Bioinformatics Lab

Ivan Gesteira Costa, Mingbo Cheng, James Nagai, Mina Shaigan, Martin Manolov

Institute for Computational Genomics

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

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
        - single cell approaches
        - computational epigenetic

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-2023/

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

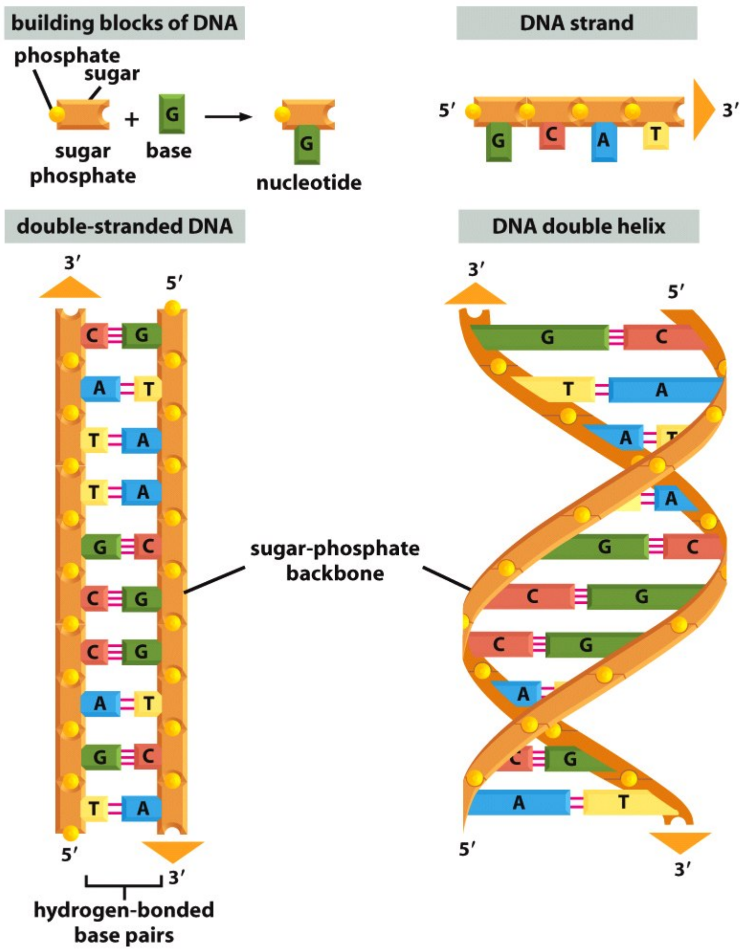
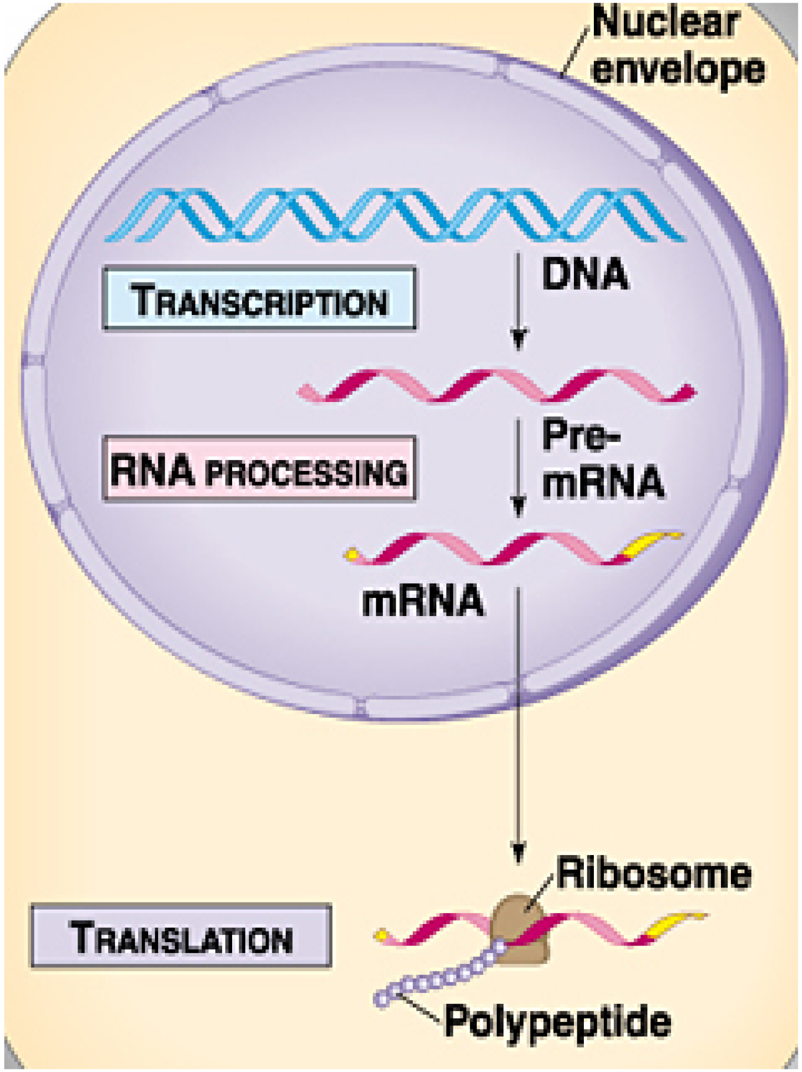


Figure 4-3 Molecular Biology of the Cell 5/e (© Garland Science 2008)

