Genome Informatics

An introduction

Overview

- Course info
- Biological complexity, bioinformatics and genomics definitions
- Molecular biology basics
- Gene sequencing

Course syllabus

- What will we cover in the course:
  - Algorithms for genomic sequence processing
    - Algorithms for string matching
    - Indexing structures for string search
    - Indexing structures for genome-scale string search
    - Algorithms for string similarity (alignment)
    - Dynamic programming
    - Graph algorithms for genome assembly
  - ~~HMM, Evolutionary genetics, phylogenetic trees~~
  - Methods for statistical inference used in genomic data processing
    - Probability distribution
    - Hypothesis testing
    - Quantification and normalization methods
  - (Biological sequence and reference databases: NCBI, EBI ENA, DNA Data Bank of Japan)

Literature

- Dan Gusfield: **Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology,** Cambridge University Press
- Pavel Pevzner, Neils Jones: **An Introduction to Bioinformatics Algorithms (Computational Molecular Biology)**, MIT Press
- R. Durbin, S. Eddy, A. Krogh, G. Mitchinson: **Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids** , Cambridge University Press
- Veli Mäkinen, Djamal Belazzougui, Fabio Cunial, Alexandru I. Tomescu: **Genome-Scale Algorithm Design: Biological Sequence Analysis in the Era of High-Throughput Sequencing**, Cambridge University press

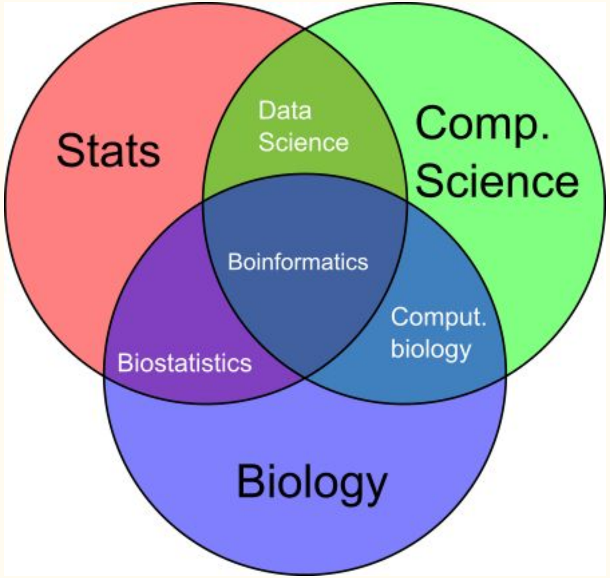
Grading

- Exam will have both theoretical and practical part
    - 40% on the exam
    - 60% during the semester
        - 20% essay on the certain topic
        - 40% project/lab exercises/test

What is bioinformatics

Bioinformatics, n. The science of information and information flow in biological systems, esp. of the use of computational methods in genetics and genomics. *(Oxford English Dictionary)*

Bioinformatics - using statistical and computing methods that aim to solve biological problems.



What is bioinformatics

"I do not think all biological computing is bioinformatics, e.g. mathematical modelling is not bioinformatics, even when connected with biology-related problems. In my opinion, bioinformatics has to do with management and the subsequent use of biological information, particular genetic information."

-- Richard Durbin

Bioinformatics in practice: Develops methods and software tools for storing, retrieving, organizing and analyzing biological data.

Biological complexity of life

“It must be admitted that the biological examples which it has been possible to give in the present paper are very limited.

This can be ascribed quite simply to the fact that **biological phenomena are usually very complicated**. Taking this in combination with the relatively elementary mathematics used in this paper one could hardly expect to find that many observed biological phenomena would be covered.

It is thought, however, that the imaginary biological systems which have been treated, and the principles which have been discussed, should be of some help in interpreting real biological forms.”

– Alan Turing, The Chemical Basis of Morphogenesis, 1952

Genomics 101

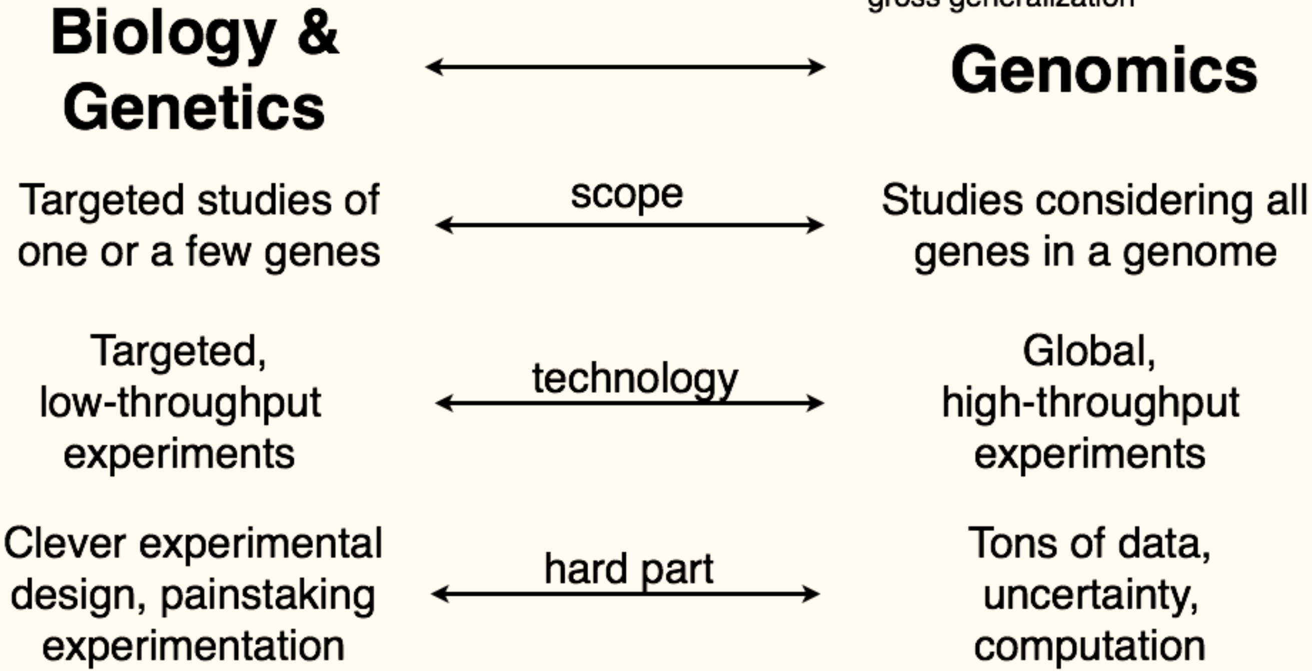
Genome: “The complete set of genes or genetic material present in a cell or organism.” *(Oxford English Dictionary)*

- “Blueprint” or “recipe” of life
- Human genome - 3 billions of base-pairs (A, C, T, G) letters
  - Can be imagined as a sting 3 billion letters long

Genomics: “The branch of molecular biology concerned with the **structure**, **function**, **evolution**, and **mapping of genomes**.

Genomics: contrast with biology & genetics*

* Everything on this slide is
a
gross generalization



Biology & Genetics		Genomics
Targeted studies of one or a few genes	scope	Studies considering all genes in a genome
Targeted, low-throughput experiments	technology	Global, high-throughput experiments
Clever experimental design, painstaking experimentation	hard part	Tons of data, uncertainty, computation

Used with permission of Ben Langmead

Human Genome Project

- Reference genome for human species
- To be used as a coordinate system

Human genome project

- World’s largest collaborative biological project
- From 1990 - 2003, price over 3B $
- Mapped around 22.000 genes
- Opened some issues due to genome size

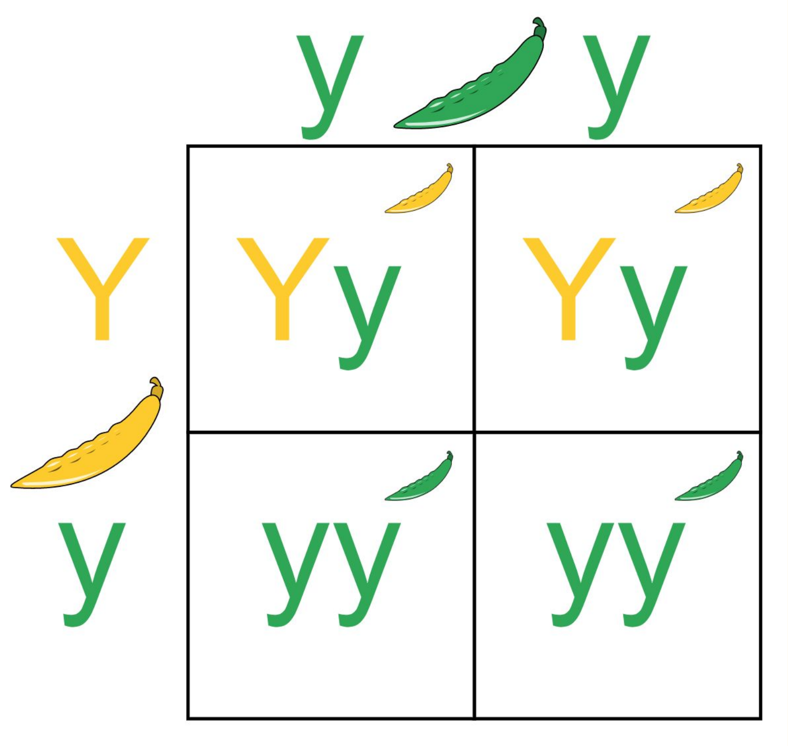
Biology 101 - genotype vs phenotype

The **genotype** is the part of the genetic makeup of a cell, and therefore of an organism or individual, which determines one of its characteristics (phenotype).

A **phenotype** (from Greek *phainein* , meaning 'to show ', and *typos* , meaning 'type') is the composite of an organism's **observable characteristics** or traits, such as its morphology, development, biochemical or physiological properties, behavior, and products of behavior (such as a bird's nest).

Rules of inheritance

- Mendelian inheritance
- Multiple alleles of the same gene
  - One allele per chromosome
  - Dominant/recessive allele

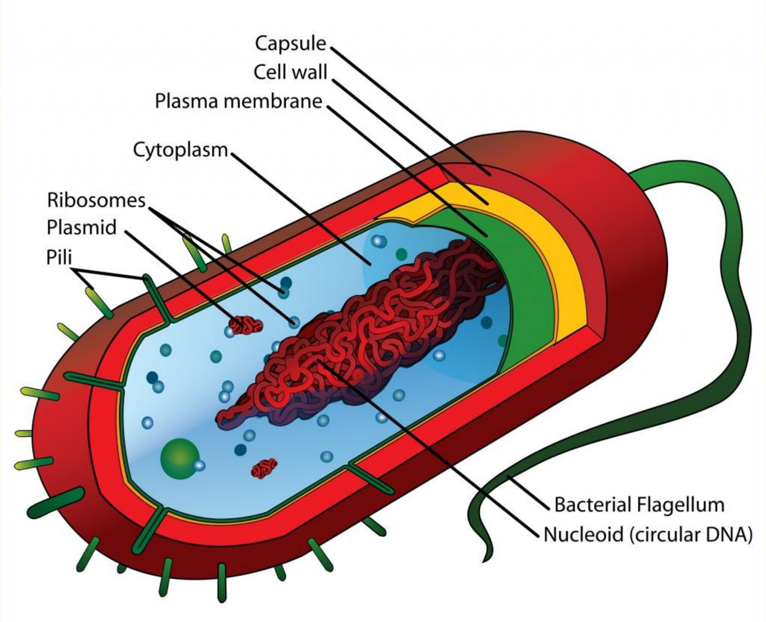


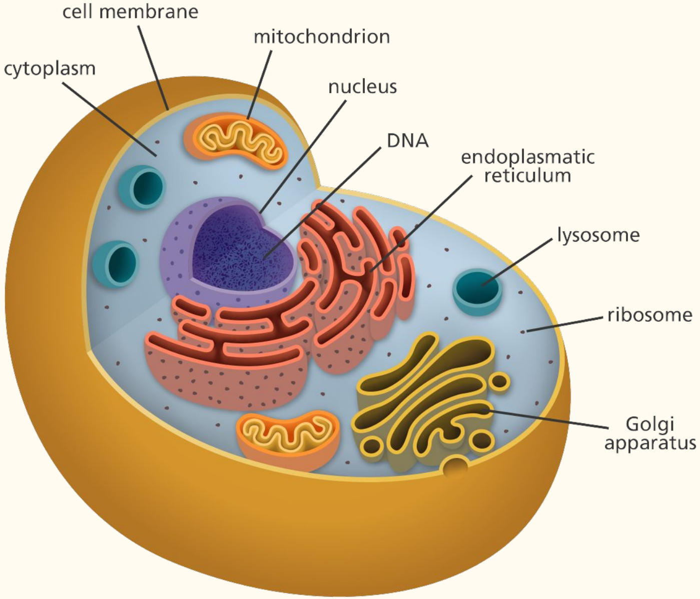
	y	y
Y	Yy	Yy
y	yy	yy

Cell

Fundamental working units of every living system.

