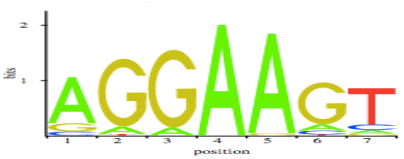
PU.1 Position Weight Matrix (PWM) →

A	5	1	1	7	7	3	1
C	0	0	0	0	0	0	2
G	2	6	6	0	0	4	0
T	0	0	0	0	0	0	4

PU.1 Logo →

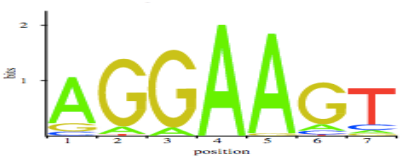


Motif Search

PU.1 PWM 

Genome `TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA`

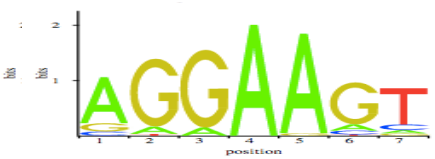
Motif Search

PU.1 PWM 

Genome **TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA**

Score 10.06

Motif Search

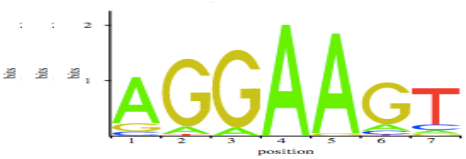
PU.1 PWM 

Genome TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA

Score 10.06

4.81

Motif Search

PU.1 PWM 

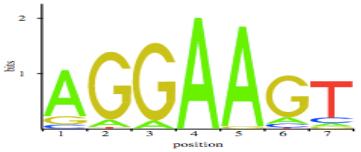
Genome **TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA**

Score 10.06 | 3.19

4.81

Motif Search

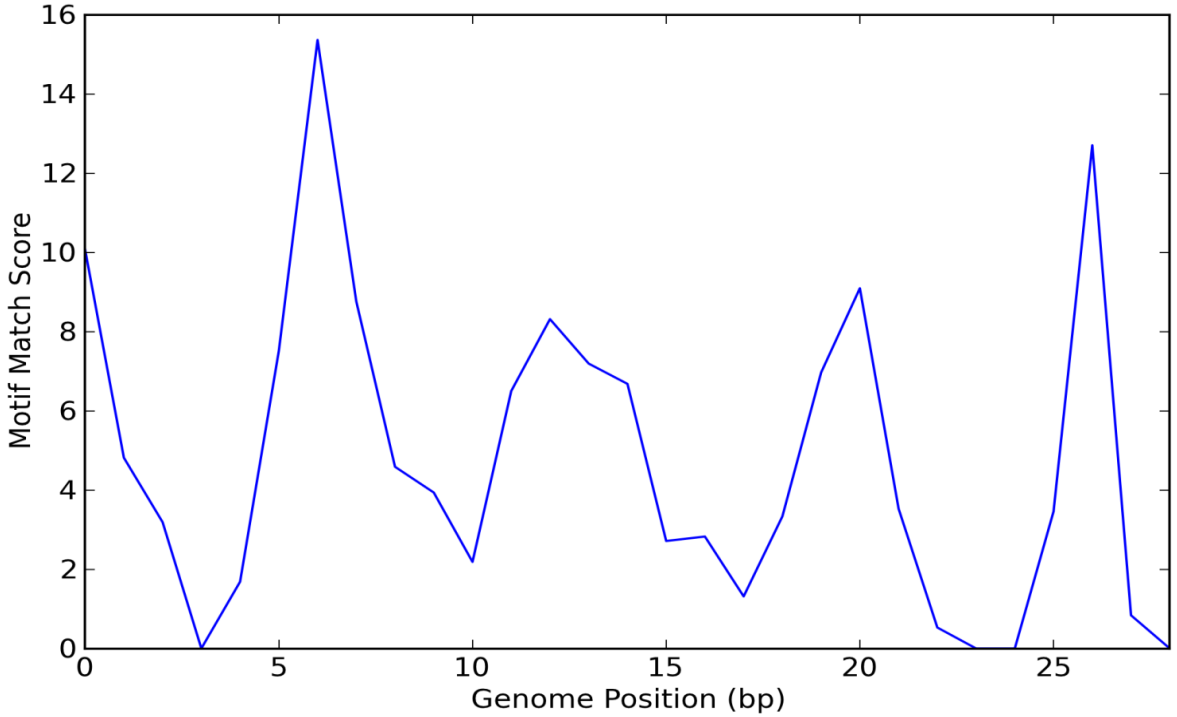
PU.1 PWM



Genome **TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA**

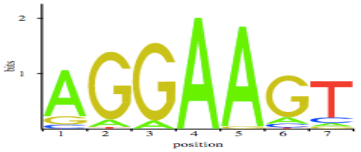
Score 10.06 | 3.19

4.81 . . .



Motif Search

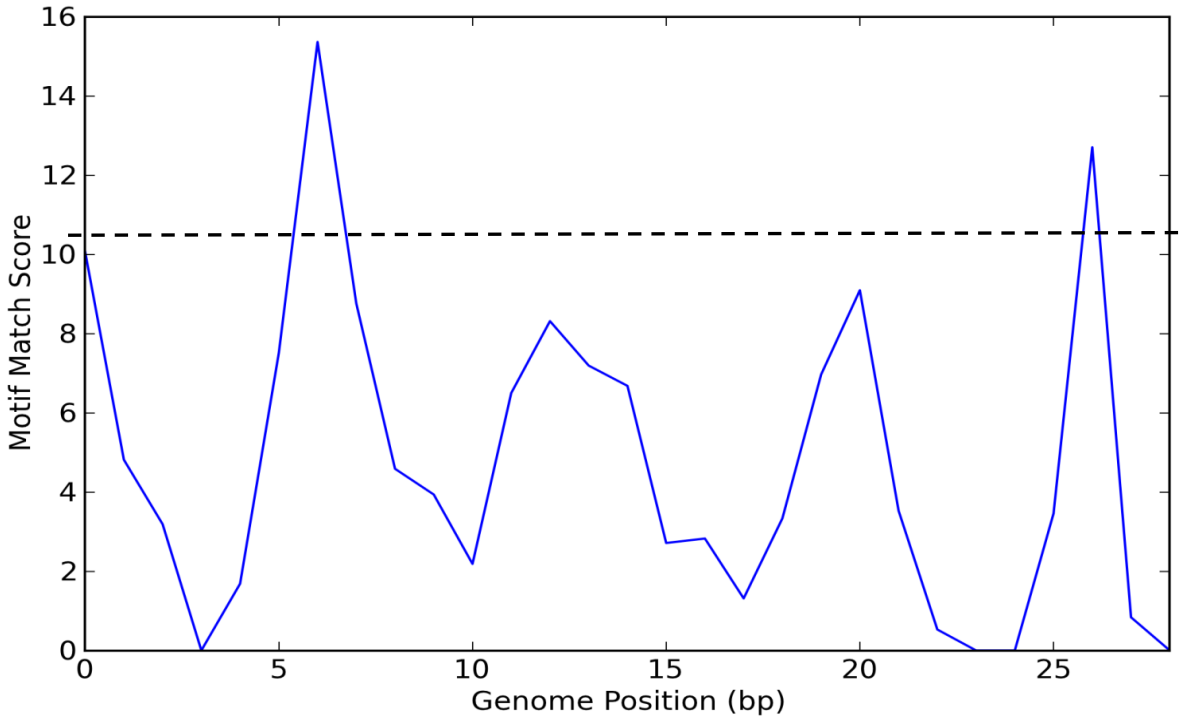
PU.1 PWM



Genome TATCTTTGGAAGTGAAACTACTATCCTGAAAGTCGAA

Score 10.06 | 3.19

4.81 . . .



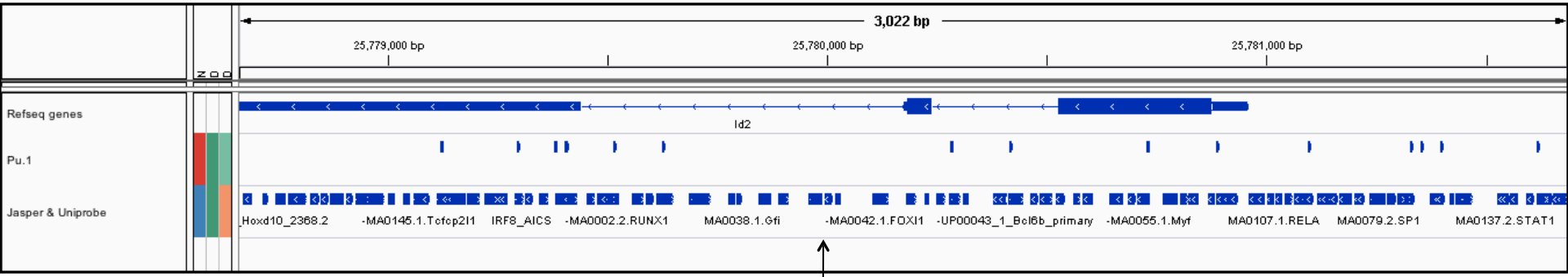
Statistical Test

FDR = 1×10^{-4}

FDR- False Discovery Rate

Example: Binding sites in ID2

Motif search for binding sites with 536 PWMs (Jaspar & Uniprobe) and FDR=0,01



> 3000 predicted binding sites

Epigenetics

Cellular Complexity

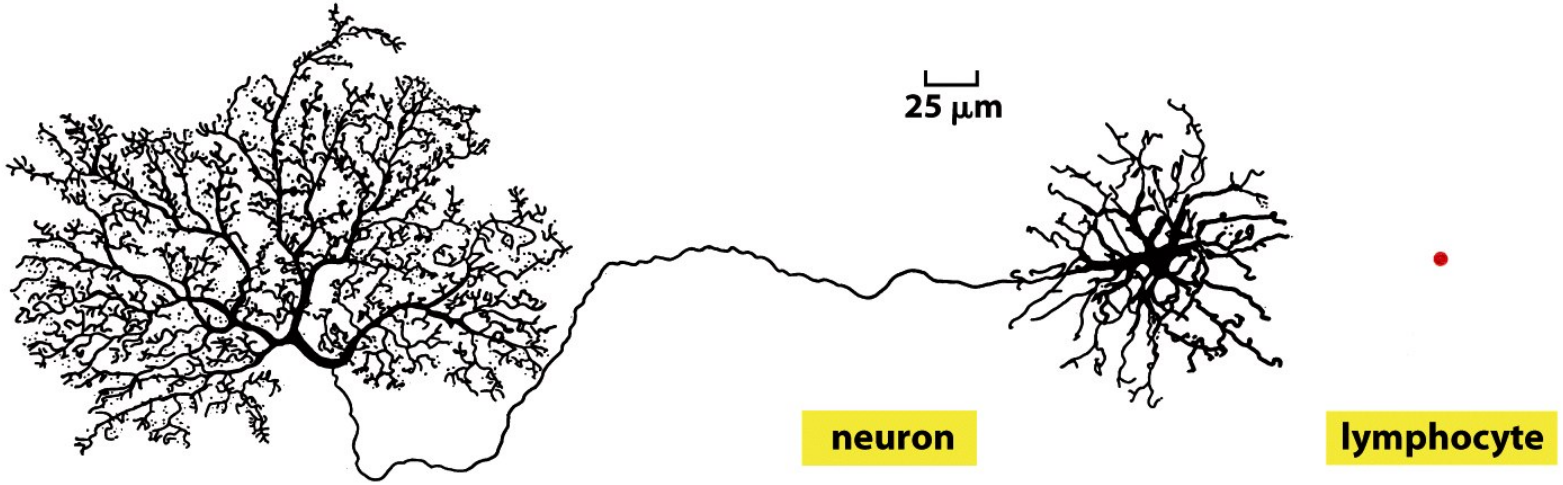


Figure 7-1 Molecular Biology of the Cell 5/e (© Garland Science 2008)

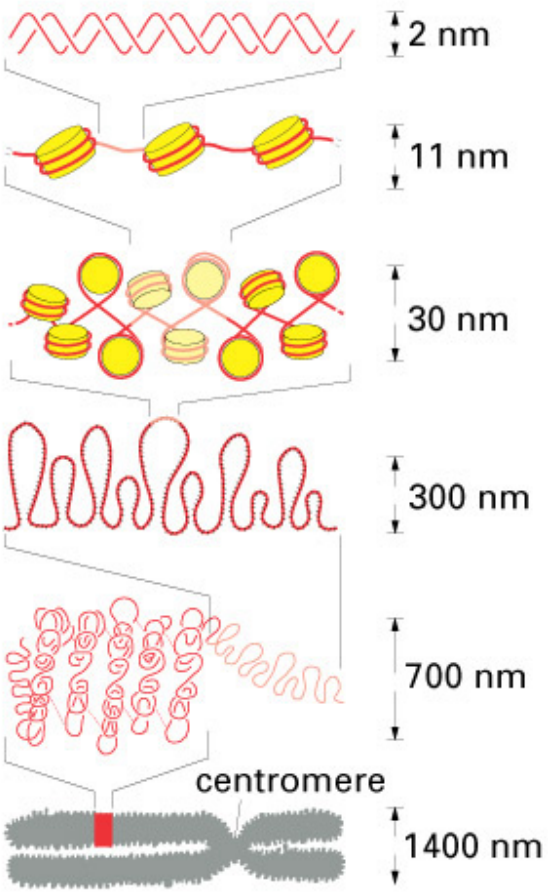
Two cells of a organism have exactly* the same DNA

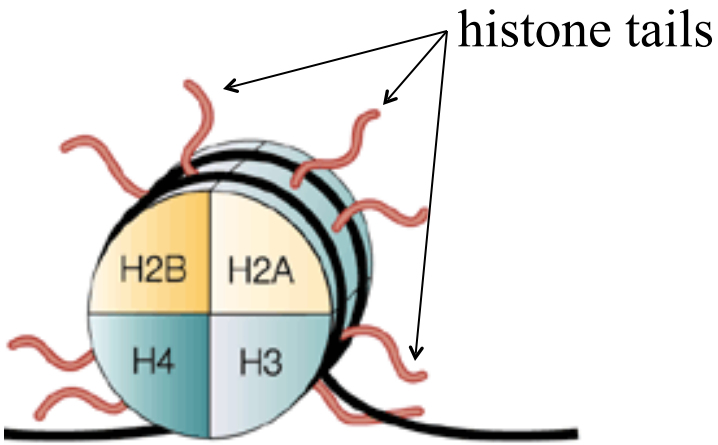
How does this differences arise?

How is cell fate remembered?