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

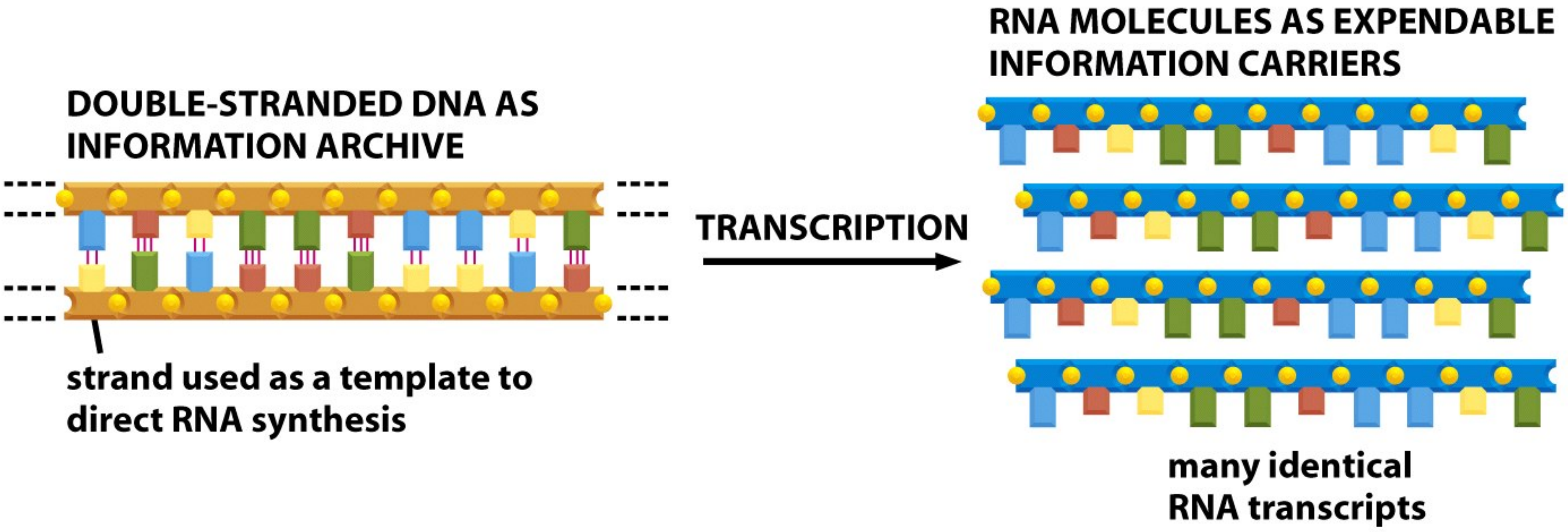
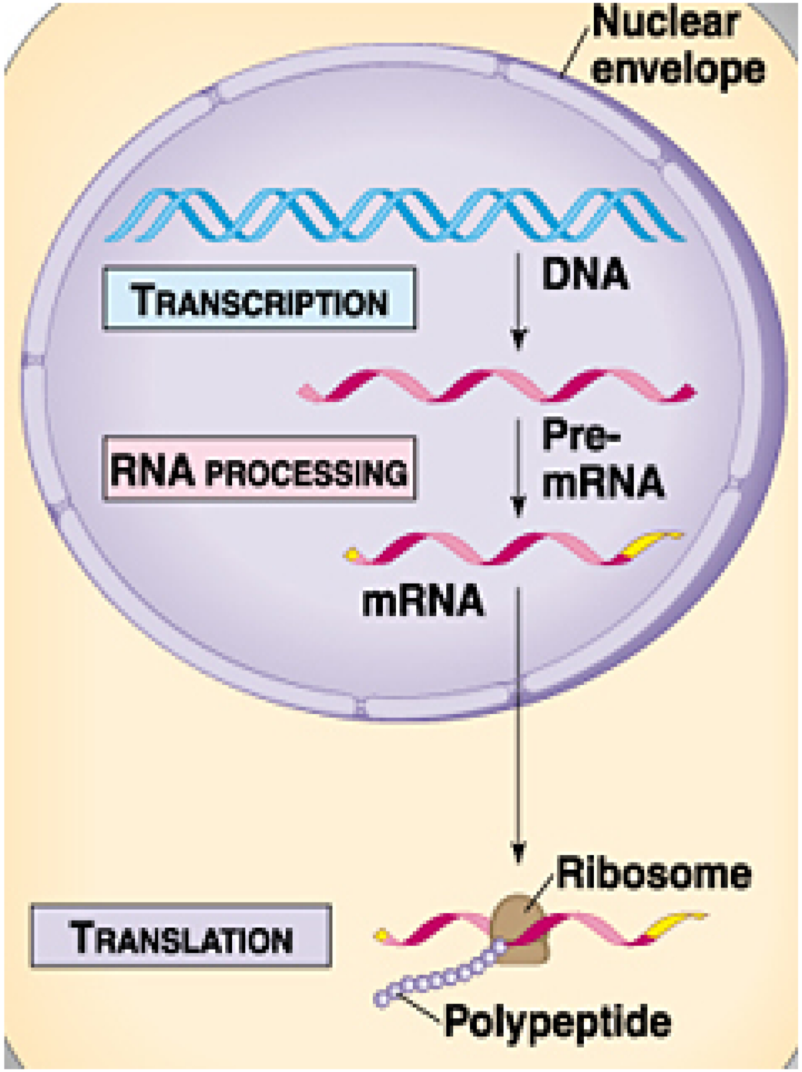

RNA MOLECULES AS EXPENDABLE INFORMATION CARRIERS
DOUBLE-STRANDED DNA AS INFORMATION ARCHIVE
TRANSCRIPTION
strand used as a template to direct RNA synthesis
many identical RNA transcripts

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



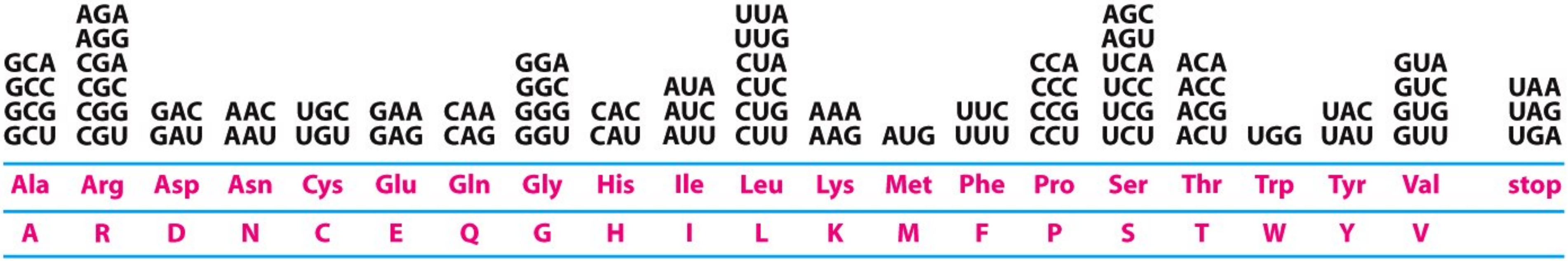
Translation

- *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

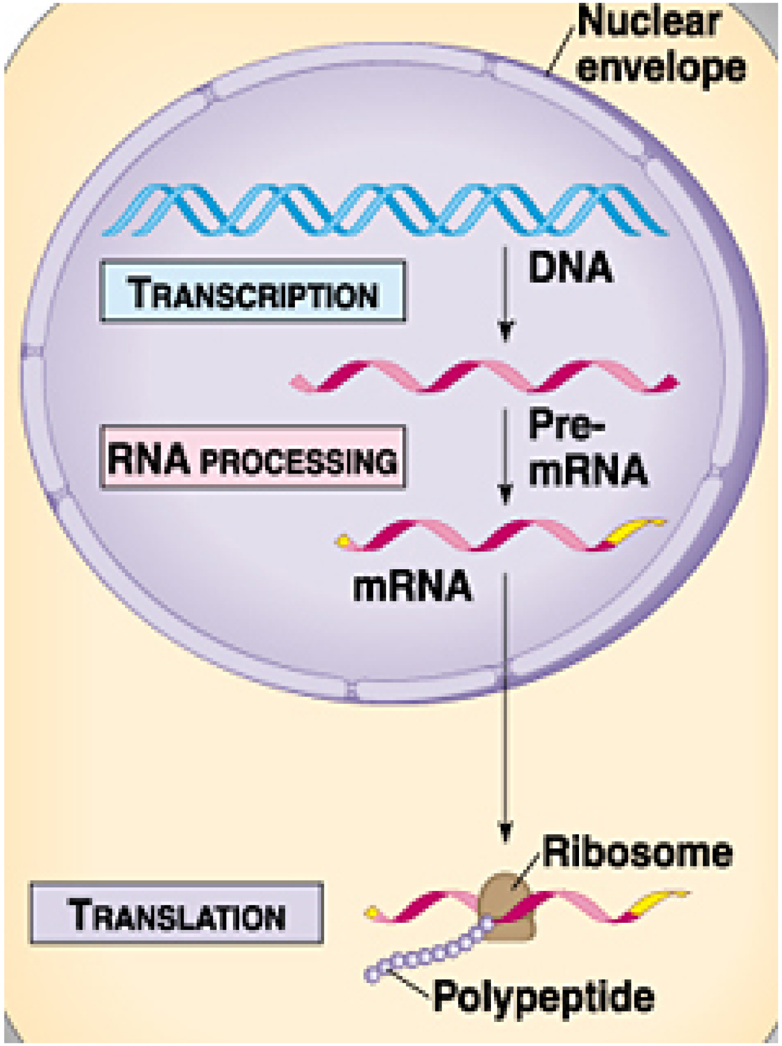


GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	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

