Bioinformatics Lab

Ivan Gesteira Costa, Mingbo Cheng, Zhijian Li, Martin Manolov

Institute for Computational Genomics

Objectives

- Hands on introduction to bioinformatics programming
- Review basic biological/computational aspects
    1. basics of molecular biology
    2. basics of sequencing
    3. basics bioinformatics problems
        - short sequences read alignment
        - gene expression quantification
        - computational epigenetic
        - single cell approaches

Objectives

- **Introduction to Bioinformatics Frameworks/Tools**
  1. **biological sequence data formats/handling**
     - **Biopython, Pysam, R/bioconductor**
  2. **bioinformatics tools**
     - **BWA (aligner), Seurat, Cell Ranger, …**

Grading/Online material

Evaluation:

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

Extra-work for media informatics:

- research report

References/Courses Online

http://costalab.org/teaching/bioinformatics-software-lab-2021/

Introduction to Molecular Biology

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

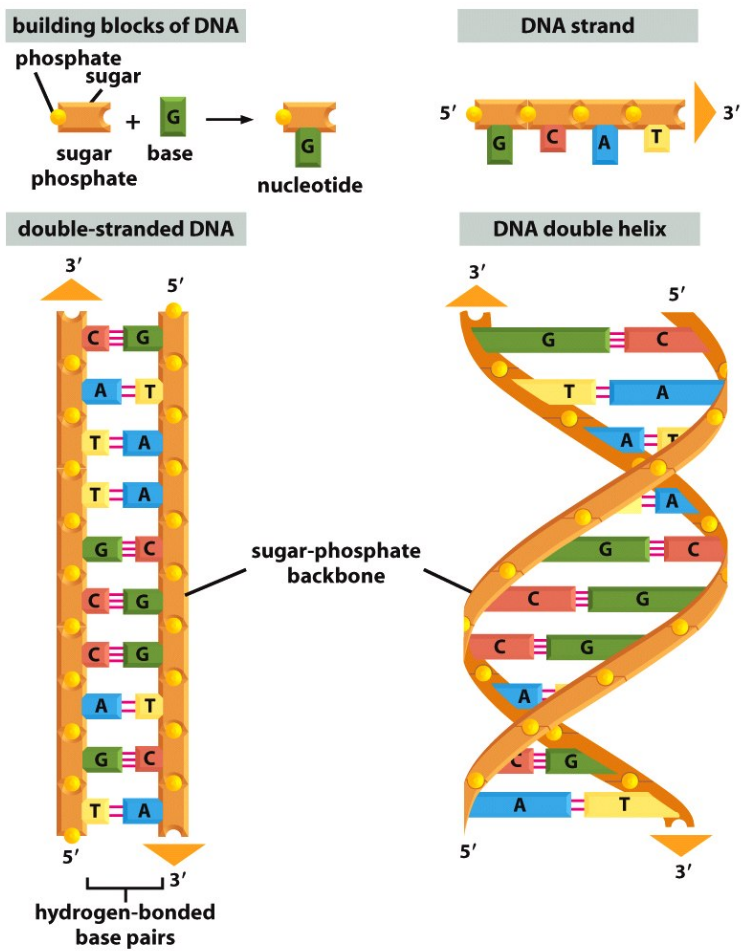
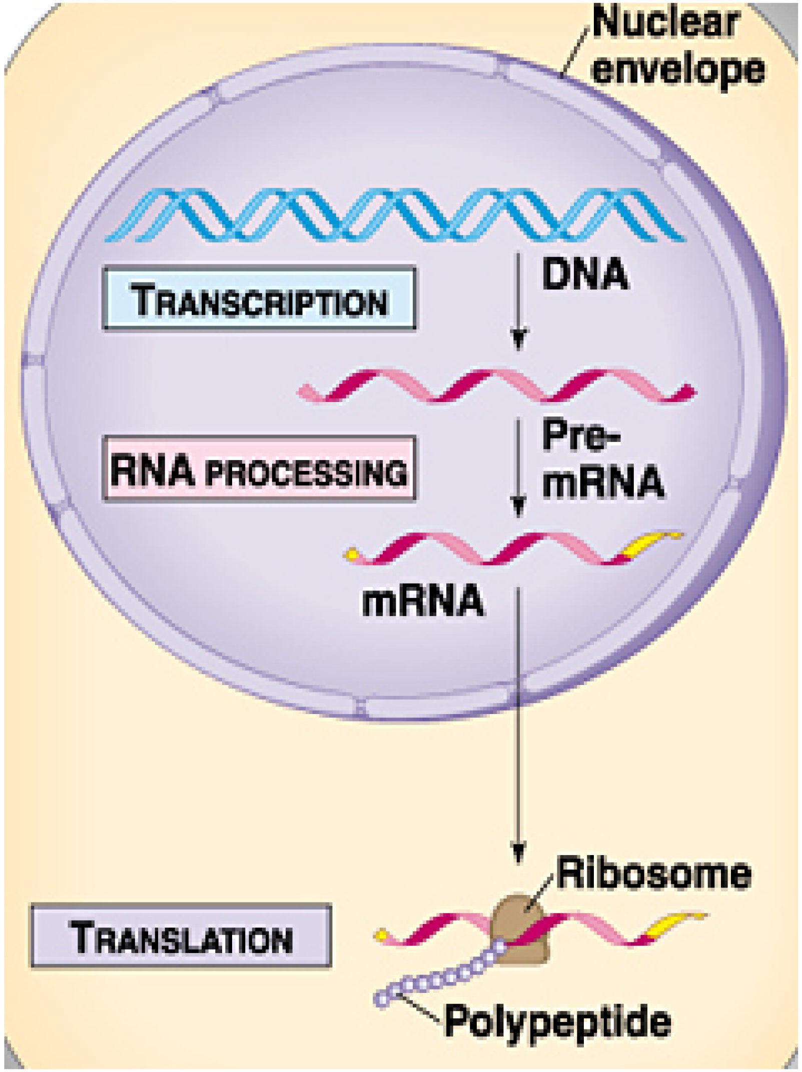


Figure 4-3 Molecular Biology of the Cell 5/e (© Garland Science 2008)

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

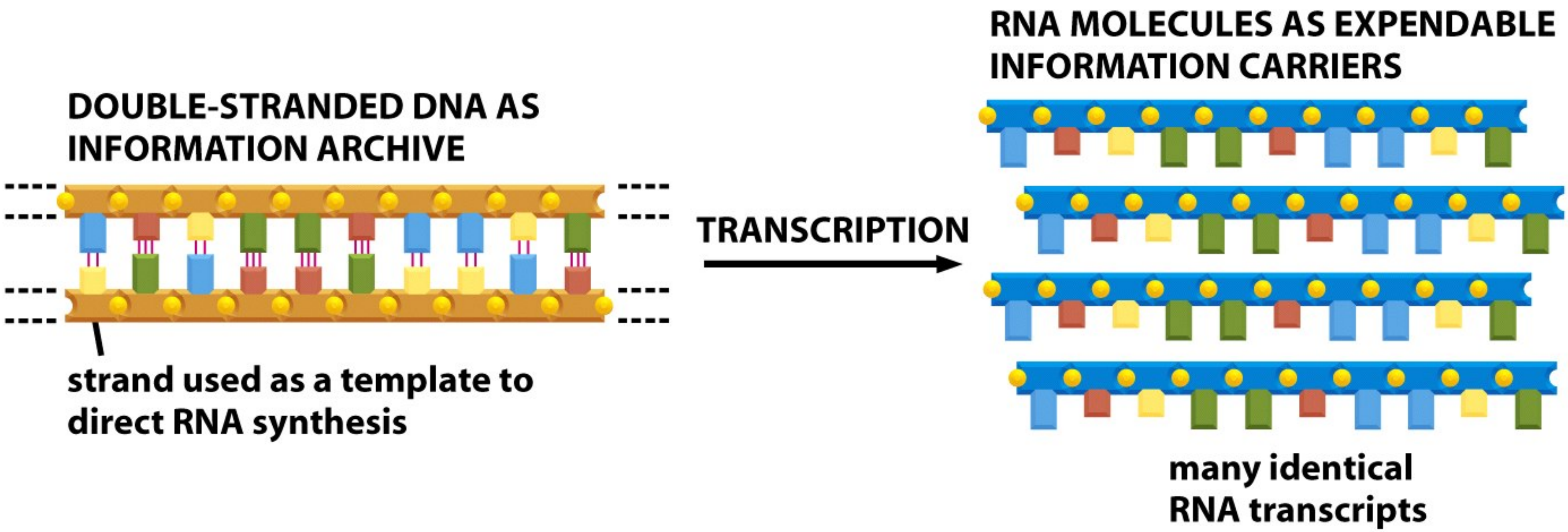
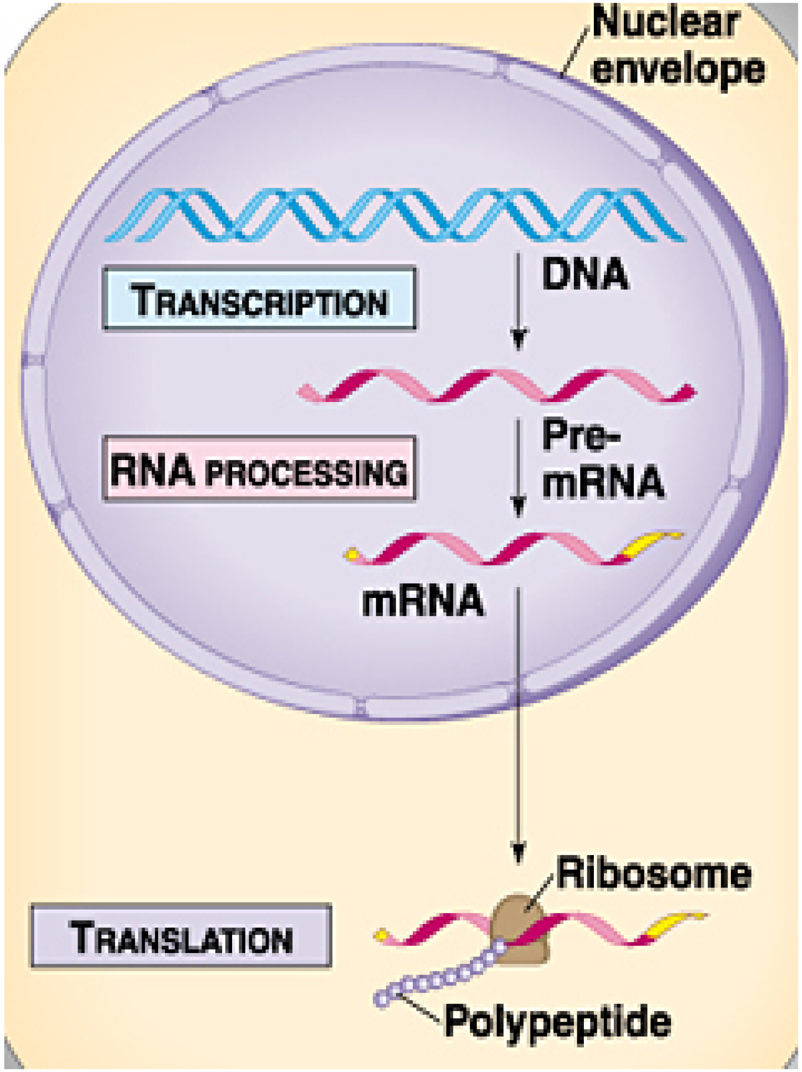


Figure 1-5 Molecular Biology of the Cell 5/e (© Garland Science 2008)