*** with exception of somatic mutations and rearrangements of immunological loci**

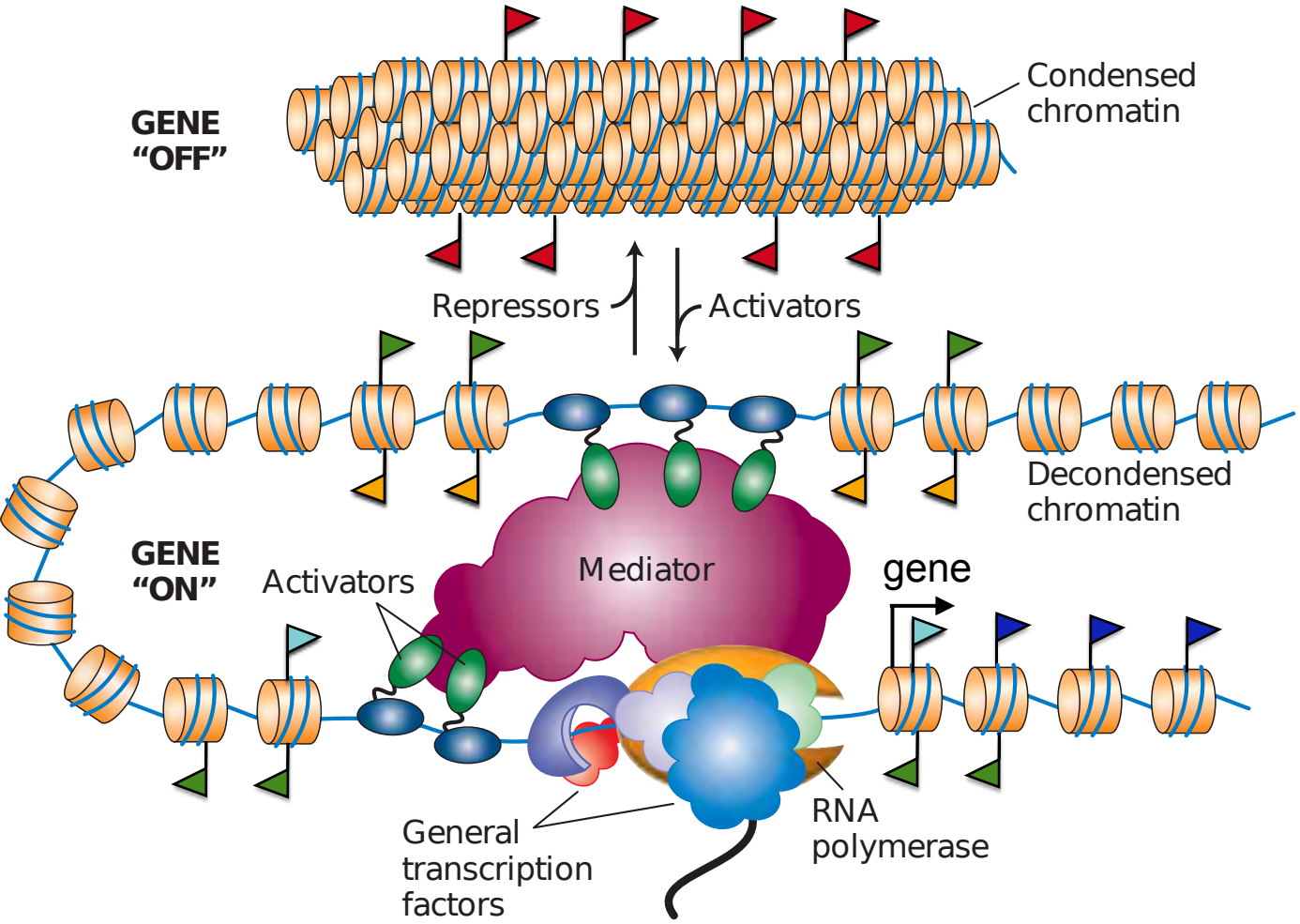
Epigenetics & Histones





Modification in histone tails
- change strength of DNA binding
- recruit transcription factors

Chromatin Structure and Regulation



Histone Code

Transcription

H3K79me2, H3k36me3

Active Regions

H3K27ac, H3K9ac

Active Promoters

H3K4me3

Active Enhancers

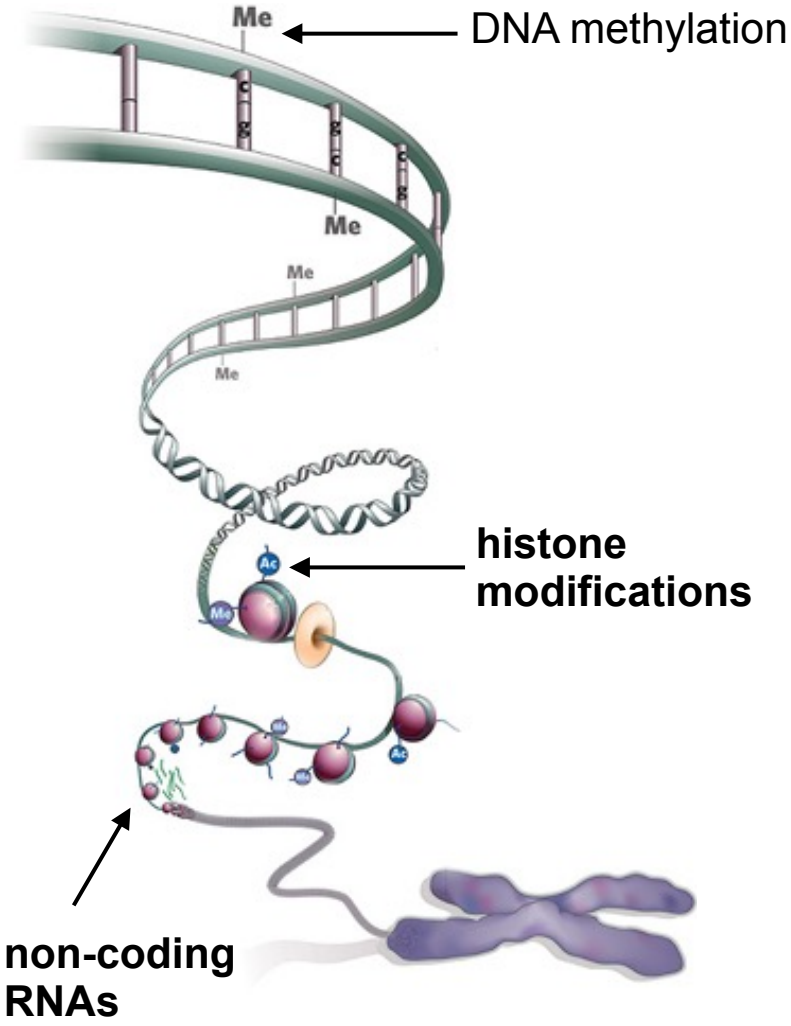
H3K4me1

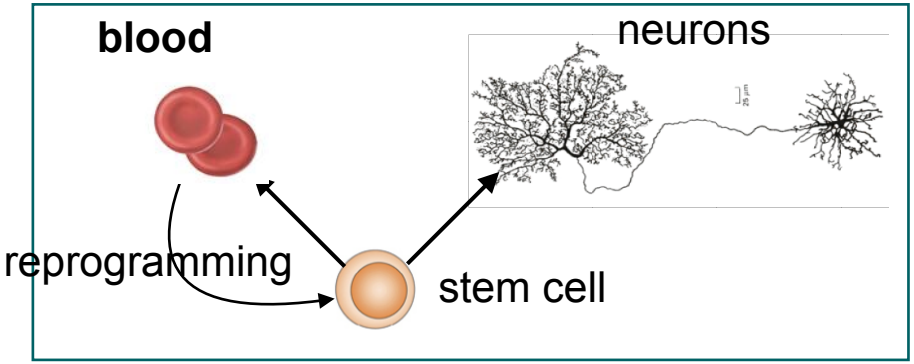
Repressed regions

H3K27me3, H3K9me3

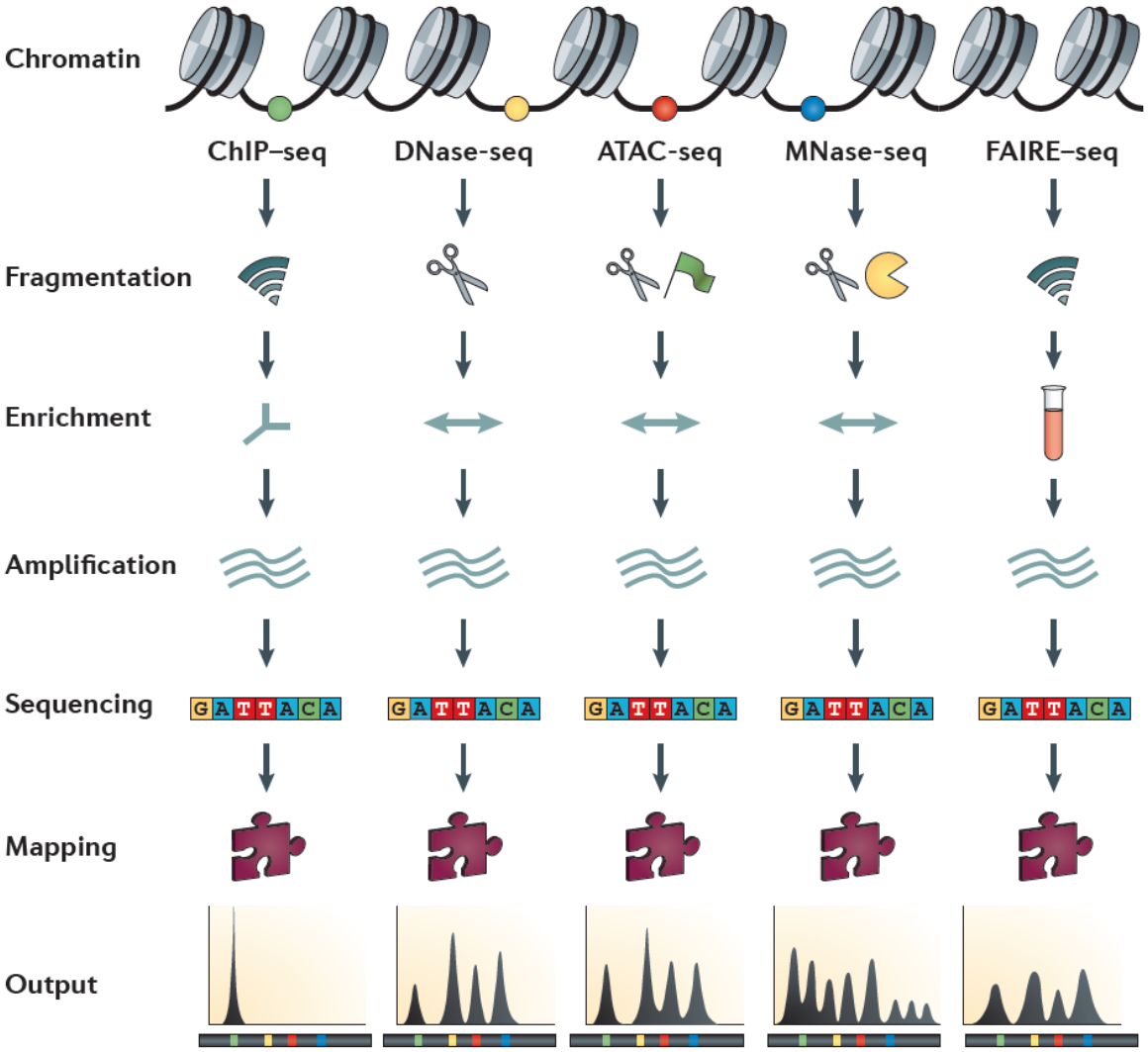
Adapted from Lodish, B. et al. (2004) 5th ed.

Epigenetics





Protein-DNA interactions with Next Generation Sequencing



Source: Meyer, C.A. *and Liu X.S.* (2014). *Nature Reviews Genetics*.

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

Sequencing

Sequencing

Read the bases of a particular DNA/RNA sequence

Applications:

- sequence DNA of known and unknown organism
- detect variants on patients
- sequence the RNA of a cell
- detect location of proteins interacting with DNA

Problem:

- only short DNA sequences (<1.000 bs) can be read

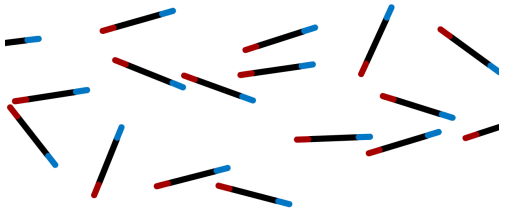
Solution: break DNA in several small pieces and bioinformatics

Next Generation Sequencing

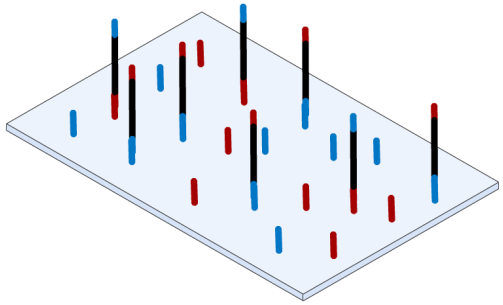
- NGS take advantage of parallelization
  - reads millions/billions of reads for a time
  - shorter reads (50-100 bps)
  - higher error rates (0.1-1%)
- commercial products:
  - 454
  - SOLiD
  - Solexa (Illumina)

Illumina Flow Cell - NGS Sequencing

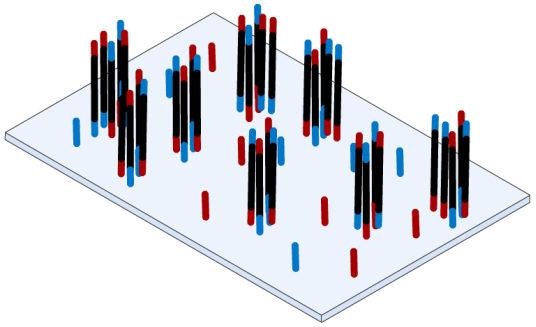
1- fragment sample DNA, insert adapters, attach to flow cell



2- use (bridge) PCR to copy fragments (close to origin)



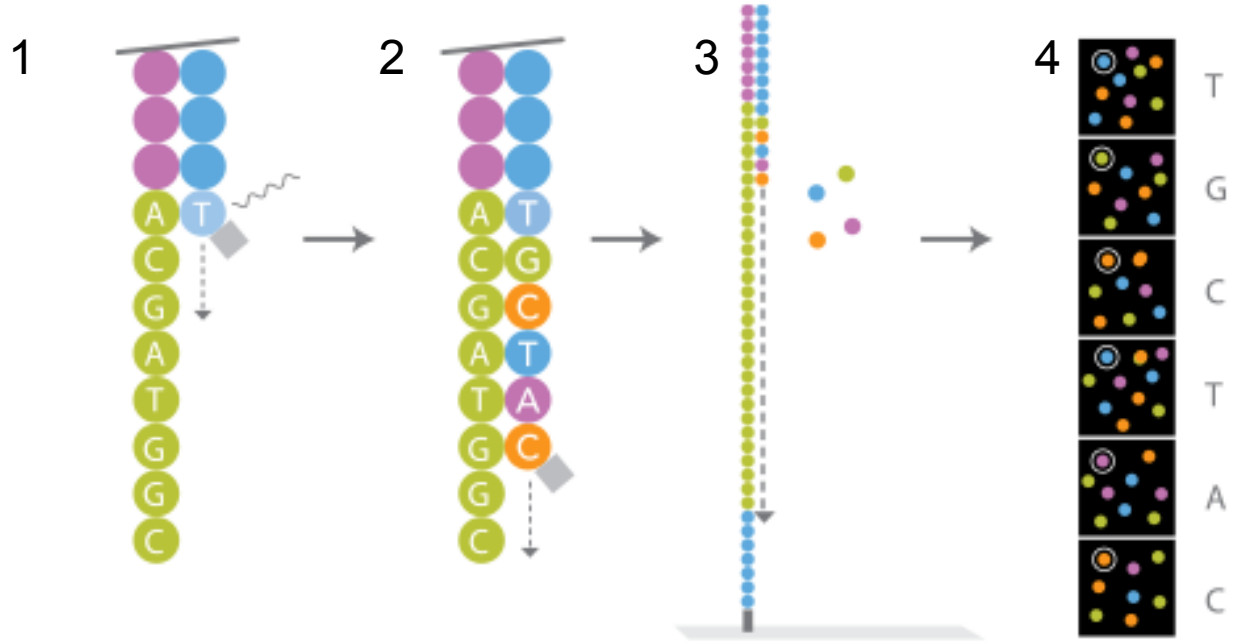
3- clusters of single stranded DNA (200m clusters with 2k DNA strands



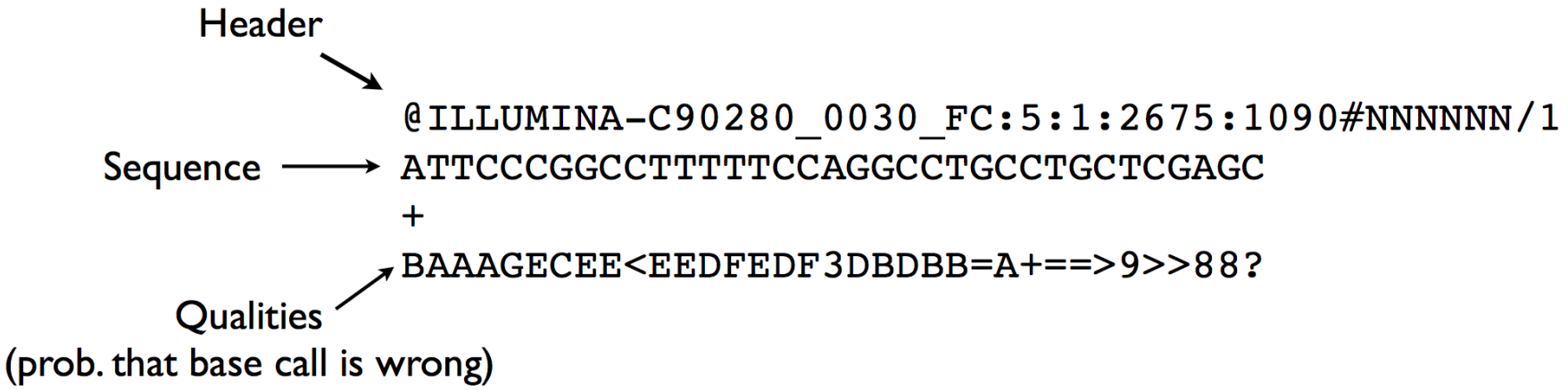
See video http://www.wellcome.ac.uk/Education-resources/Education-and-learning/Resources/Animation/WTX056051.htm