- Prokaryotic (bacteria)
- Eukaryotic (higher organisms - animals, plants)

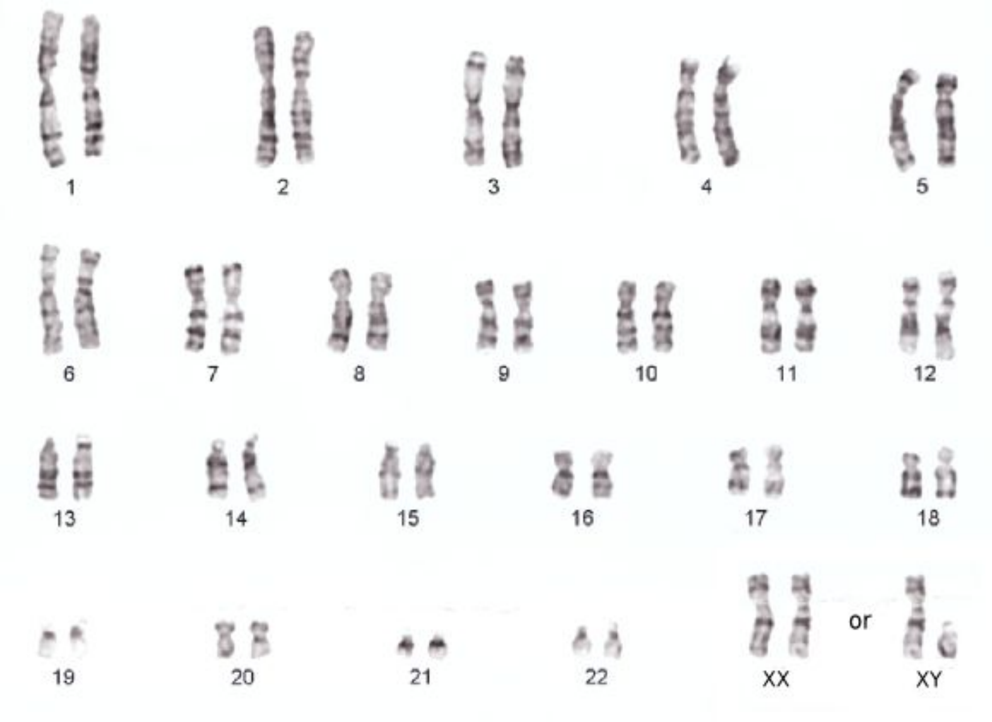




Genome

- Set of all pairs of chromosomes

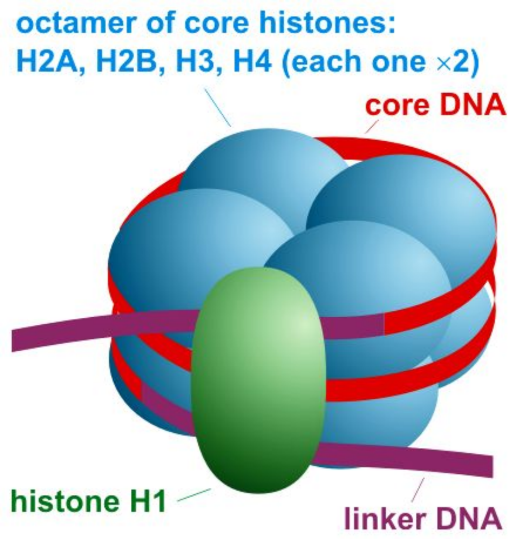
- Human genome:
  - 23 pair of chromosomes
  - 22 autosomes
  - 1 sex chromosome (X and/or Y)
  - 3 billion base-pairs

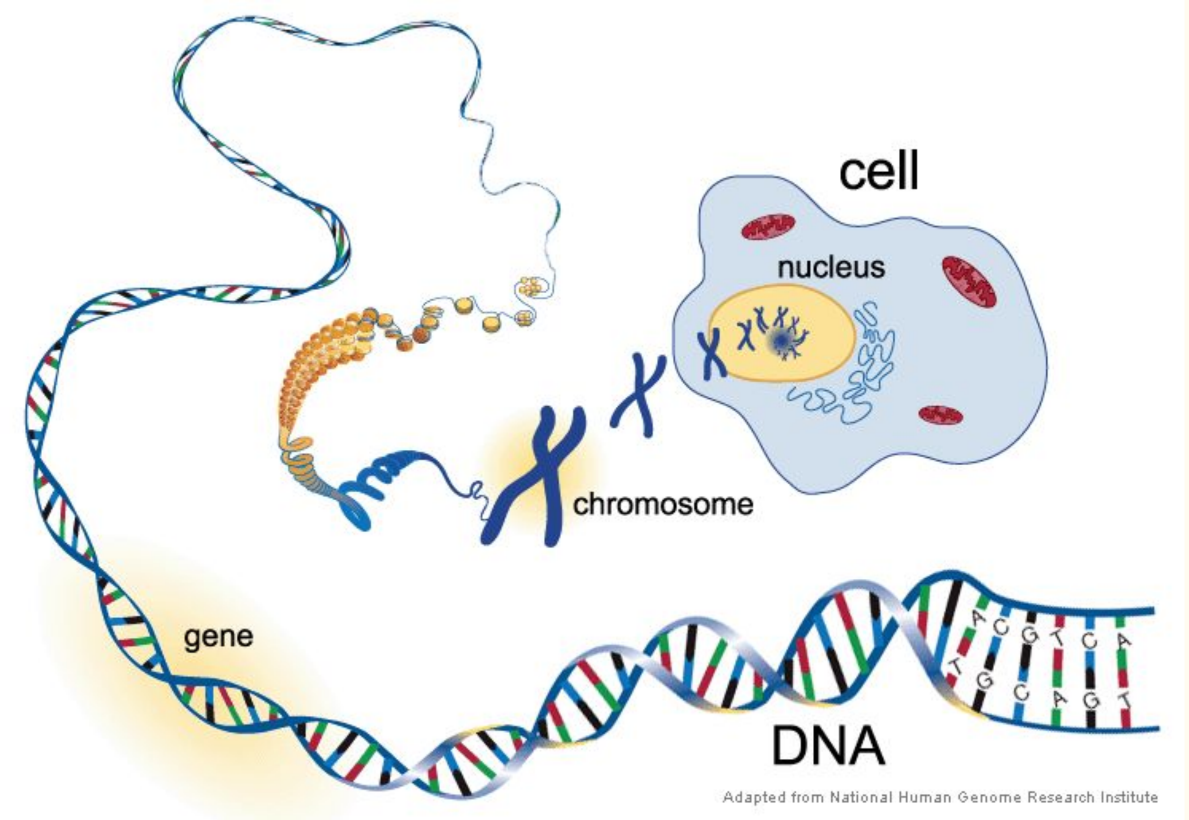


Karyotype

How DNA is packed

- Chromatin - tightly packed DNA
- A **nucleosome** is a basic unit of DNA packaging in eukaryotes, consisting of a segment of DNA wound in sequence around eight histone protein cores.
- Current model



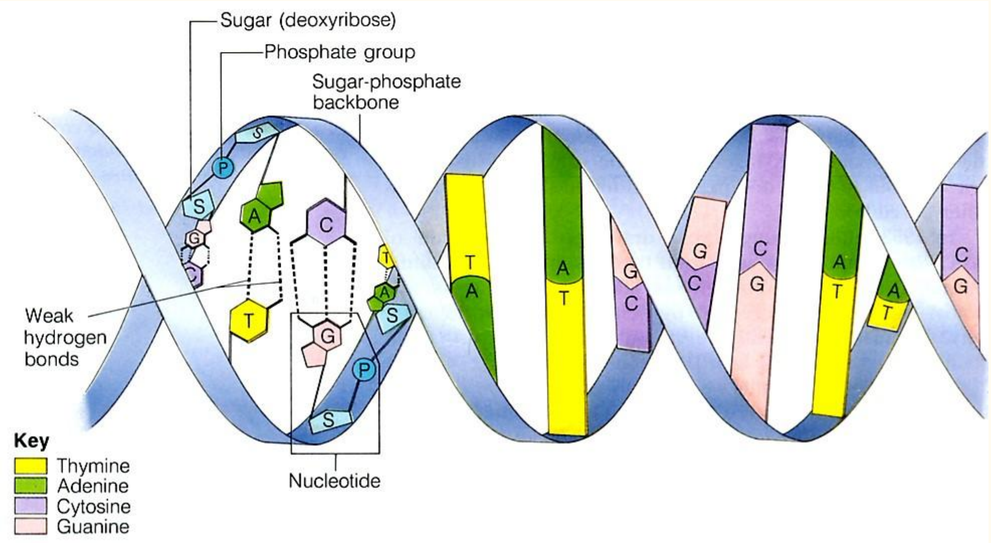


Adapted from National Human Genome Research Institute

DNA - code of life

- DNA (deoxyribonucleic acid) - double stranded molecule
- Double helix structure
- More stable, redundant information - complementary chain
- Base pairs (complementary bases)
  - A - T (adenine and thymine)
  - C - G (cytosine and guanine)
- Nucleobase vs nucleotide

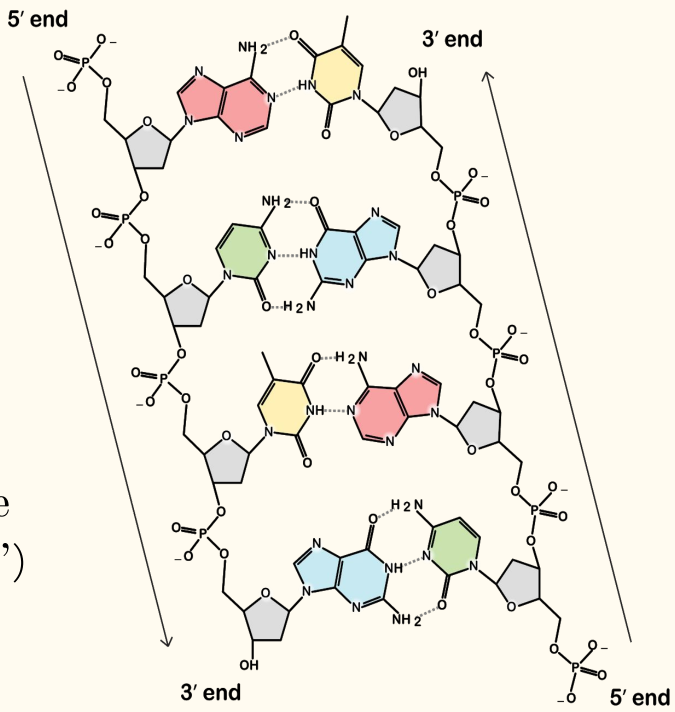
Nucleotide = nucleobase + sugar + phosphate group

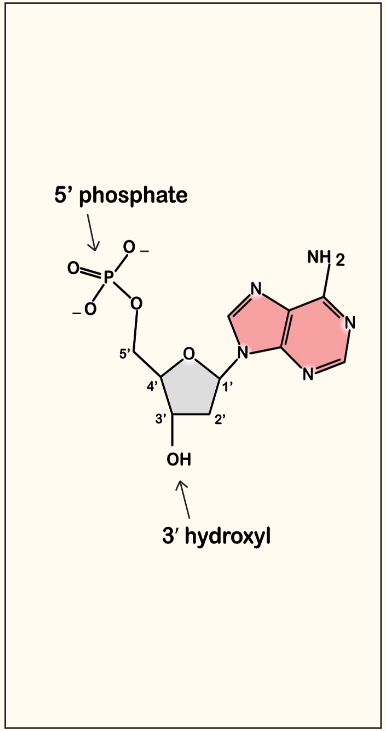


DNA - structure

- Consists of:
  - Phosphate group
  - Sugar (deoxyribose)
  - Nitrogen base
- Hydrogen bonds
- Forward and reverse strand
- DNA direction:
  - 5’ head and 3’ tail
  - Transcribed from 5’ to 3’ end
- In bioinformatics we write just one strand (by convention from 5’ to 3’)

5’ ACTG *3’*
↕
3’ TGAC *5’*
(reverse complement)





Nucleotide

DNA - discovery

1952-1953 James D. Watson and Francis H. C. Crick deduced the double helical structure of DNA from X-ray diffraction images by Rosalind Franklin (provided by M. Wilkins) and data on amounts of nucleotides in DNA.

“Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid”

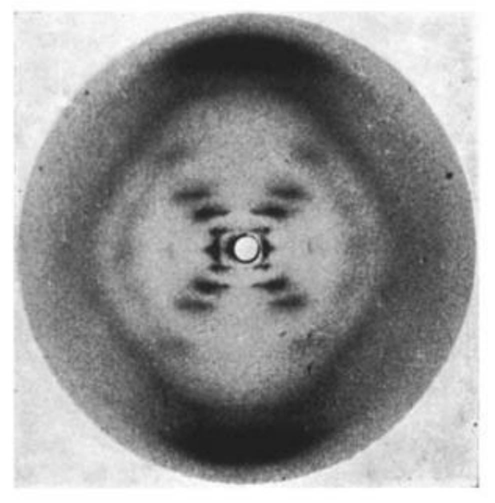
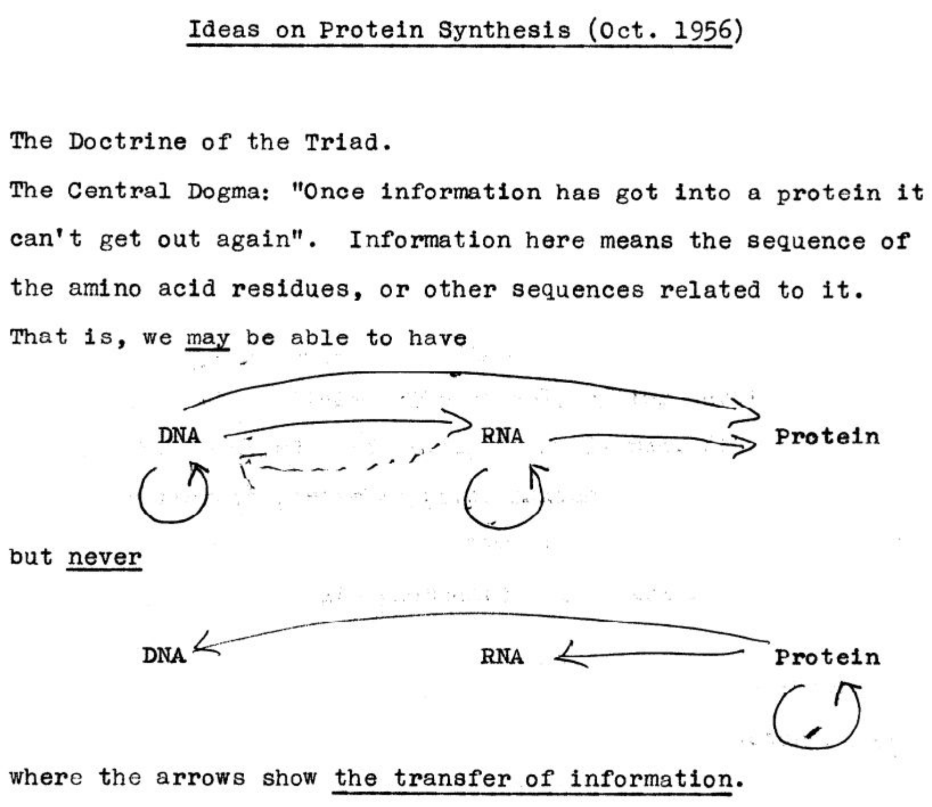


Photo 51

Central dogma of molecular biology



Ideas on Protein Synthesis (Oct. 1956)

The Doctrine of the Triad.

The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it.

That is, we may be able to have

DNA
RNA
Protein

but never

DNA
RNA
Protein

where the arrows show the transfer of information.

Central dogma of molecular biology

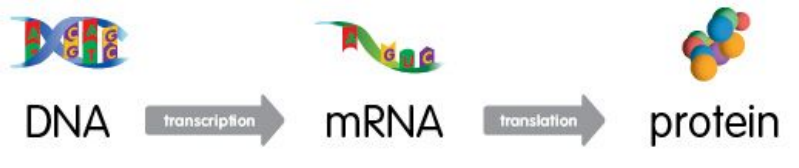
DNA ----> RNA -----> Protein

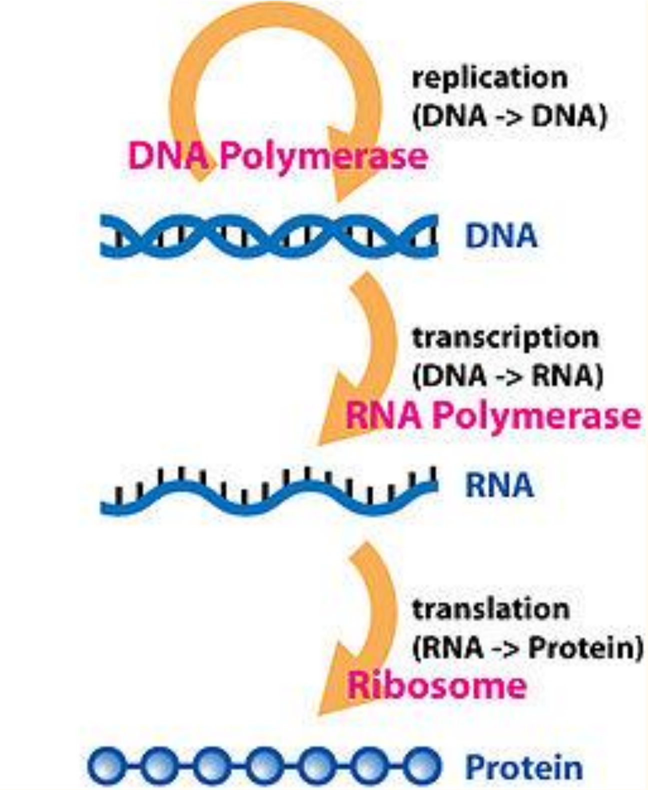
Transcription: DNA ->RNA

- particular segment of DNA is copied into RNA (especially mRNA) by the enzyme RNA polymerase.

Translation: RNA -> Protein

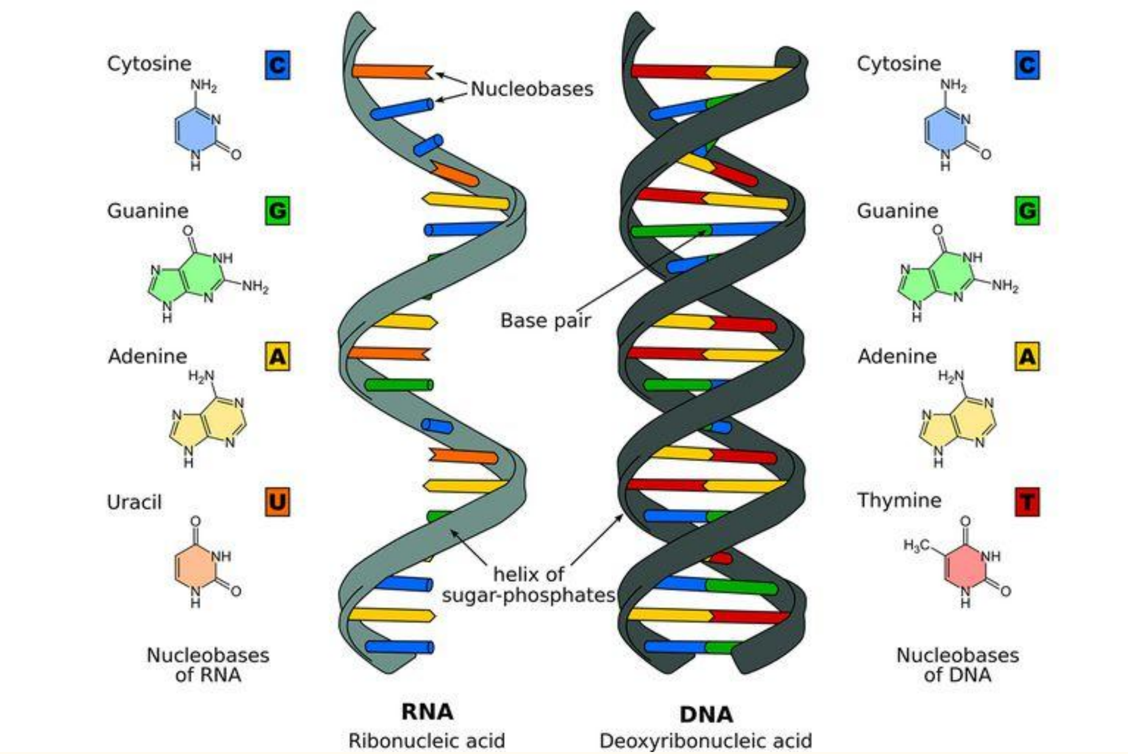
- process in which ribosomes synthesize proteins after the process transcription of DNA to RNA in the cell's nucleus.





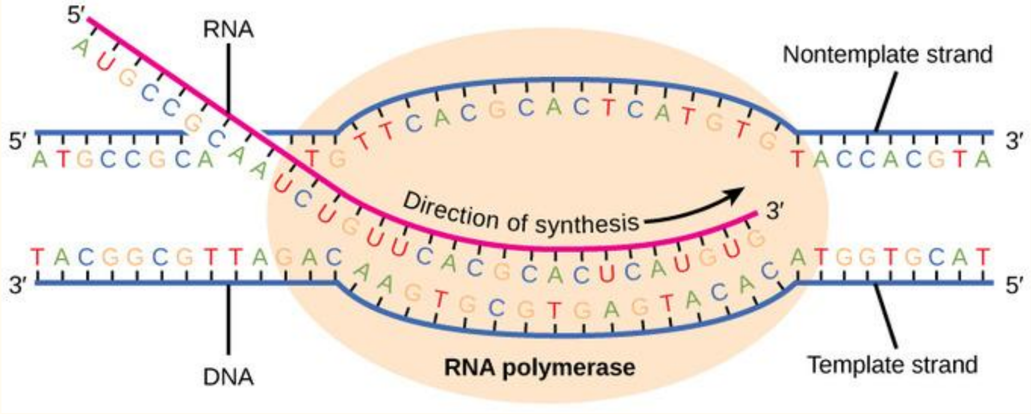
RNA

- Single stranded
- Sugar:
  - ribose (instead of deoxyribose)
- Uracil instead of Thymine

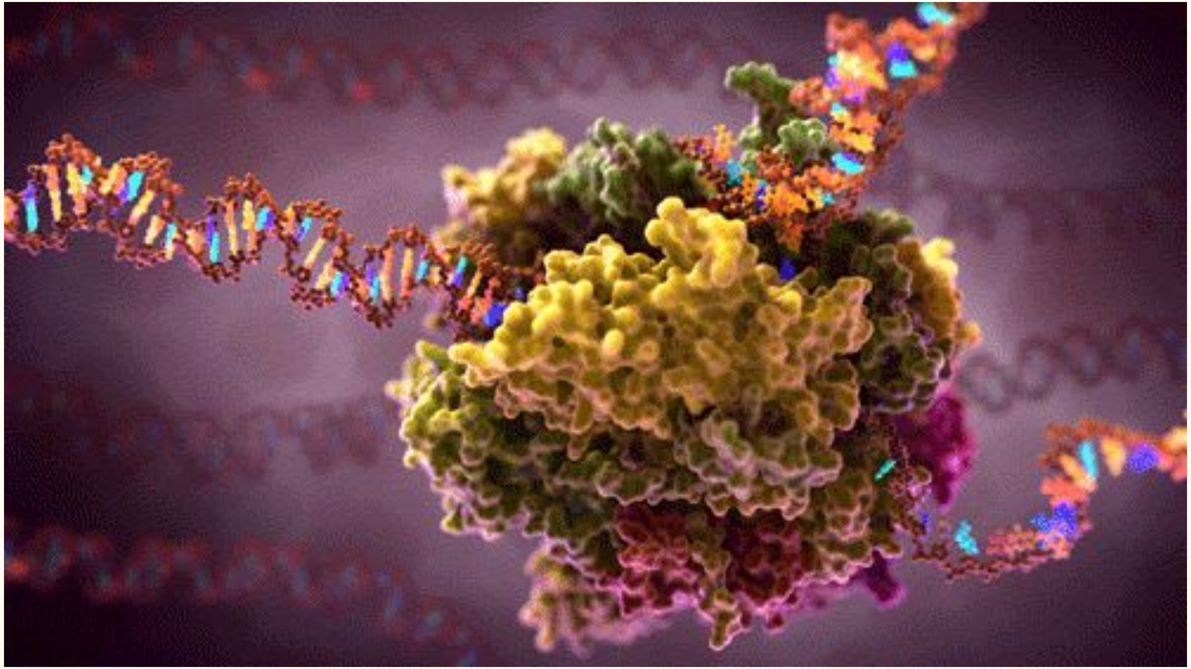


Transcription

- Template (noncoding) strand
  - One which is transcribed by RNAP (RNA polymerase)
- Nontemplate (coding) strand
  - Not transcribed
  - Same sequence as newly created RNA