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

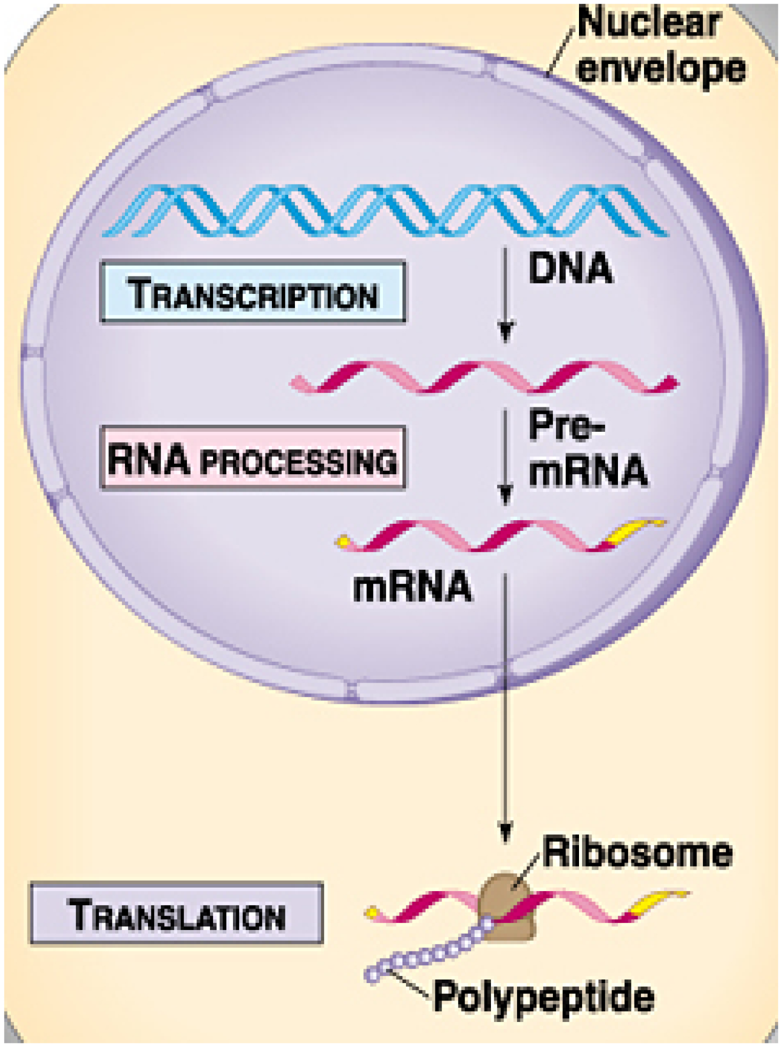
Genetic Code - Translation

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

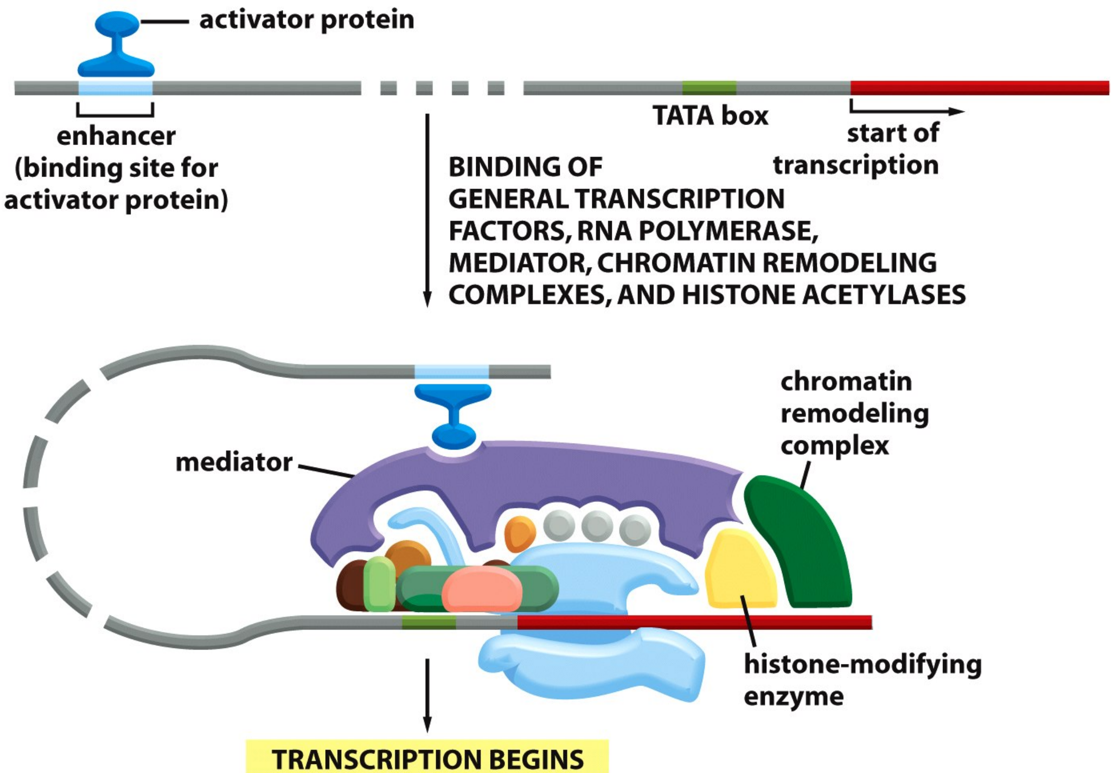


Figure 6-19 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Gene Expression

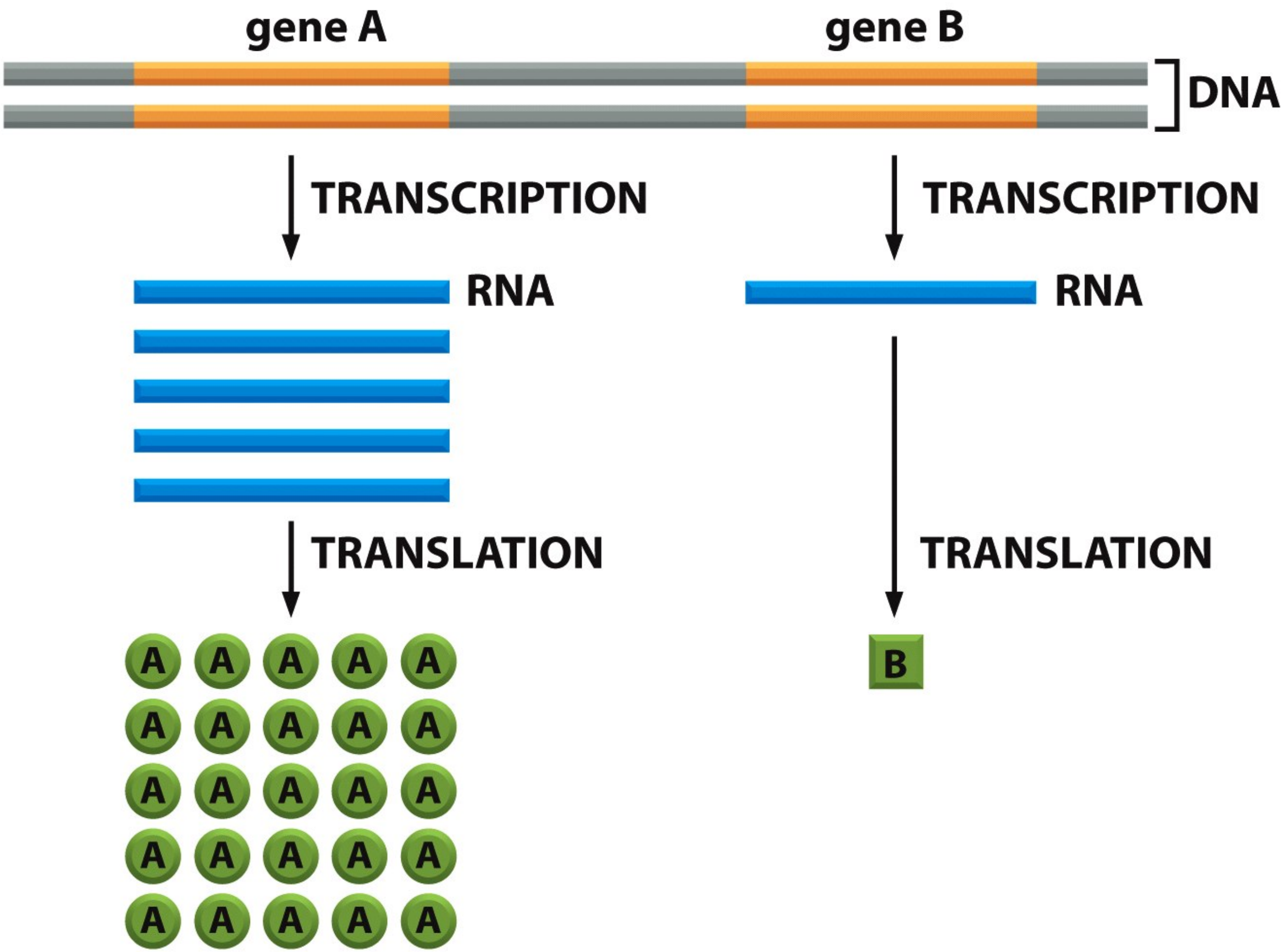
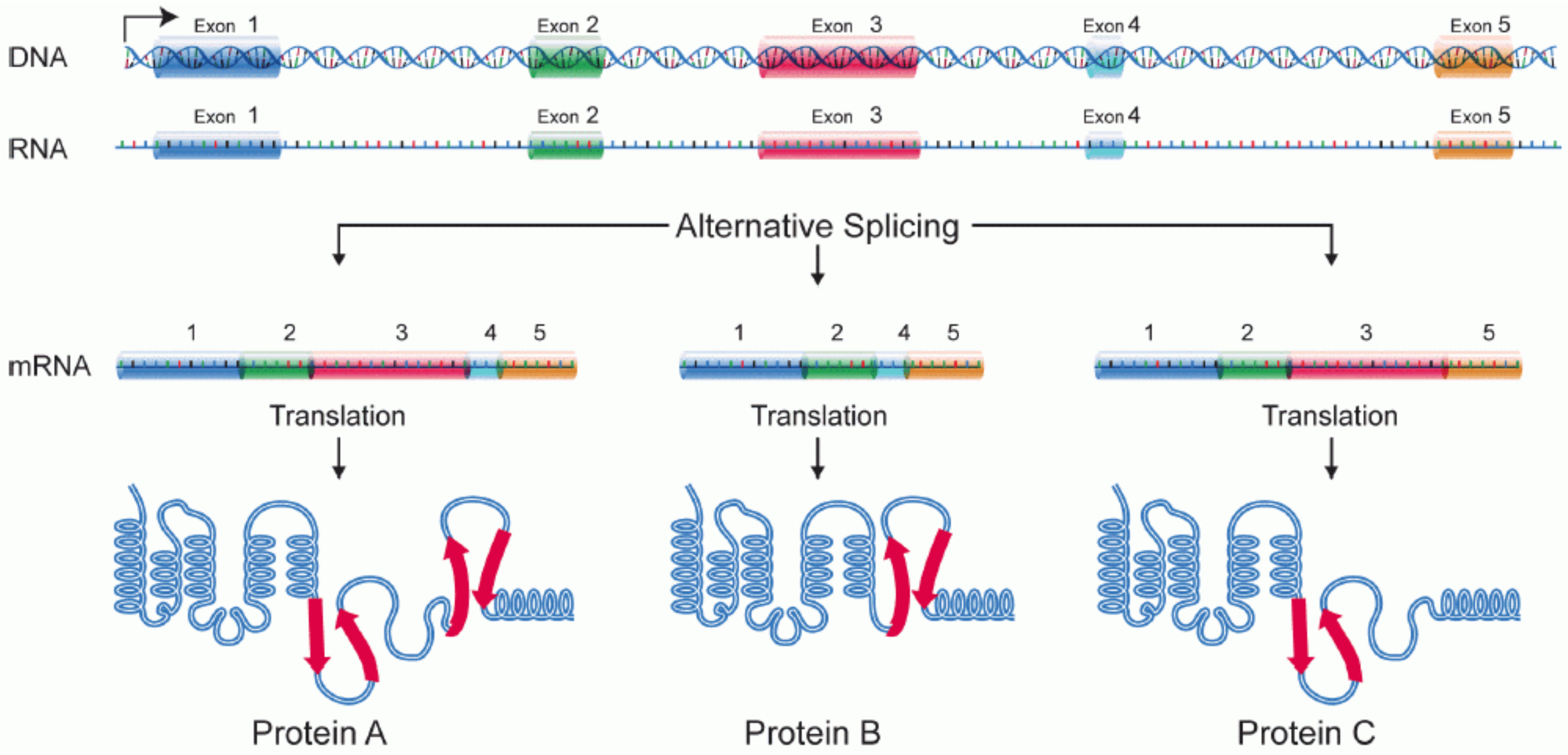


Figure 6-3 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Gene / Alternative Splicing



Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

Cellular Complexity

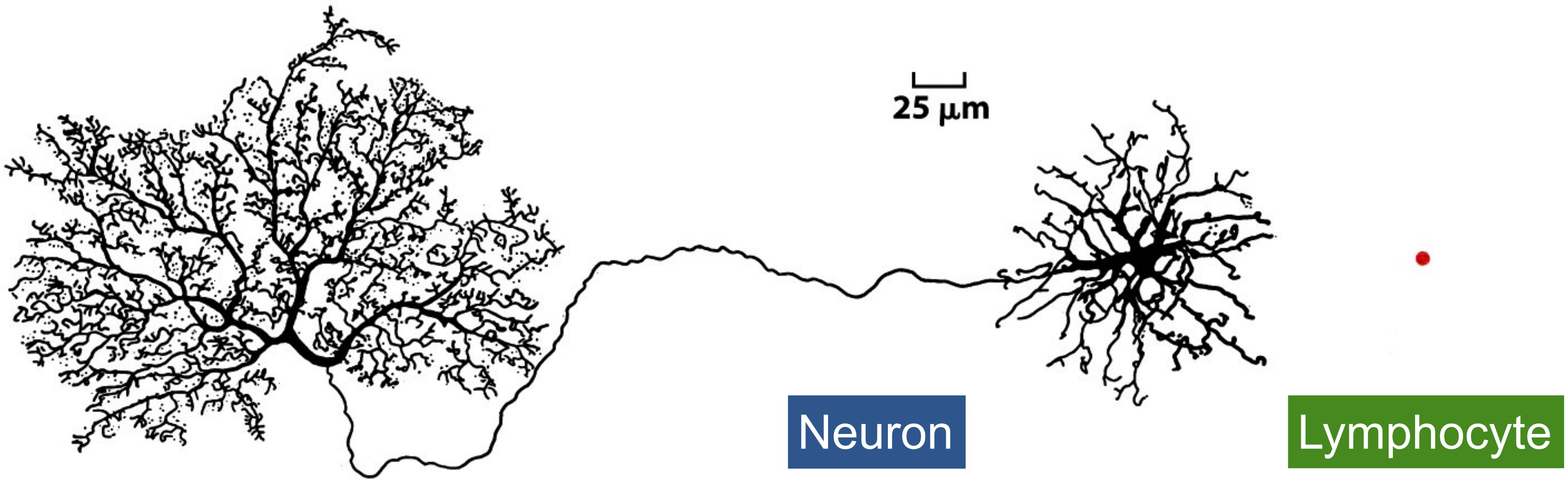


Figure 7-1 Molecular Biology of the Cell 5/e (© Garland Science 2008)

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**

Cellular Complexity & Gene Expression

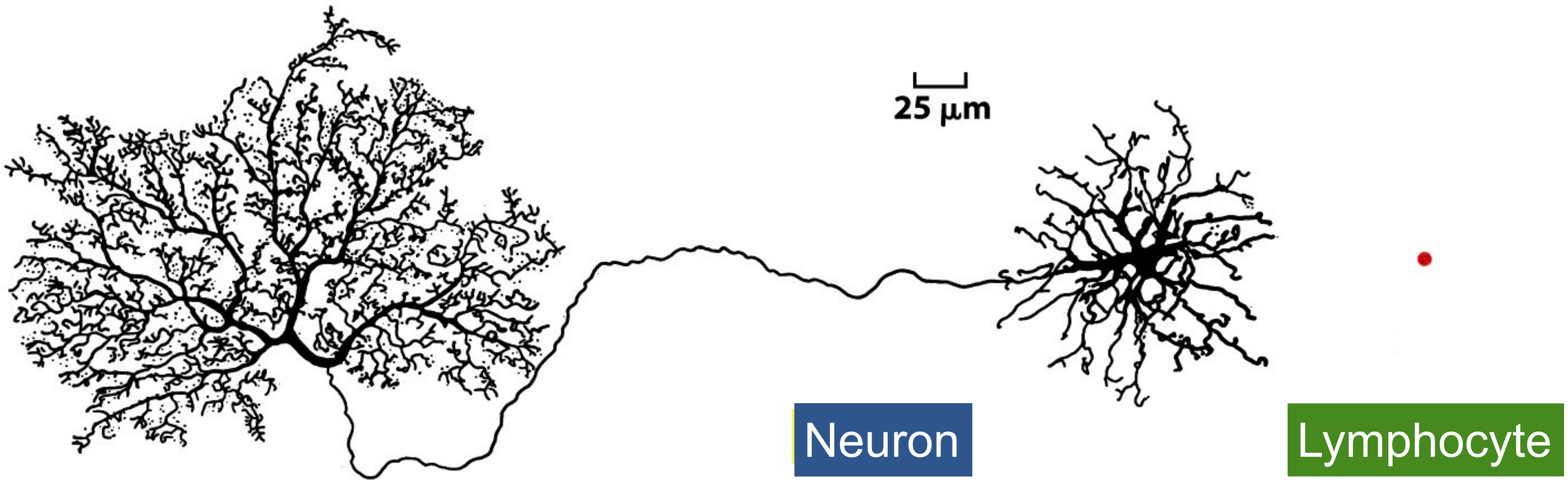
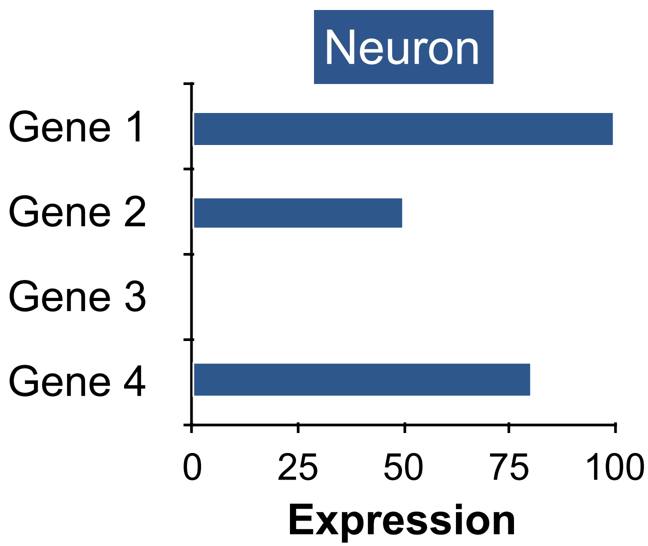
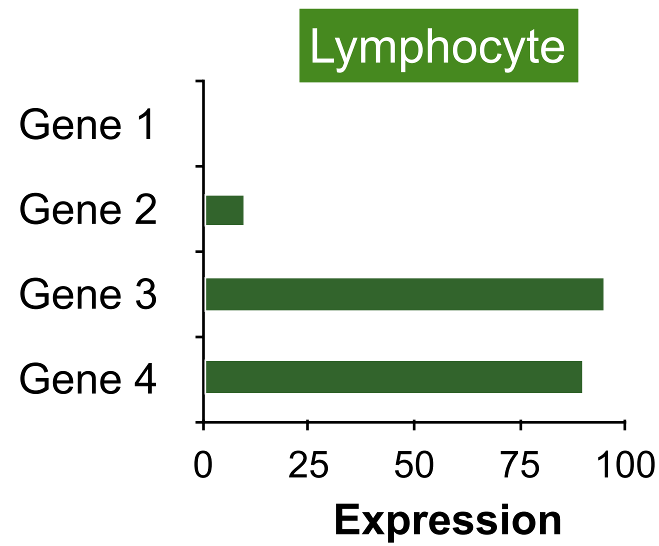


Figure 7-1 Molecular Biology of the Cell 5/e (© Garland Science 2008)





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 or open chromatin

Problem:

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

Solution:

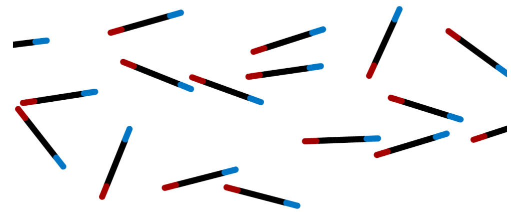
break DNA in several small pieces and use **bioinformatics**

Next Generation Sequencing

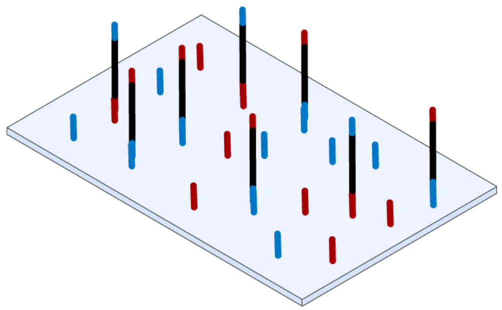
- NGS take advantage of parallelization
  - reads millions/billions of reads for a time
  - short reads (50-100 bps)
  - moderate error rates (0.1%)
- commercial products:
  - 454
  - SOLiD
  - Solexa (Illumina)

Illumina Flow Cell - NGS Sequencing

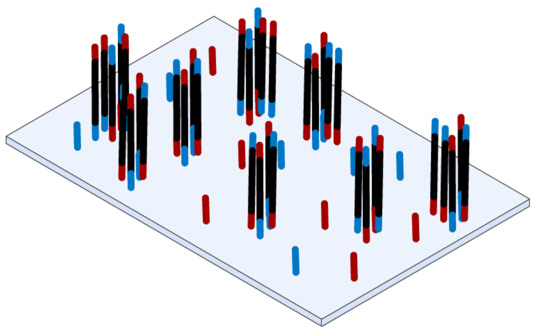
1- fragment sample DNA, insert adapters, attach to flow cell



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



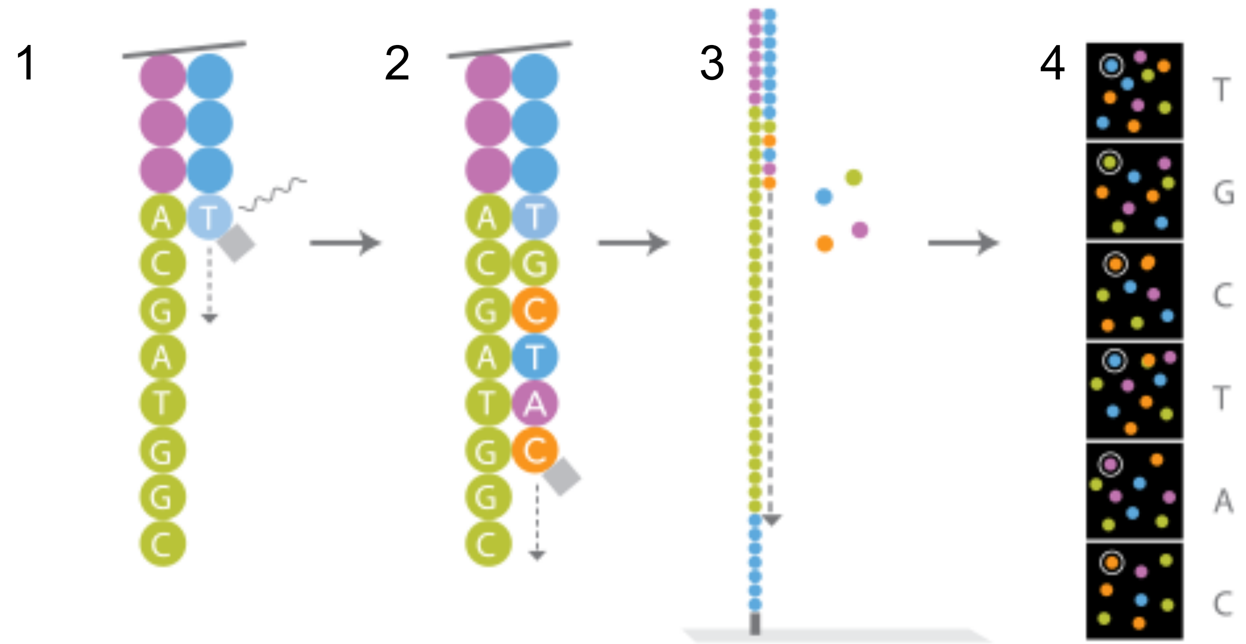
3- clusters of single stranded DNA (200m clusters with 2k DNA strands



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

Header

Sequence

```
@ILLUMINA-C90280_0030_FC:5:1:2675:1090#NNNNNN/1
ATTCCCGGCCTTTTTCCAGGCCTGCCTGCTCGAGC
+
BAAAGECEE<EEDFEDF3DBDDBB=A+==>9>>88?
```

Qualities
(prob. that base call is wrong)

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

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

<u>Phred-scale</u>

Q = -10 * log10 P

P = 10^(-Q/10)