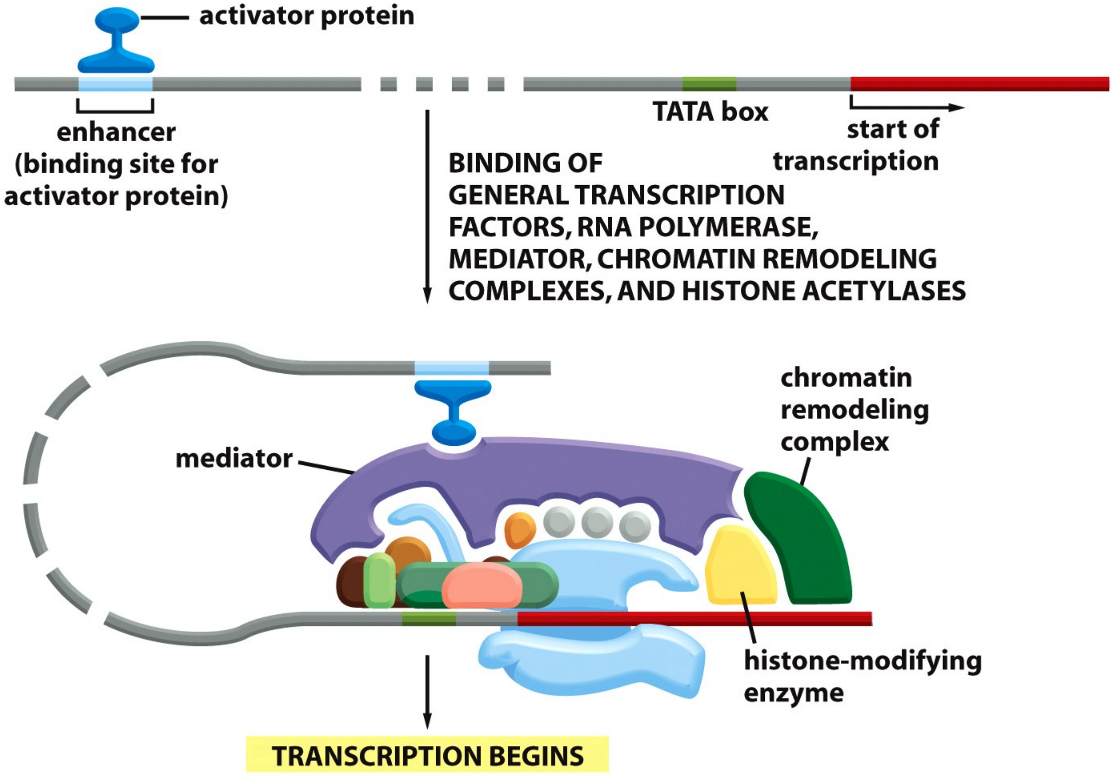

activator protein
enhancer (binding site for activator protein)
TATA box
start of transcription
BINDING OF GENERAL TRANSCRIPTION FACTORS, RNA POLYMERASE, MEDIATOR, CHROMATIN REMODELING COMPLEXES, AND HISTONE ACETYLASES
mediator
chromatin remodeling complex
histone-modifying enzyme
TRANSCRIPTION BEGINS

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

Gene Expression

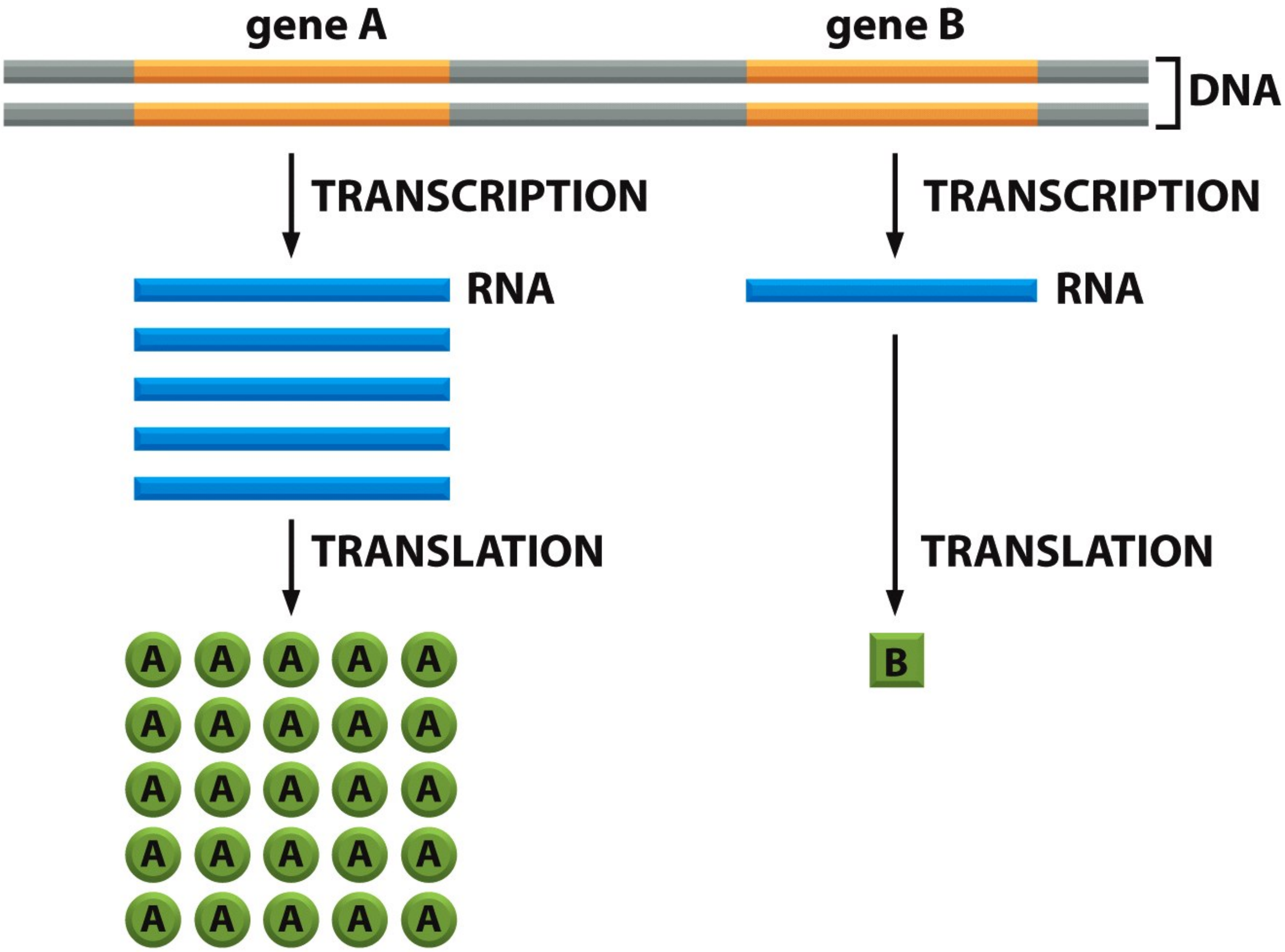


Figure 6-3 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Cellular Complexity