Illumina Flow Cell - NGS Sequencing

- **Iterative evaluation process:**
  1. **add RT-bases, polymerases integrate them**
  2. **wash away all not integrated elements**
  3. **take picture of flow cell to determine current base by dye**
  4. **derive reads from pictures**



Sequencing Results



One character encodes a number using ascii table (0-255)

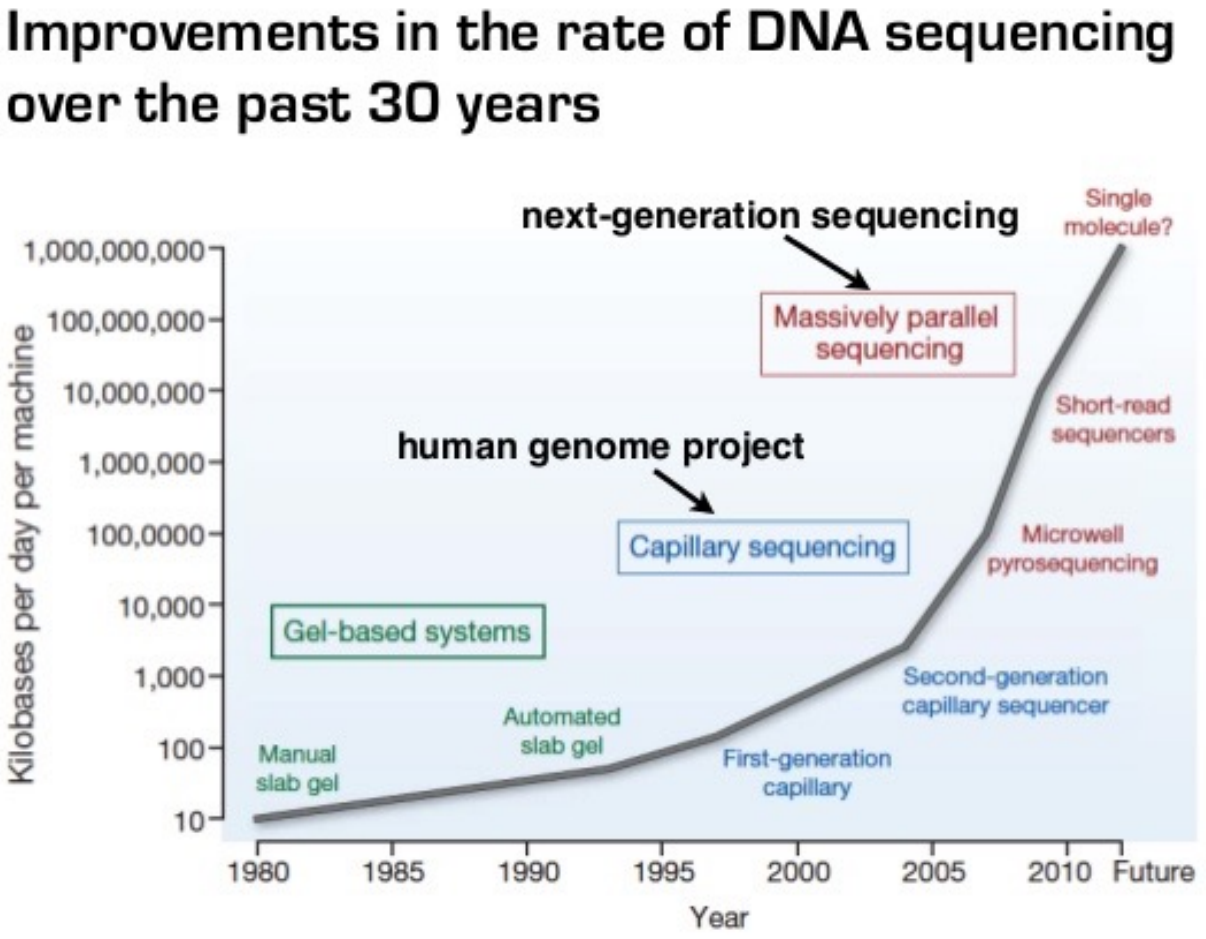
This number (*Q*) can be converted to *P*

Phred-scale

$Q = -10 * \log10\ P$

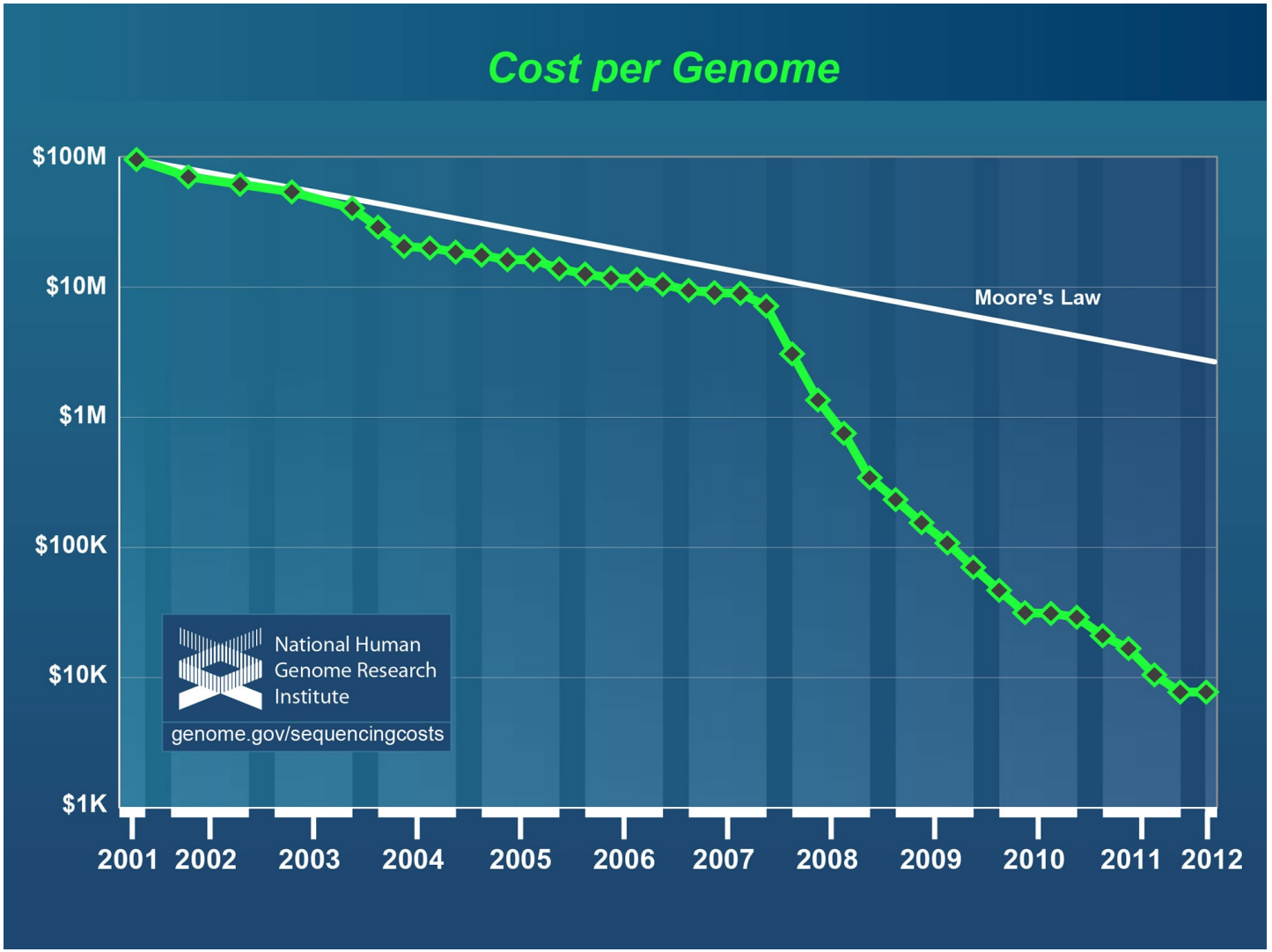
$P = 10^{(-Q/10)}$

Next Generation Sequencing



Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).

Sequencing Costs



Sequence Alignment

Sequence Alignment

NGS

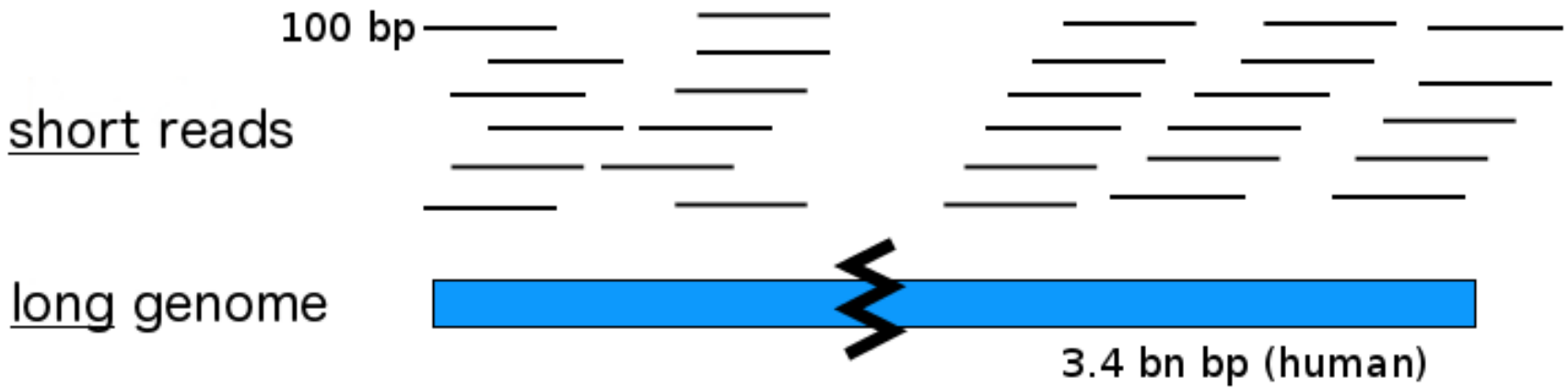
- reads from DNA fragments
- position in genome is unknown
- solution: alignment

DNA Sequencing

- de-novo assembly
  - construct unknown reference sequence from scratch
- resequencing / mapping
  - reference sequence given (applies to human- and mouse-studies)
  - build sequence that is similar but not necessarily identical to reference sequence

Alignment Problem

- a large reference sequence is given (genome)
  - up to billions of base pairs
- short reads (<200bps)
- find most probable position of the read in the genome (by inexact string matching)



Pitfals

- (Unknown) divergent of sample and reference genome
- Repeats in the genome (larger than read size)
- Recombinations
- Poor genome reference quality
- Sequencing/read errors

Algorithms - Alignment

Alignment/Mapping is a typical inexact string match problem

Algorithmic Solutions: ?

Algorithms - Alignment

Alignment/Mapping is a typical inexact string match problem

Algorithmic Solutions:

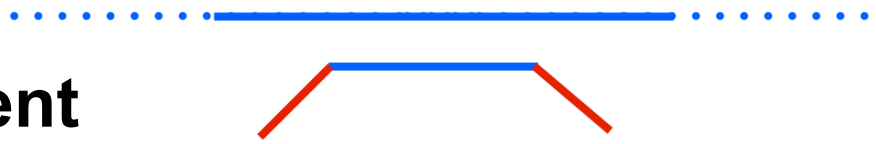
- **Smith & Waterman - dynamic programming (quadratic time/memory)**

Algorithms - Alignment

Alignment/Mapping is a typical inexact string match problem

Algorithmic Solutions:

- **Smith & Waterman - dynamic programming (quadratic time/memory)**
- **Blast - k-mer search for seeding followed by dynamic programming**
  - **large memory requirement**
  - **local alignment**



Algorithms - Alignment

Short read alignment is a special problem

- **reference sequence is large and fixed**
- **query sequence (reads) are short and many**

Solution: ?

Algorithms - Alignment

Short read alignment is a special problem

- **reference sequence is large and fixed**
- **query sequence (reads) are short and many**

Solution: ?

1. Use a data structure to represent reference

- **k-mer hash table (>40GB)**
- **suffix trees (> 4GB)**

Algorithms - Alignment

Short read alignment is a special problem

- **reference sequence is large and fixed**
- **query sequence (reads) are short and many**

Solution: ?

1. Use a data structure to represent reference

- **k-mer hash table (>40GB)**
- **suffix trees (> 4GB)**

2. Find candidate (k-mer) hits on genome (>100)

Algorithms - Alignment

Short read alignment is a special problem

- **reference sequence is large and fixed**
- **query sequence (reads) are short and many**

Solution: ?

1. Use a data structure to represent reference

- **k-mer hash table (>40GB)**
- **suffix trees (> 4GB)**

2. Find candidate (k-mer) hits on genome (>100)

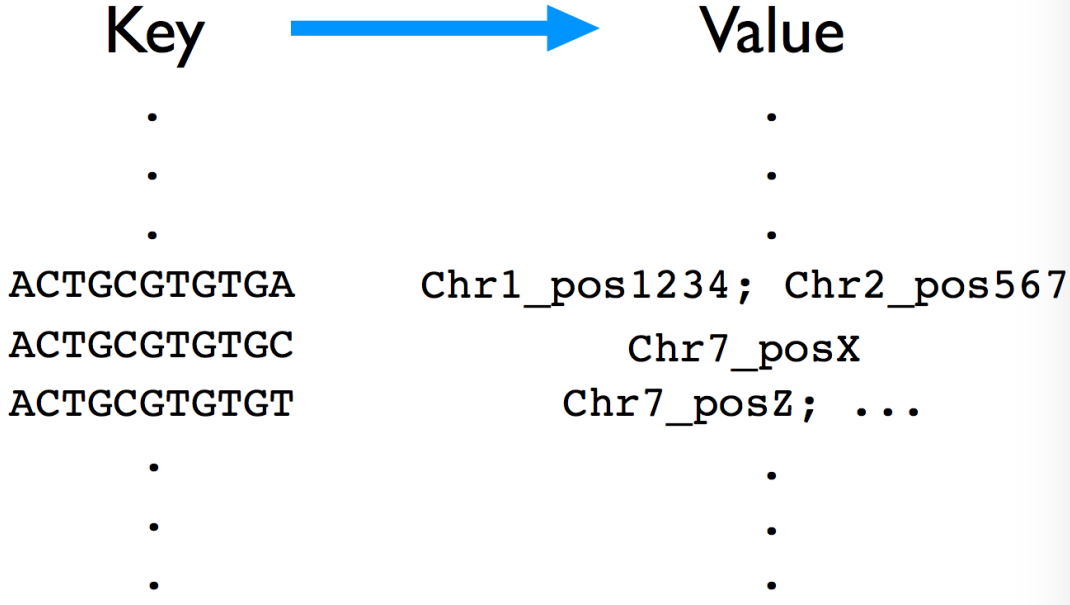
3. Improve alignment with Smith-Waterman

Methods work on linear time (query sequence)

Hash based algorithm

Lookups in hashes are *fast!*

1. Index the reference using *k*-mers.
2. Search reads vs. hash *k*-mers
3. Perform alignment of entire read around seed
4. Report best alignment