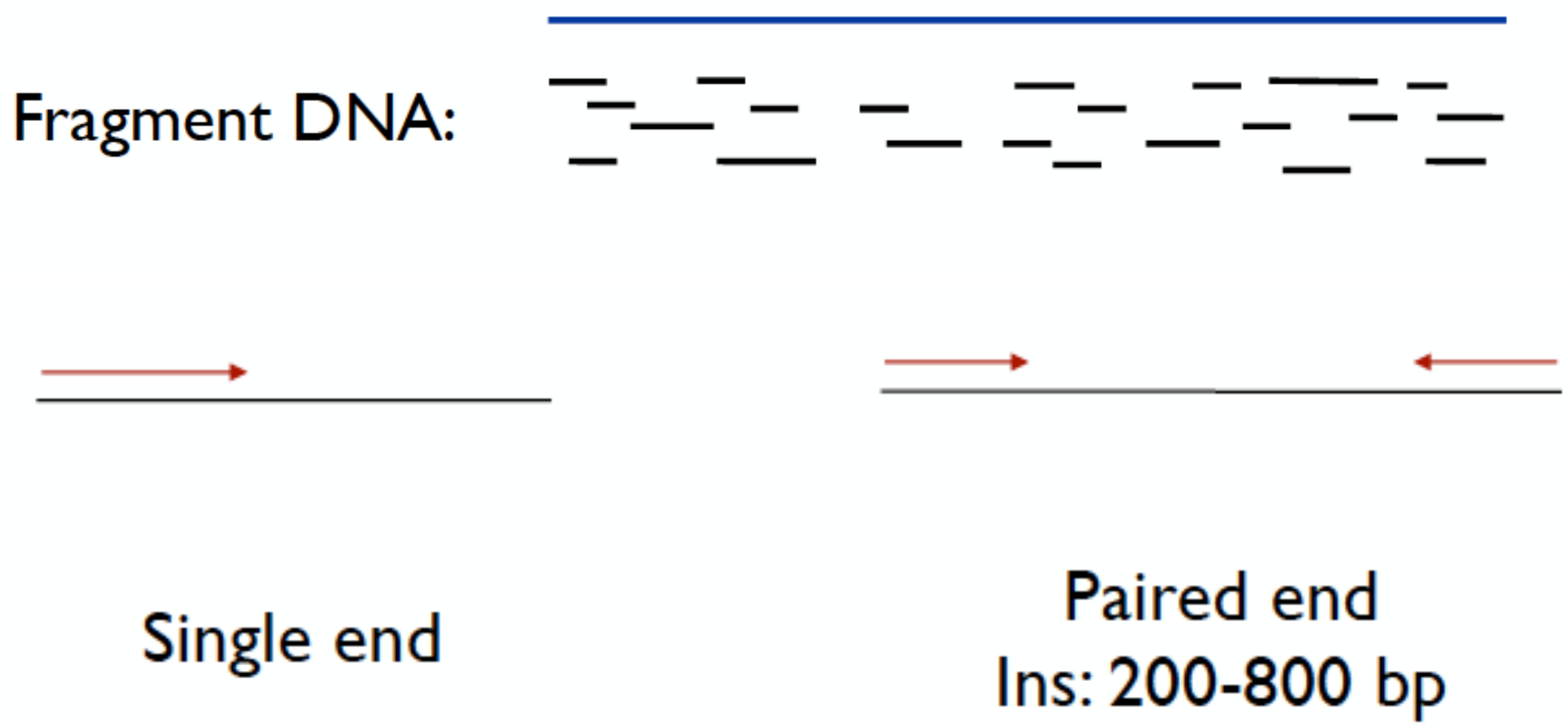
Sequencing Results / Phred scores

Uses letters/symbols to represent numbers:

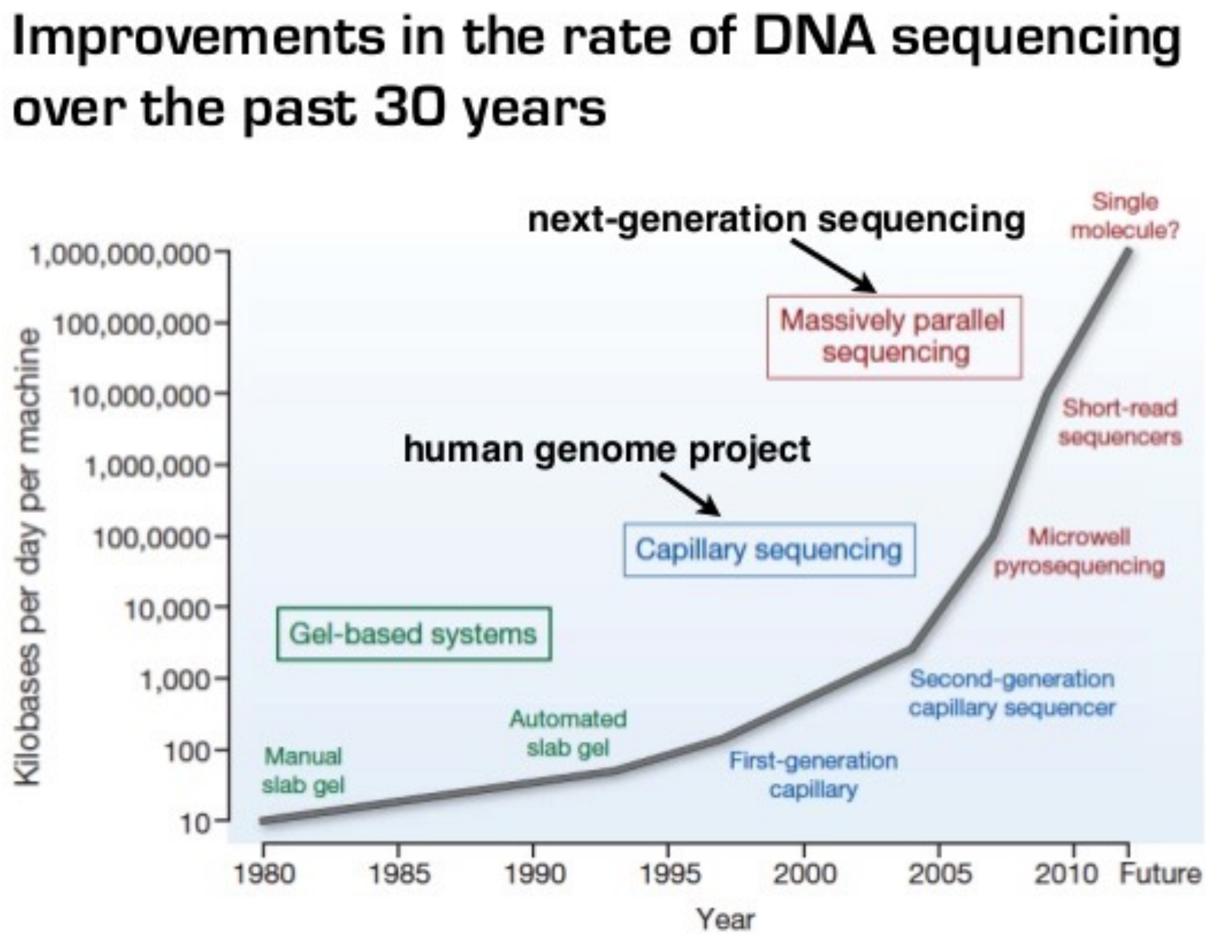
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ

Q0	Q10	Q20	Q30	Q40
bad	*maybe*	*ok*	*good*	*excellent*

Read Types

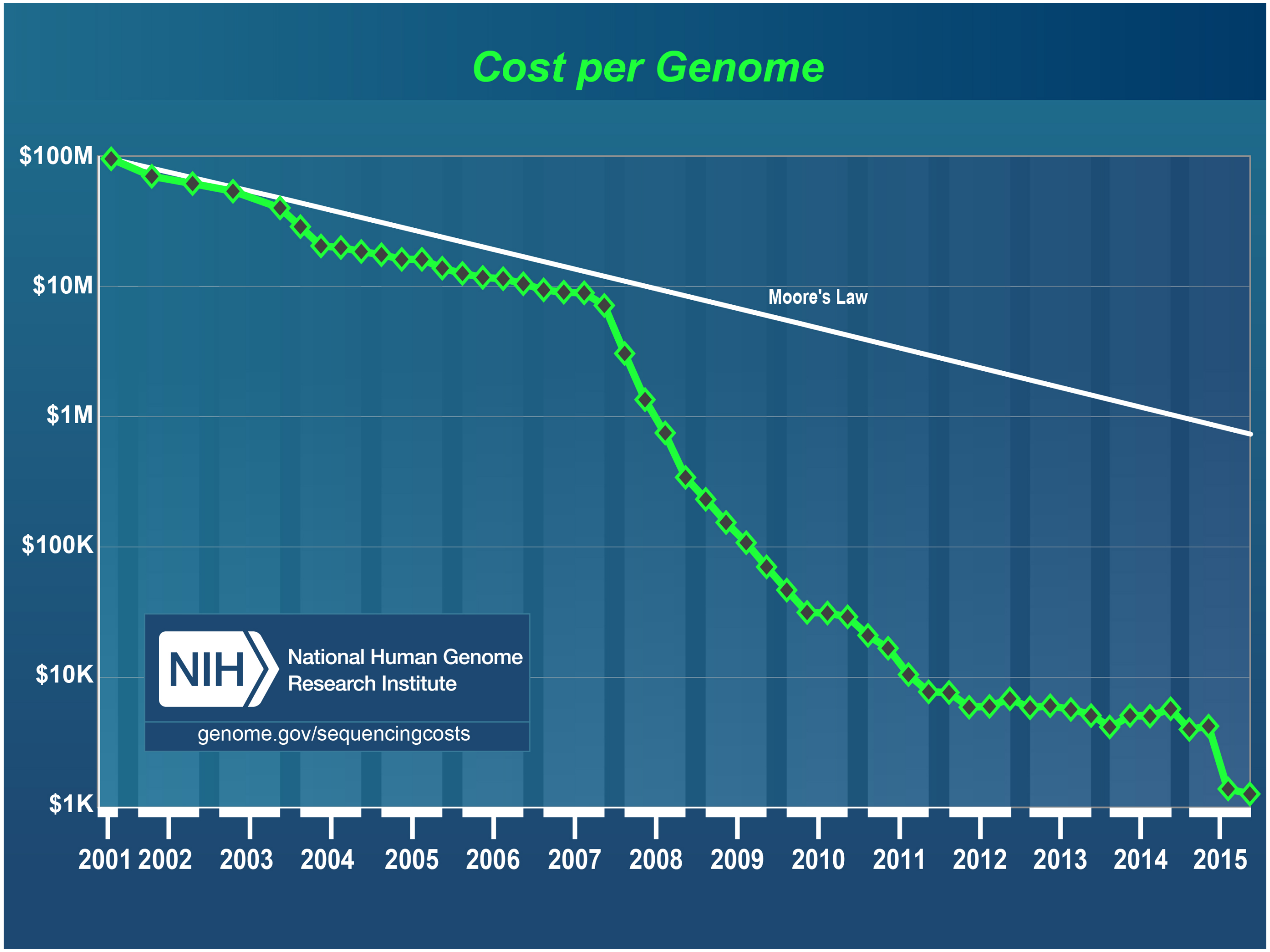


Next Generation Sequencing



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

Sequencing Costs



Sequence Alignment

Sequence Alignment

NGS

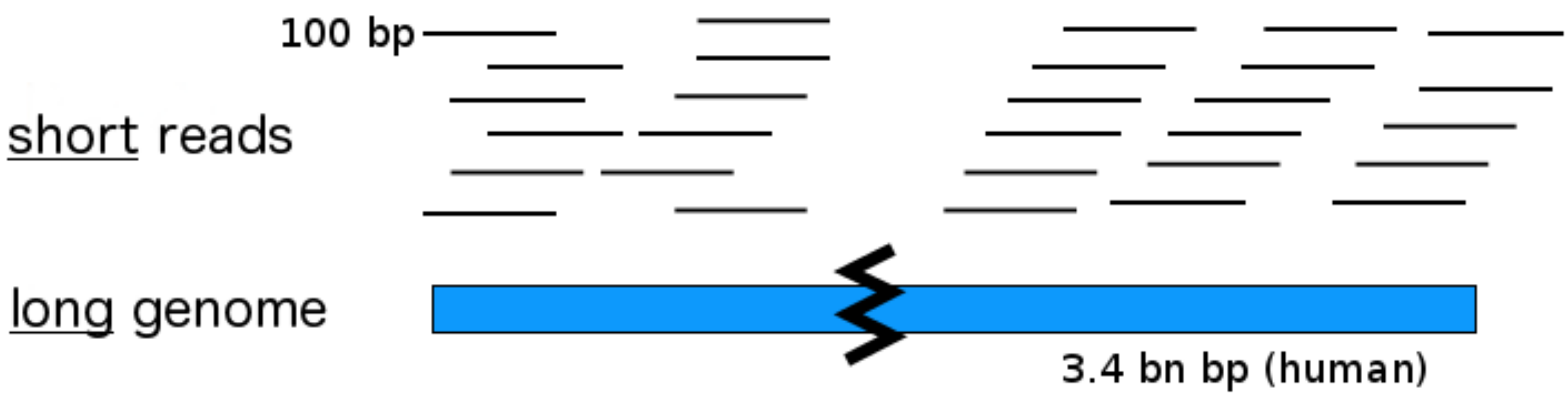
- reads from DNA fragments
- position in genome is unknown
- solution: alignment

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
- millions of 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:

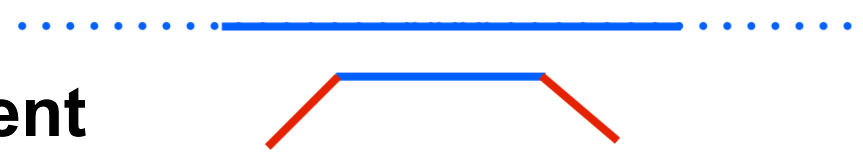
- **Smith & Waterman - dynamic programming (quadratic time/memory)**

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**



11th and 13th most cited papers ever!!!

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 for k=8)**
  - **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 for k=8)
    - 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 for k=8)
    - 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!*

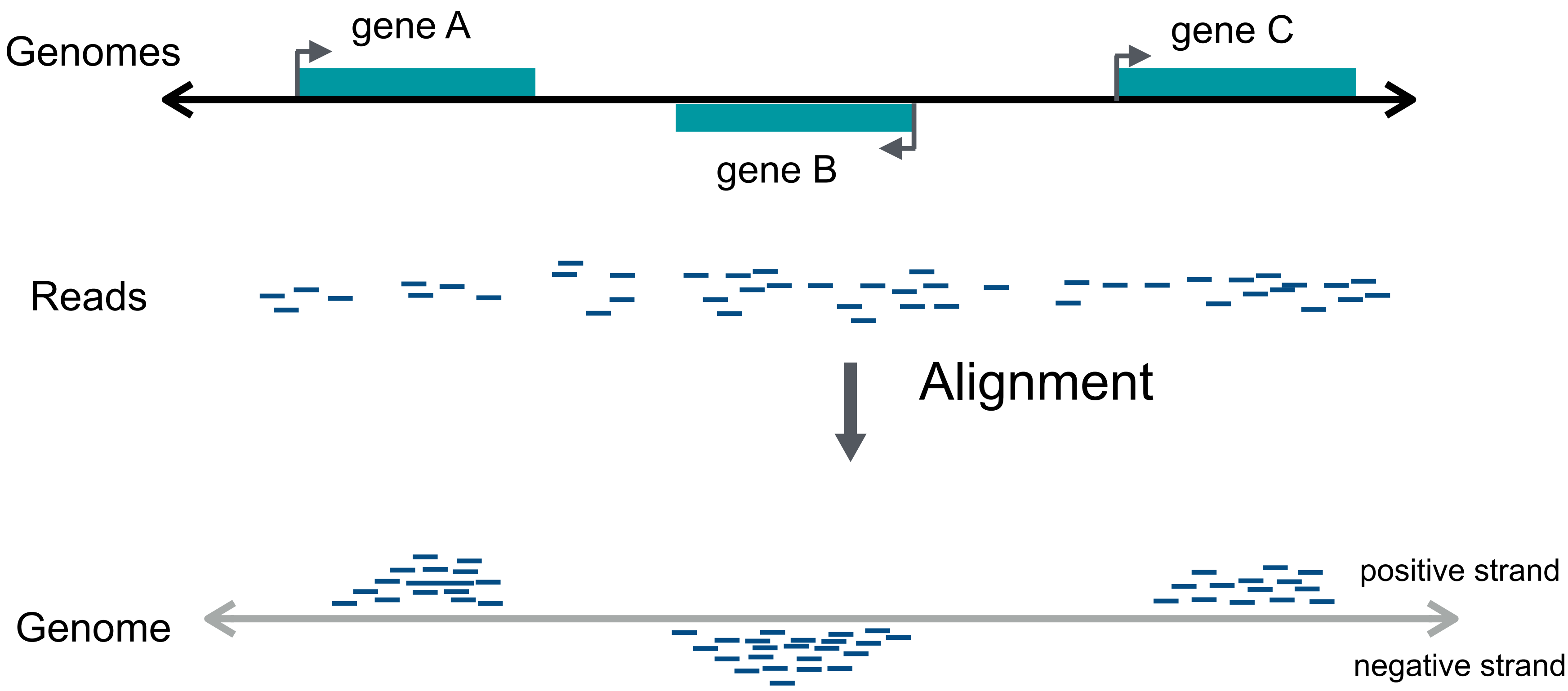
1. Index the reference using *k*-mers.
2. Search reads vs. hash *k*-mers
3. Perform alignment of entire read around seed
4. Report best alignment

Key	Value
⋮	⋮
`ACTGCGTGTGA`	`Chr1_pos1234; Chr2_pos567`
`ACTGCGTGTGC`	`Chr7_posX`
`ACTGCGTGTGT`	`Chr7_posZ; ...`
⋮	⋮