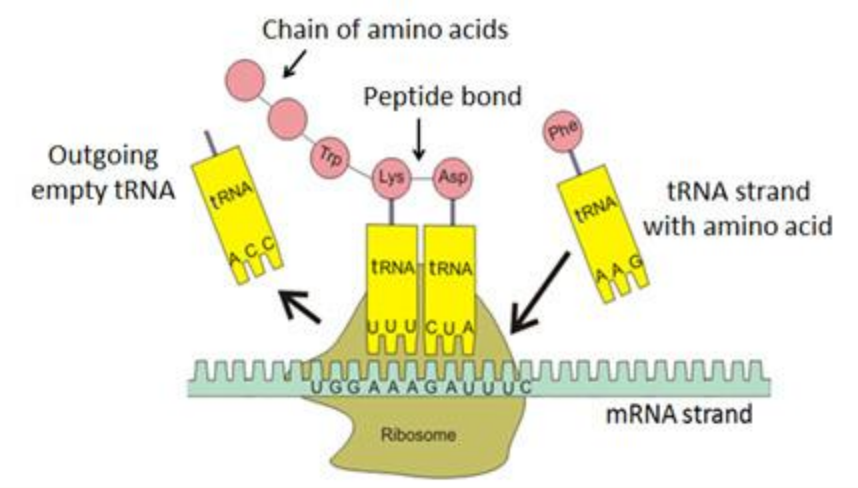


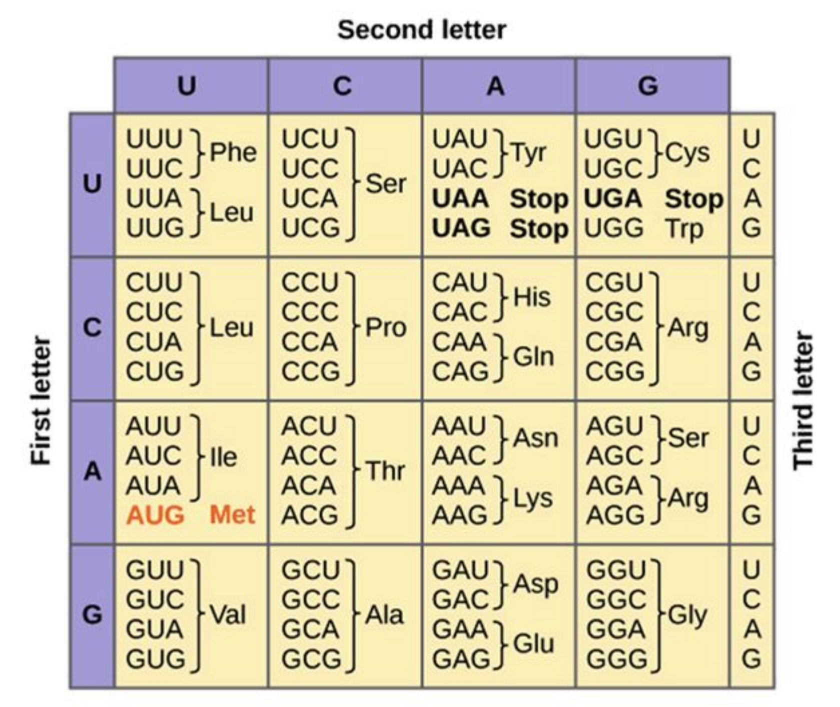
Transcription



Translation

- Occurs in ribosome
- Each triplet of nucleotides (codon) codes for specific amino-acid
  - “Letters of protein code”
  - 20 amino-acid (some redundancy)



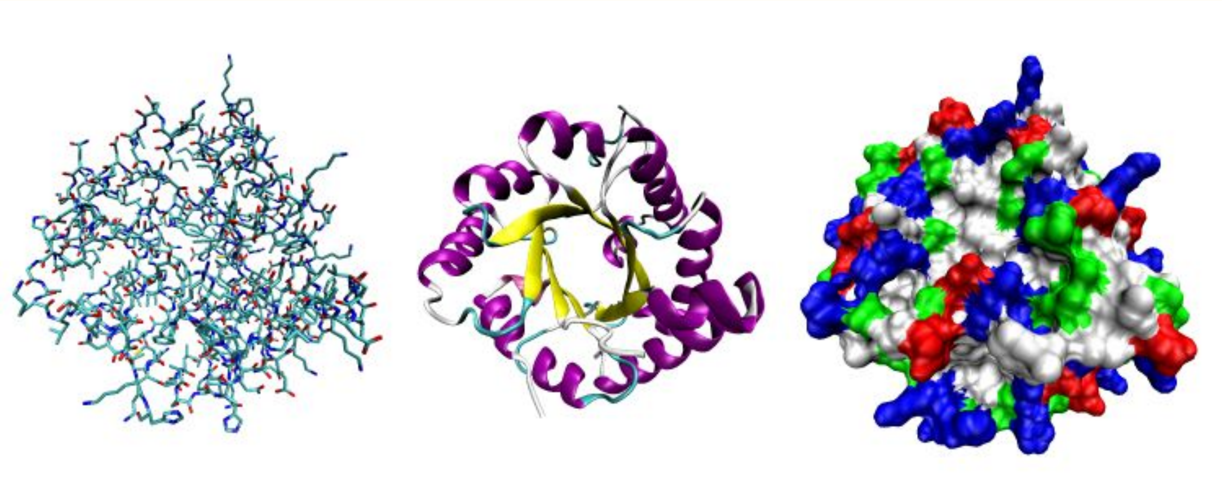


Second letter

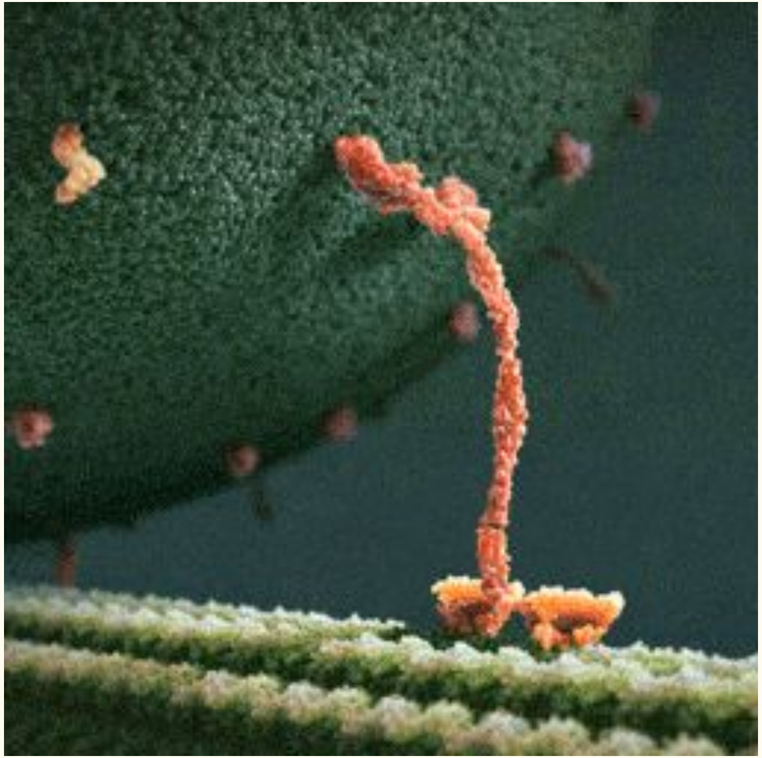
First letter	U	C	A	G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC	UCC	UAC	UGC	C
	UUA Leu	UCA	**UAA Stop**	**UGA Stop**	A
	UUG	UCG	**UAG Stop**	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Gln	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC	ACC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

Proteins

- Building blocks of life
  - Various functions in the organism (transportation, regulation, metabolism, DNA replication)
- Long chains of amino-acids, that also fold into complicated 3D structures
  - We often distinguish protein primary, secondary, tertiary and quaternary structure