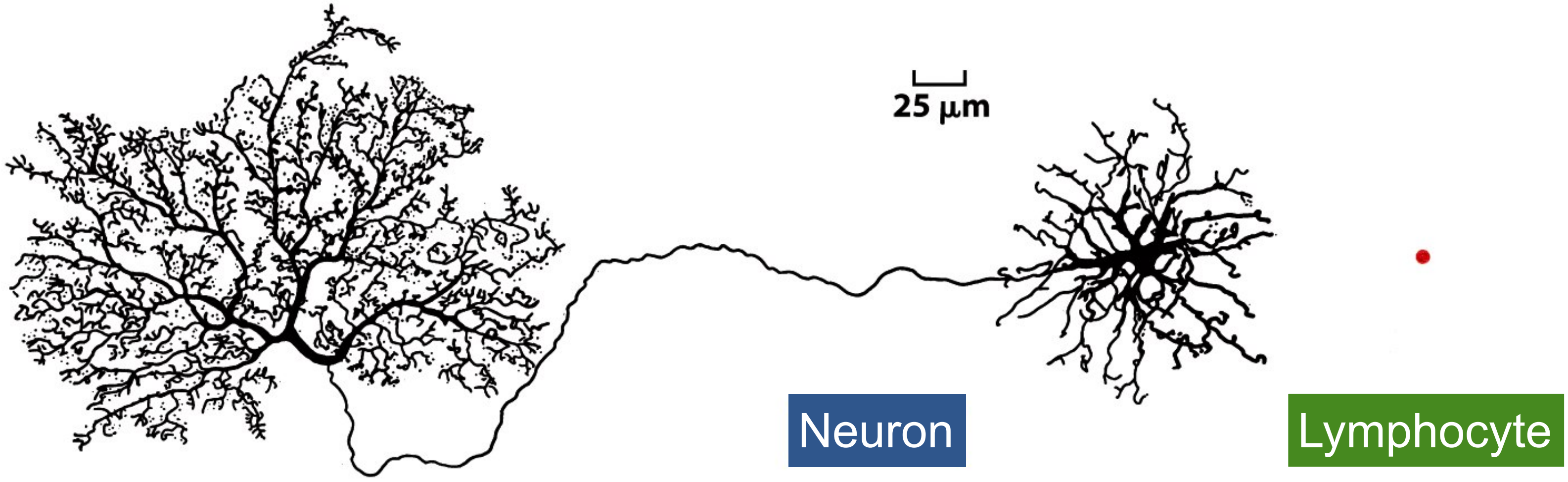


Figure 7-1 Molecular Biology of the Cell 5/e (© Garland Science 2008)

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

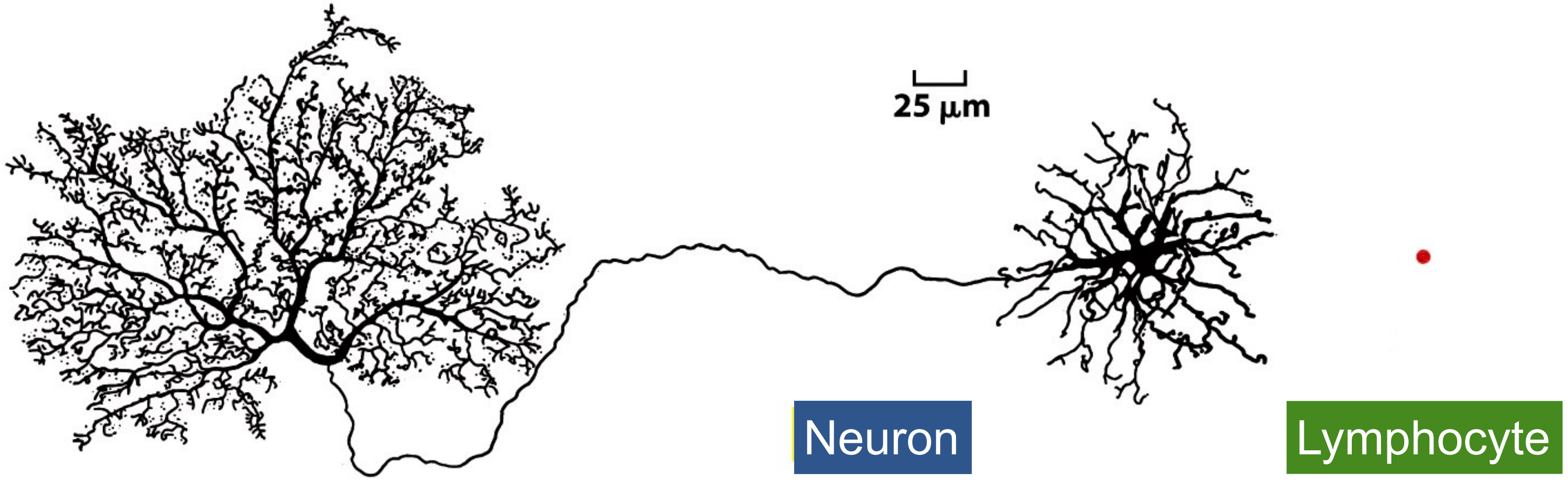
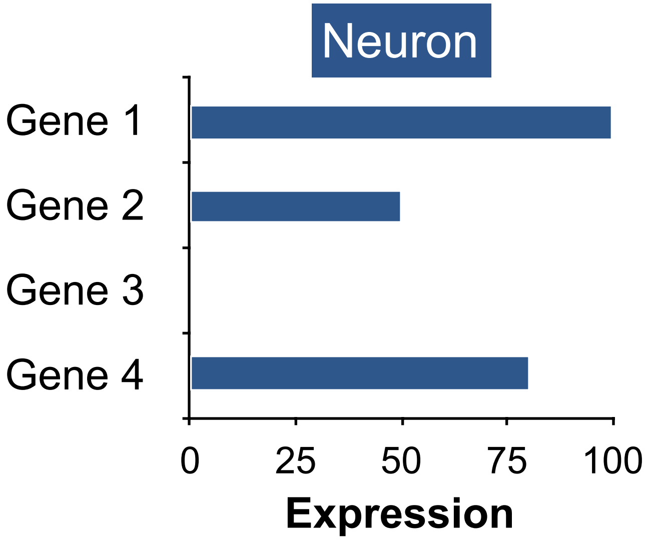
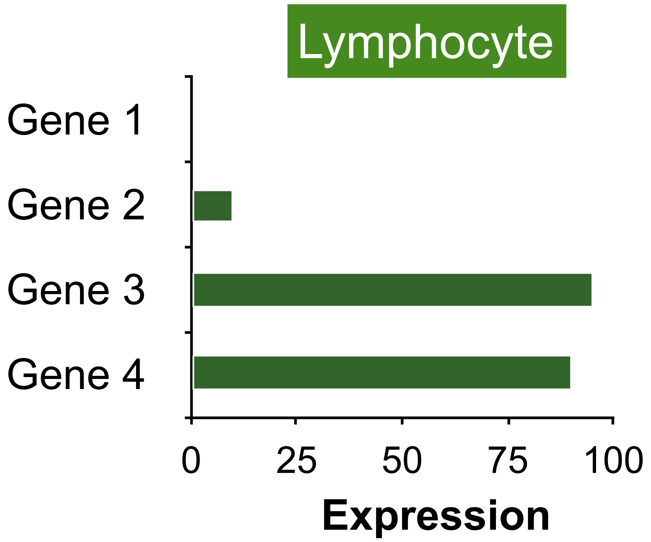


Figure 7-1 Molecular Biology of the Cell 5/e (© Garland Science 2008)





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:

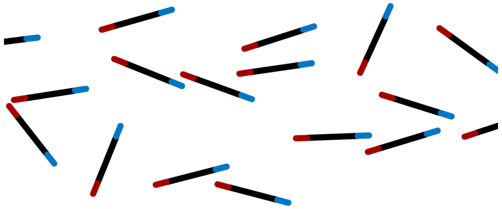
break DNA in several small pieces and use **bioinformatics**

Next Generation Sequencing

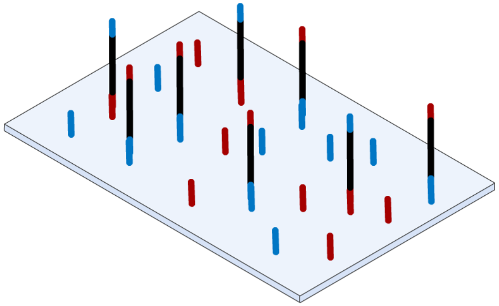
- NGS take advantage of parallelization
  - reads millions/billions of reads for a time
  - short reads (50-300 bps)
  - moderate error rates (0.1%)
- commercial products:
  - 454
  - SOLiD
  - Solexa (Illumina)