Also known as *Seed and extend*

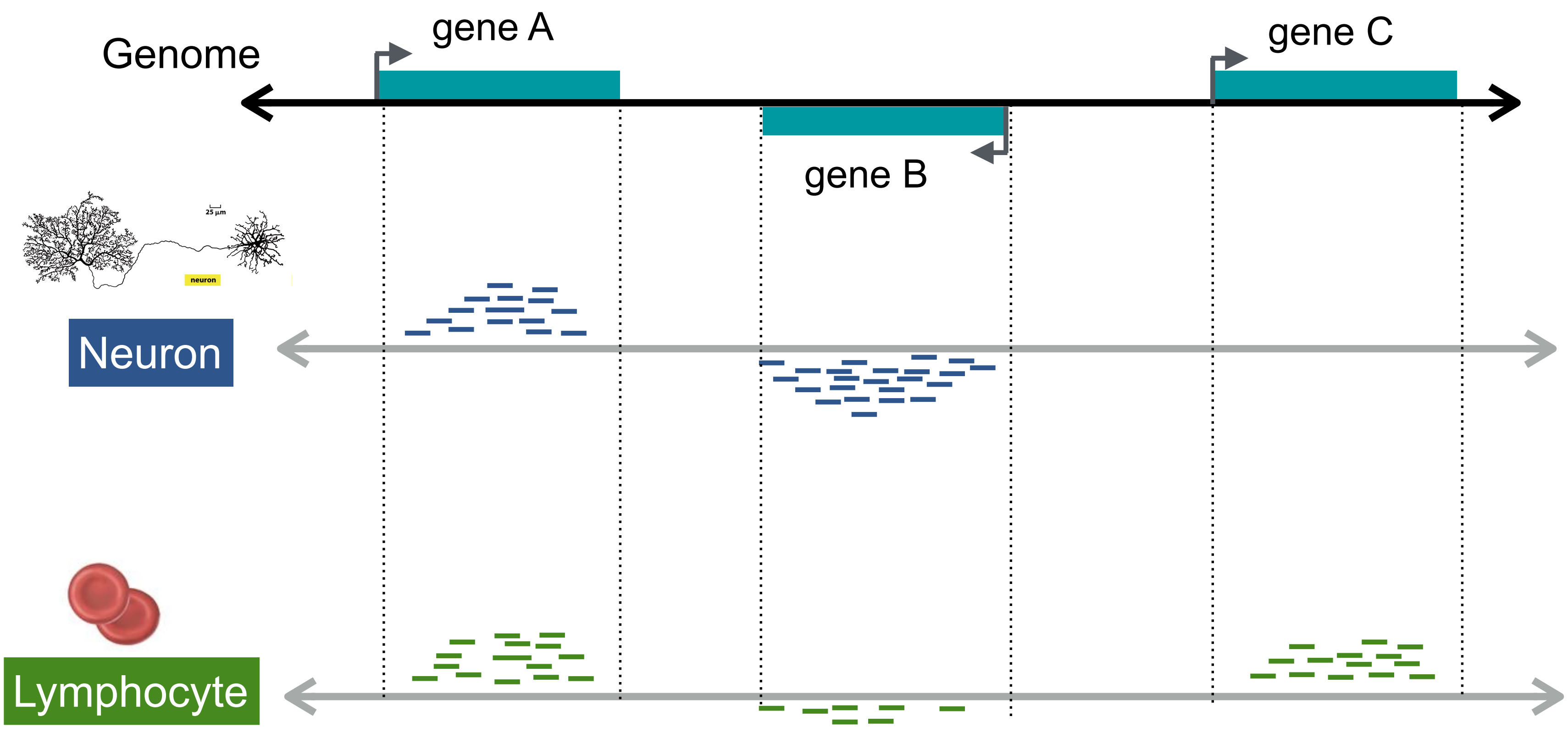
RNA sequencing / Alignment Results

- **Position and strand of reads aligned to the genome**



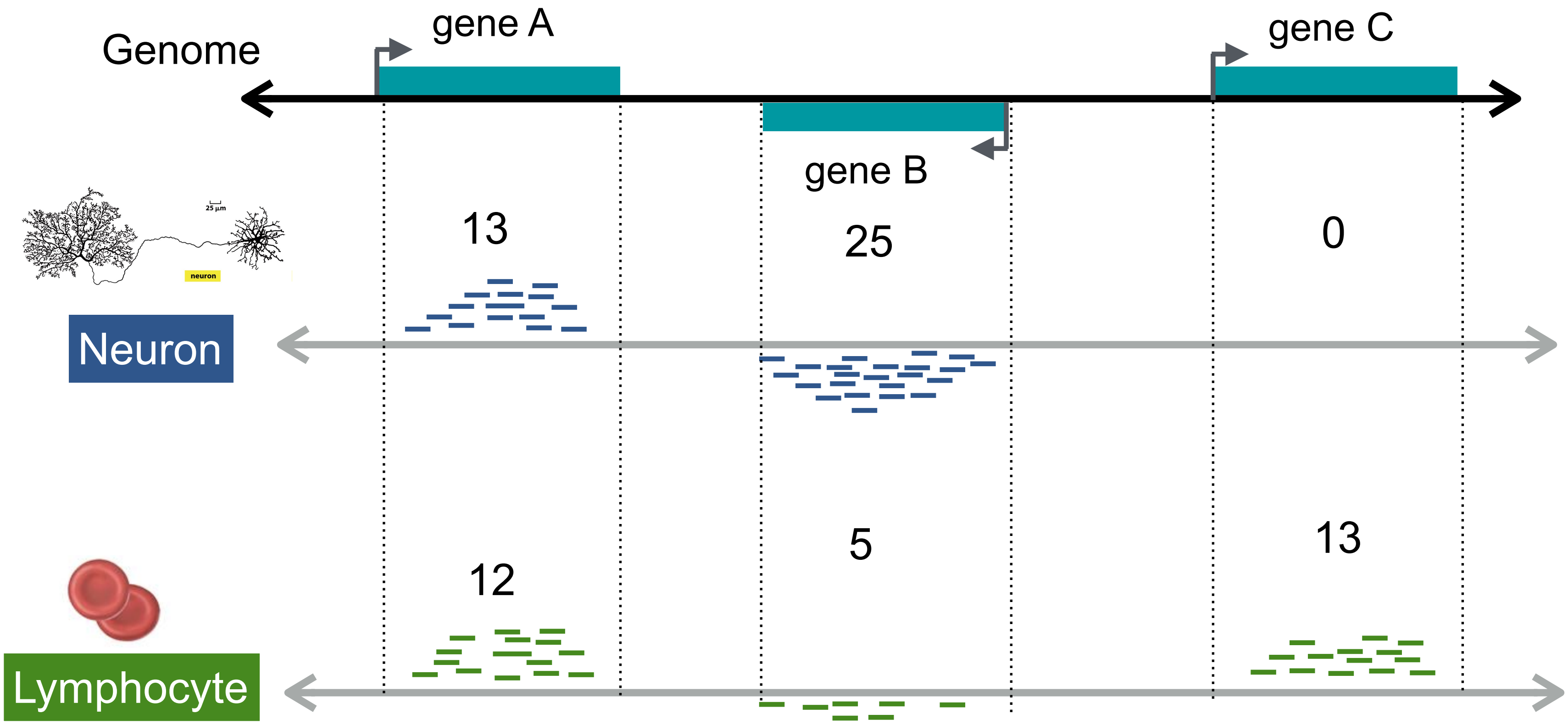
Gene Quantification

- Perform sequencing for each cell (neuron, lymphocyte)
- Align reads to genome

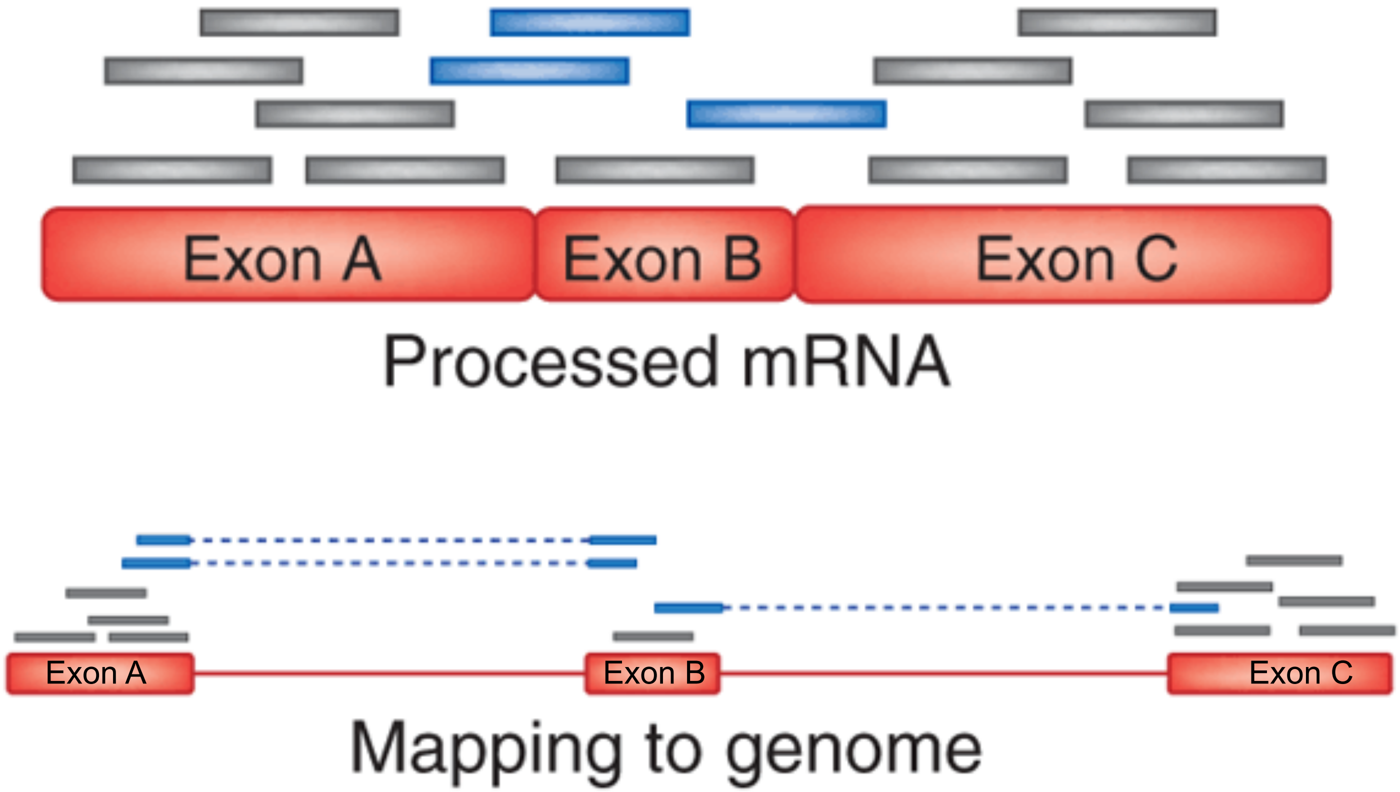


Gene Quantification

- Perform sequencing for each cell (neuron, lymphocyte)
- Align reads to genome
- Count number of reads inside genes (using known genes annotation)

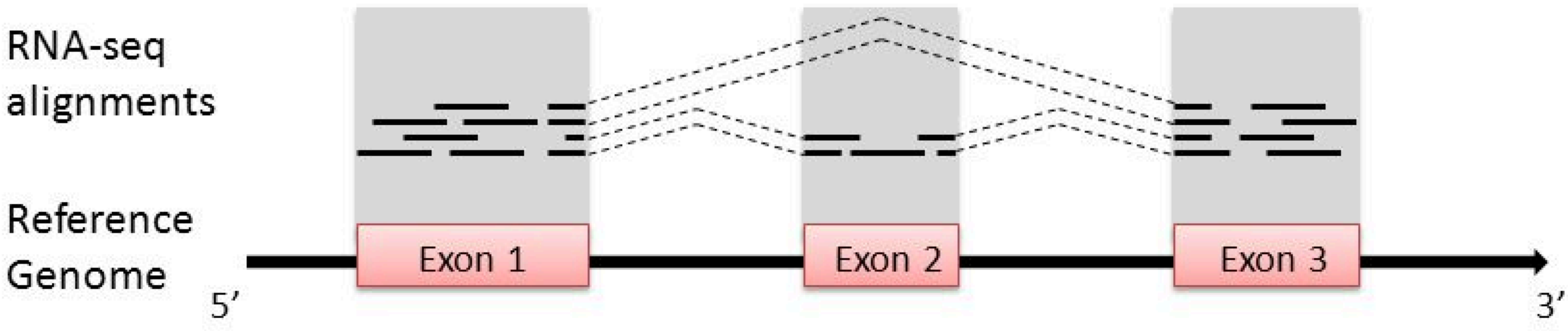


Alignment - Split Read Mapping (RNA-Seq)



- **reads needs to be split within intros when mapped to genome (special aligners / STAR)**

Quantification - Gene vs. Transcript vs. Exon



Counting Strategies

Gene Level - 17 reads
Exon level - exon 1 (8 reads), exon 2 (3 reads), exon 3 (6 reads)
Transcript Level - Exons 1,2 & 3 (10 reads) and exon 1 & 3 (7 reads) *
* complex computational methods required (TopHAT)

Quantificaiton - Normalization

- **Correct for:**
  - Genes having distinct size
  - Sequencing efficiency differs between cell (usually same RNA quantity provided for sequencing)

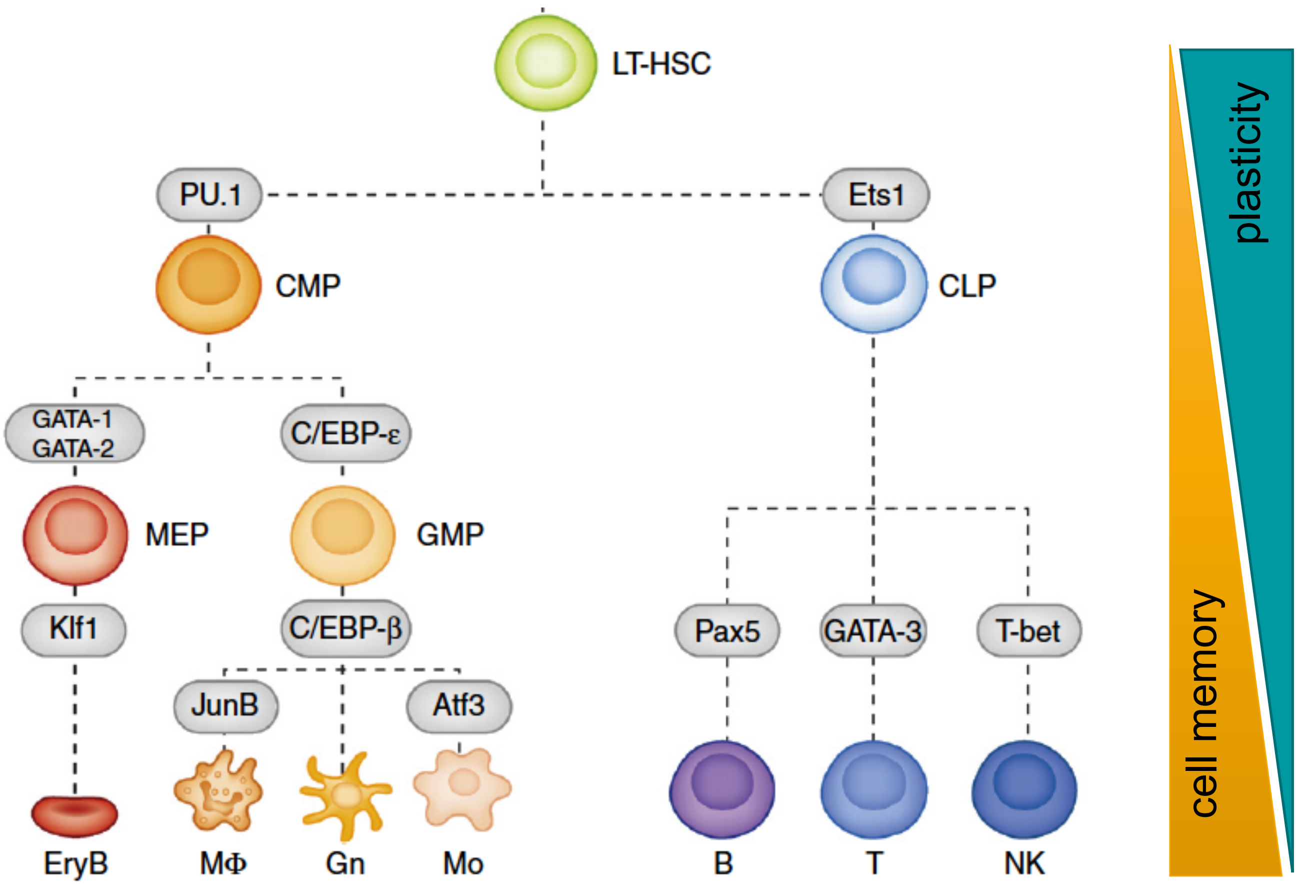
	Cell A	*Cell B*	*…*
GeneA (1kb)	*20*	*15*	*30*
GeneB (2kb)	*100*	*300*	*10*
GeneC (1.5kb)	*10*	*20*	*100*
Gene D (3kb)	*300*	*200*	*100*
Total Library	*430*	*535*	*240*

$$\text{Reads per kilobase million (RPKM)} = \#reads * \frac{gene\ size}{1.000} * \frac{total\ library}{1.000.000}$$