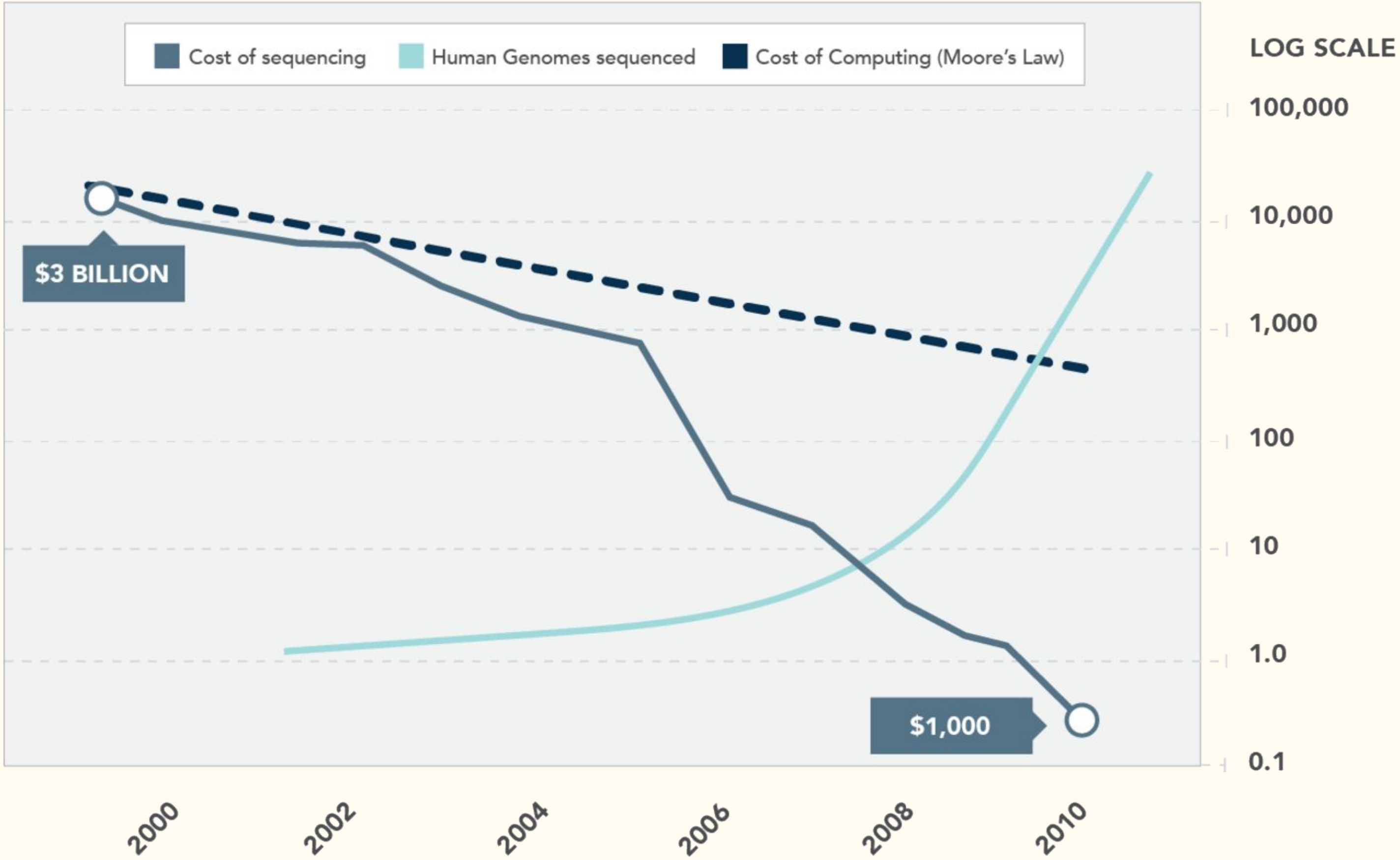
Proteins



Genome sequencing

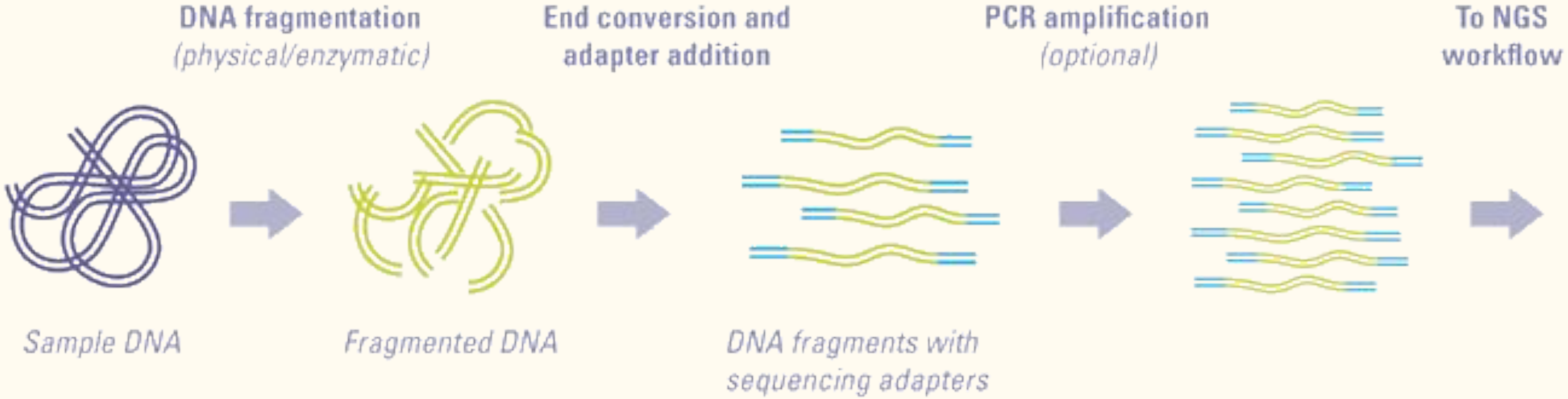
- Digitalization of genome
- Human Genome Project (1990-2003), 3B $
- Sanger sequencing
  - Long (took 13 years)
  - Costly (3B$ for one human genome)
- Currently NGS (next-generation sequencing, second generation sequencing) in use
  - Illumina
  - Around 1000$ and 1 day needed to sequence the genome
- Also third generation sequencing in use
  - Longer read-length (up to 50k base)
  - Oxford nanopore, PacBio
  - Higher error rate

GROWTH OF DNA SEQUENCING

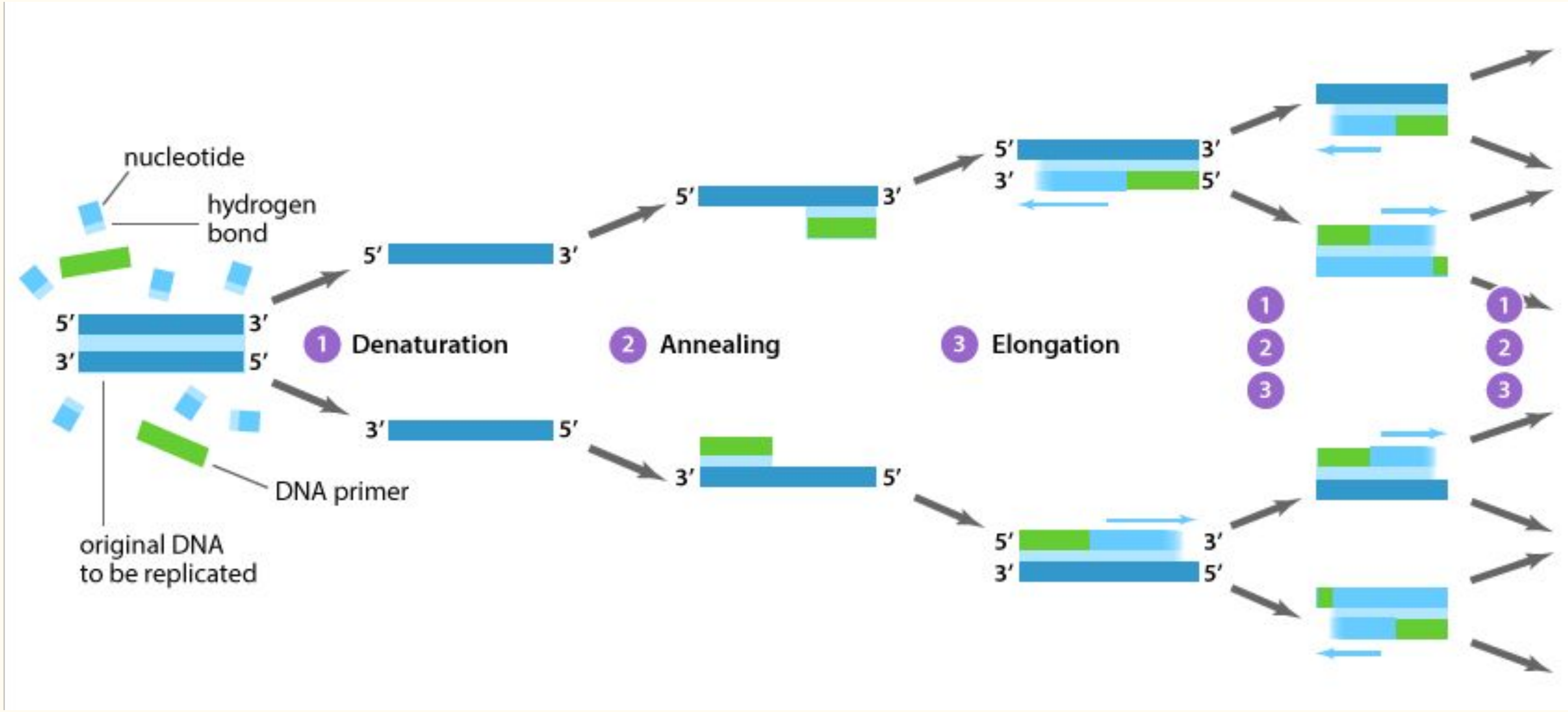


NGS sequencing

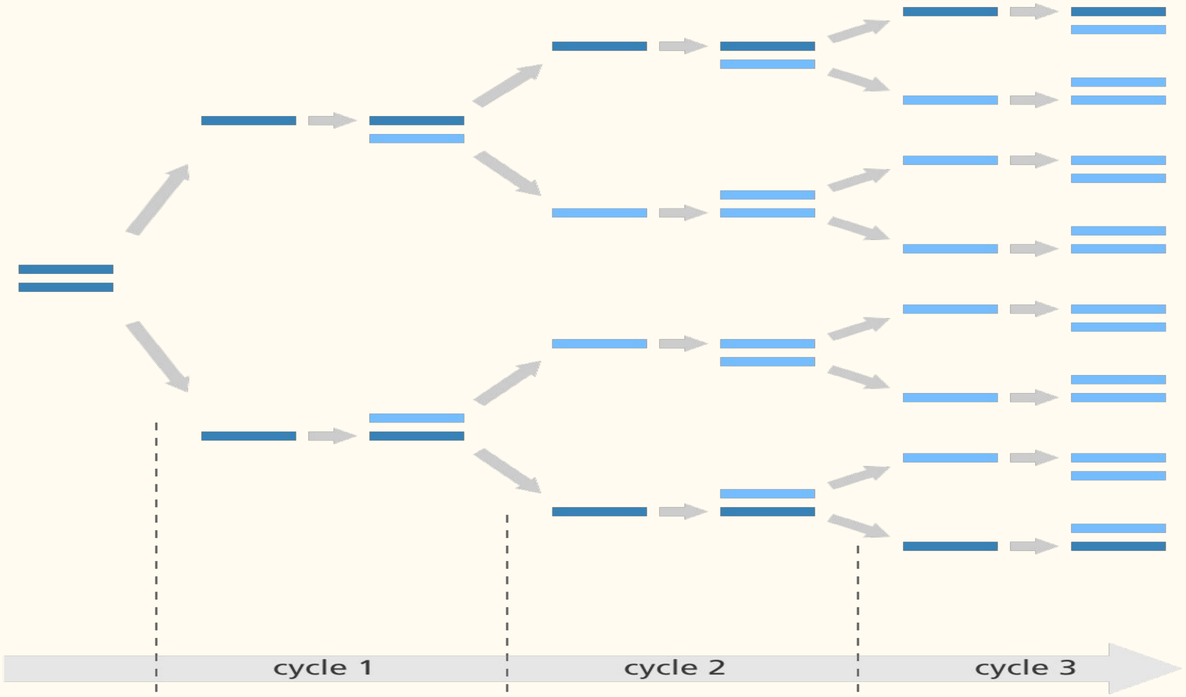
- Read - DNA fragment after reading it in sequencer
- Typical whole genome sequencing experiment:
  - 200-500 million reads
  - 50-150 bases (letters long)



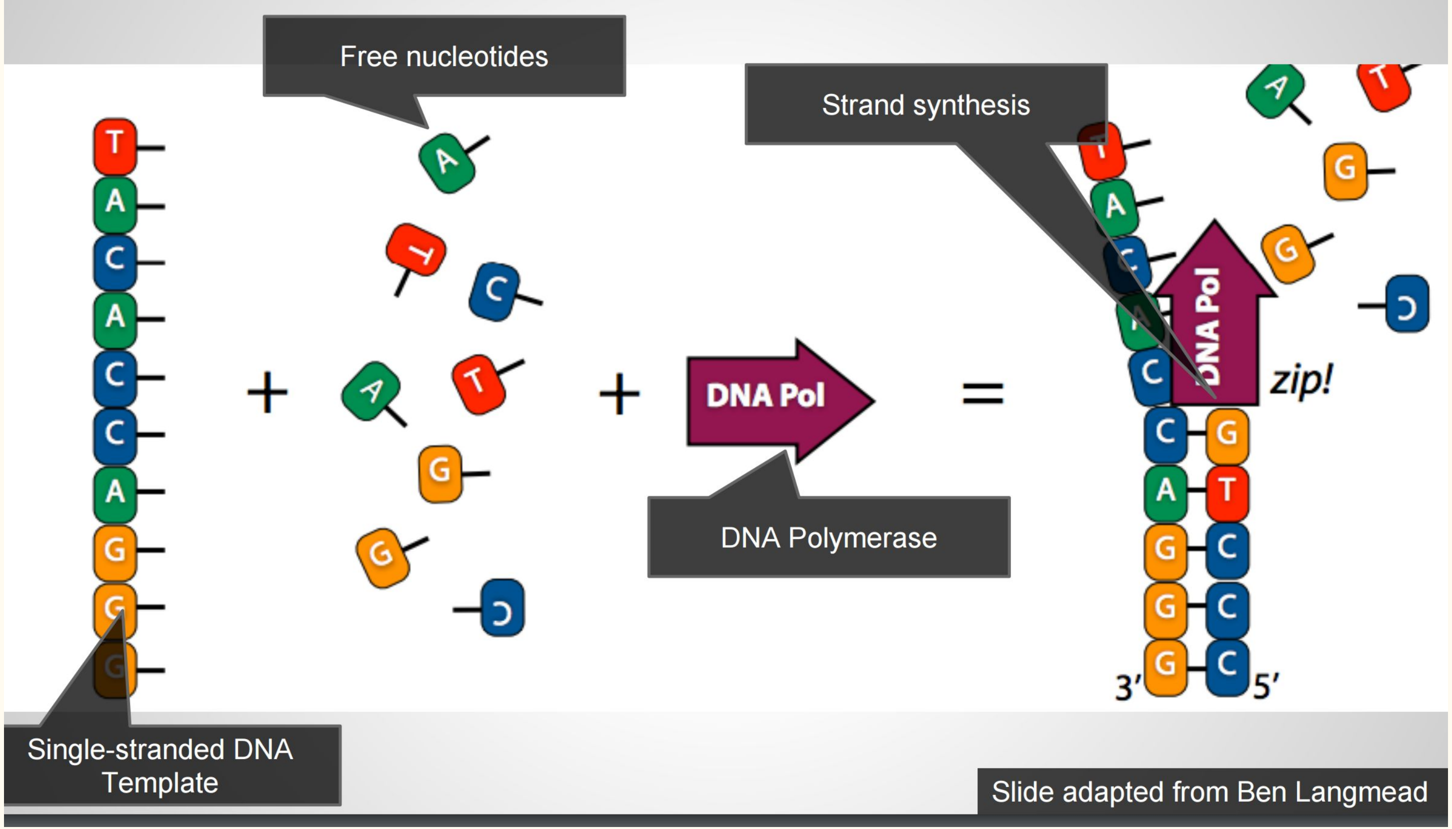
Sequencing - PCR (polymerase chain reaction)