Also known as *Seed and extend*

Computational Epigenomics

Institute for Computational Genomics
RWTH AACHEN UNIVERSITY

Computational Epigenomics - Problems

1. Motif Search (already seen)

- search transcription factor binding sites in genomic sequences

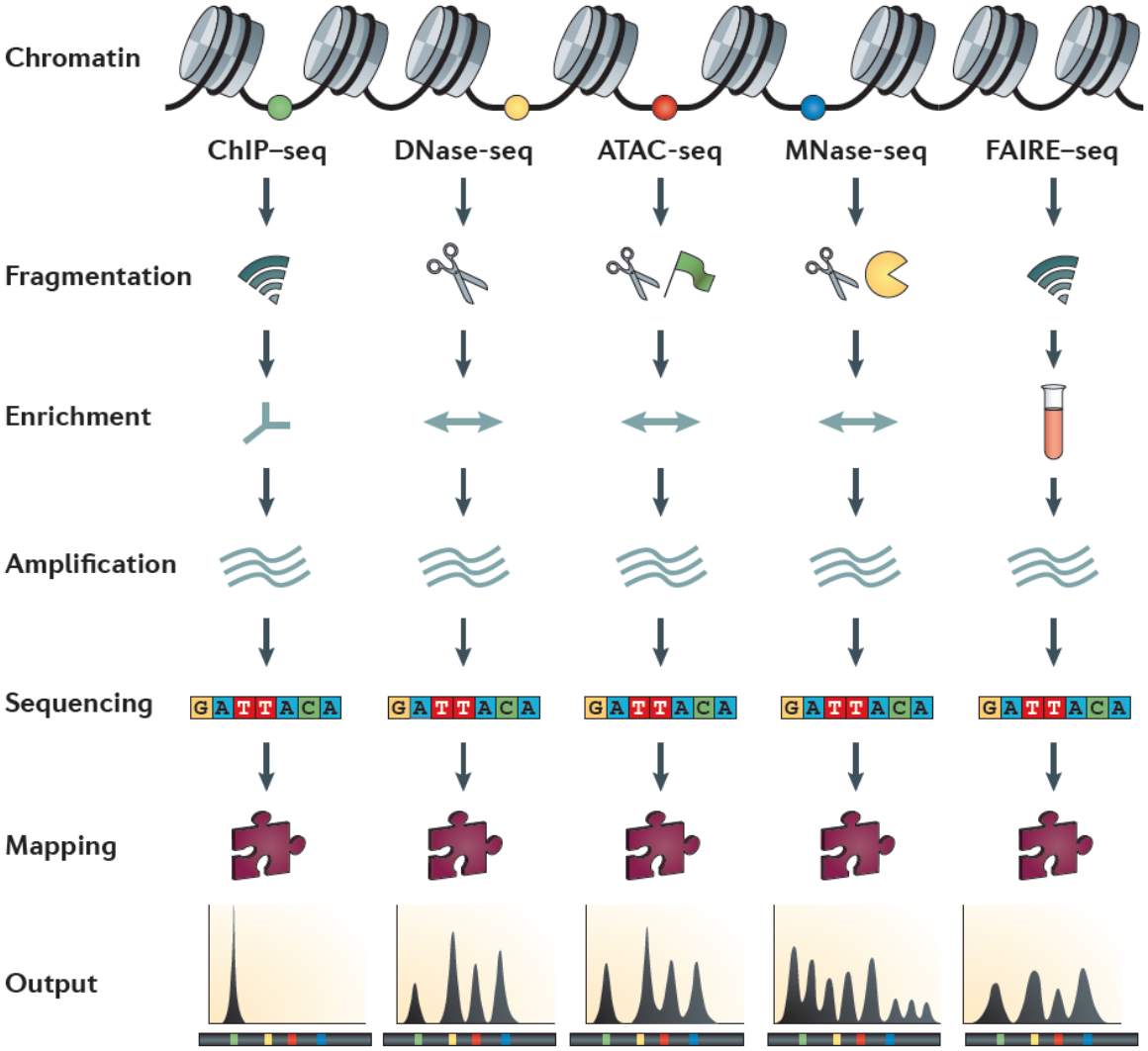
2. Peak Calling

- find regions with high number of ChIP-Seq signals

3. Digital Footprinting

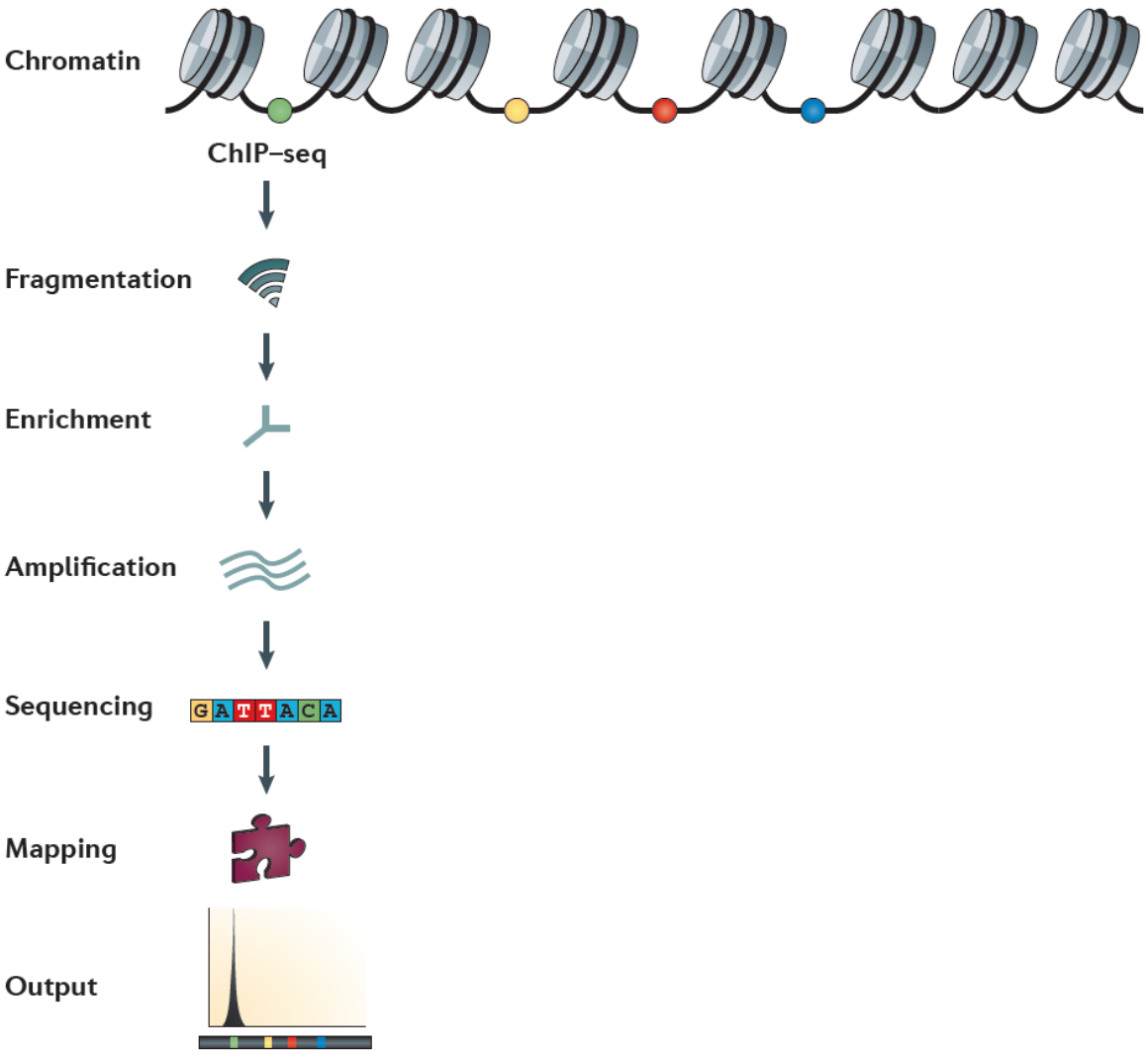
- find genomic regions with depletion of DNase-seq signals

DNA - Protein interactions with NGS



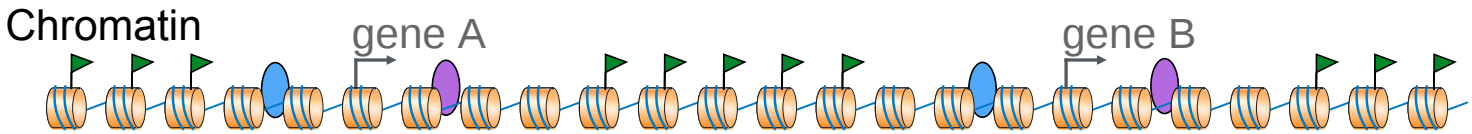
Source: Meyer, C.A. *and Liu X.S.* (2014). *Nature Reviews Genetics*.

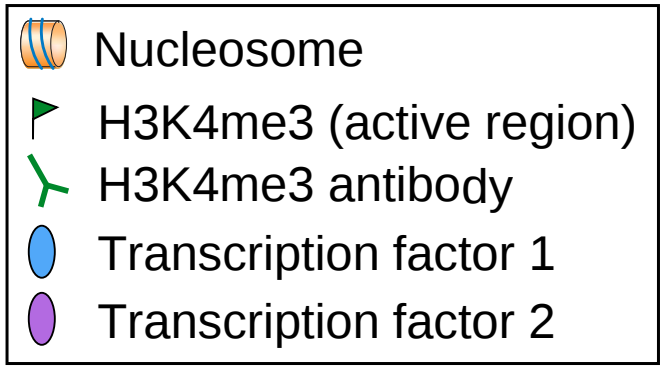
DNA - Protein interactions with ChIP-Seq



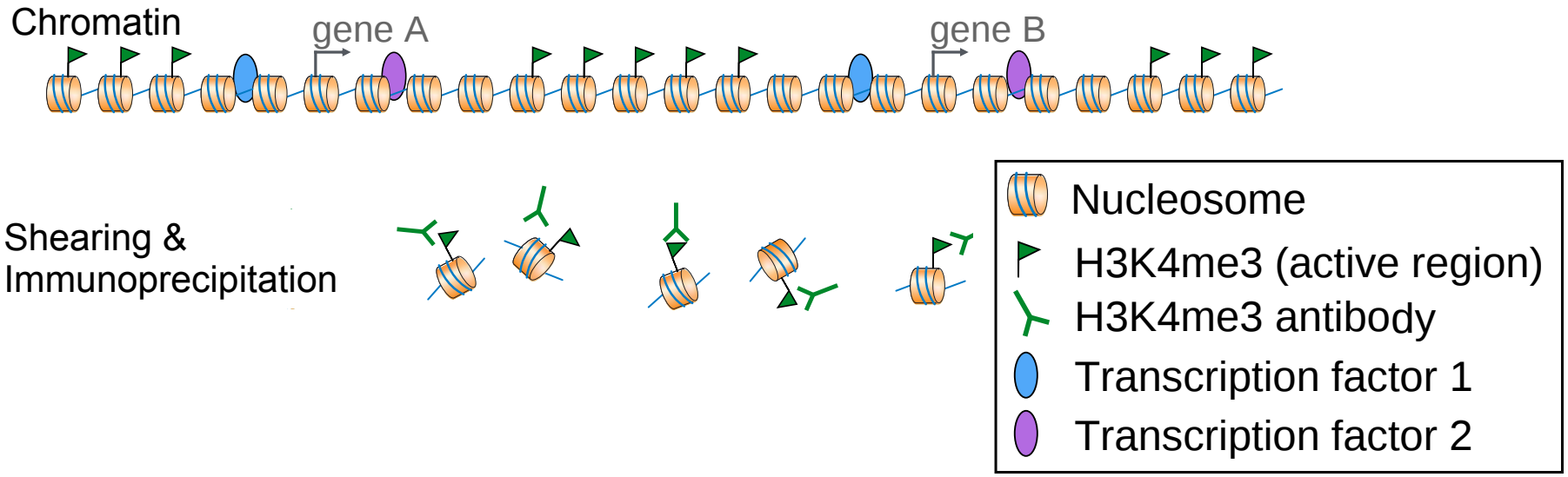
Source: Meyer, C.A. *and Liu X.S.* (2014). *Nature Reviews Genetics*.

DNA - Protein interactions with ChIP-Seq

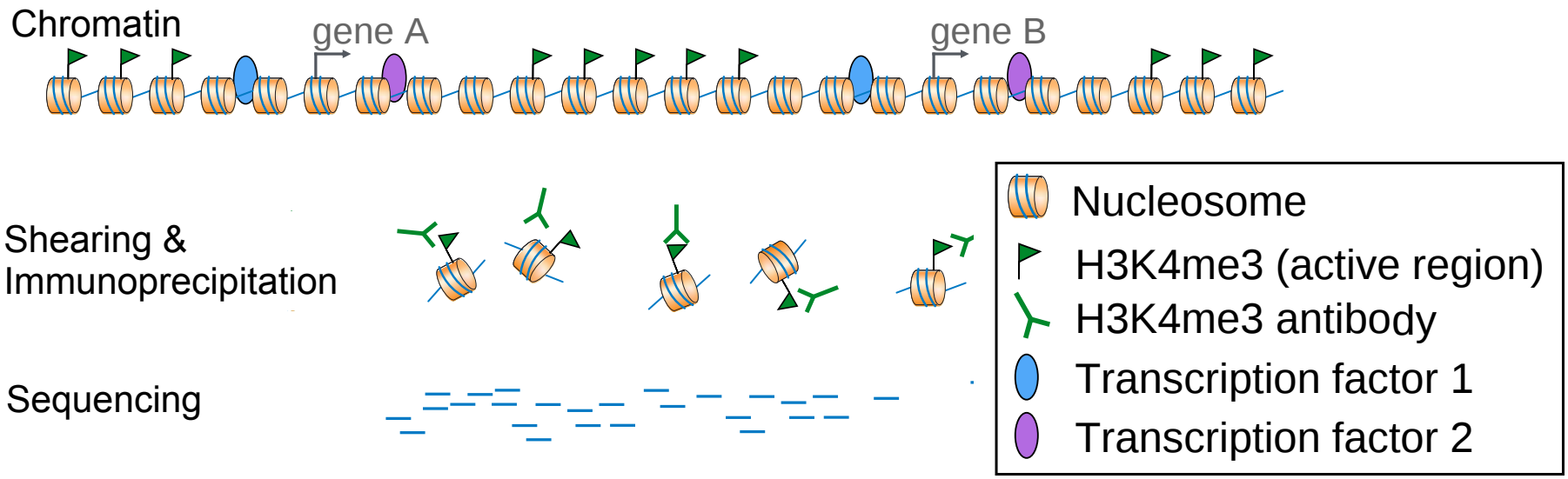




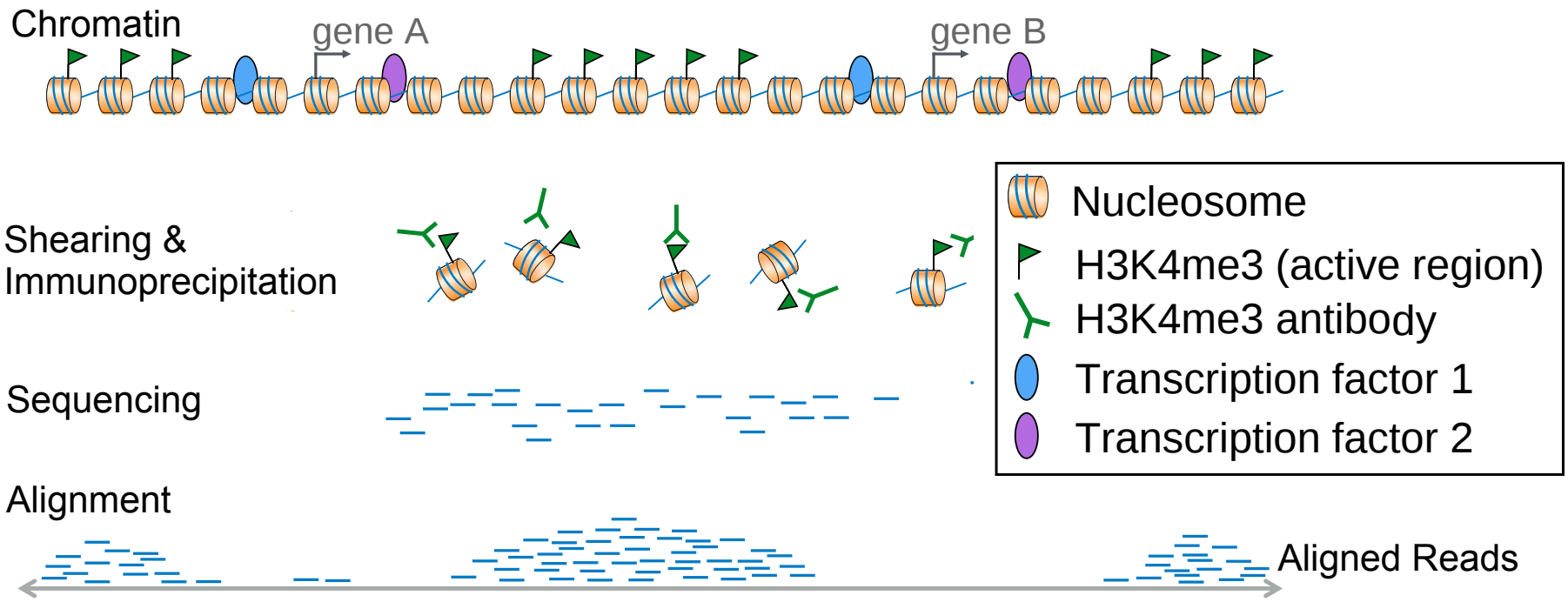
DNA - Protein interactions with ChIP-Seq