Computational Epigenomics

Cell Differentiation

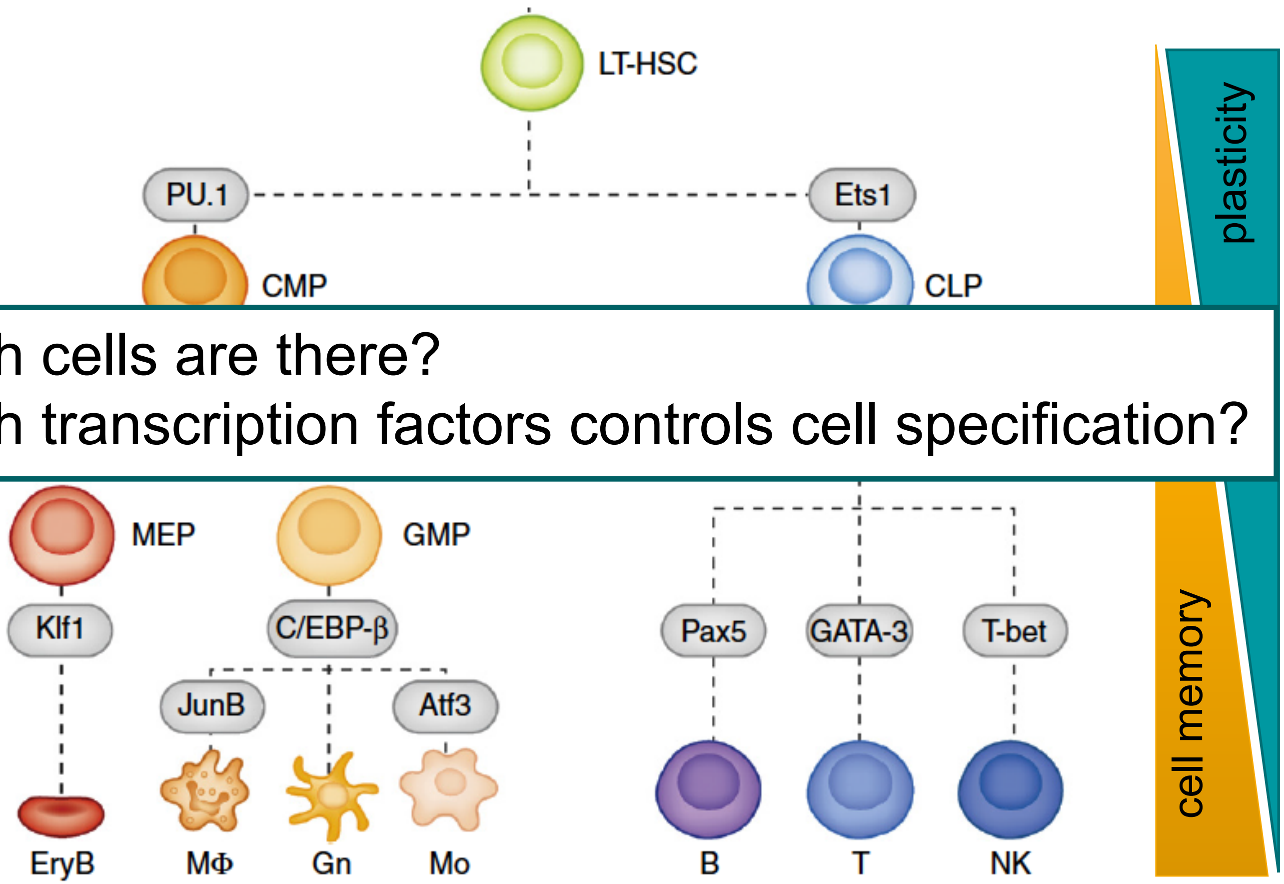
Hematopoiesis



Source: Amit (2016), *Nature Immunoloy.*

Cell Differentiation

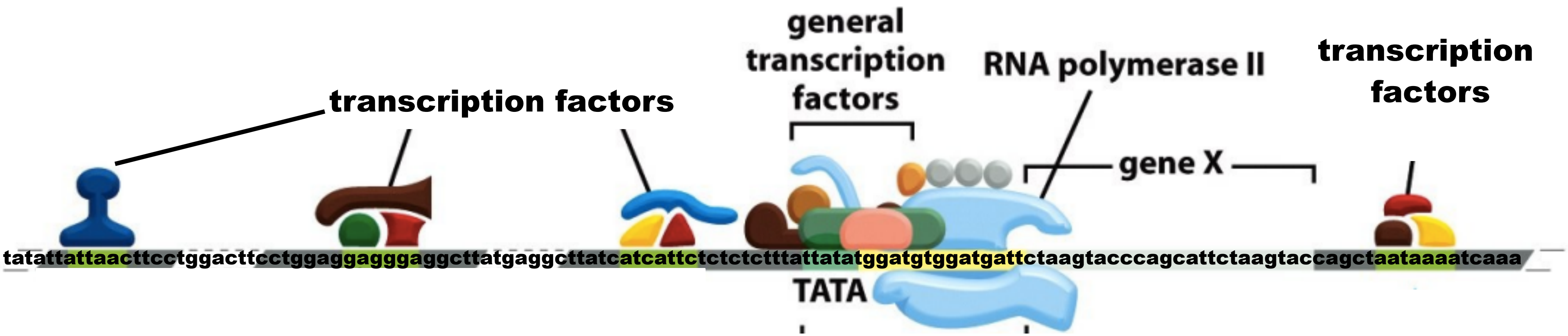
Hematopoiesis



Which cells are there?
Which transcription factors controls cell specification?

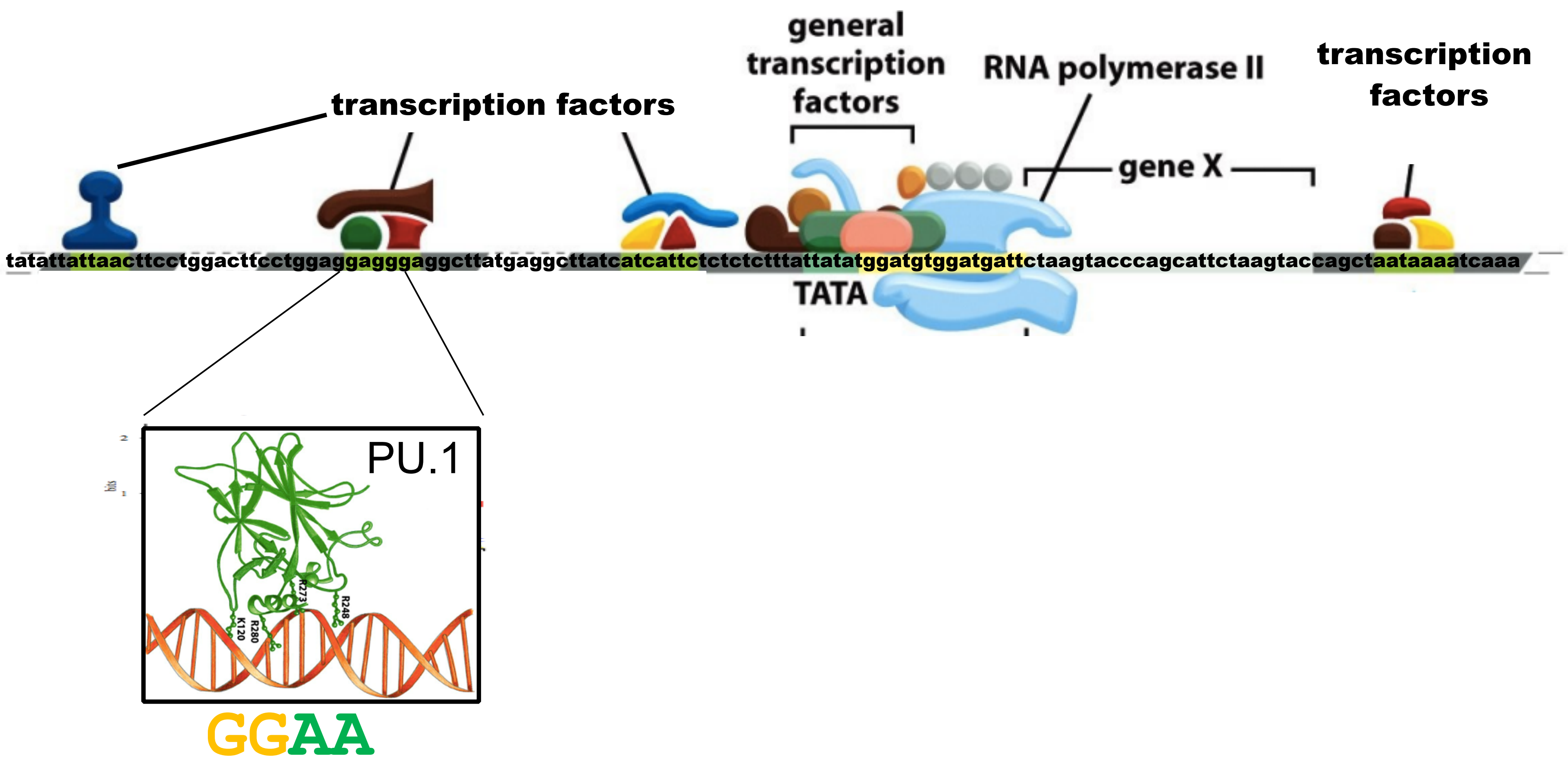
Source: Amit (2016), *Nature Immunoloy.*

Regulatory Control – Transcription Factor Binding



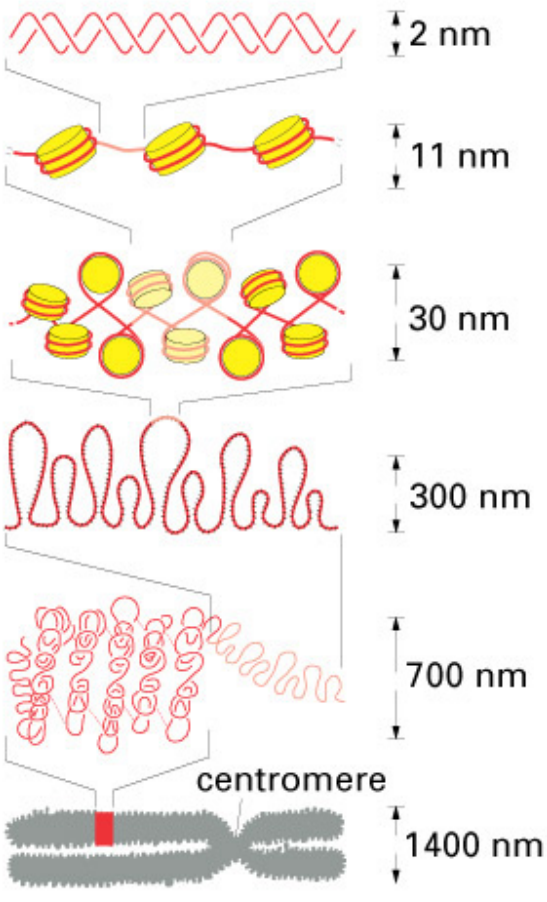
Source: Alberts, B. et al. (2008) Garland Science, 5th ed.

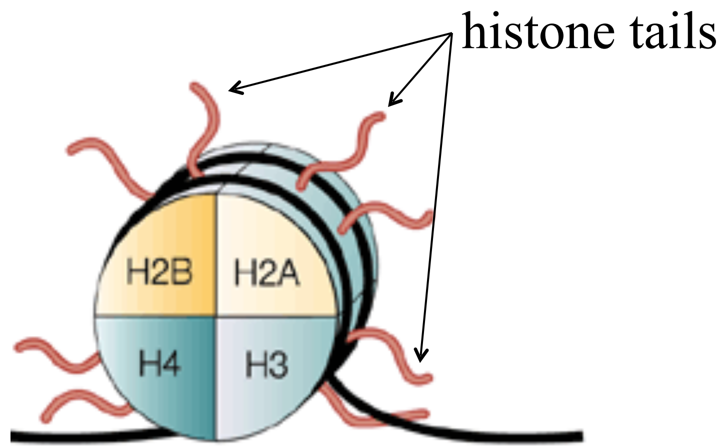
Regulatory Control – Transcription Factor Binding



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

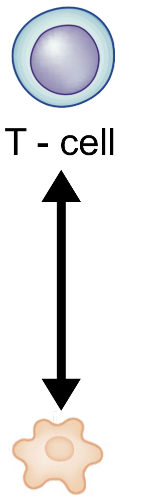
Epigenetics & Histones

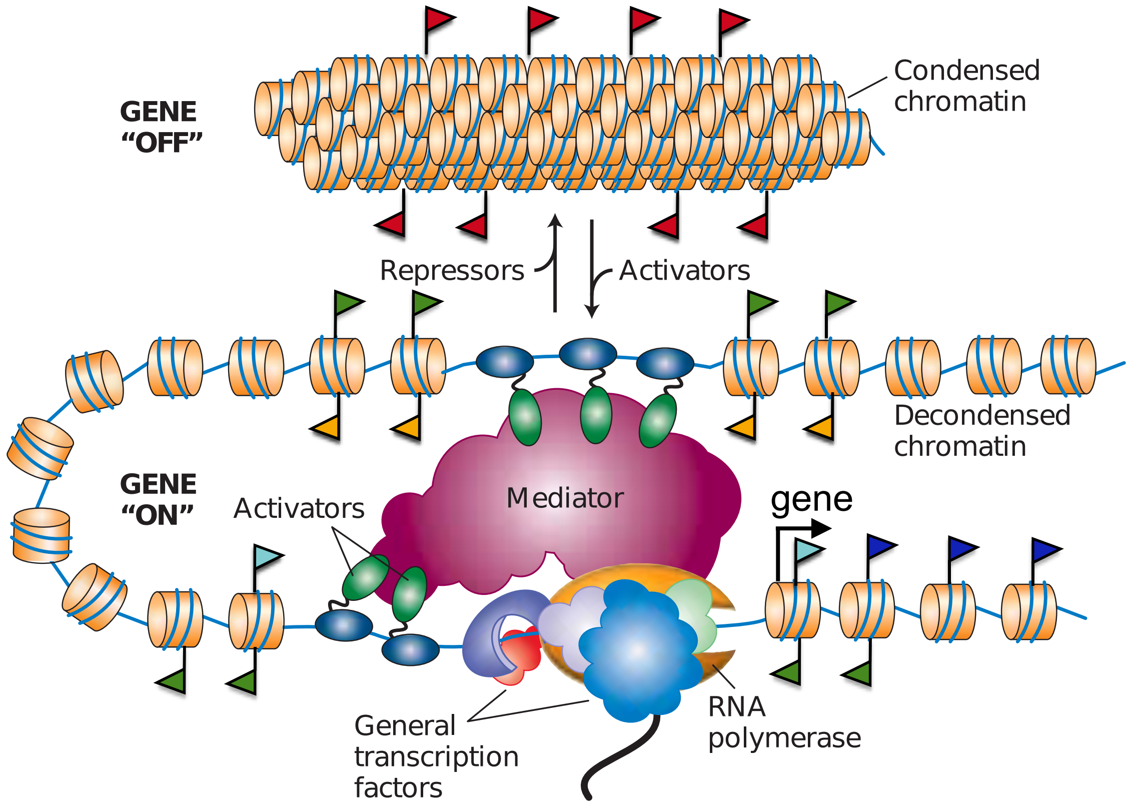




Modification in histone tails
- change strength of DNA binding
- recruit transcription factors

Chromatin, Regulation and Cellular Memory



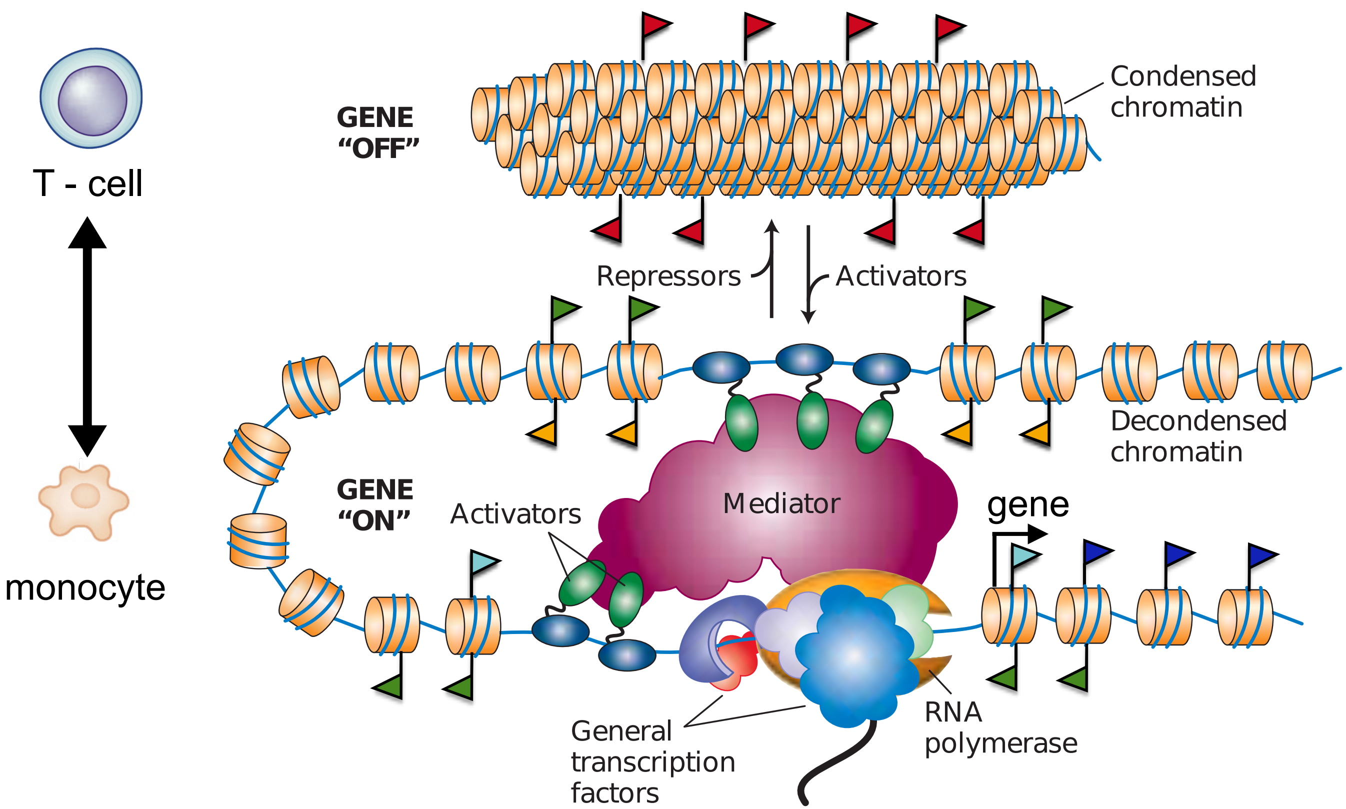


Adapted from Lodish, B. et al. (2004) 5th ed.