Sequencing - PCR (polymerase chain reaction)

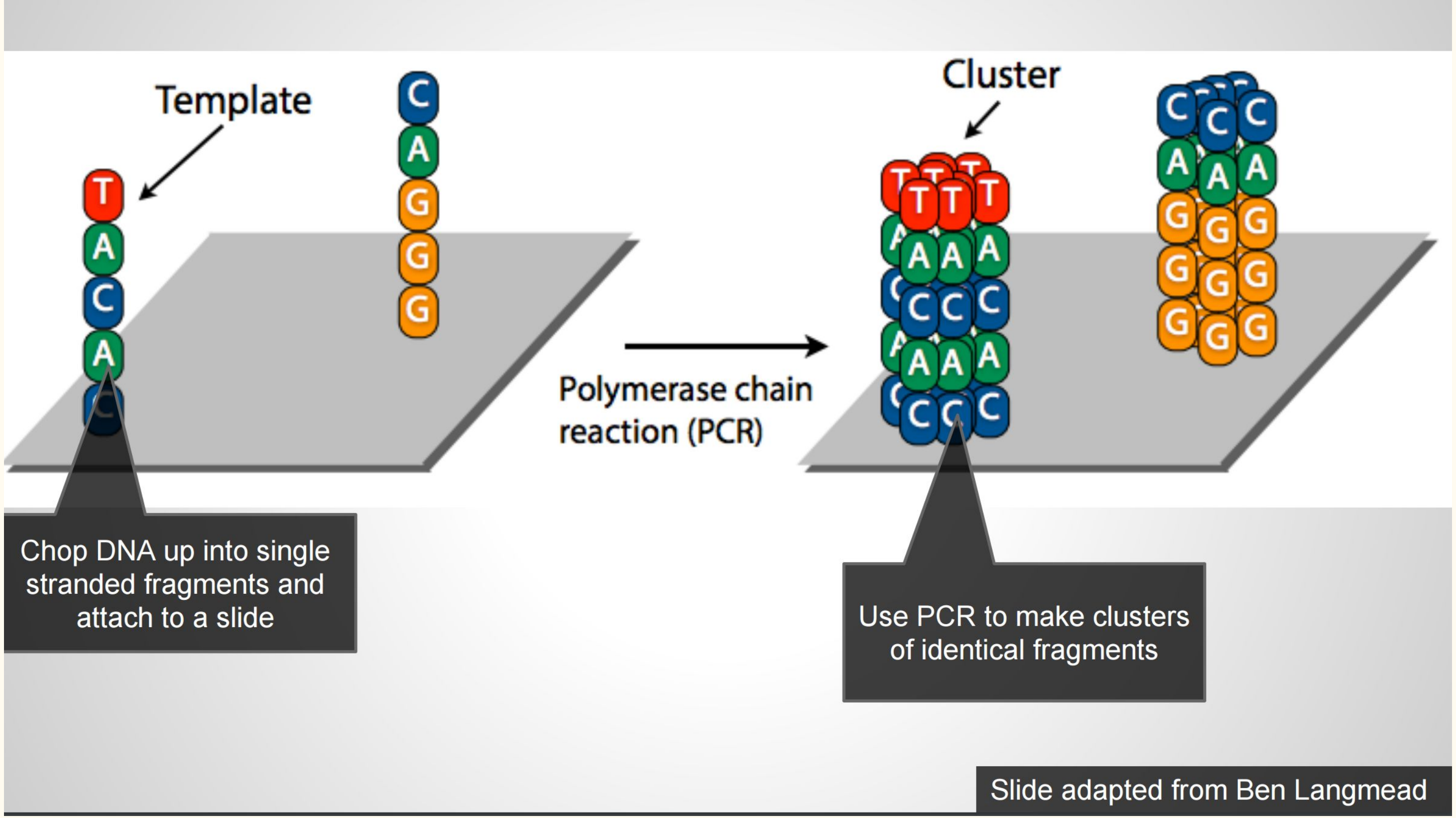


Sequencing (Illumina)



Slide adapted from Ben Langmead

Sequencing (Illumina)

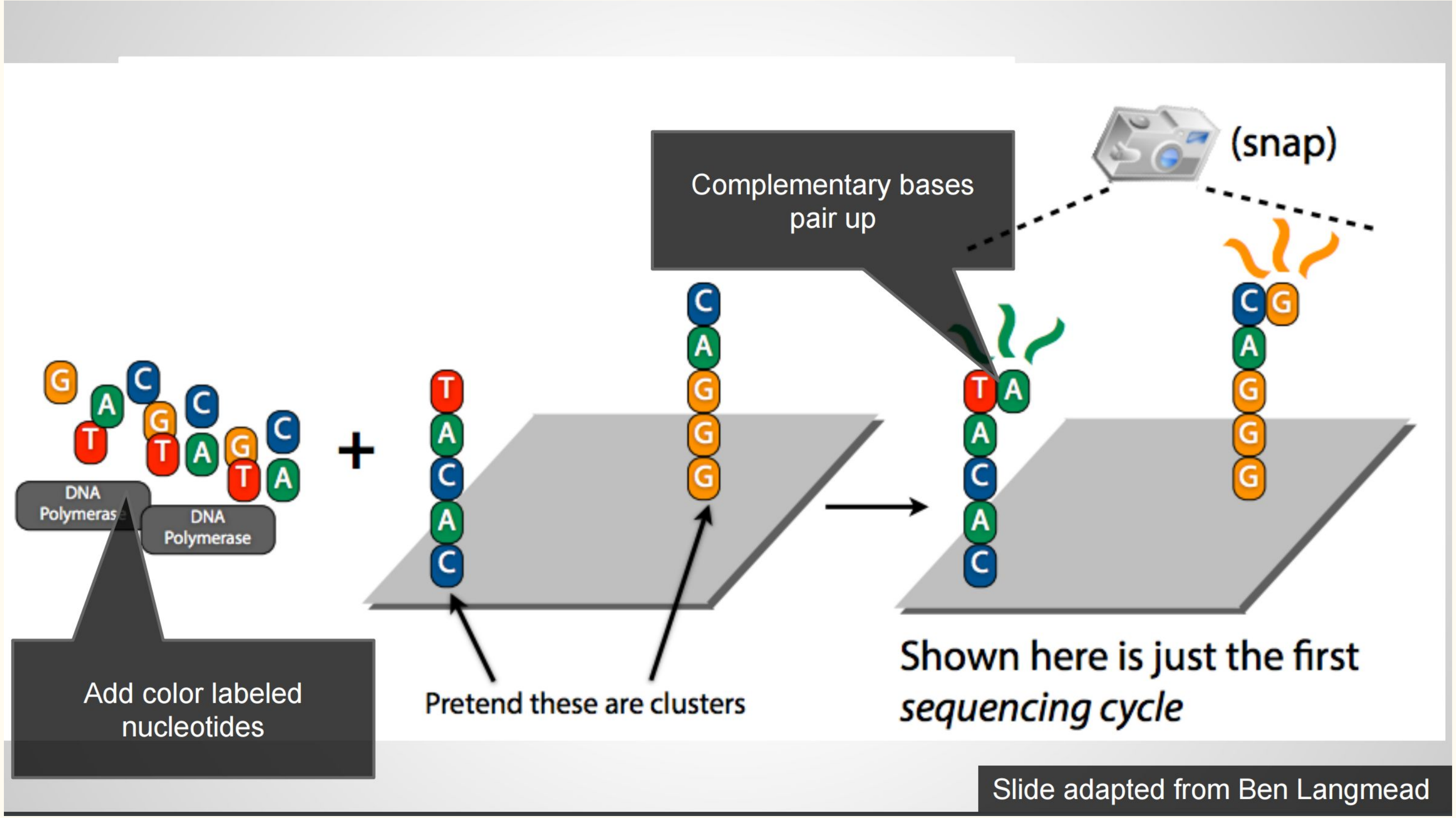


Chop DNA up into single stranded fragments and attach to a slide

Use PCR to make clusters of identical fragments

Slide adapted from Ben Langmead

Sequencing (Illumina)



Complementary bases pair up

(snap)