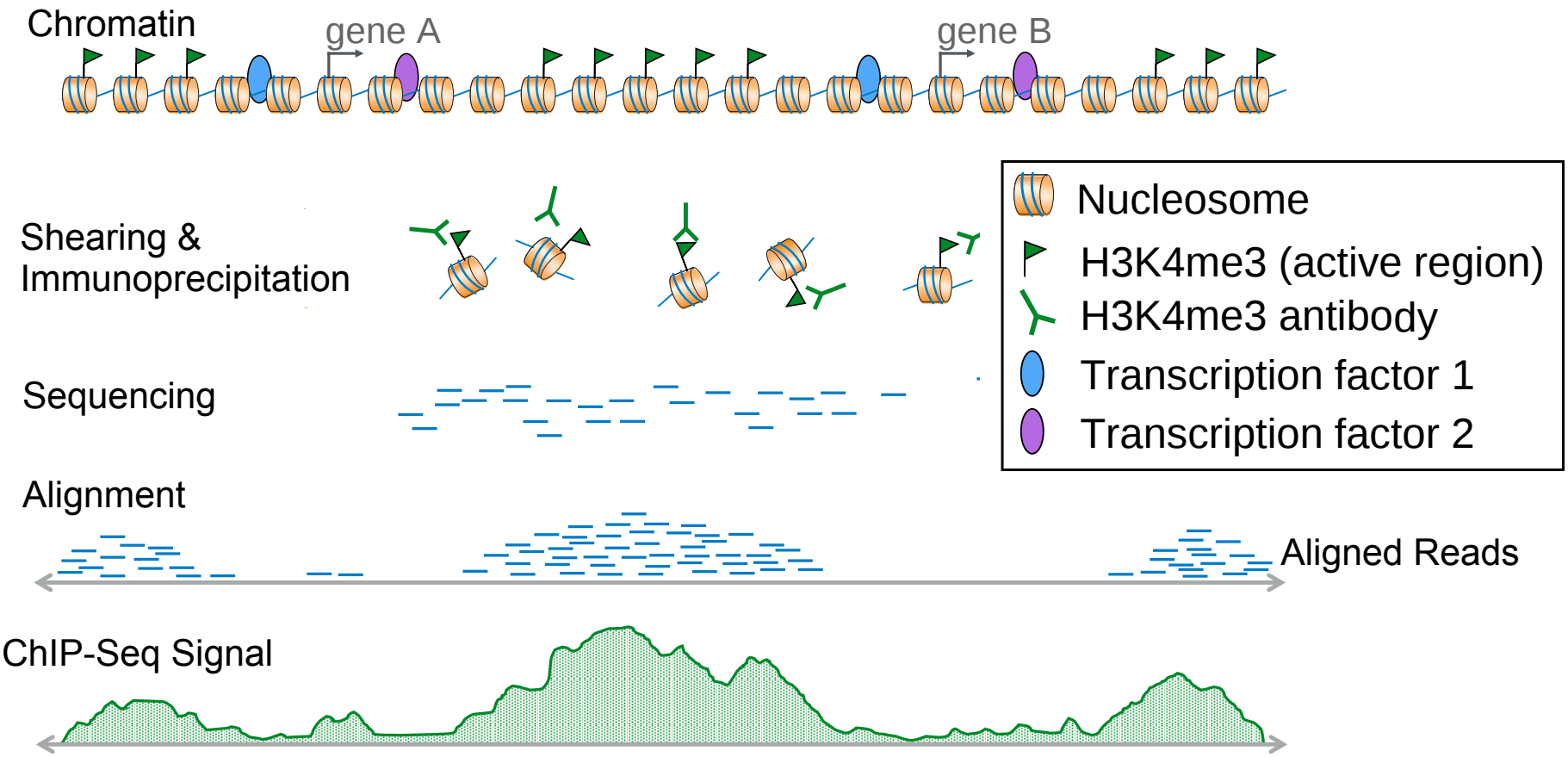
DNA - Protein interactions with ChIP-Seq



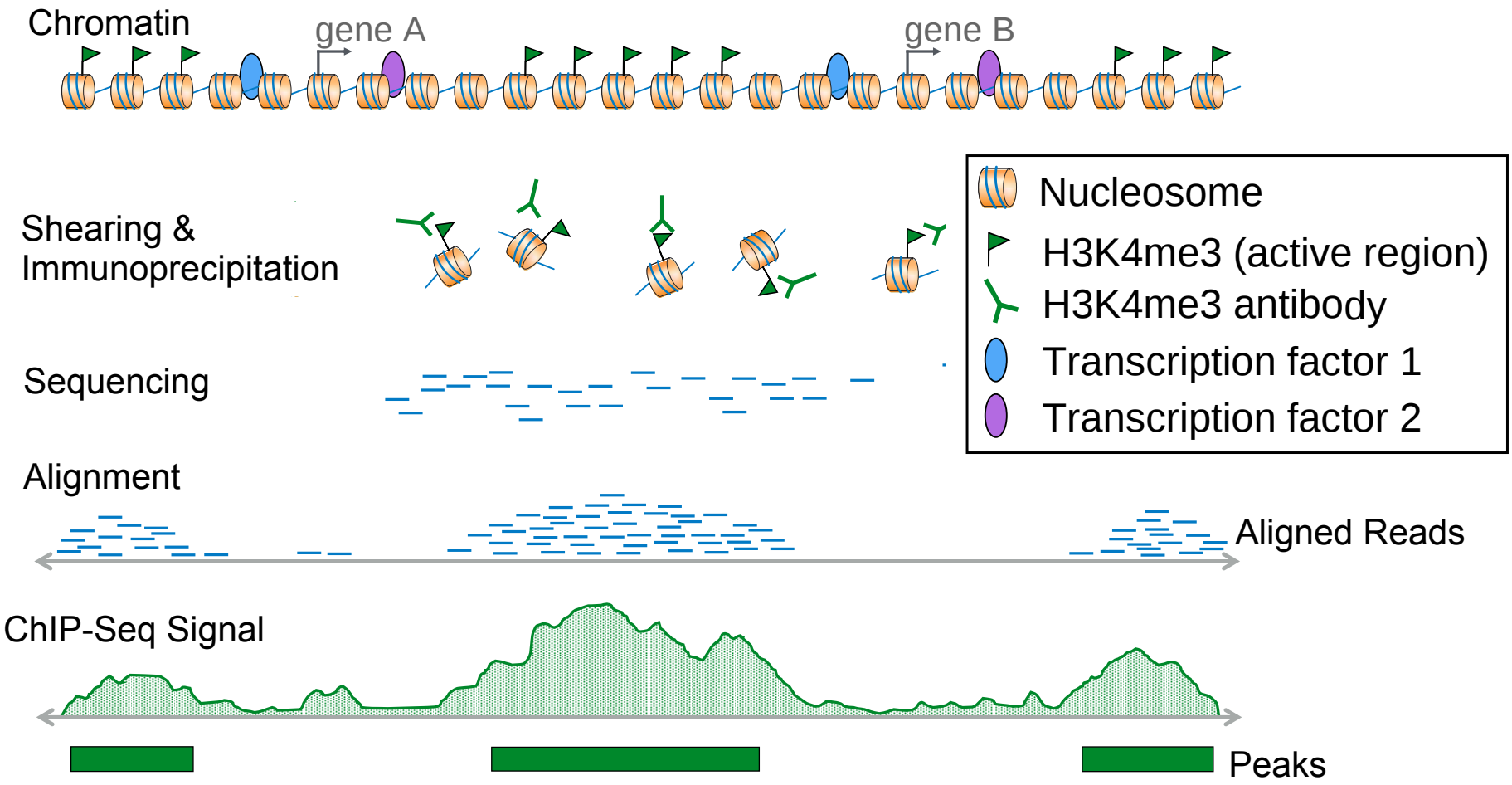
DNA - Protein interactions with ChIP-Seq



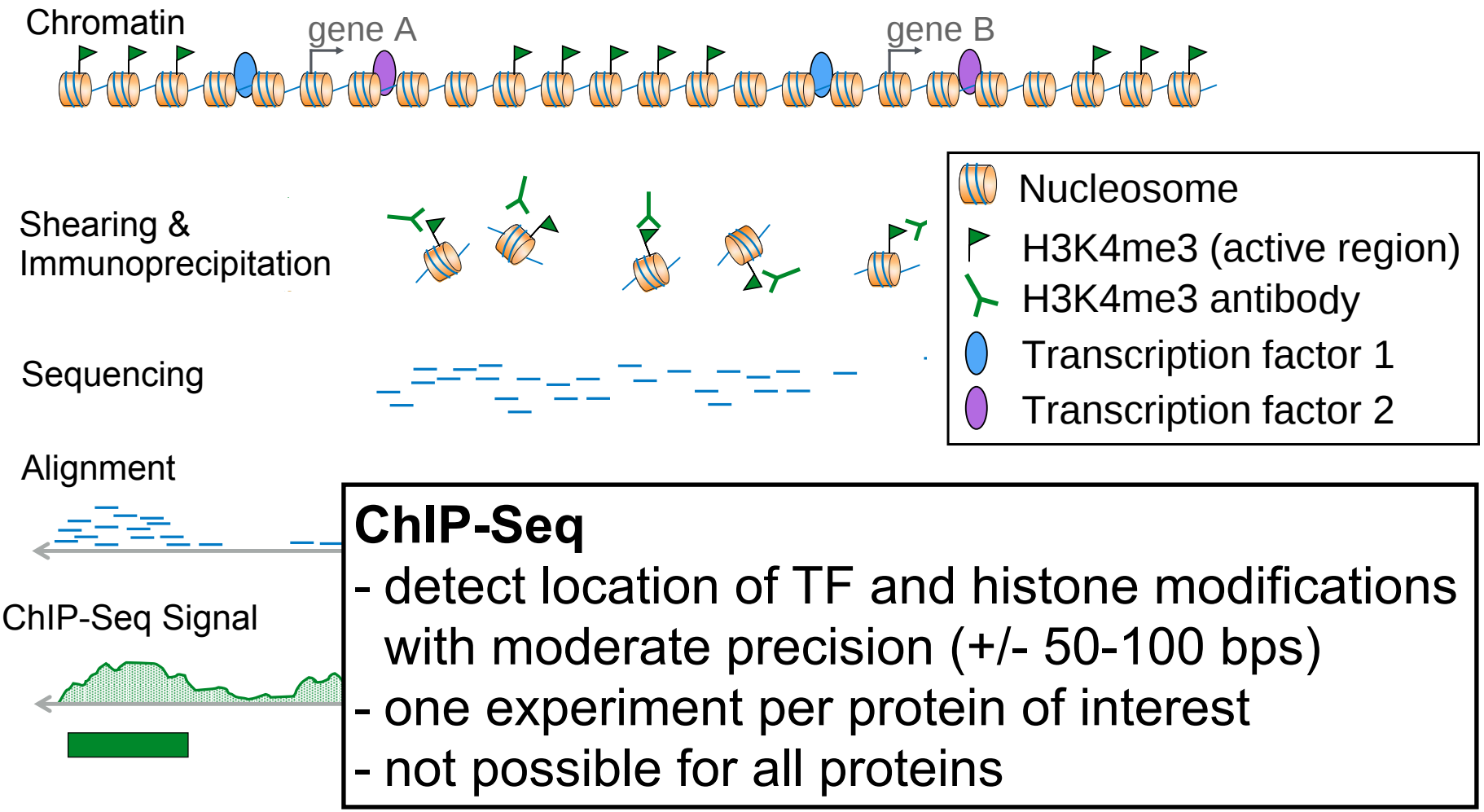
DNA - Protein interactions with ChIP-Seq



DNA - Protein interactions with ChIP-Seq



DNA - Protein interactions with ChIP-Seq



ChIP-Seq

- detect location of TF and histone modifications with moderate precision (+/- 50-100 bps)
- one experiment per protein of interest
- not possible for all proteins

Peak Calling

Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.

Aligned Reads

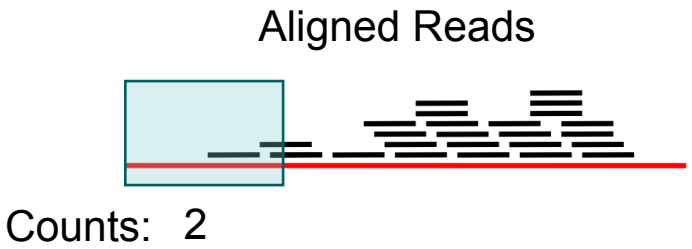
Example of a simple peak caller :

1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change

See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.



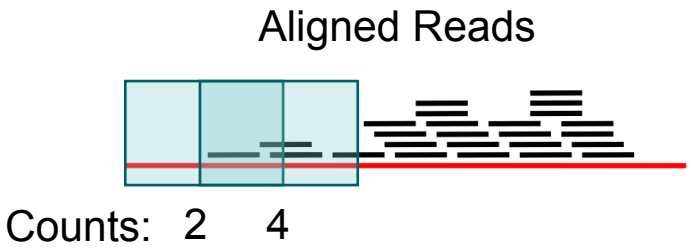
Example of a simple peak caller :

1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change

See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.



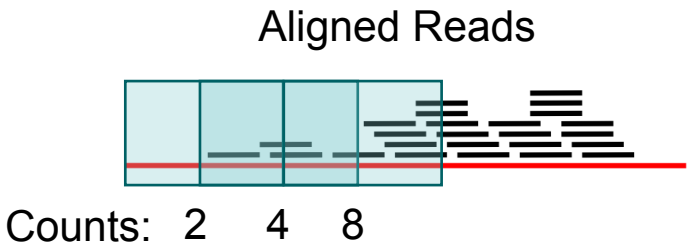
Example of a simple peak caller :

1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change

See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.



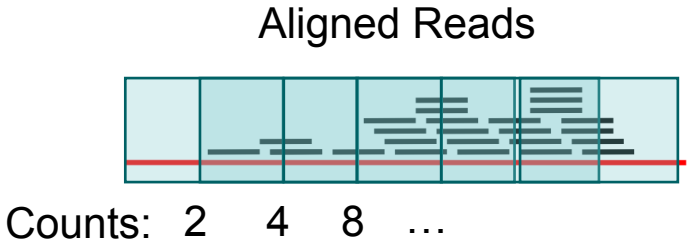
Example of a simple peak caller :

1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change

See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.



Example of a simple peak caller :

1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change

See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

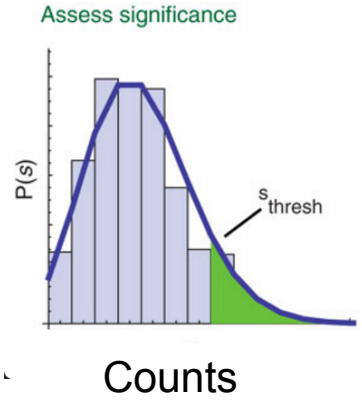
Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.

Aligned Reads

Counts: 2 4 8 …

Example of a simple peak caller :

1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change

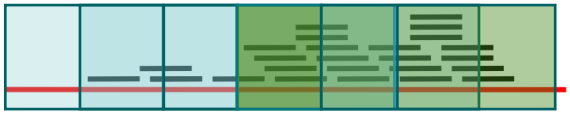


See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.

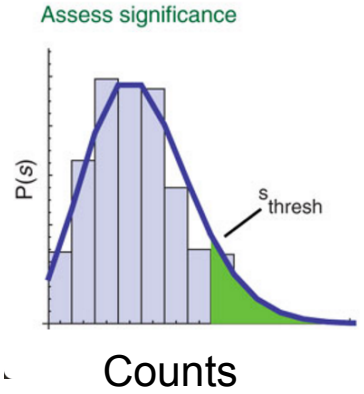
Aligned Reads



Counts: 2 4 8 …

Example of a simple peak caller :

1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change



See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

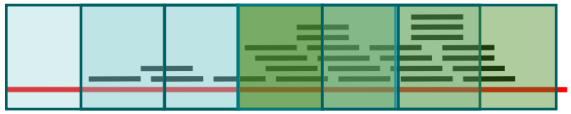
Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.

Problems:

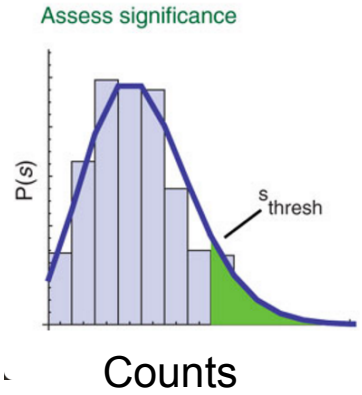
- which window size to use?
- distinct proteins have distinct peak sizes
- proper quantification of read counts require several further steps: fragment size estimation, CG bias correction, mappability, …

We will see examples in the next lectures ….