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

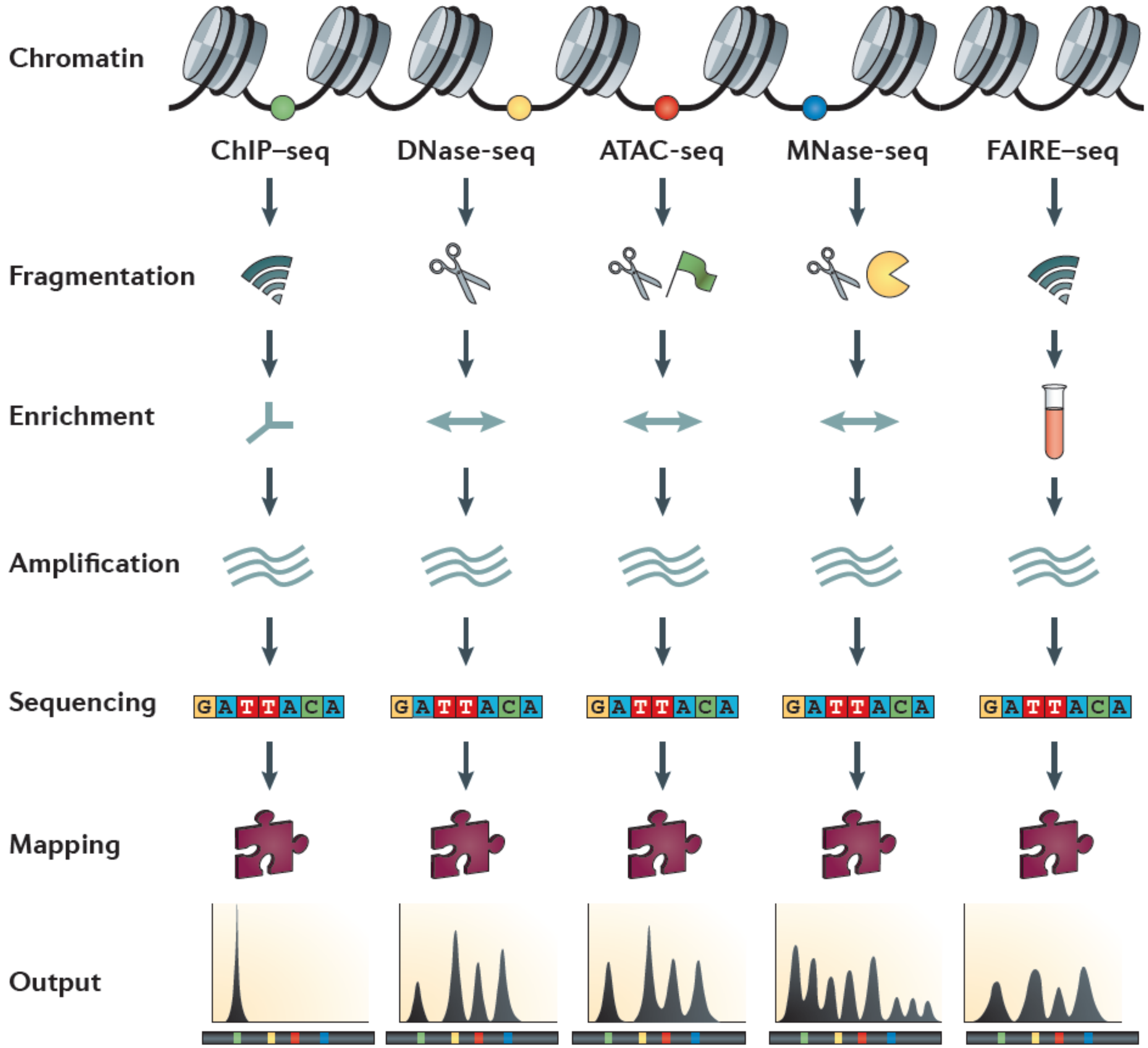
Chromatin & Histone Code



Histone Code

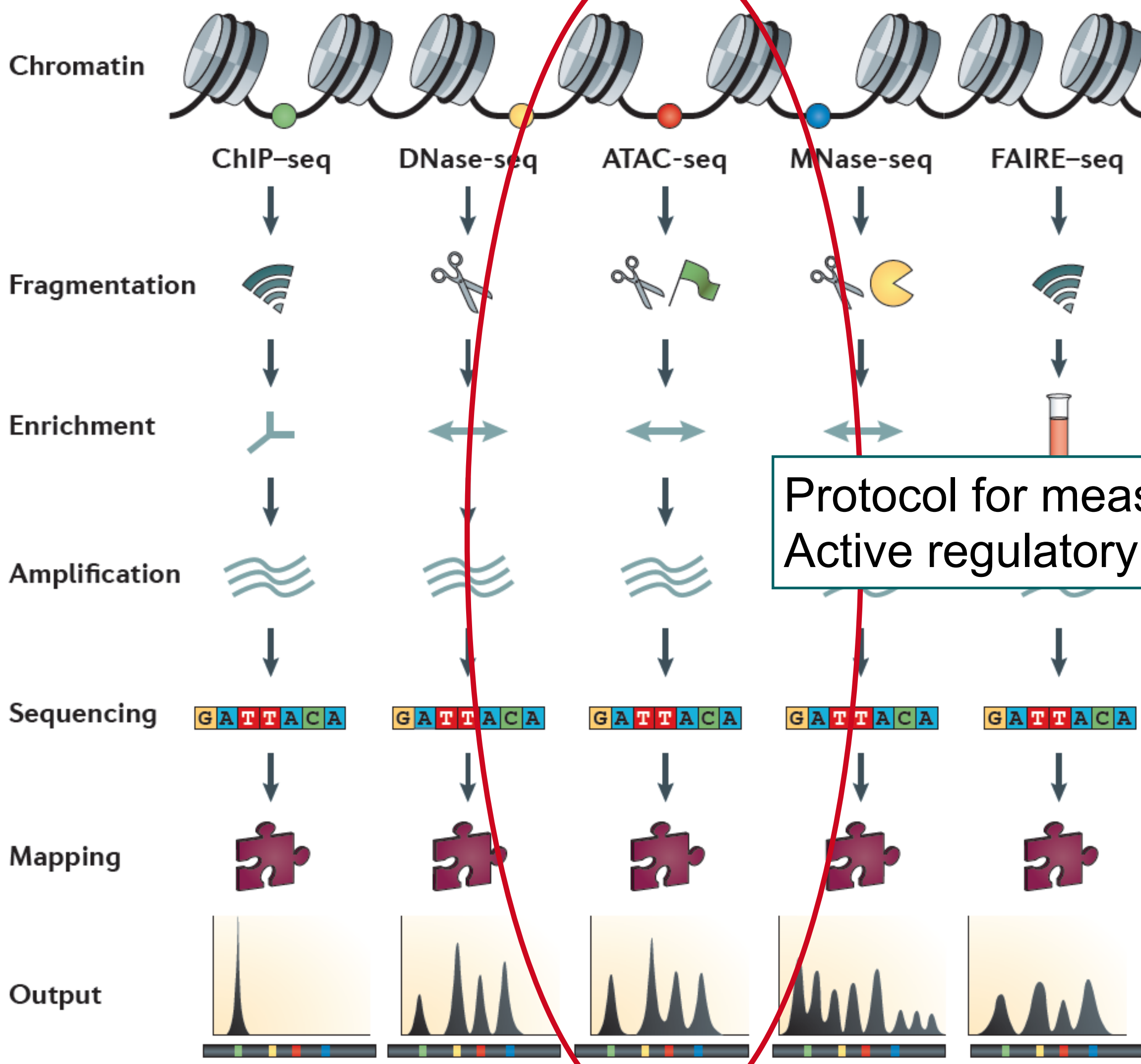
- **Transcription**
  H3K79me2, H3k36me3
- **Active Regions**
  H3K27ac, H3K9ac
- **Active Promoters**
  H3K4me3
- **Active Enhancers**
  H3K4me1
- **Repressed regions**
  H3K27me3, H3K9me3

Chromatin with Next Generation Sequencing



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

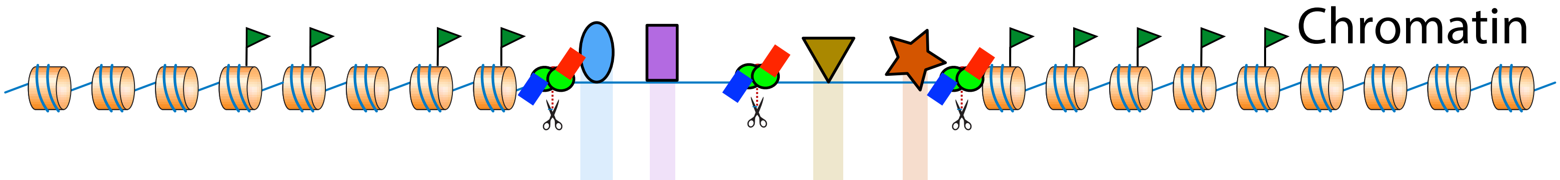
Chromatin with Next Generation Sequencing



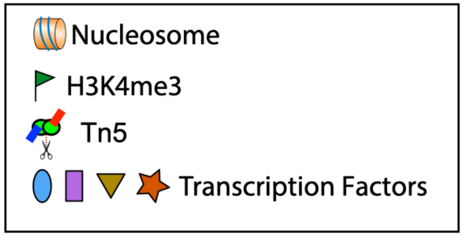
Protocol for measuring open chromatin
Active regulatory regions