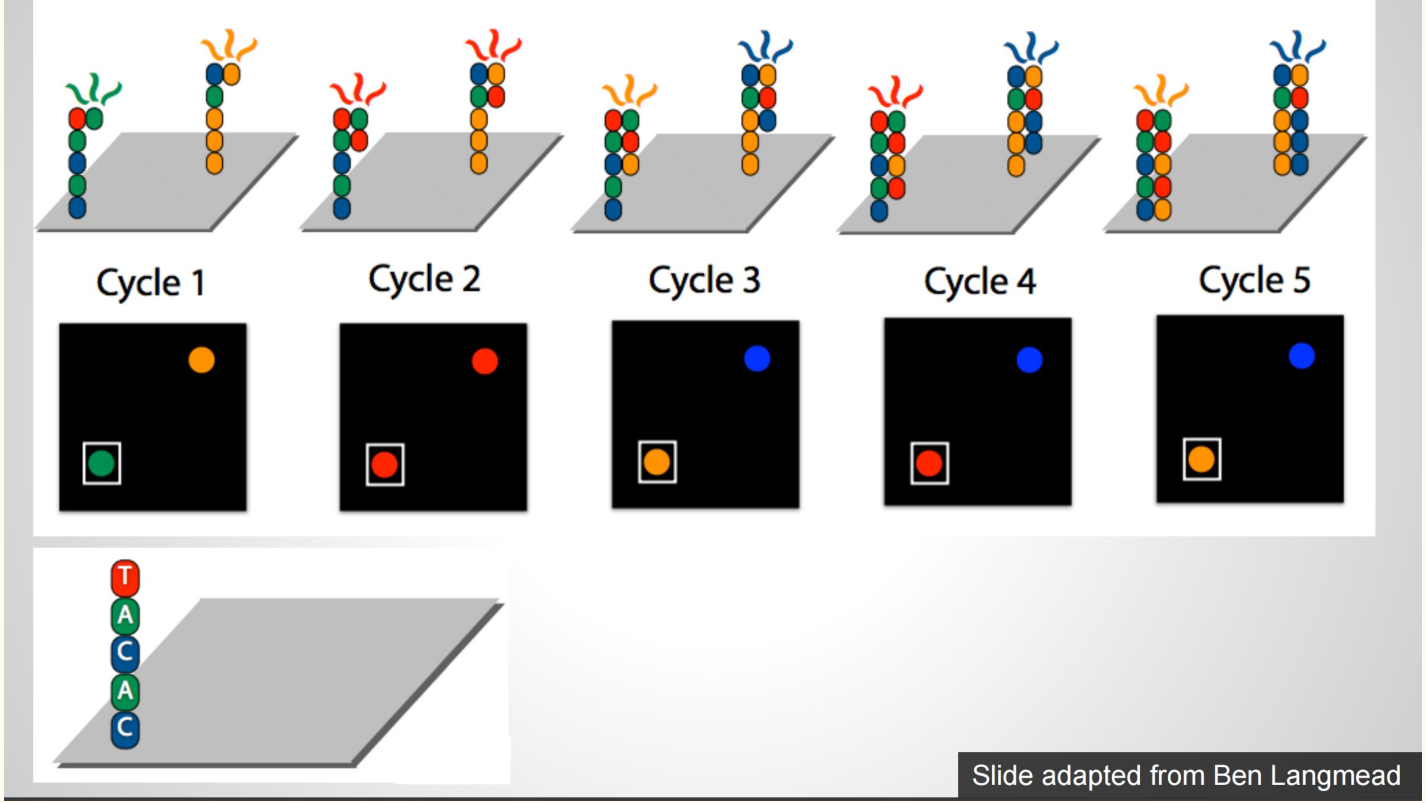
Add color labeled nucleotides

Pretend these are clusters

Shown here is just the first *sequencing cycle*

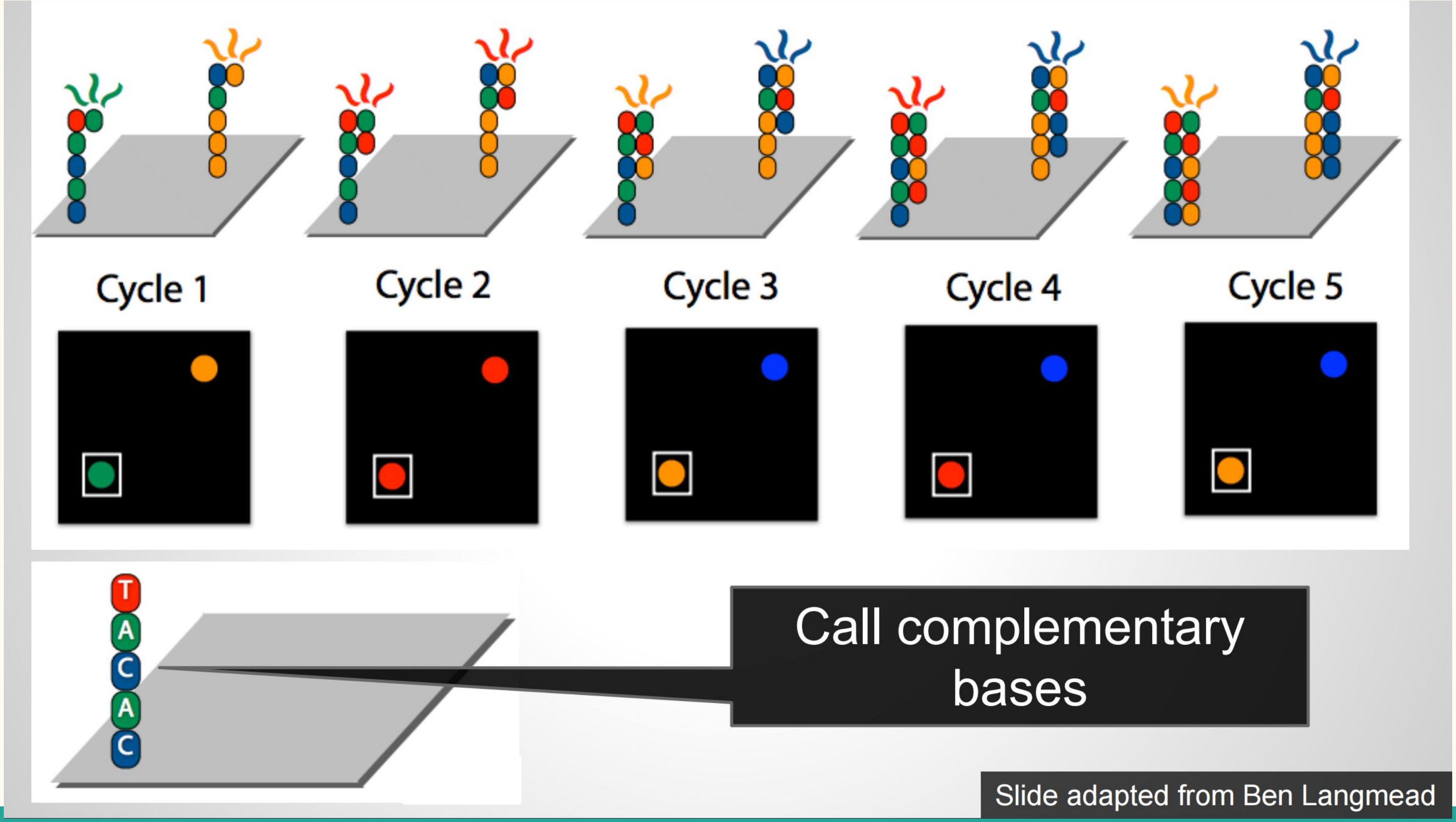
Slide adapted from Ben Langmead

Sequencing (Illumina)



Slide adapted from Ben Langmead

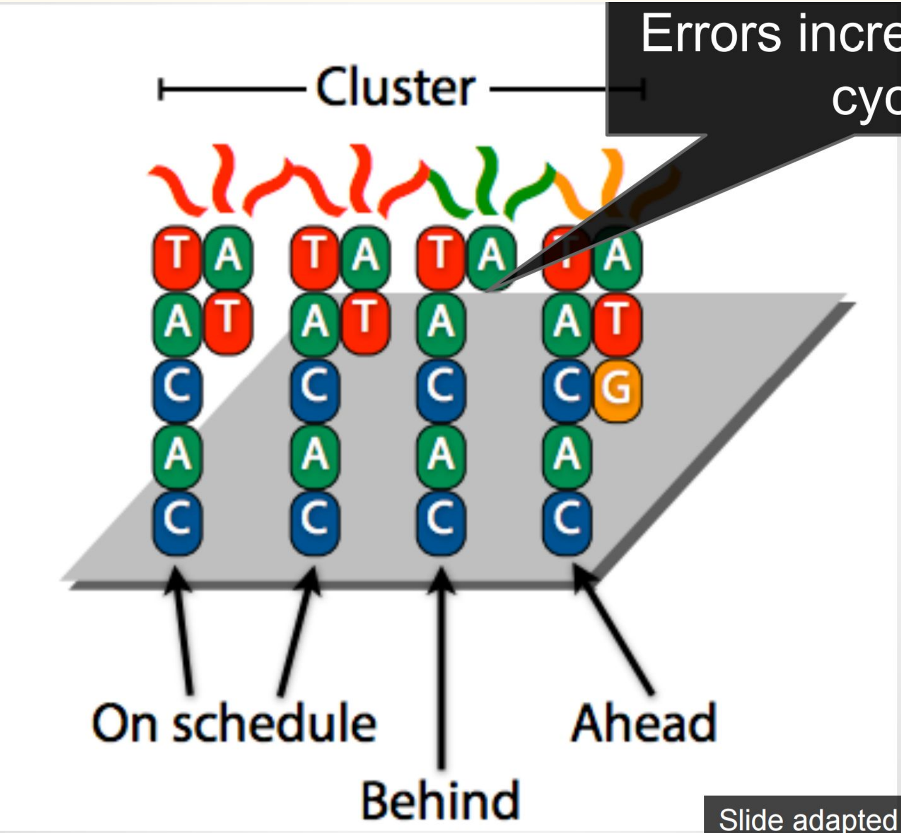
Sequencing (Illumina)



Call complementary bases

Slide adapted from Ben Langmead

Sequencing error

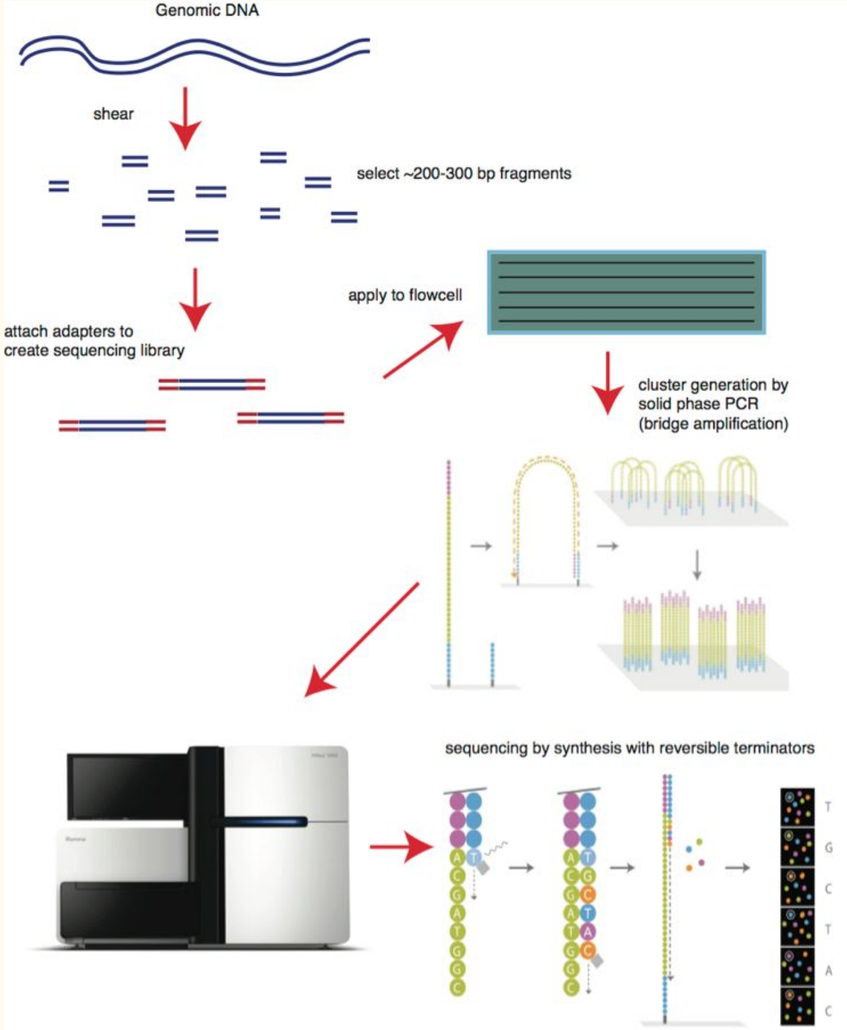


Errors increase in later cycles

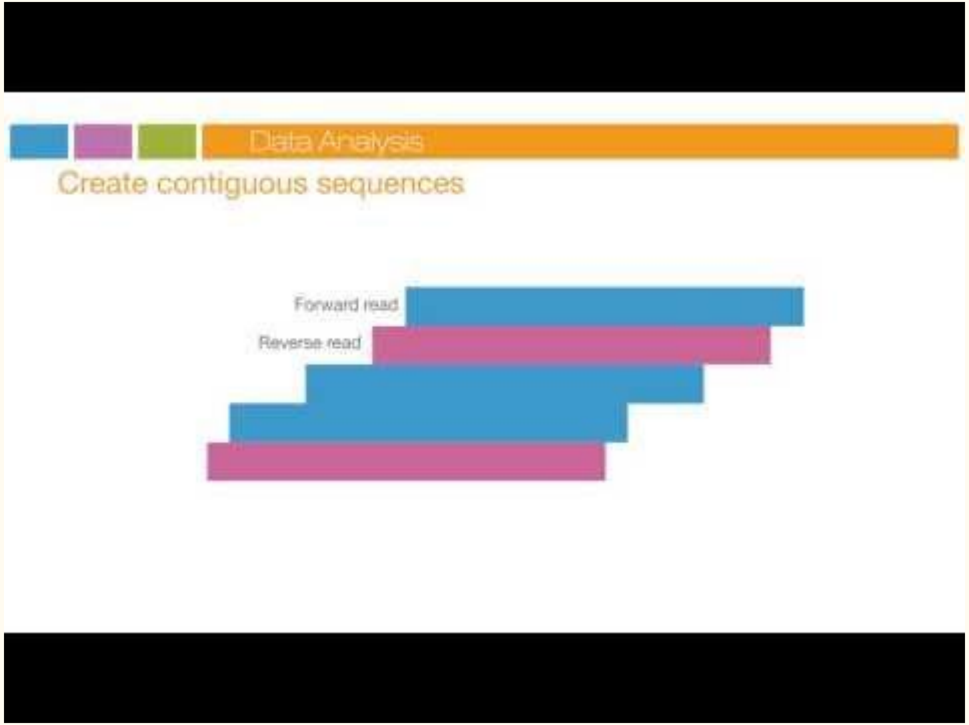
Slide adapted from Ben Langmead

Sequencing (sum up)

1. Shearing (fragmentation of the genome)
2. Attaching adapters
3. PCR amplification (optional)
4. Attaching template to surface/flowcel
5. PCR/bridge amplification (cluster creation)
6. Adding fluorescent bases and taking a picture after each cycle (repeat this many times)
7. Stack up images and read the sequence