Illumina Flow Cell - NGS Sequencing

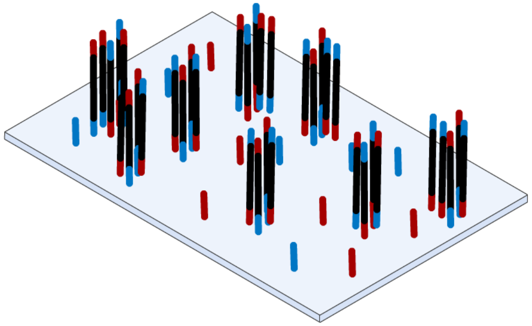
1- fragment sample DNA, insert adapters, attach to flow cell



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



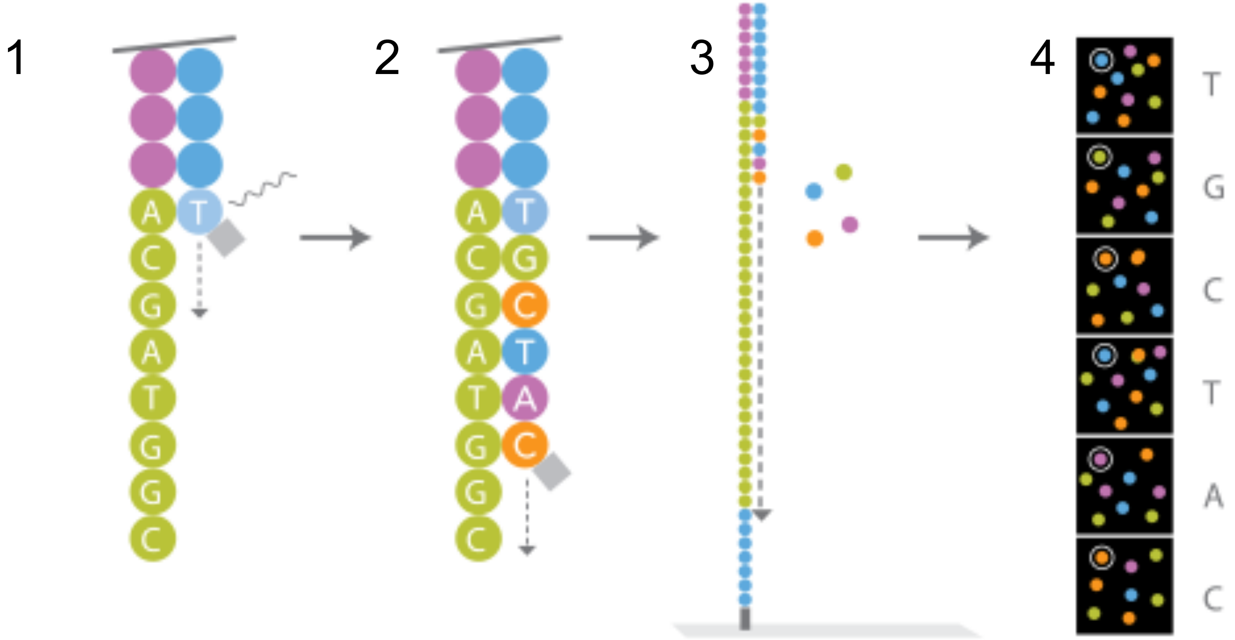
3- clusters of single stranded DNA (200m clusters with 2k DNA strands



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<EEDFEDF3DBDBB=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*

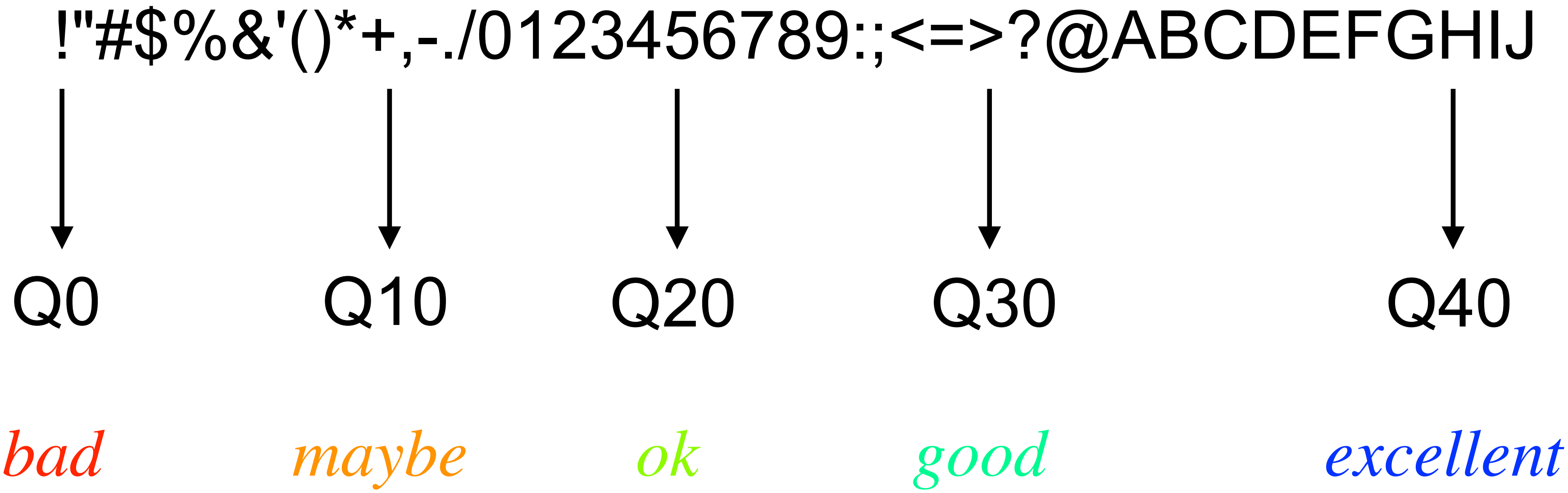
<u>Phred-scale</u>

$$Q = -10 * \log10 P$$

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

Sequencing Results / Phred scores

Uses letters/symbols to represent numbers:

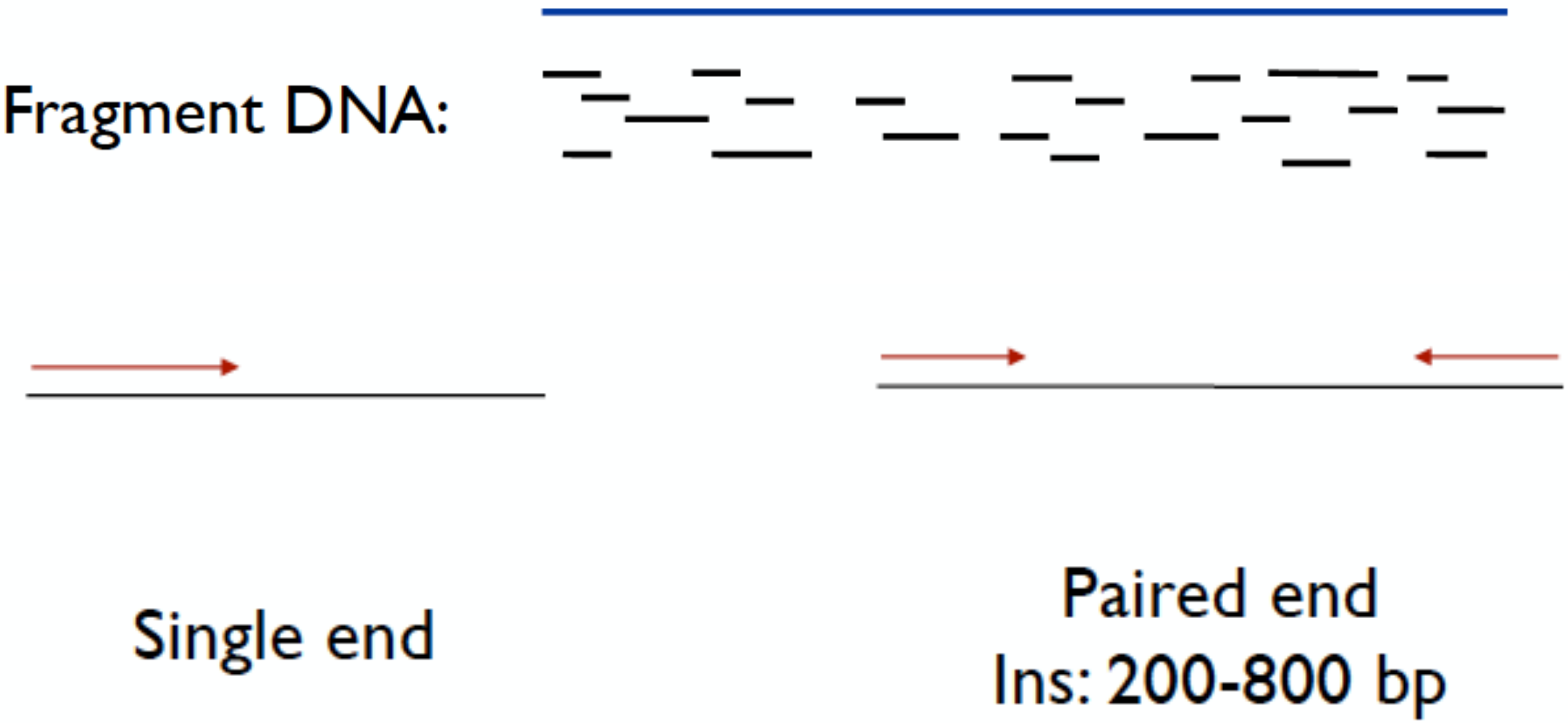


!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ

Q0 Q10 Q20 Q30 Q40

bad *maybe* *ok* *good* *excellent*

Read Types



Fragment DNA:

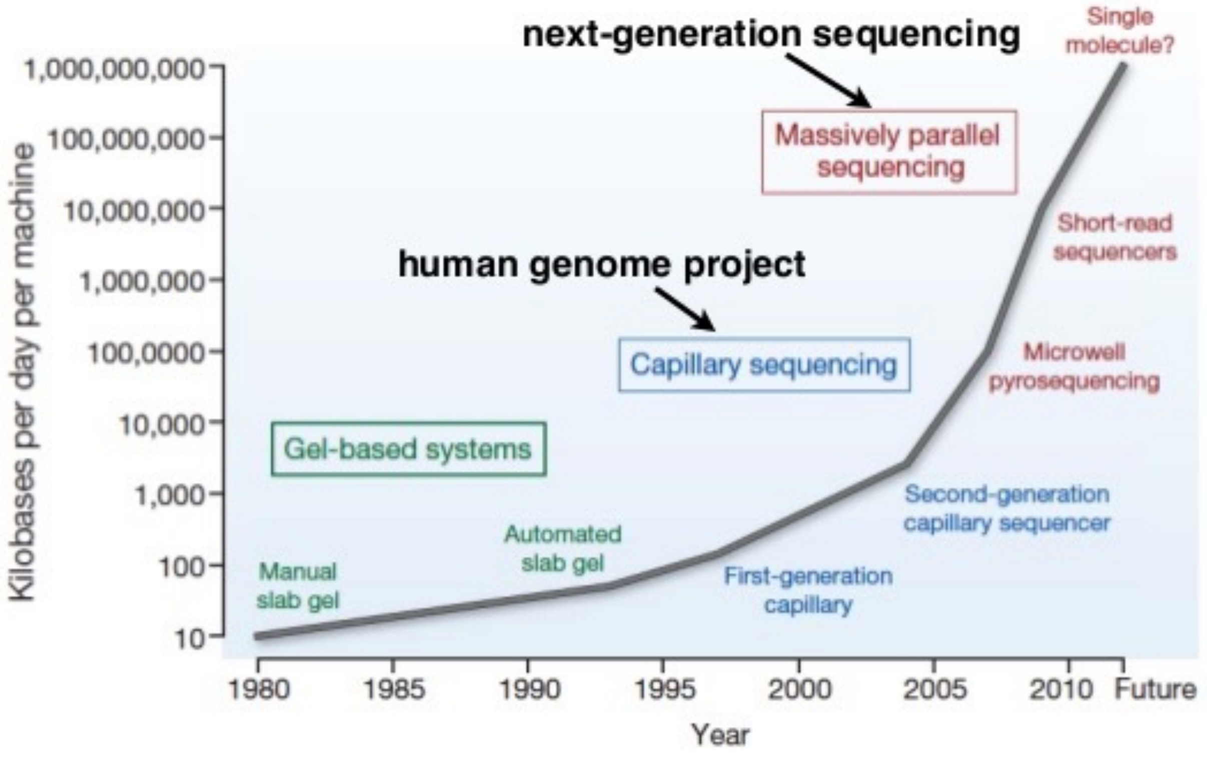
Single end

Paired end

Ins: 200-800 bp

Next Generation Sequencing

Improvements in the rate of DNA sequencing over the past 30 years

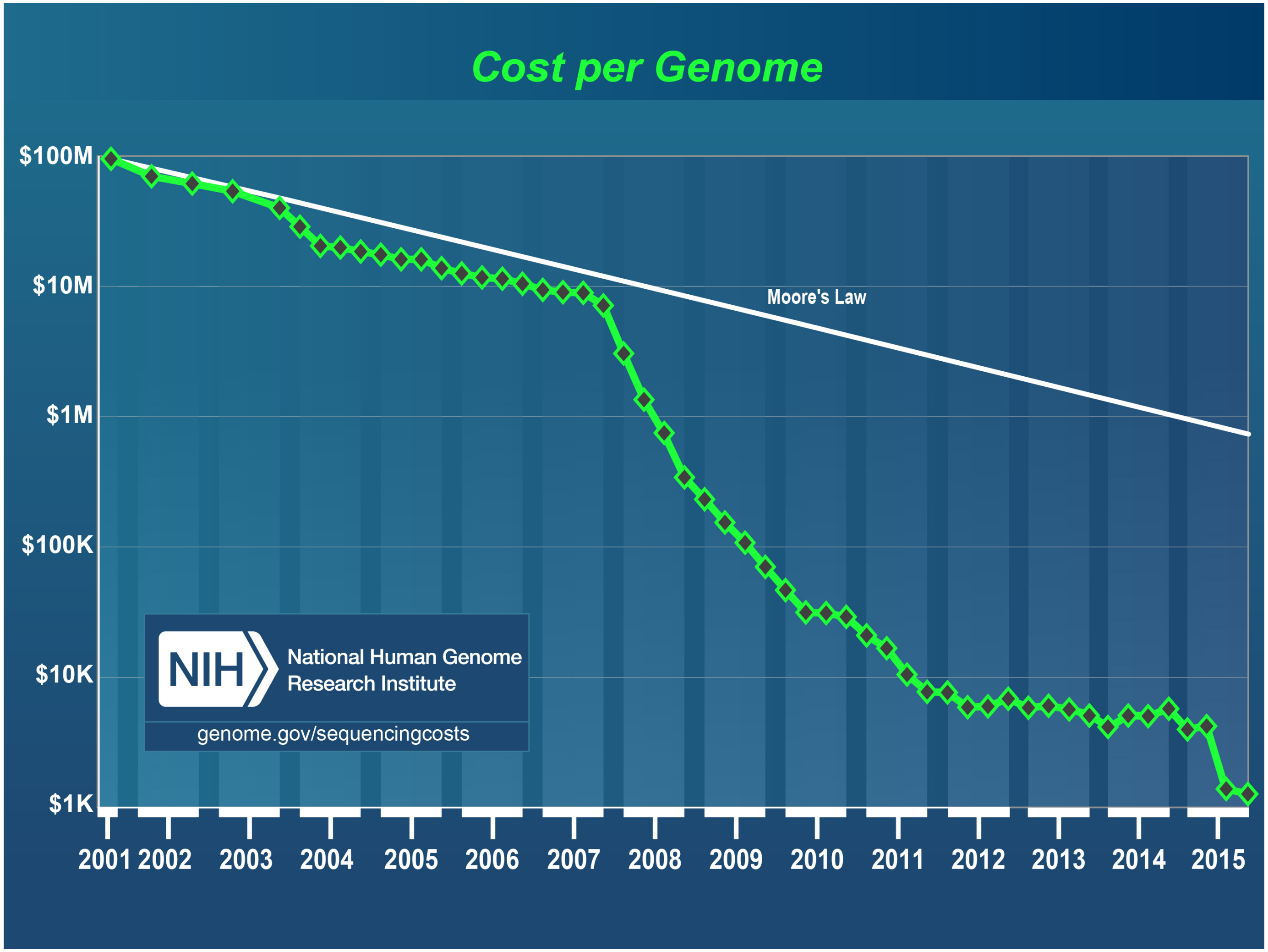


next-generation sequencing

human genome project

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

Sequencing Costs



Sequence Alignment

Sequence Alignment

NGS

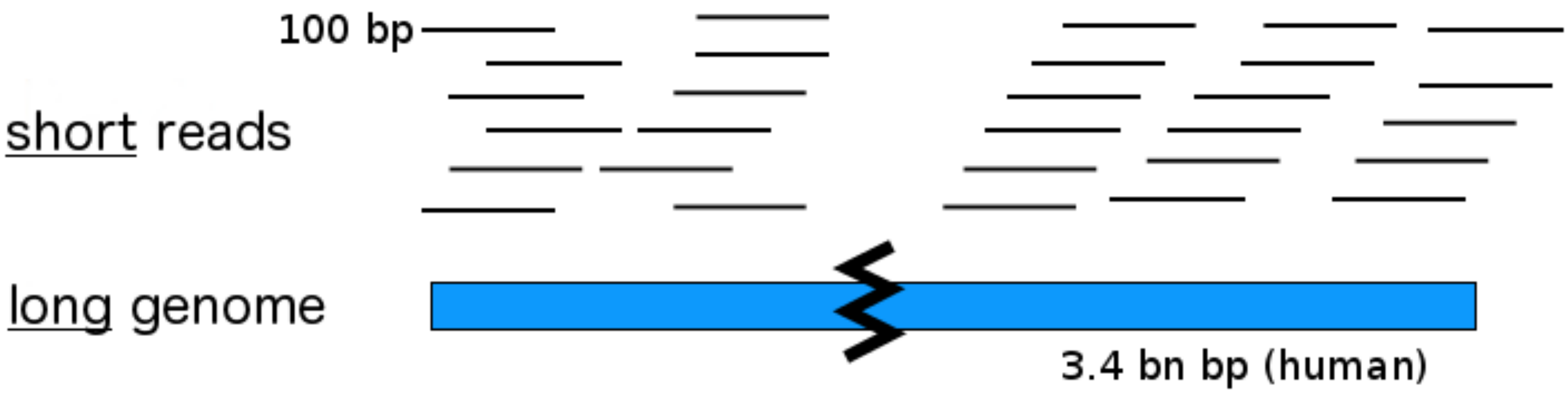
- reads from DNA fragments
- position in genome is unknown
- solution: alignment

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:

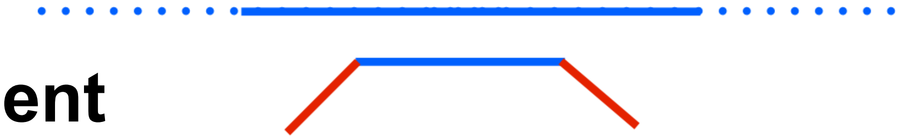
- **Smith & Waterman - dynamic programming (quadratic time/memory)**

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!*

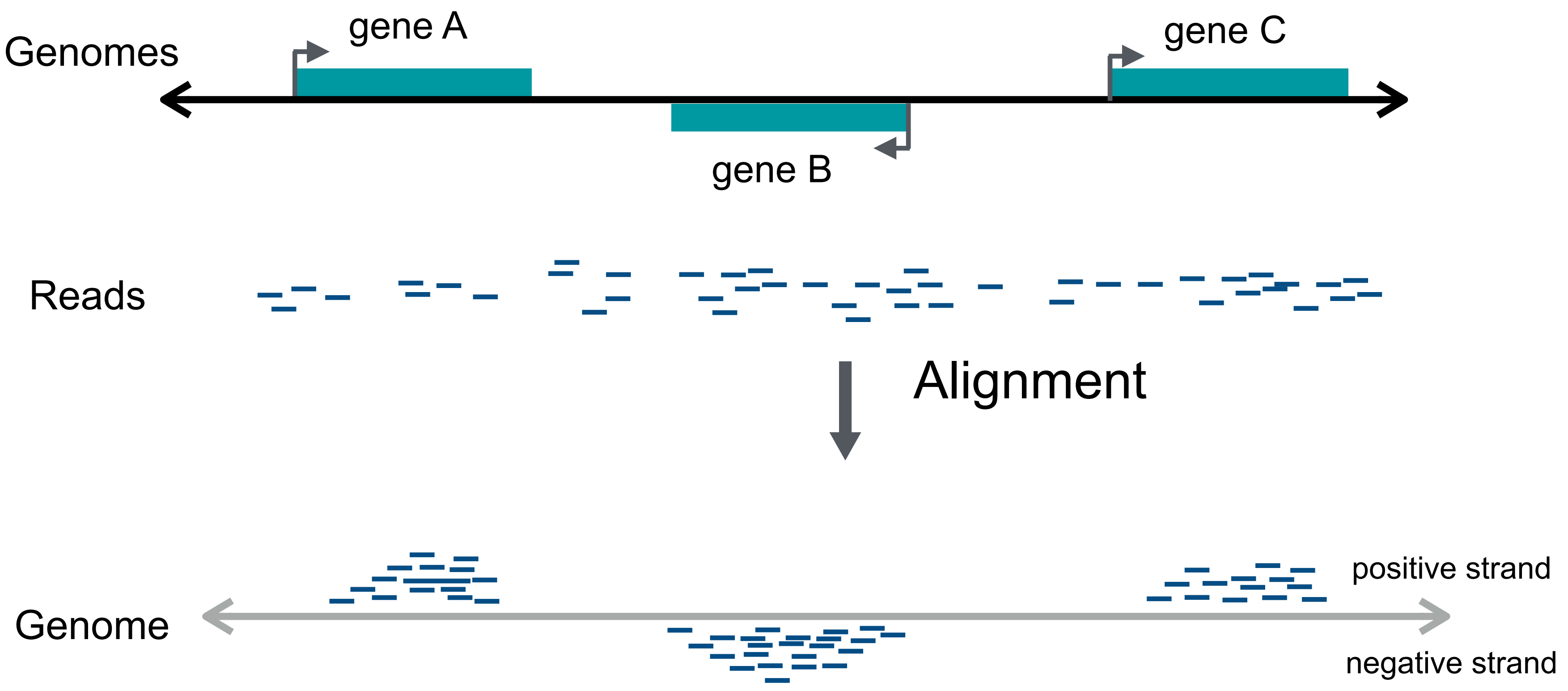
1. Index the reference using *k*-mers.
2. Search reads vs. hash *k*-mers
3. Perform alignment of entire read around seed
4. Report best alignment