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

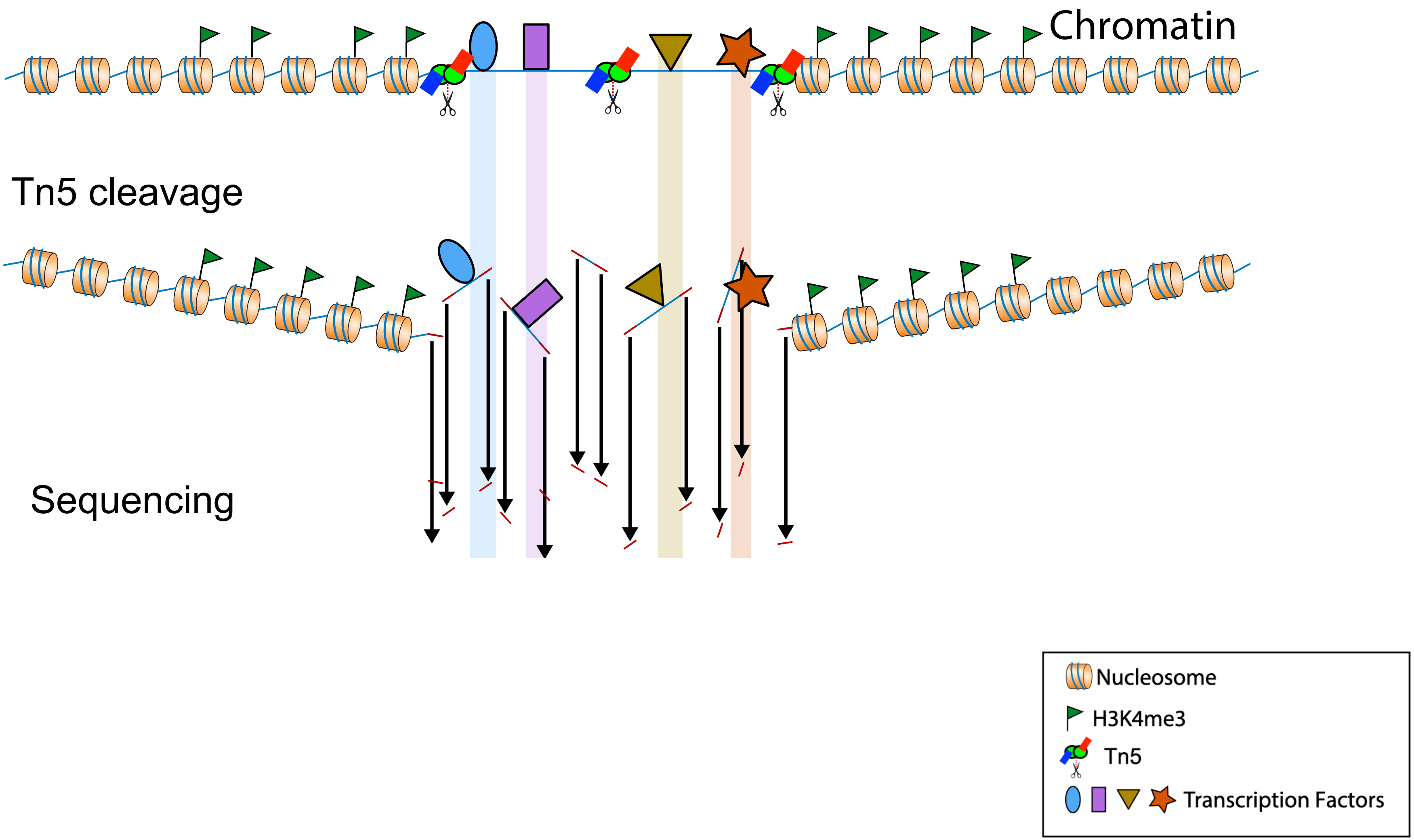
Open Chromatin with ATAC-seq



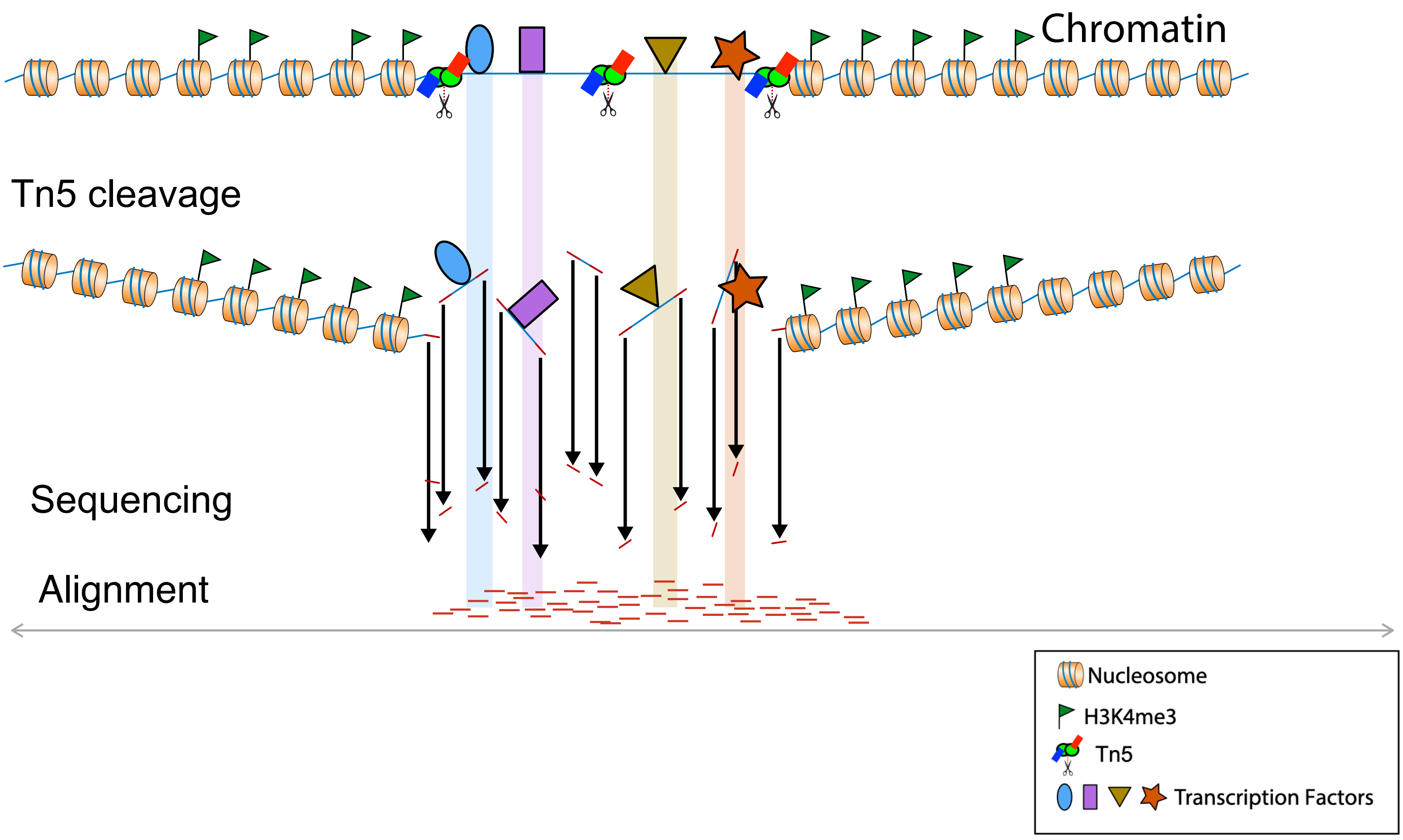
Tn5 cleavage



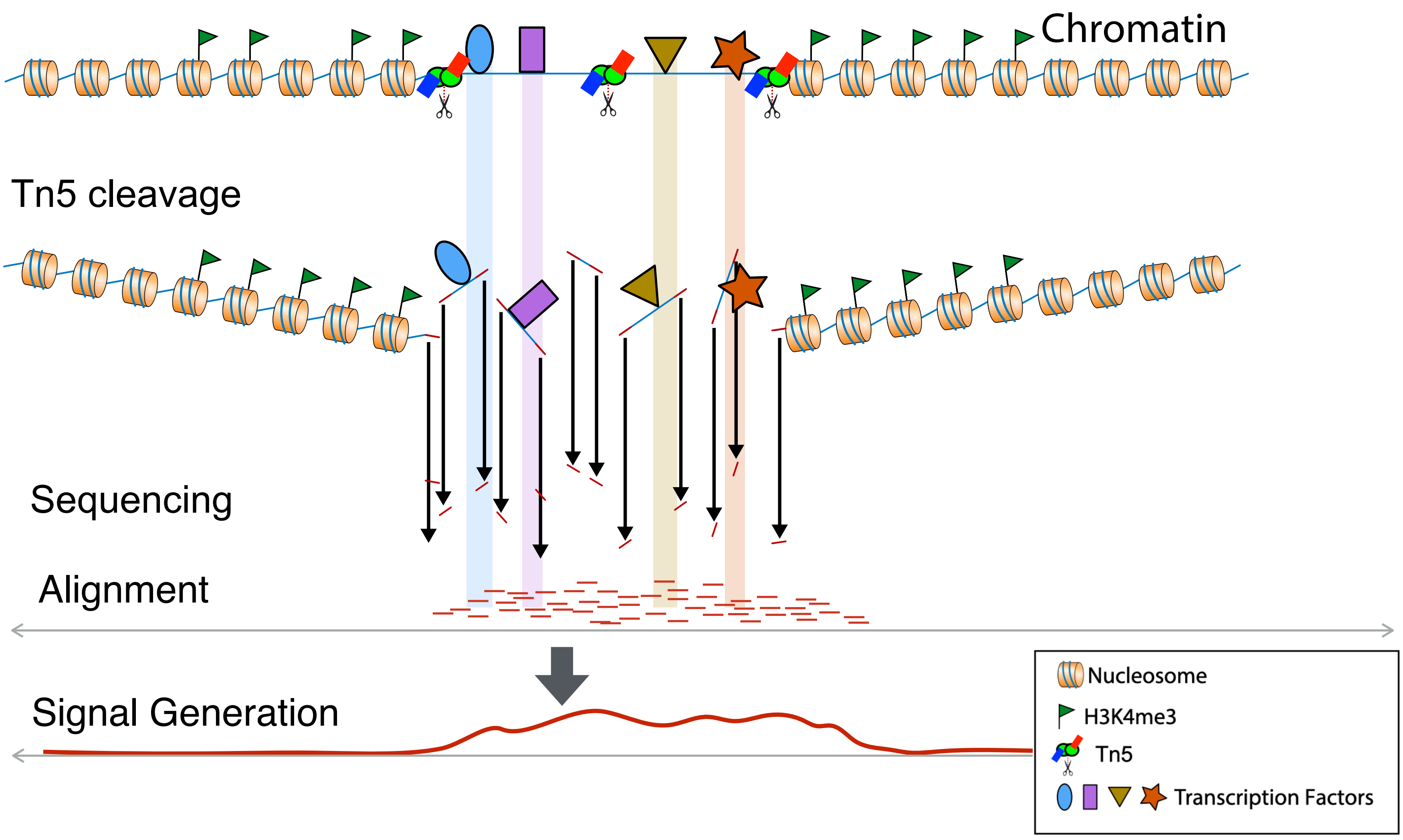
Open Chromatin with ATAC-seq



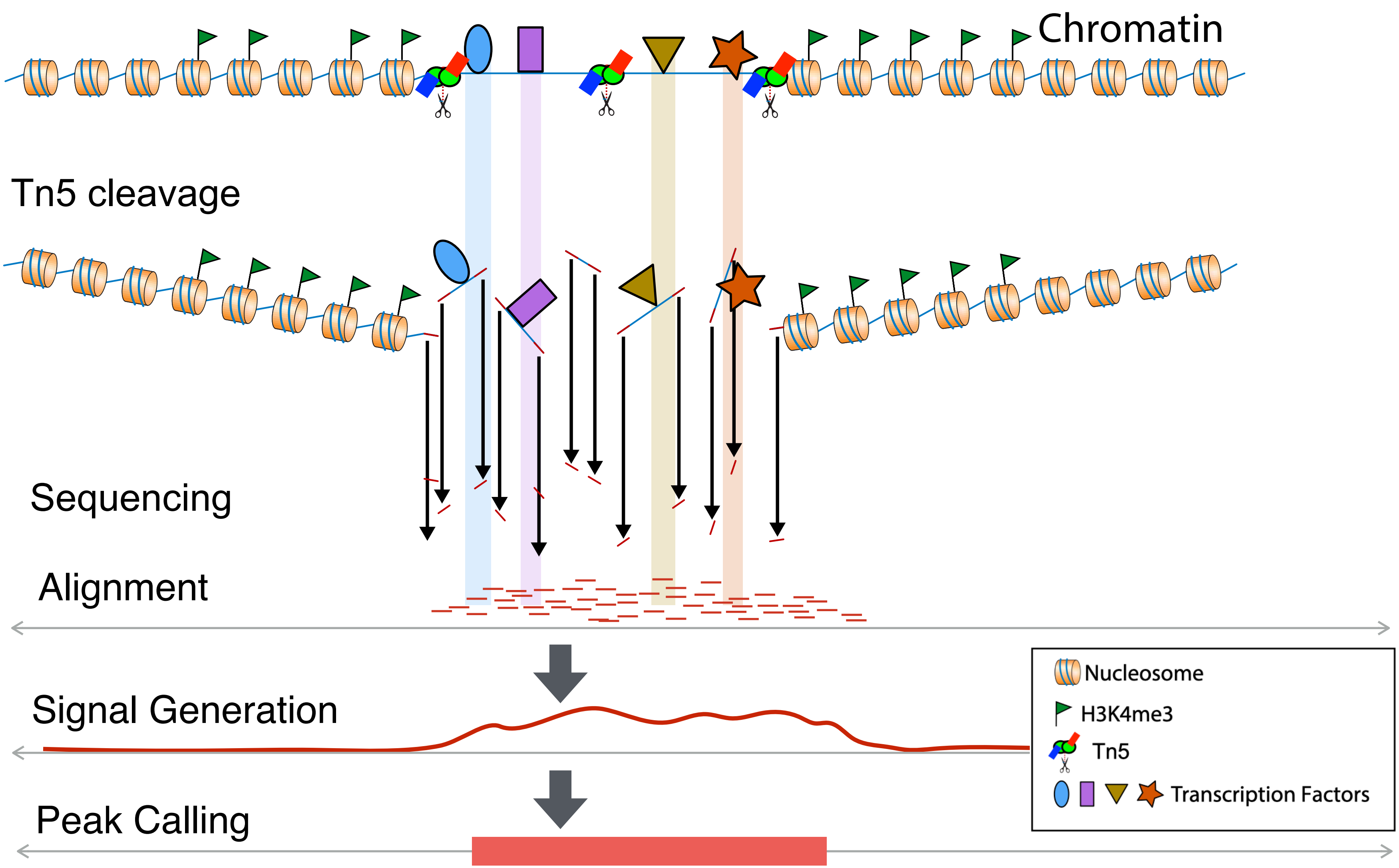
Open Chromatin with ATAC-seq



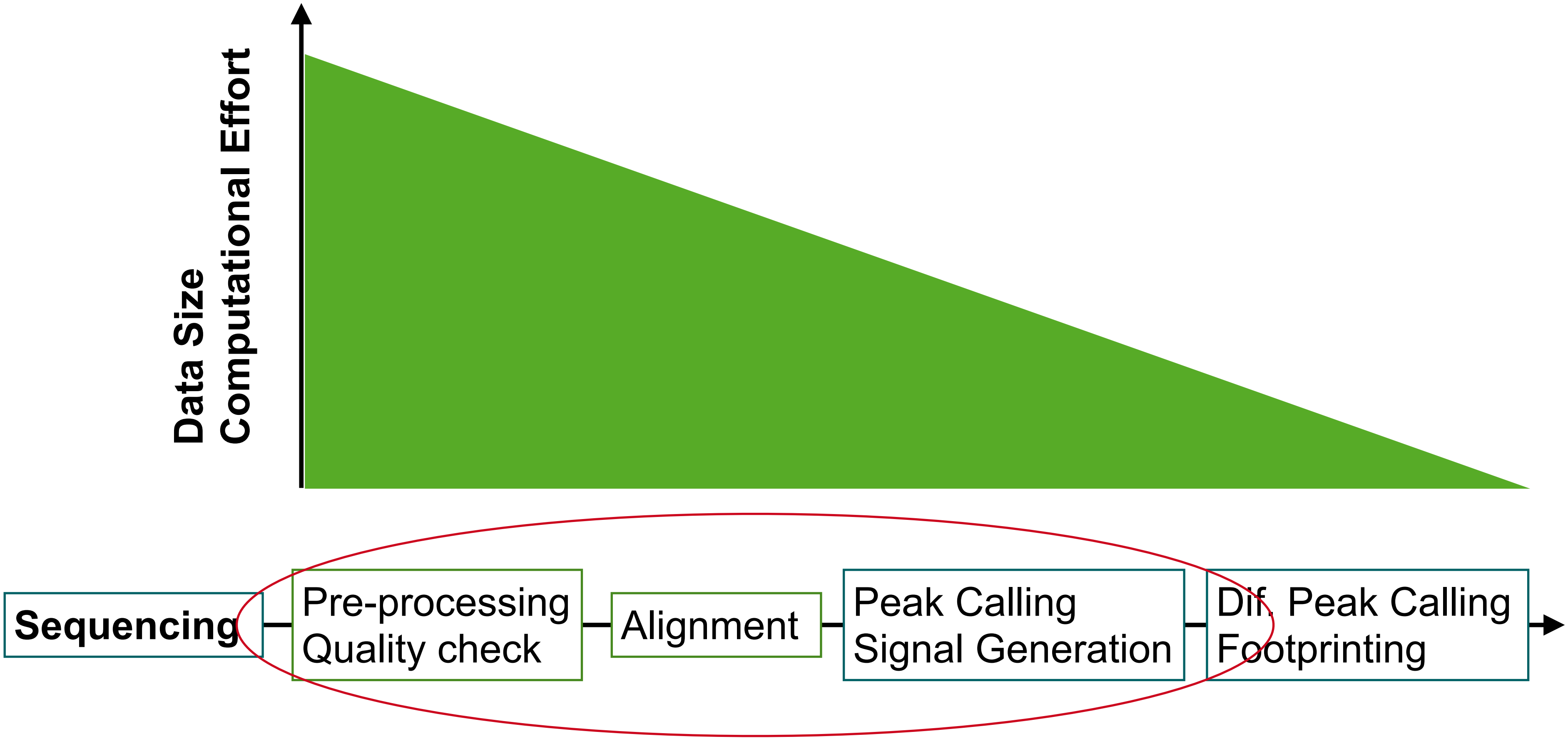
Open Chromatin with ATAC-seq



Open Chromatin with ATAC-seq



Bioinformatics Pipeline / ATAC-seq



Adapted from Rasmussen:
http://www.cbs.dtu.dk/courses/27626/programme.php

Institute for Computational Genomics
RWTH AACHEN UNIVERSITY

Peak Calling

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

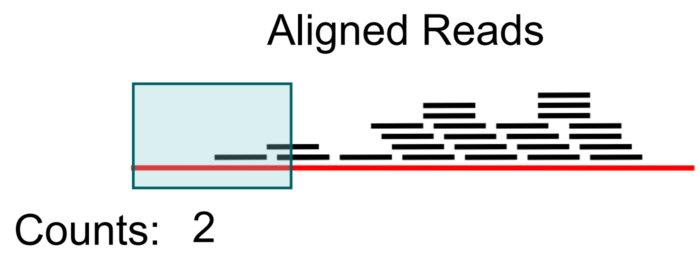
Example of a simple peak caller :

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.



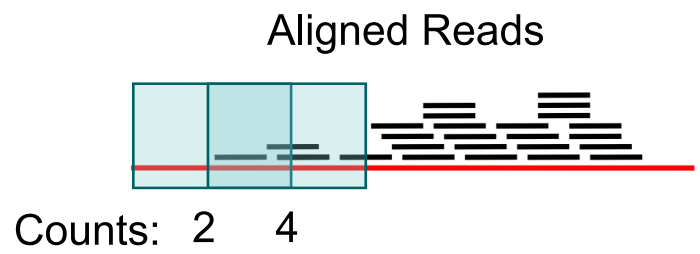
Example of a simple peak caller :

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.



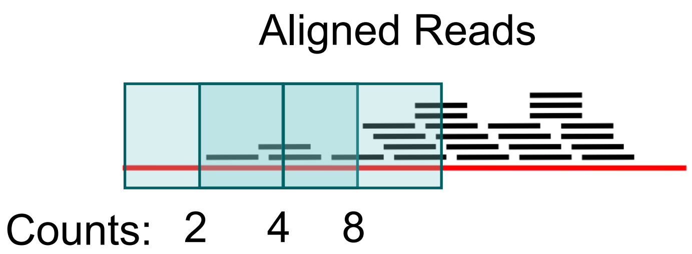
Example of a simple peak caller :

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.



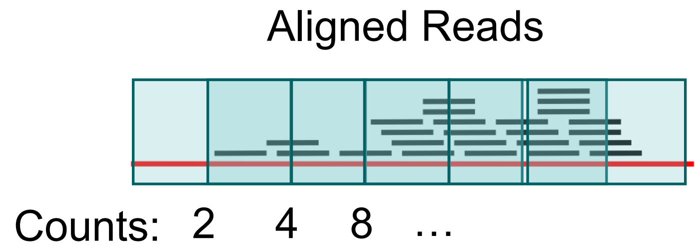
Example of a simple peak caller :

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.



Example of a simple peak caller :

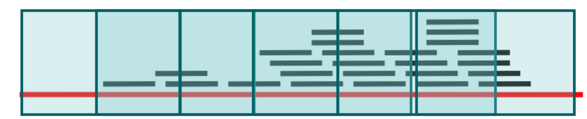
1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change

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.

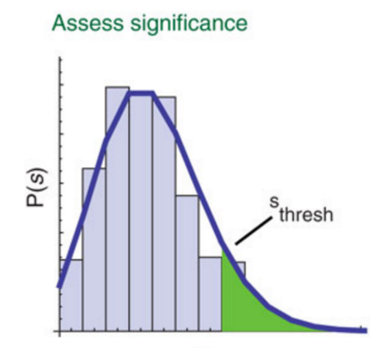
Aligned Reads



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

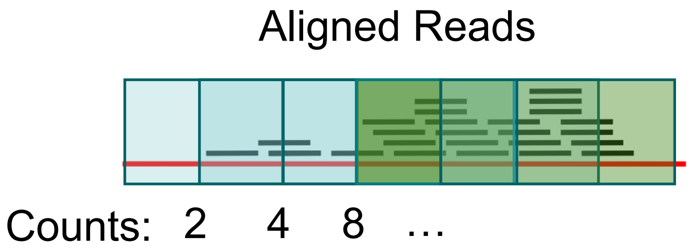


Counts

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

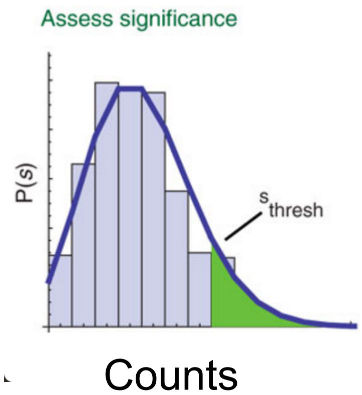
Peak Calling

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.

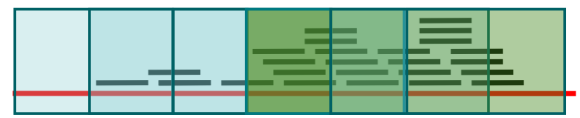
Problems:

- which window size to use?
- proper quantification of read counts require several further steps: CG bias correction, duplicated reads, mappability, **fragment size**, …

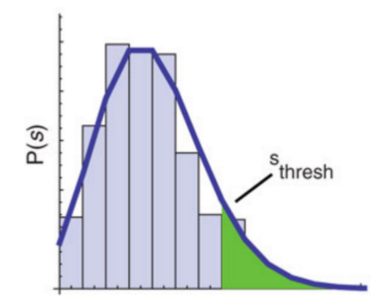
E

if the number of reads in higher than expected by change

Aligned Reads



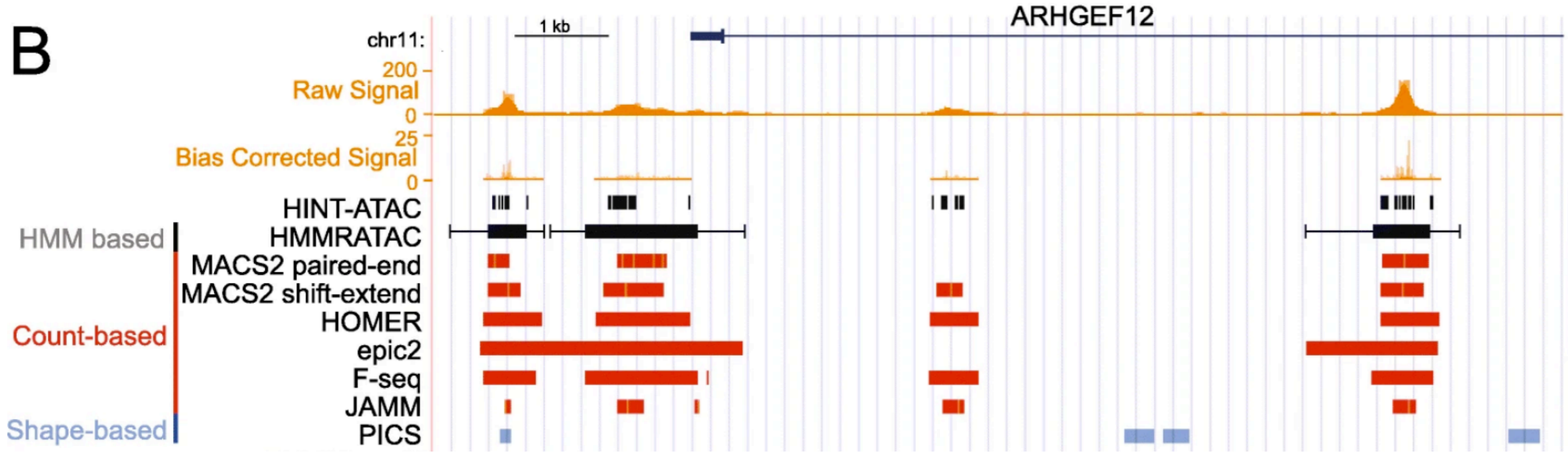
unts: 2 4 8 …



Counts

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 in ATAC-seq



- MACS2
  - most frequently used
- HMMRATAC
  - ATAC-seq specific peak caller
  - ignores reads from large fragments / linker cleavage sites

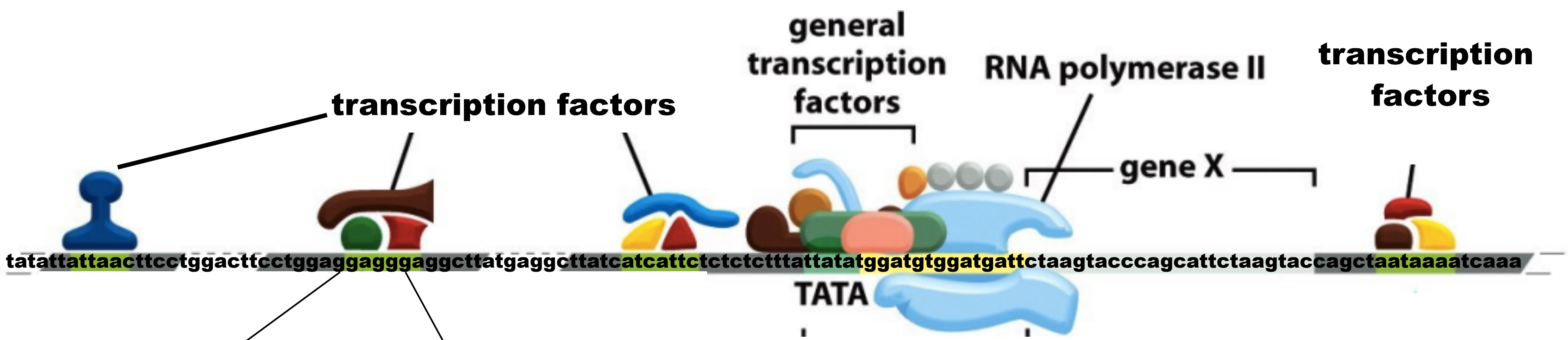
Source: Yan, Genome Biology, 2020.