Aligned Reads



unts: 2 4 8 …



See

http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling - Example for Transcription Factors

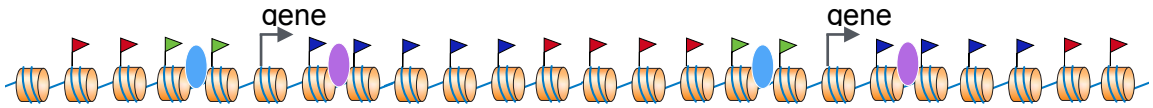
Example of analysis of ChIP-Seq for transcription factors (small peaks)

Histone code

- H3K79me2 - Transcribed
- H3K27ac - Active
- H3K27me3 - Repressed

Transcription Factors

- PU.1
- IRF4

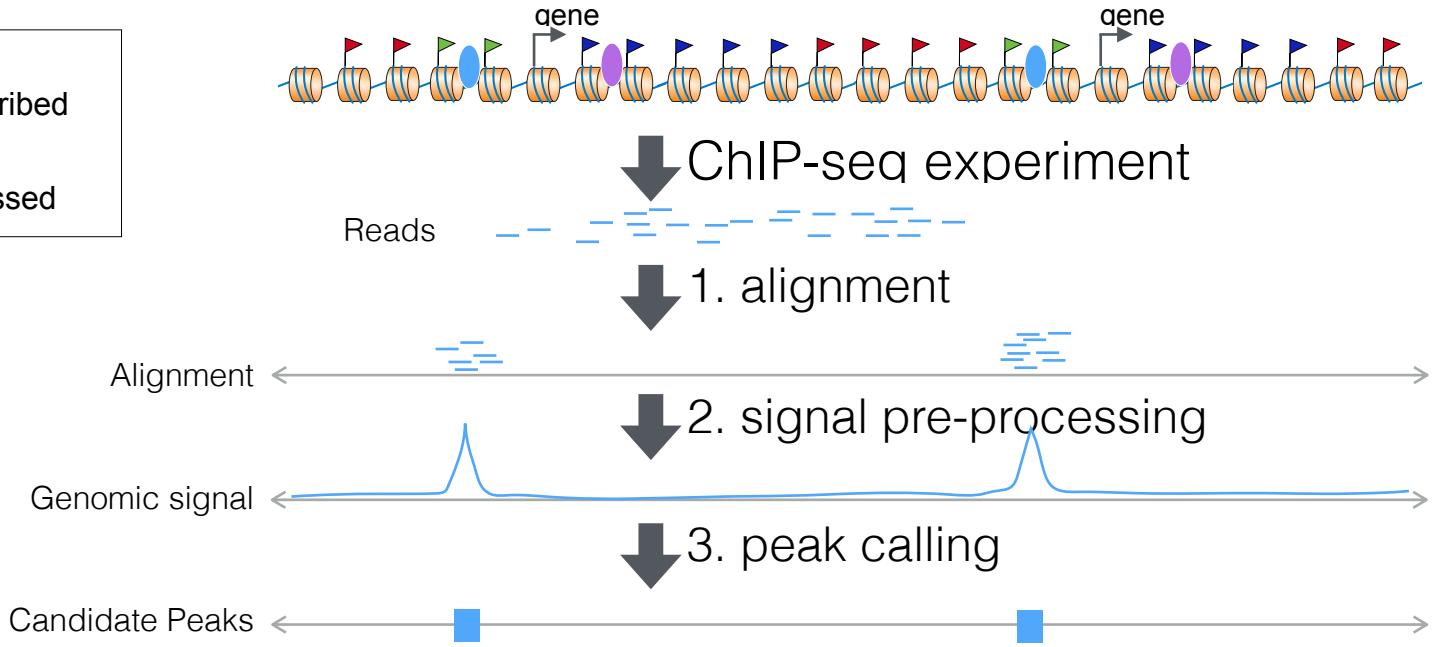


Peak Calling - Example for Transcription Factors

Example of analysis of ChIP-Seq for transcription factors (small peaks)

Histone code
- H3K79me2 - Transcribed
- H3K27ac - Active
- H3K27me3 - Repressed

Transcription Factors
- PU.1
- IRF4

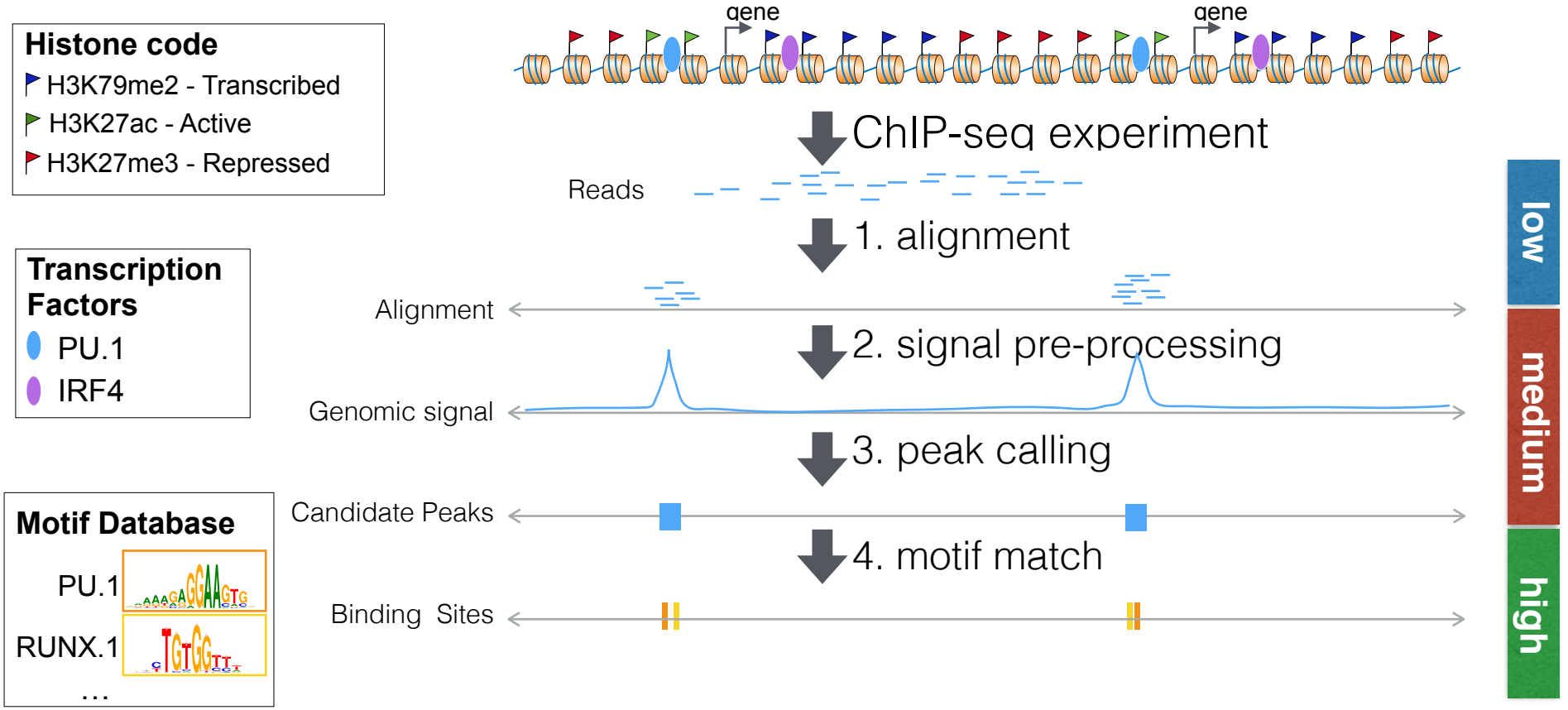


low

medium

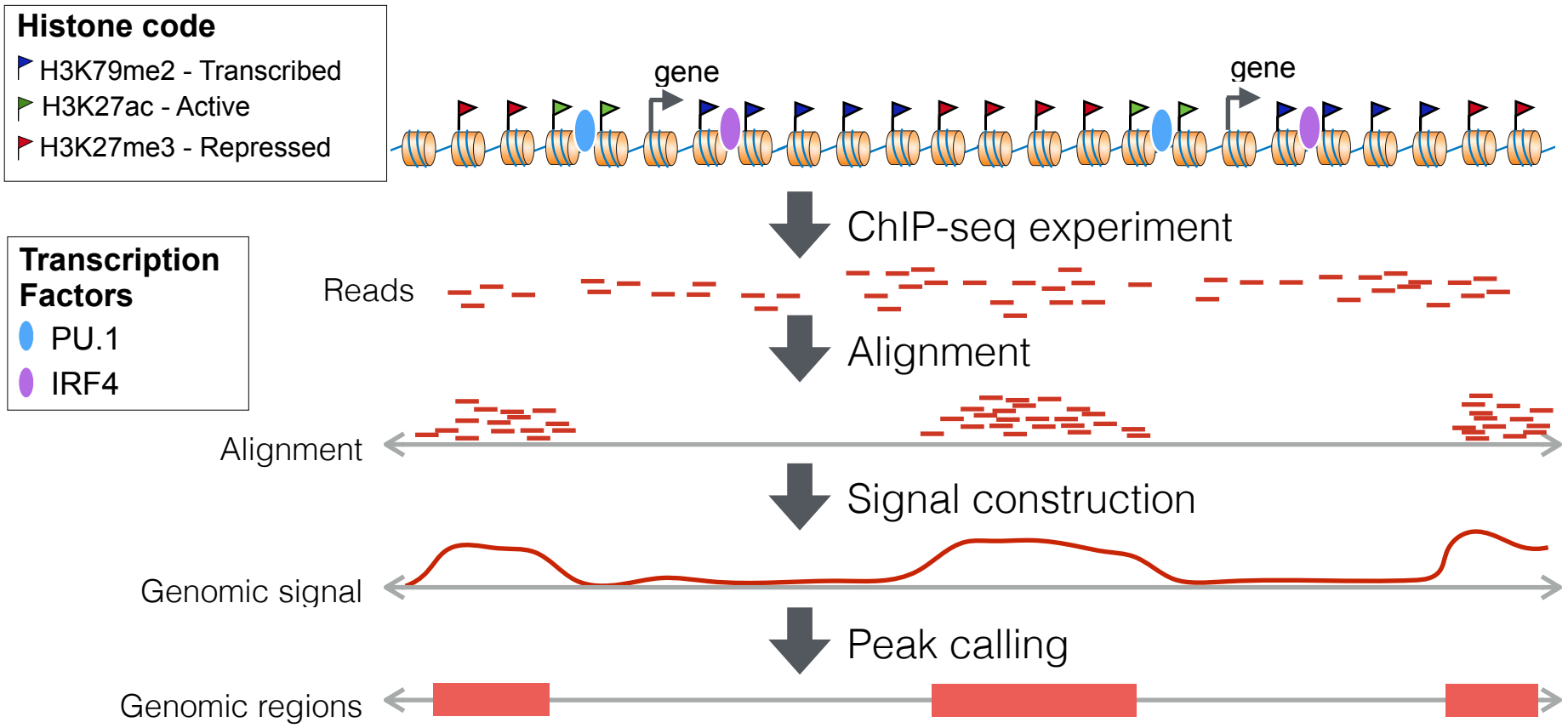
Peak Calling - Example for Transcription Factors

Example of analysis of ChIP-Seq for transcription factors (small peaks)

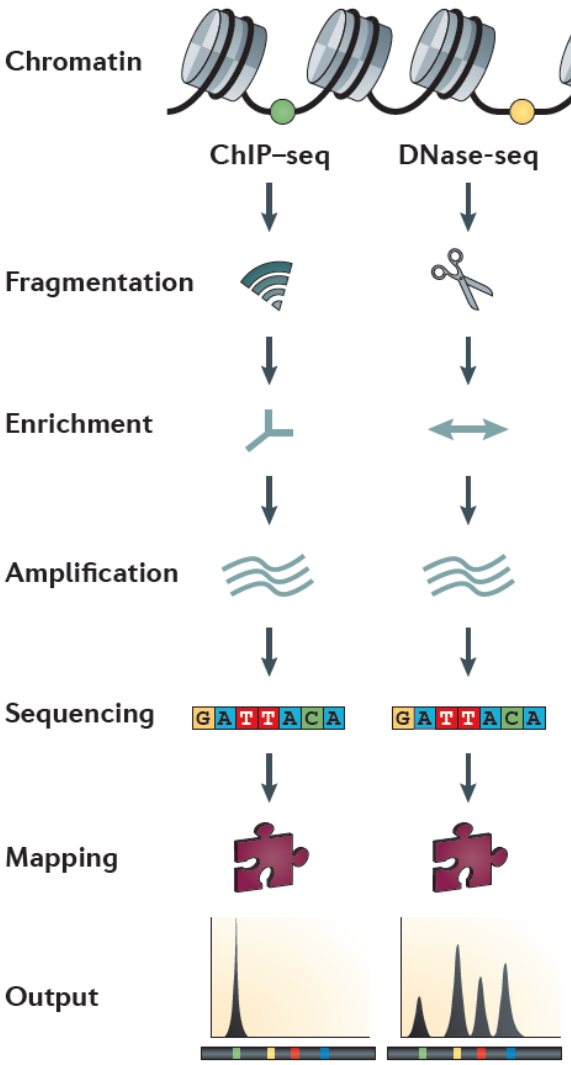


Peak Calling - Example for Histones

Example of analysis of ChIP-Seq for histones (medium to broad peaks)



DNA - Protein interactions with DNase-Seq

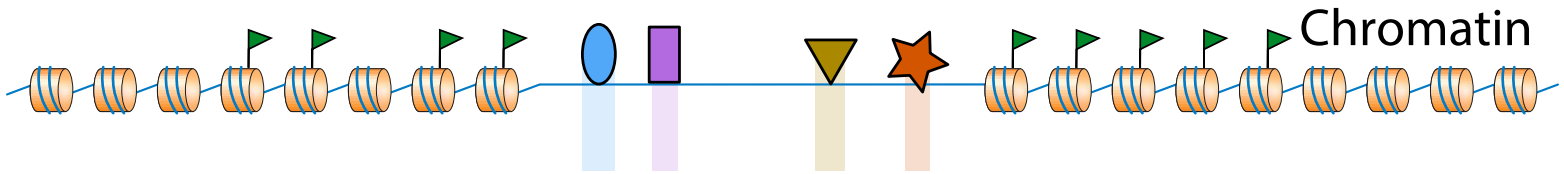


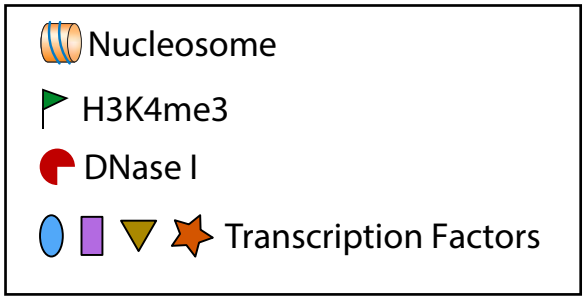
DNase-Seq

- detects open chromatin
- one experiment per cell of interest
- high resolution (single base level)
- potentially indicate binding of all TFs

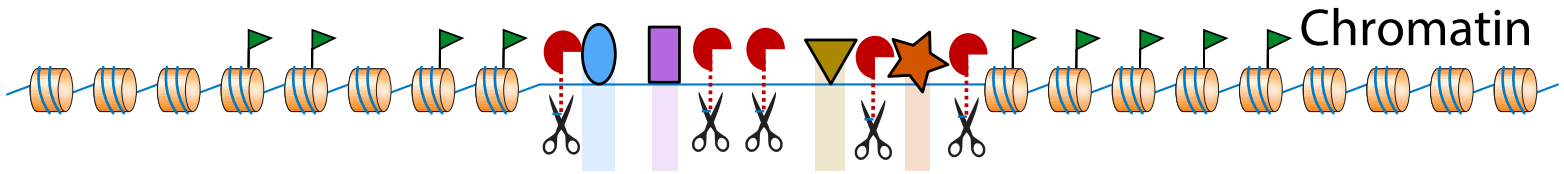
Source: Meyer, C.A. *and Liu X.S.* (2014). *Nature Reviews Genetics*.