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

Also known as *Seed and extend*

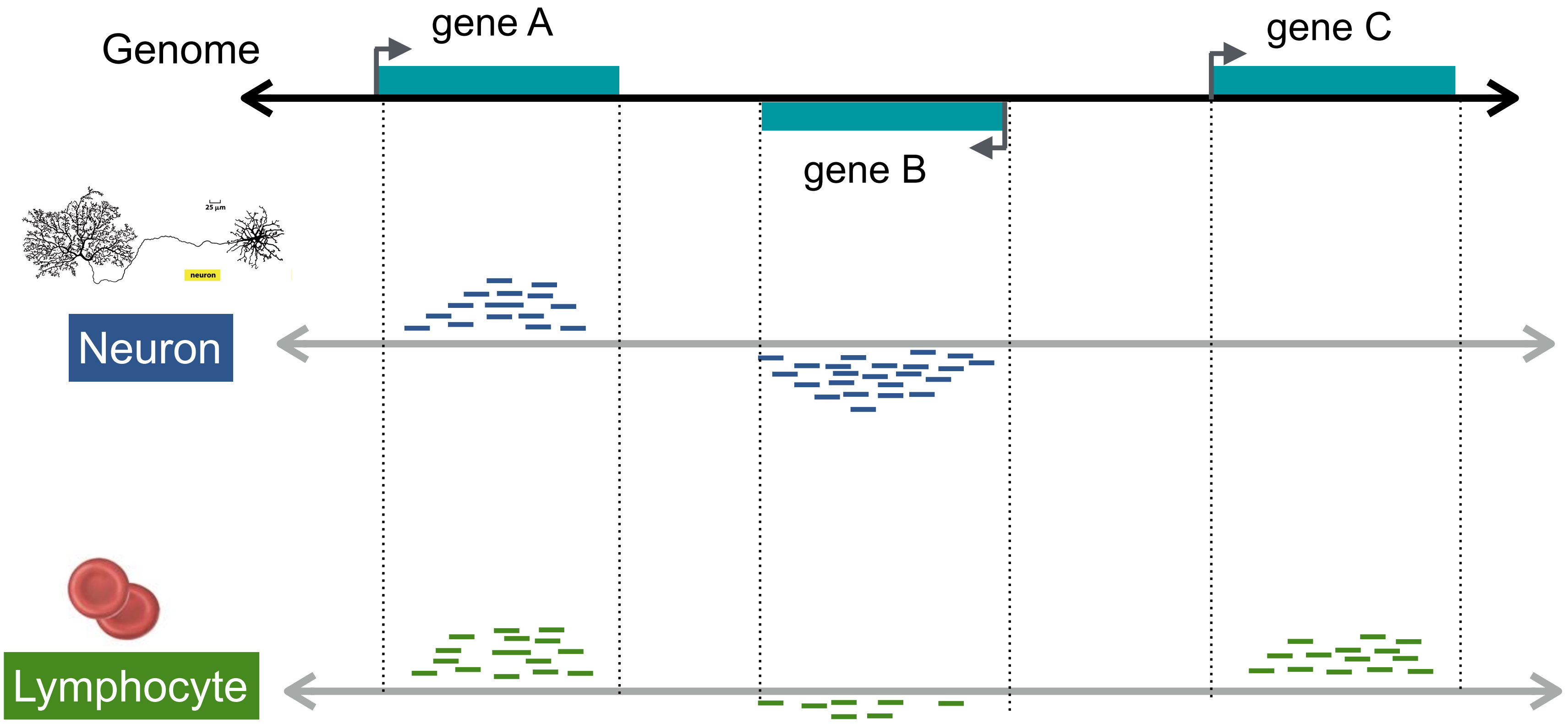
RNA sequencing / Alignment Results

- Position and strand of reads aligned to the genome



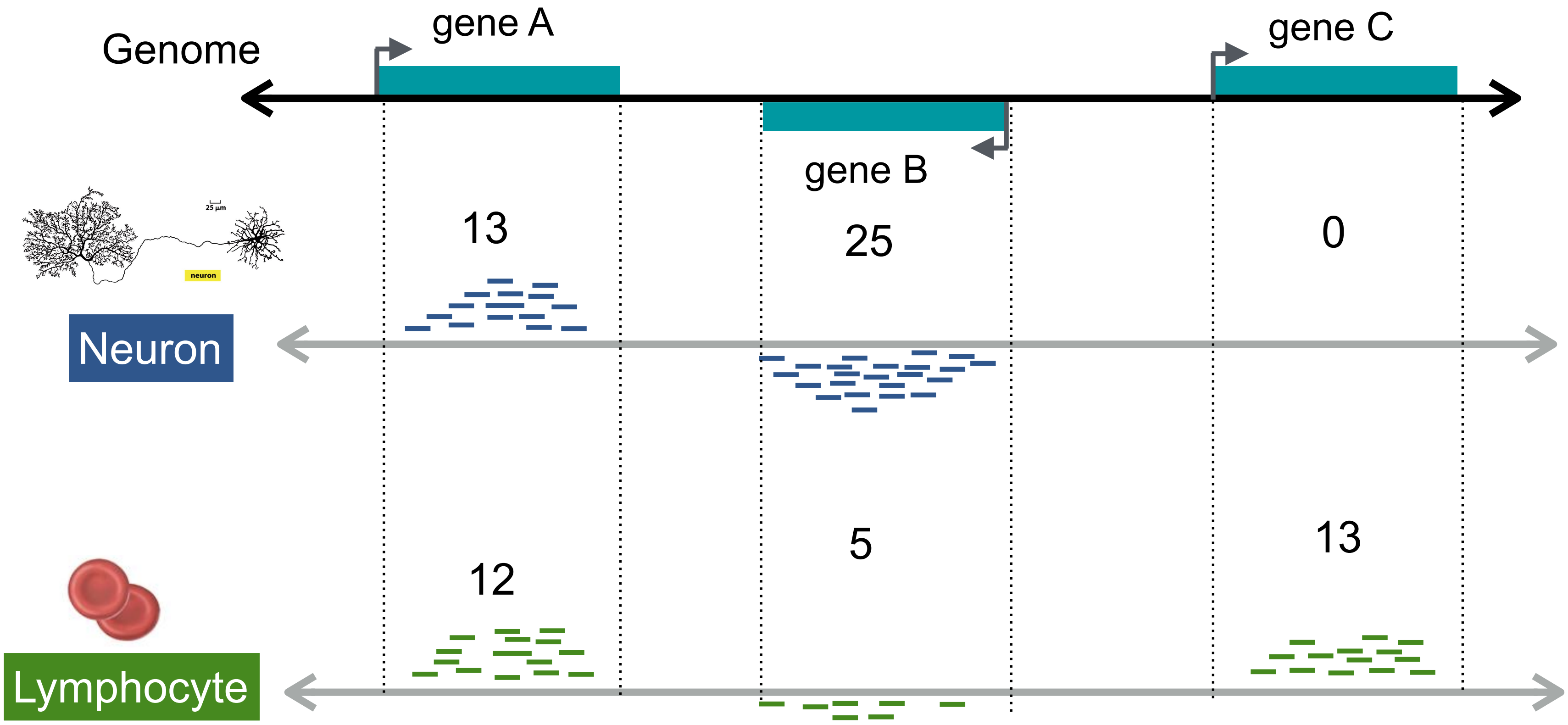
Gene Quantification

- 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)



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*

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

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
     - single cell sequencing (next)
     - computational epigenetic (next weeks)