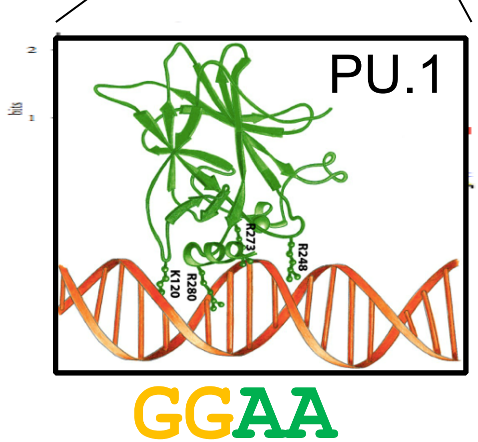
Bioinformatics Pipeline / ATAC-seq

- **Sequencing**
- Pre-processing / Quality check
- Alignment
- Peak Calling / Signal Generation
- Dif. Peak Calling / Motif search / Footprinting

Adapted from Rasmussen:
http://www.cbs.dtu.dk/courses/27626/programme.php

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!

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

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

Model for DNA-protein binding

PU.1 binding sites →

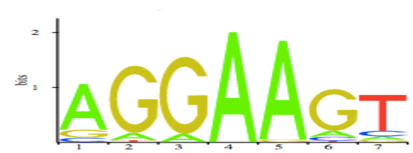
Kanno, Y. et al. (2005) Immune Cell-Specific Amplification of Interferon Signaling by the IRF-4/8-PU.1 Complex.

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

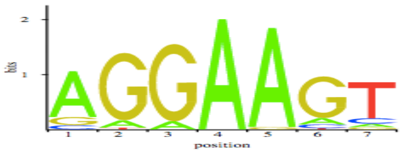
PU.1 Position Weight Matrix (PWM) →

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 →

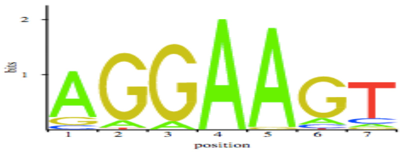


Motif Search

PU.1 PWM 

Genome **TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA**

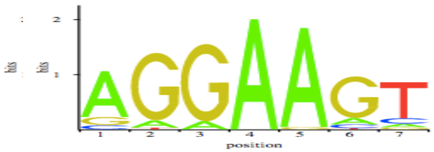
Motif Search

PU.1 PWM 

Genome **TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA**

Score 10.06

Motif Search

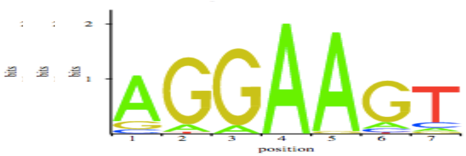
PU.1 PWM 

Genome **TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA**

Score 10.06

4.81

Motif Search

PU.1 PWM 

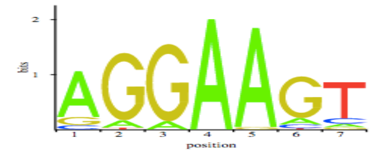
Genome **TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA**

Score 10.06 3.19

4.81

Motif Search

PU.1 PWM

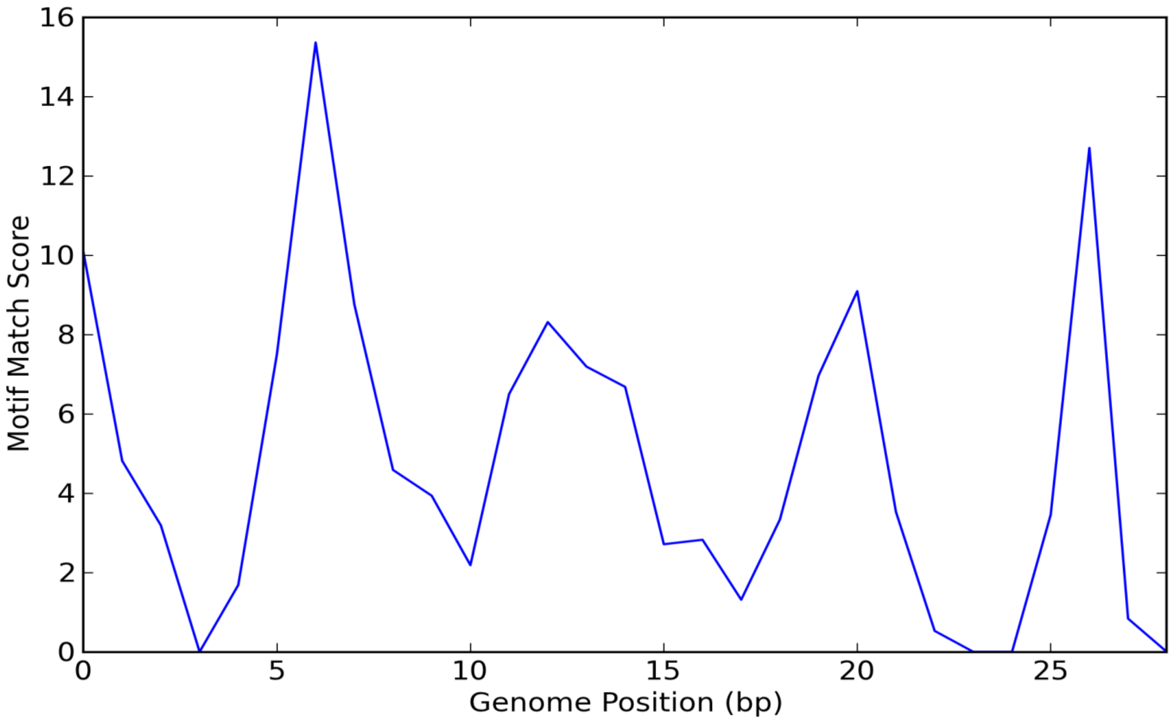


Genome **TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA**

Score 10.06 | 3.19

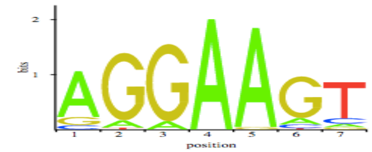
. . .

4.81



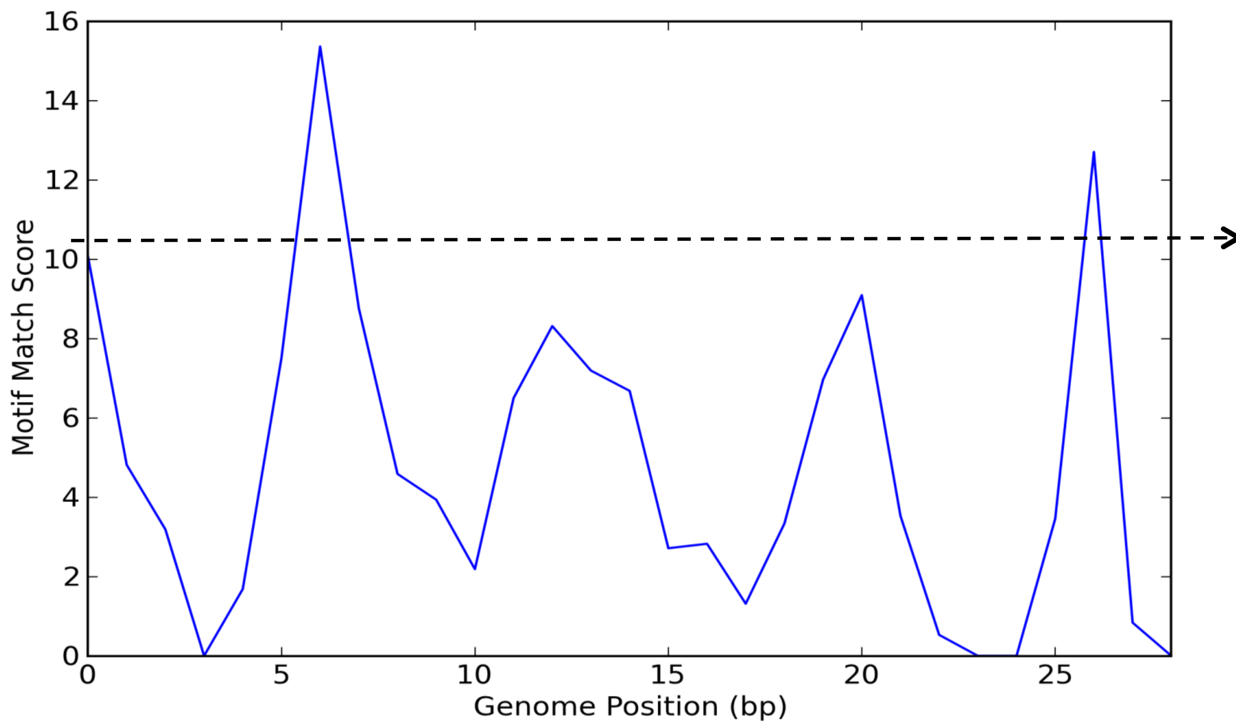
Motif Search

PU.1 PWM



Genome **TATCTTTGGAAGTGAAACTACTATCCTGAAAGTCGAA**

Score 10.06 | 3.19
4.81
. . .



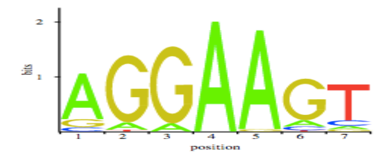
FDR = 1 x 10^{-4}

Statistical Test

FDR- False Discovery Rate

Motif Search

PU.1 PWM

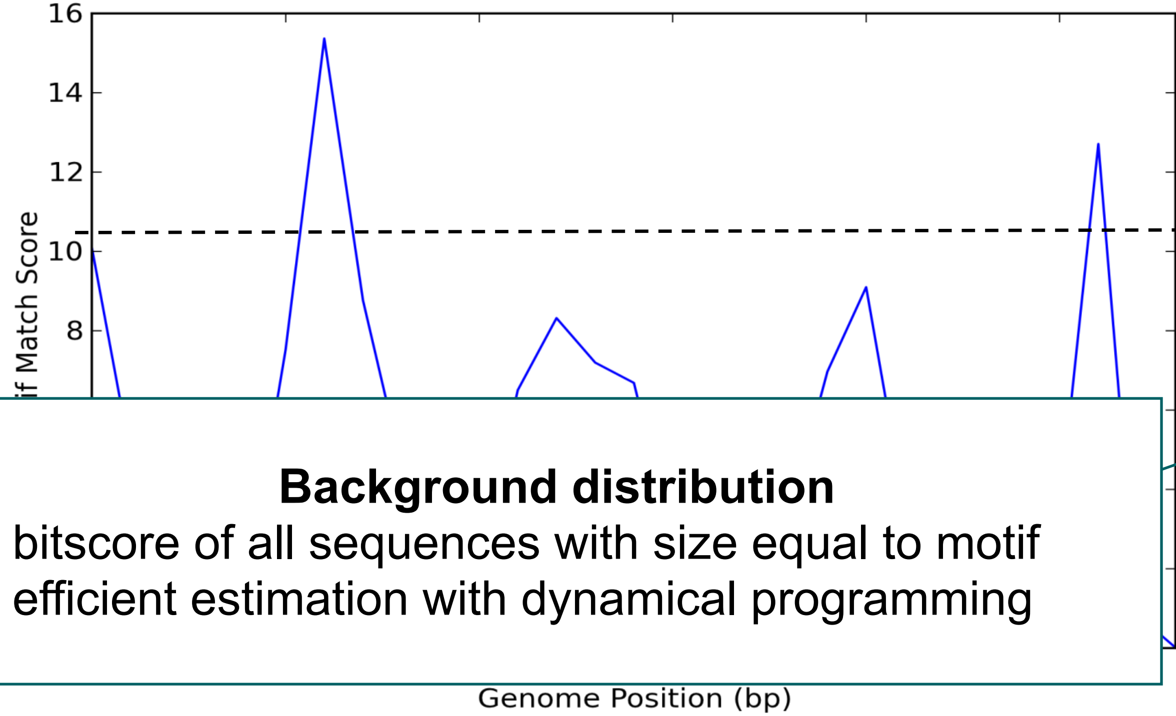


Genome TATCTTT**GGAAGT**GAAACTACTATCCT**GAAAGT**CGAA

Score 10.06 | 3.19 . . .

4.81

Statistical Test



FDR = 1 x 10^{-4}

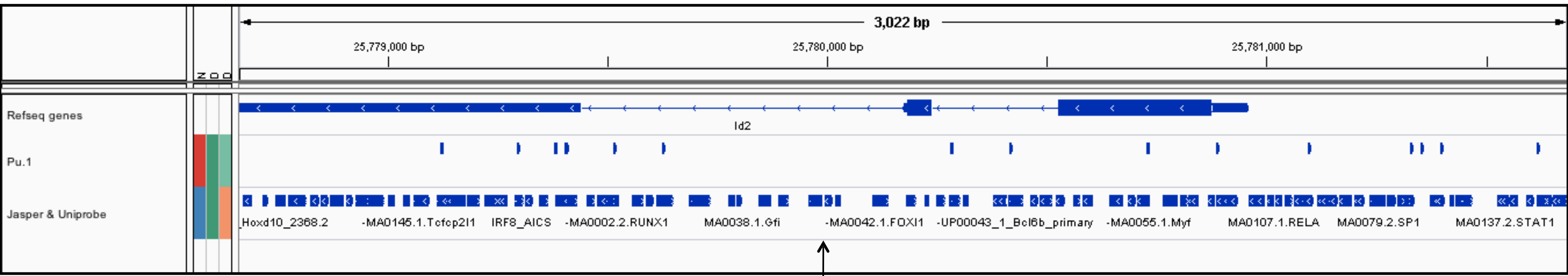
Background distribution

- bitscore of all sequences with size equal to motif
- efficient estimation with dynamical programming

FDR- False Discovery Rate

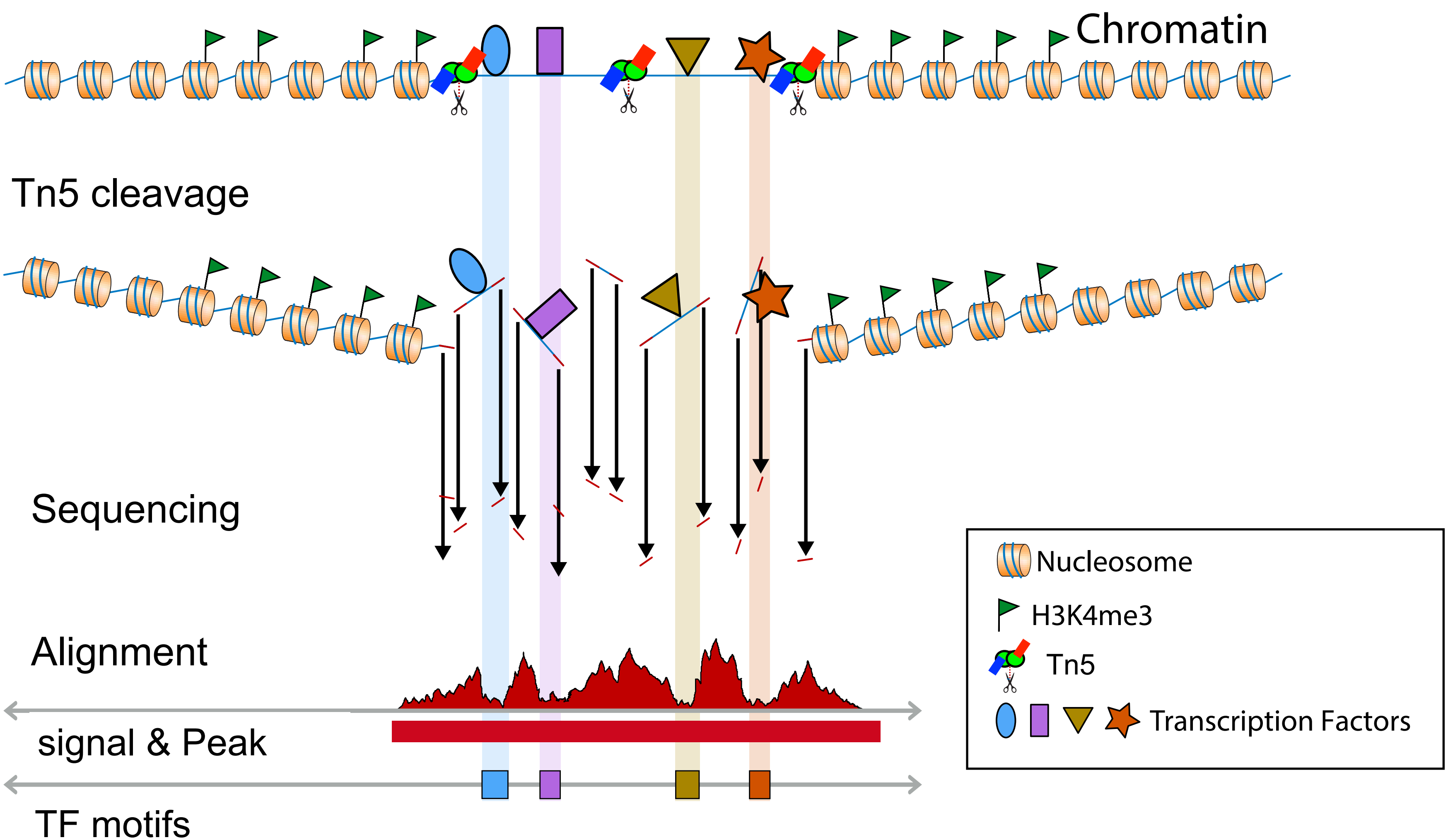
Example: Binding sites in ID2

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



> 3000 predicted binding sites

Open Chromatin and TF Binding with ATAC-seq



Resume

- Review basic biological/computational aspects
  1. basics of molecular biology
  2. basics of sequencing
  3. basics bioinformatics problems
     - short sequences read alignment
     - gene expression quantification
     - computational epigenetic
     - single cell approaches (next week)

Calendar

Today – Introduction to Bioinformatics, Next Generation Sequencing

26.04.2021 – Single cell sequencing / Practical Course

3.05.2021 – Project Description / Introduction to HPC clusters and GPUs

10.05.2021 – 5.7.2021 – Project development

12.07.2021 – Project Presentation

Communication/discord channel: https://discord.gg/jycmaCUkAj

Thank you!