Open chromatin with DNase-Seq

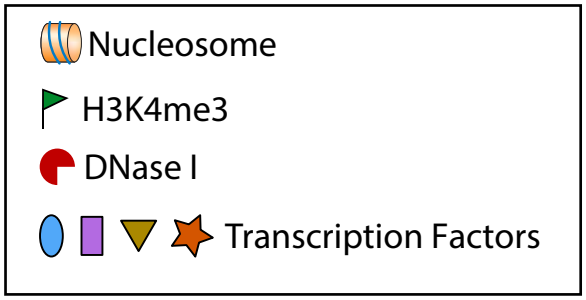




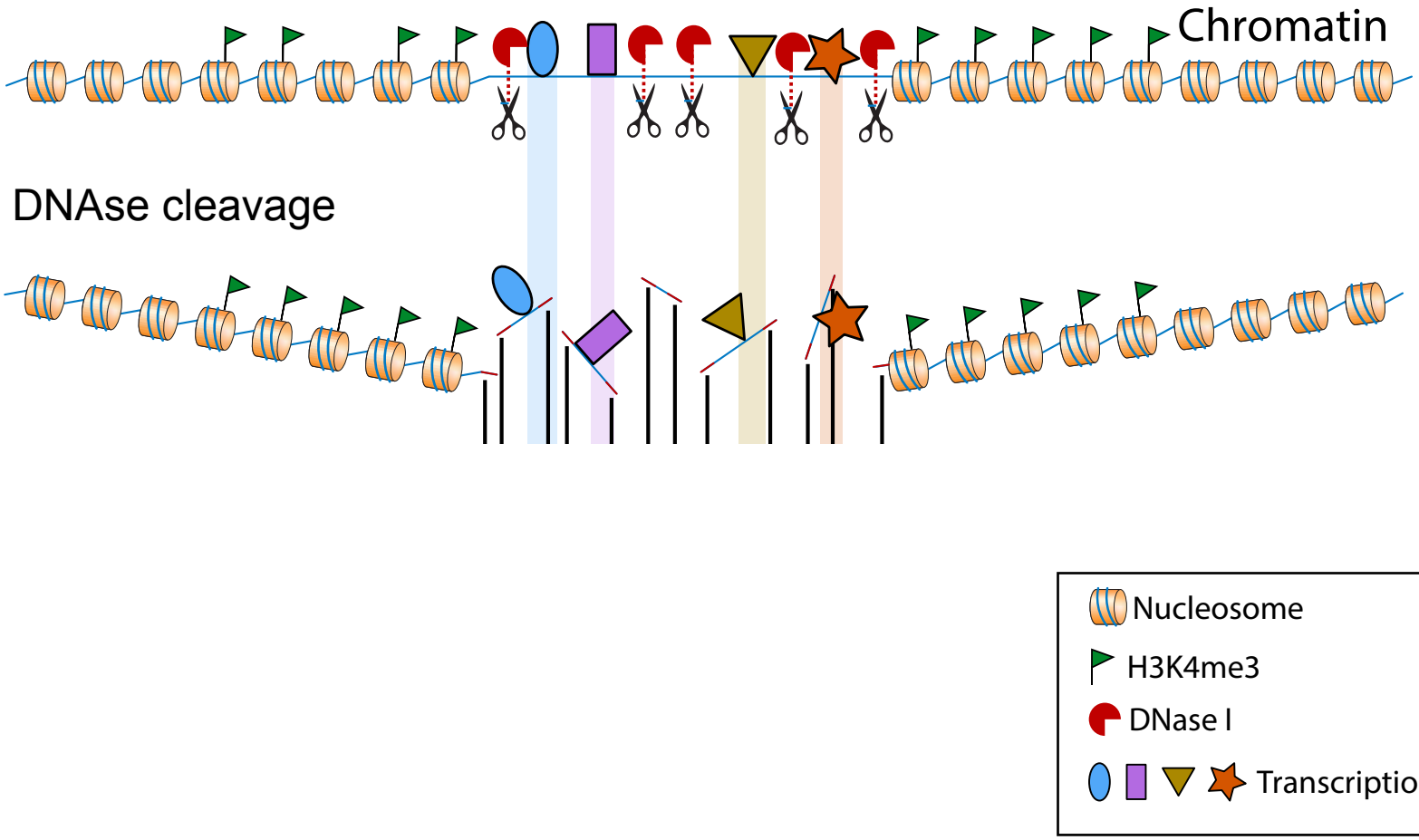
Open chromatin with DNase-Seq



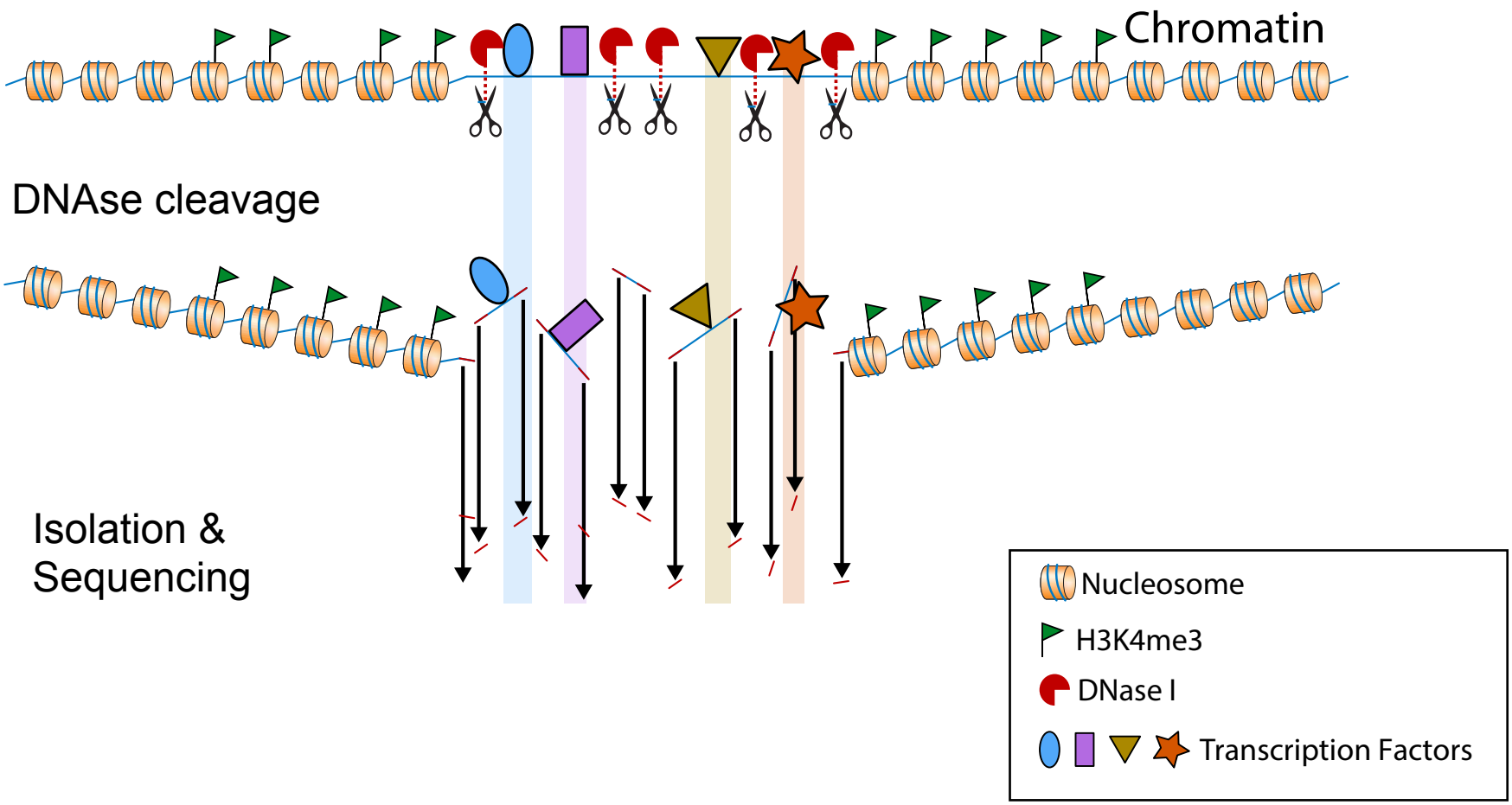
DNAse cleavage



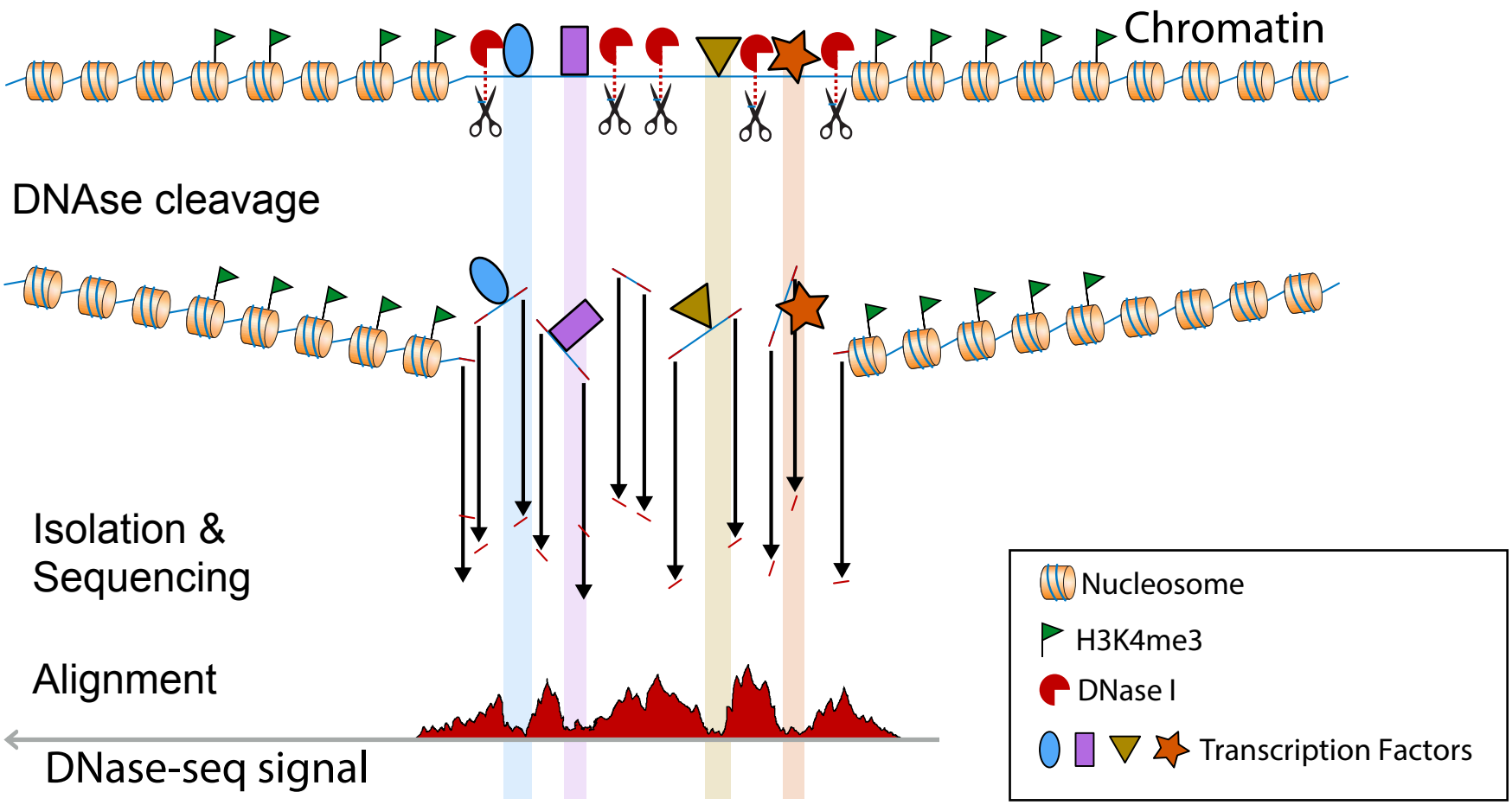
Open chromatin with DNase-Seq



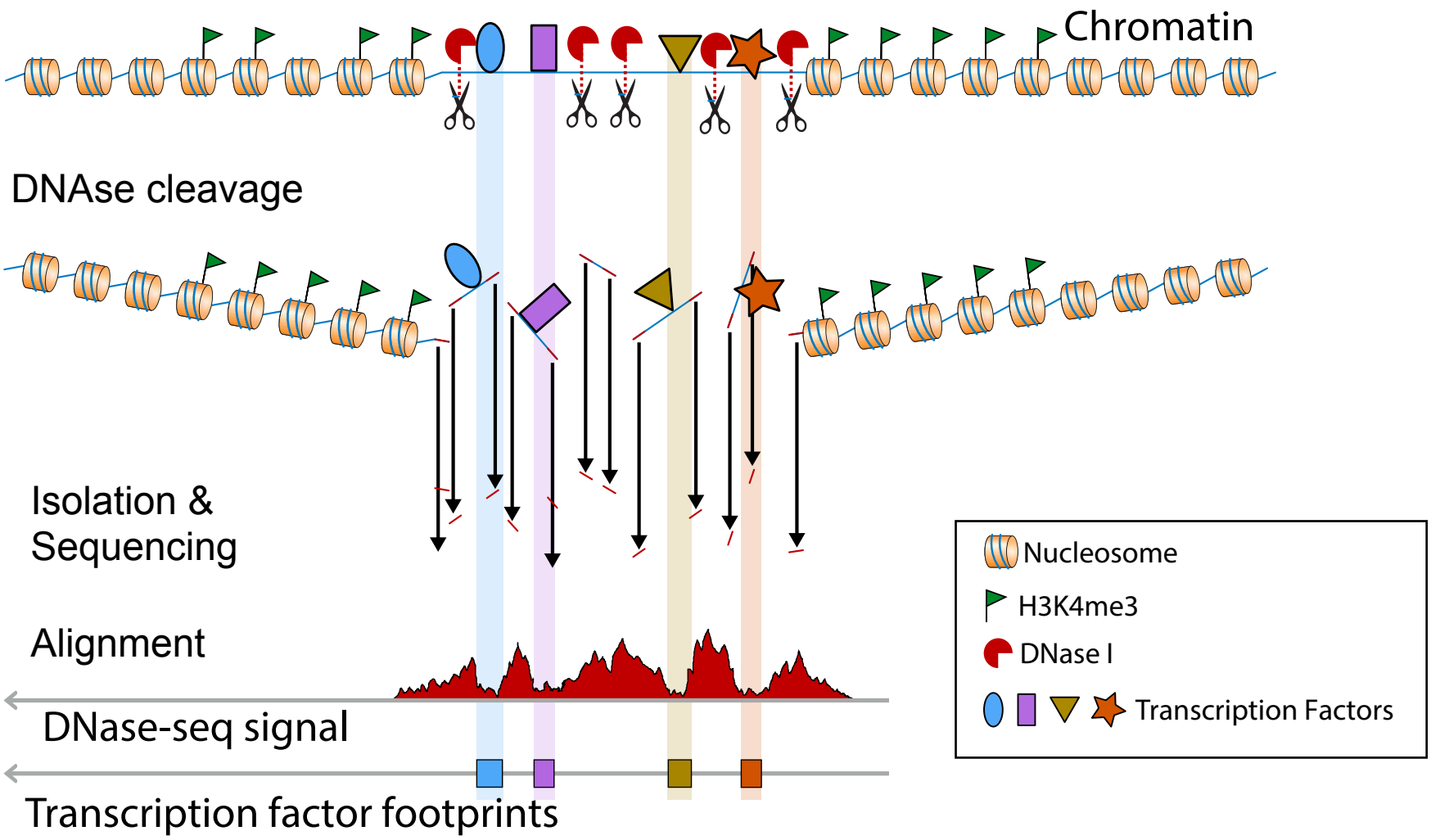
Open chromatin with DNase-Seq



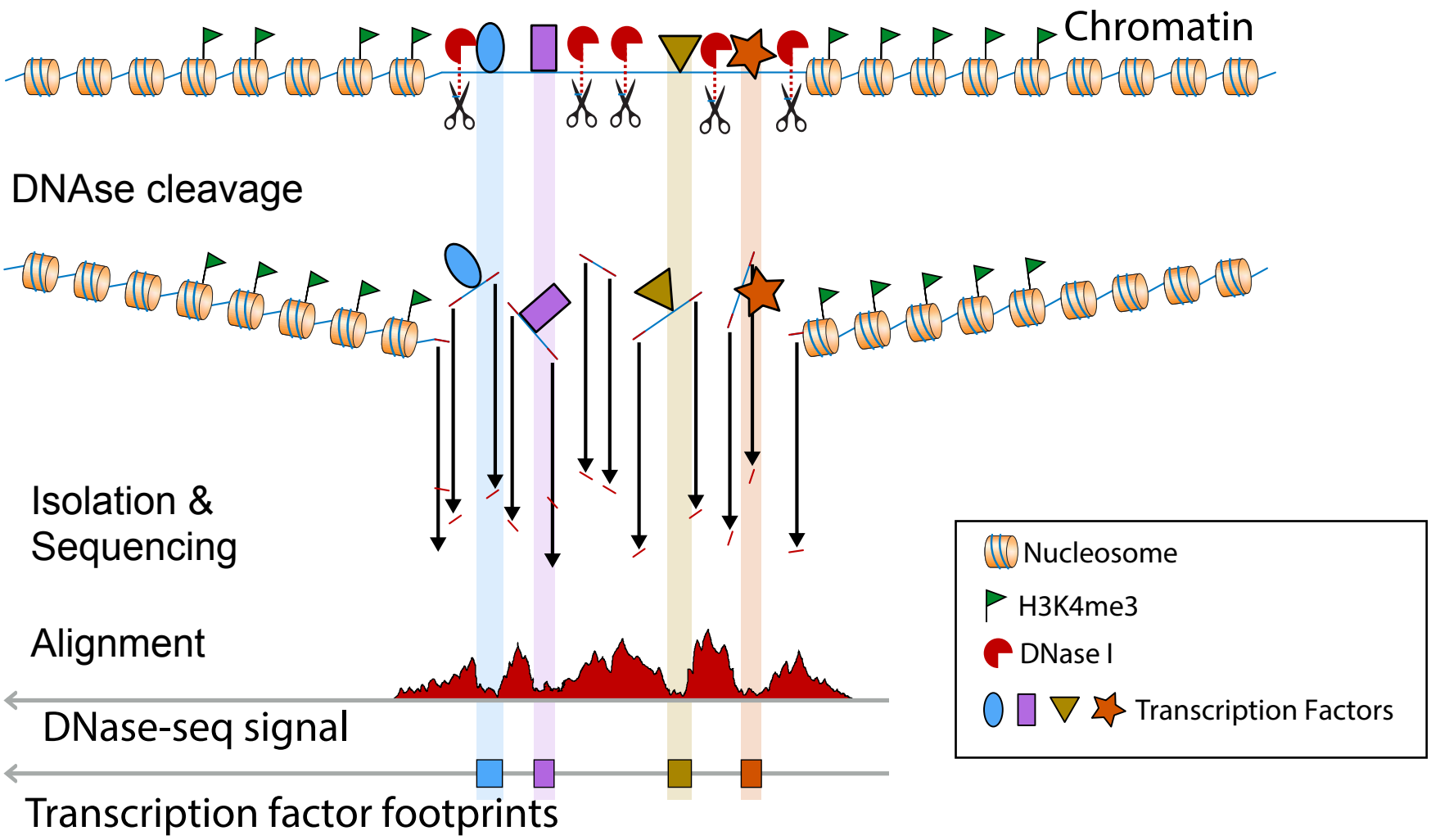
Open chromatin with DNase-Seq



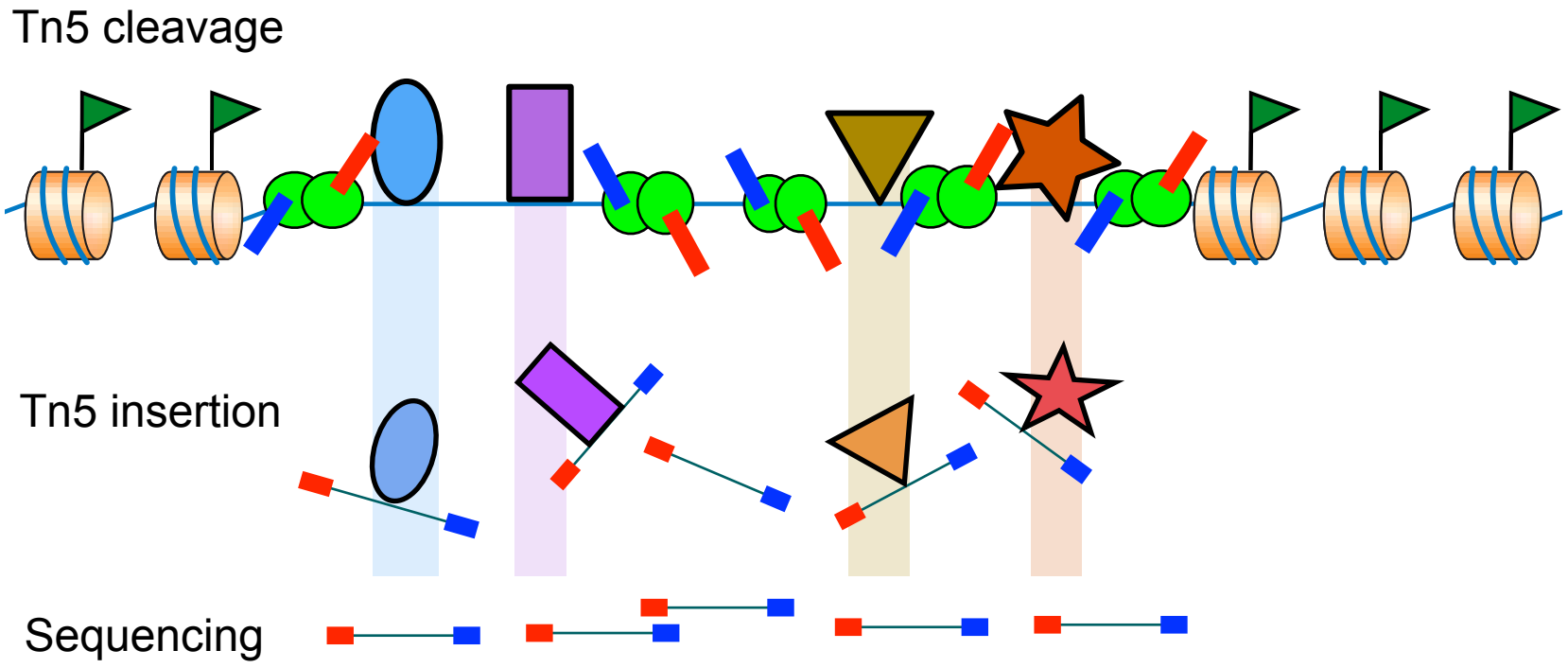
Open chromatin with DNase-Seq



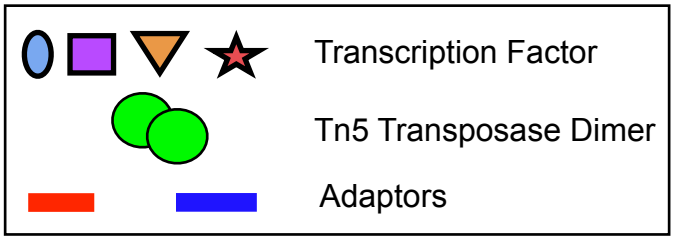
Open chromatin with DNase-Seq



Open chromatin with ATAC-seq

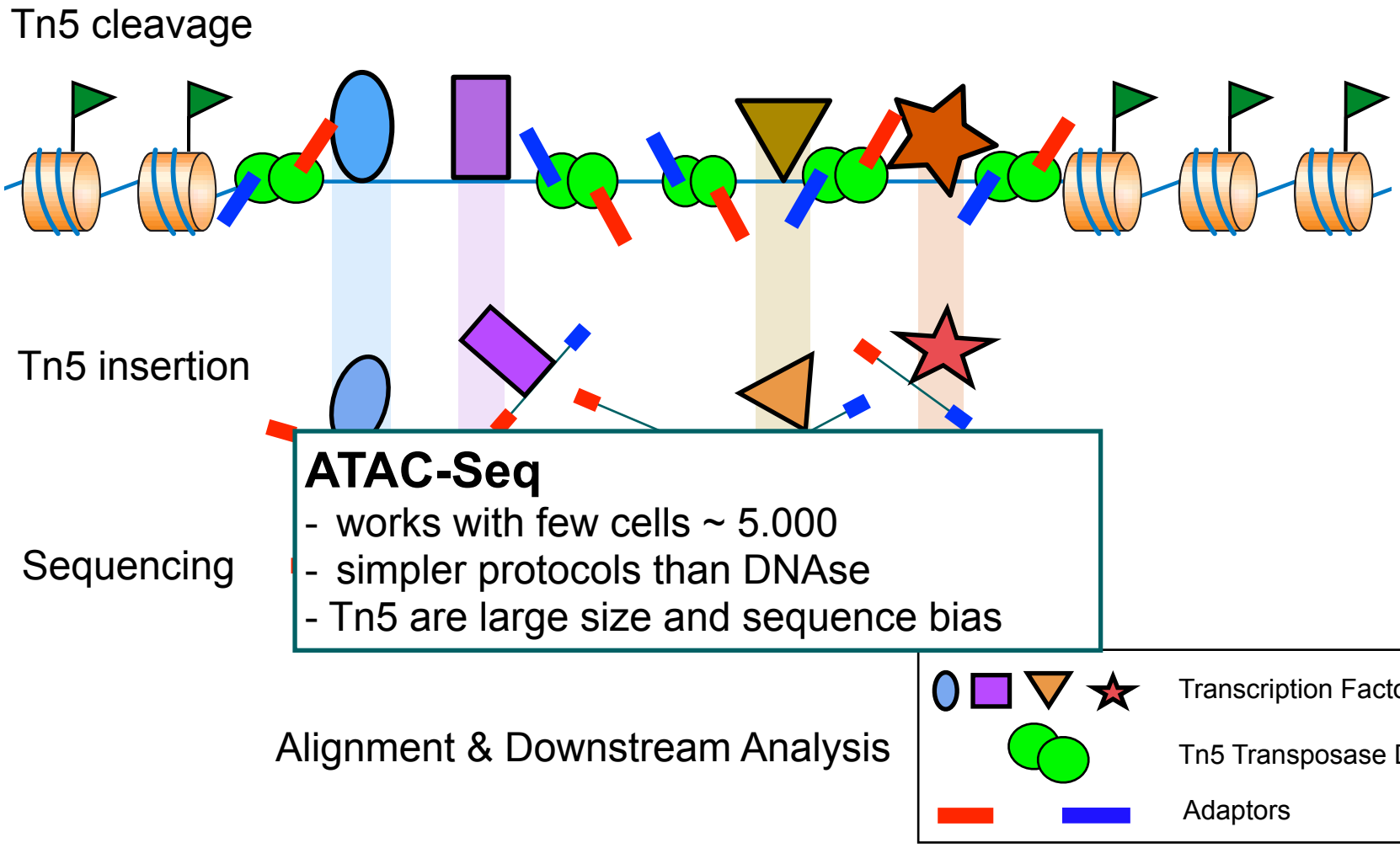


Alignment & Downstream Analysis



Li *et. al, unpublished*

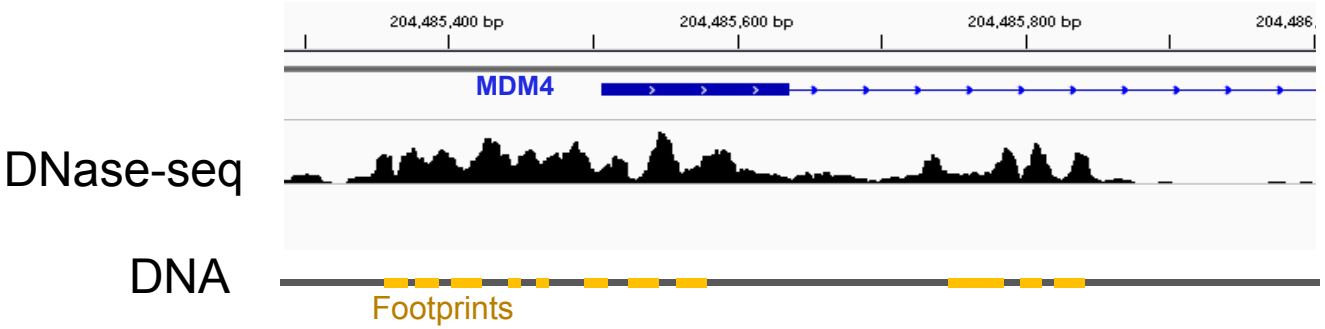
Open chromatin with ATAC-seq



Li *et. al, unpublished*

Computational Footprint Methods

Segmentation-based: search DNase-seq signal for footprint patterns



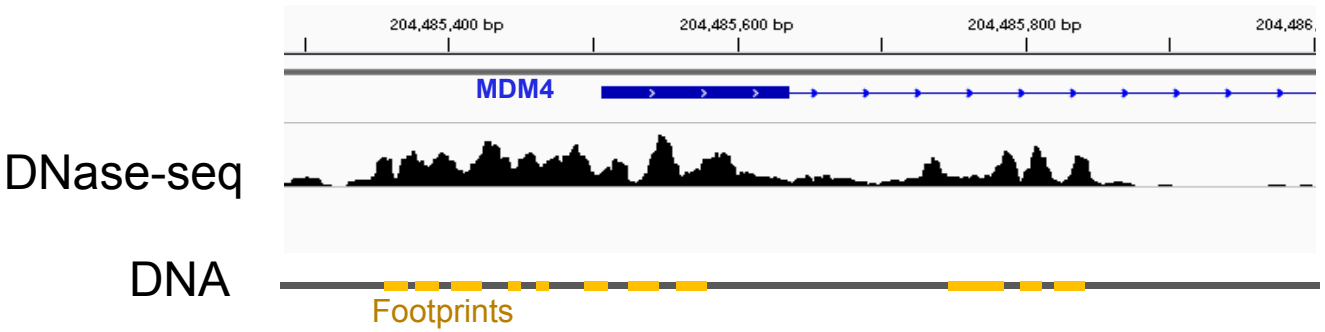
Window based:
- Neph, 2012 (ENCODE)
- Wellington (Piper, 2013)
- DNAse2TF (Sung, 2014)

HMM based:
- Boyle, 2011
- HINT (Gusmao, 2014)

*** only applicable if "true positives" TF are known**

Computational Footprint Methods

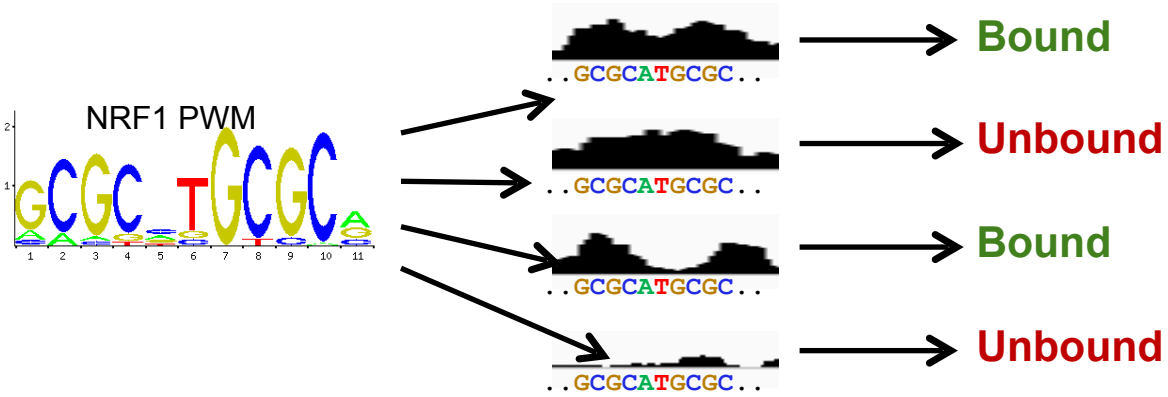
Segmentation-based: search DNase-seq signal for footprint patterns



Window based:
- Neph, 2012 (ENCODE)
- Welligton (Piper, 2013)
- DNAse2TF (Sung, 2014)

HMM based:
- Boyle, 2011
- HINT (Gusmao, 2014)

Site-centric: Classify motif-predicted binding sites as "bound" or "unbound".



Probabilistic/Bayesian Models:
- Centipede (Pique-Regi, 2012)
- Cuellar-Partida, 2012
- PIQ (Sherwood, 2013)
- FLR (Yardimci, 2014)

Supervised Classifier: *
- BinDNase (Kahara, 2015)

*** only applicable if "true positives" TF are known**

Calendar

16.04.2018 - Introduction to Biology & Bioinformatics

30.04.2018 - Example of NGS pipeline / data formats

07.05.2018 - Intro to RGT / Problem definition

… - Project development

16.07.2015 - Final Presentation

Thank you!