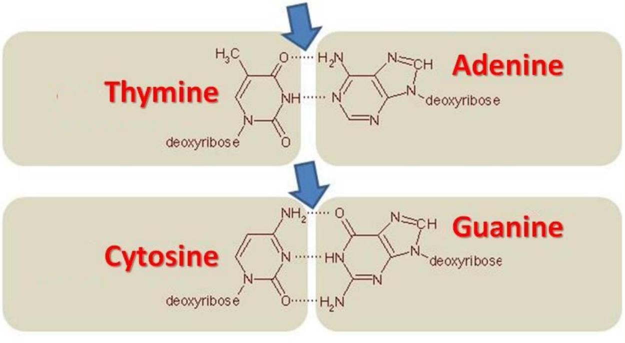
Illumina sequencing

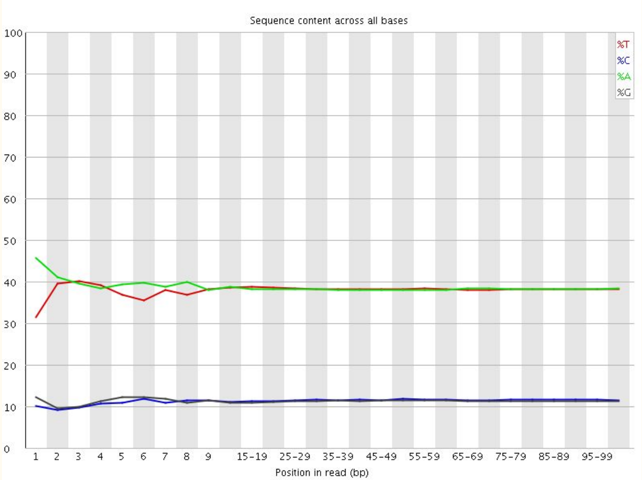


https://www.youtube.com/watch?v=womKfikWlxM

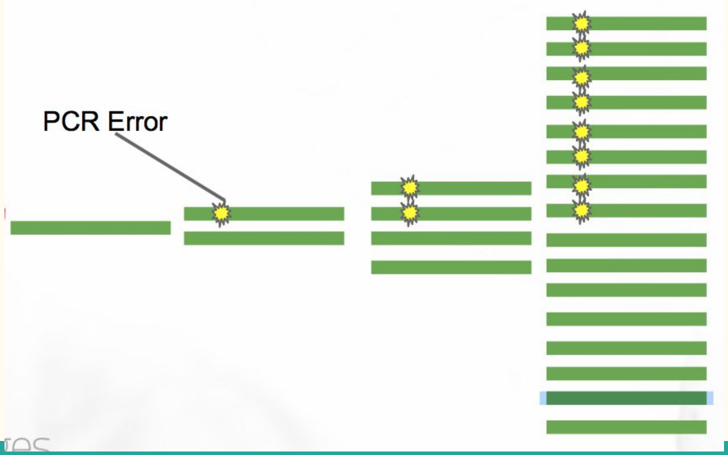
Sequencing errors

1. GC bias





2. Error propagation (1 in 10.000 error rate)



Paired-end sequencing

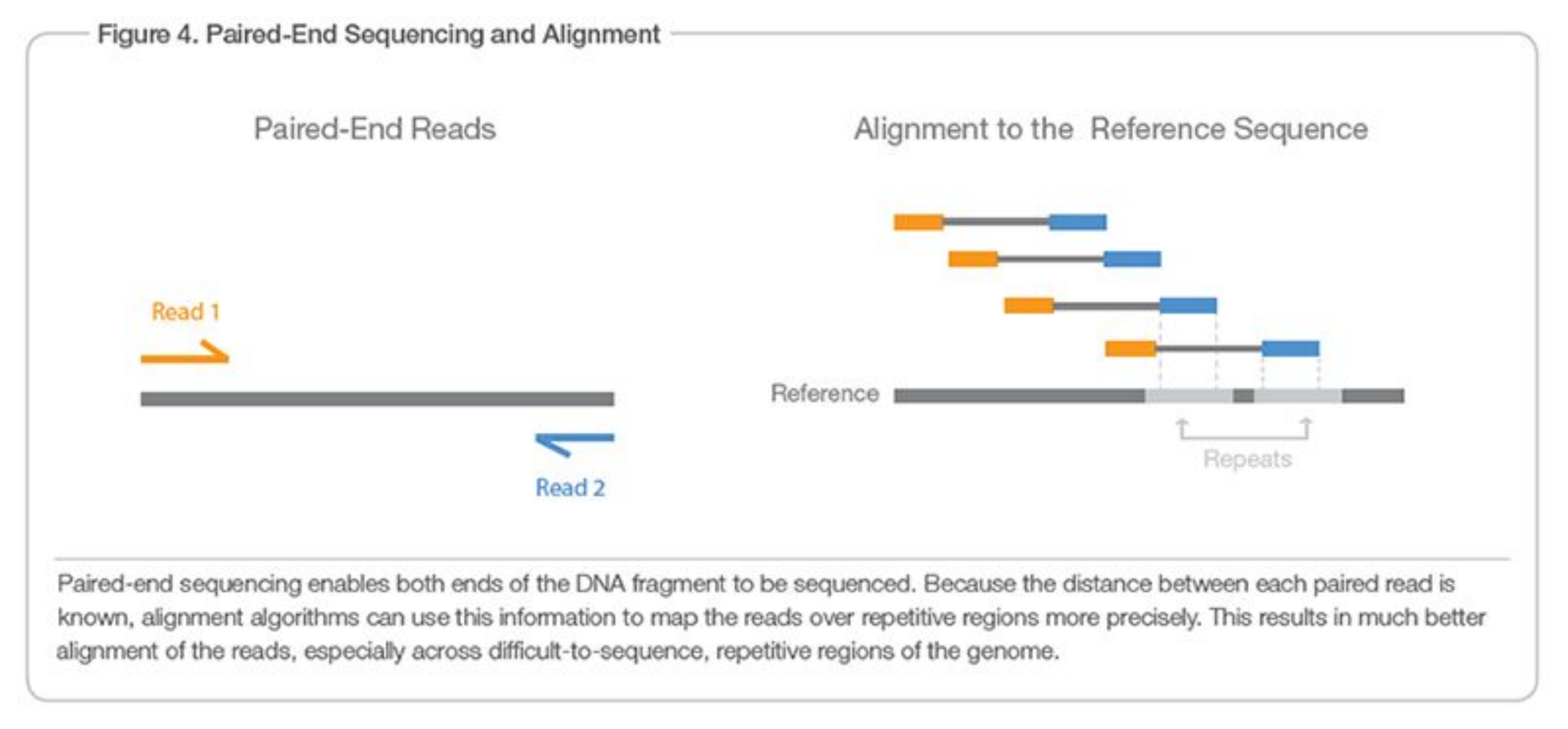


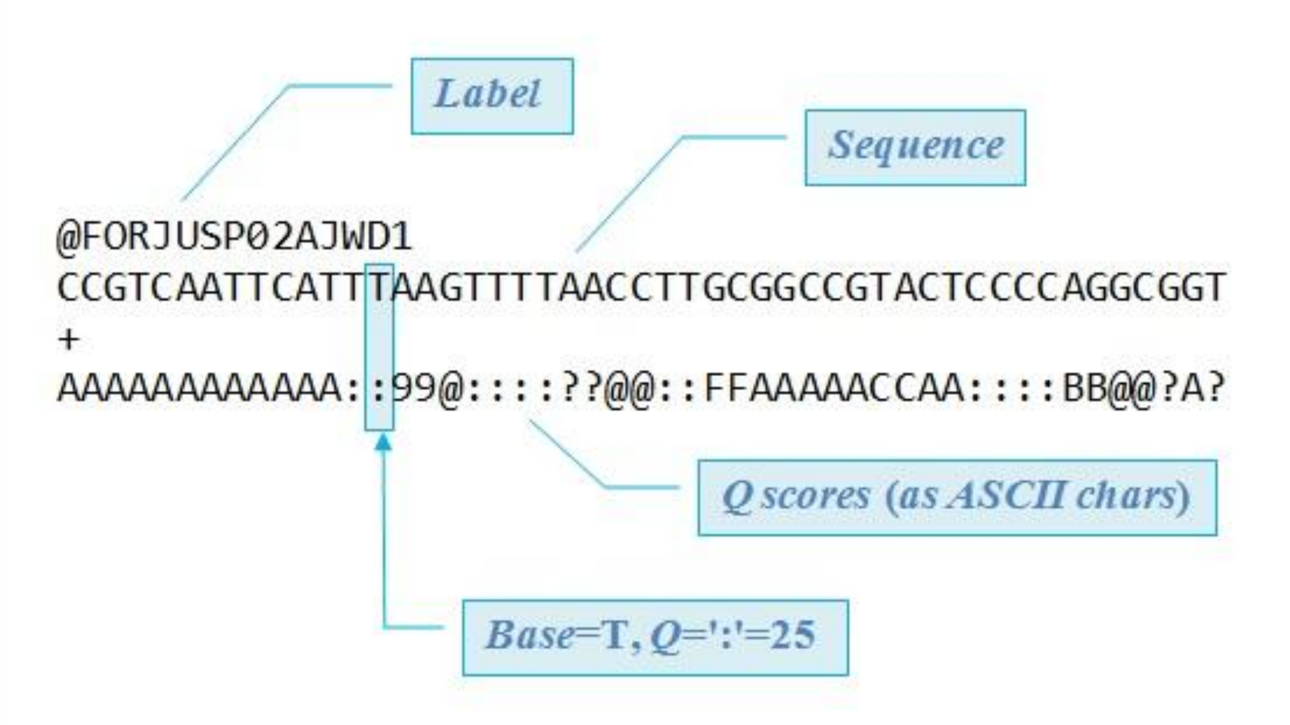
Figure 4. Paired-End Sequencing and Alignment

Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Sequencing data - FASTQ file

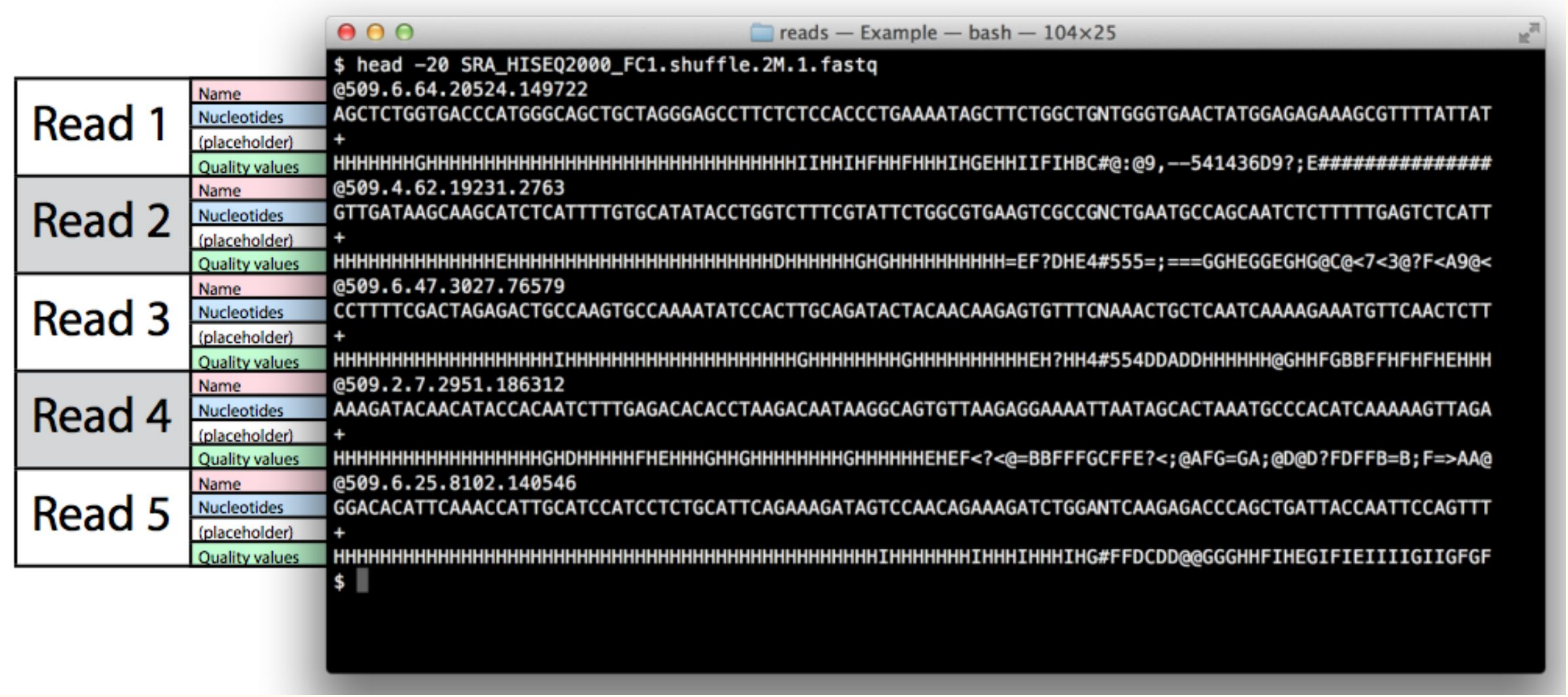
4 lines for each read

- Read id
- Read sequence
- + sign
- ASCII encoded quality



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

Sequencing data - FASTQ file



```
$ head -20 SRA_HISEQ2000_FC1.shuffle.2M.1.fastq
@509.6.64.20524.149722
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGAGCCTTCTCTCCACCCTGAAAATAGCTTCTGGCTGNTGGGTGAACTATGGAGAGAAAGCGTTTTATTAT
+
HHHHHHGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIIHHIHFHHFHHHIHGEHHIIFIHBC#@:@9,--541436D9?;E###############
@509.4.62.19231.2763
GTTGATAAGCAAGCATCTCATTTTGTGCATATACCTGGTCTTTCGTATTCTGGCGTGAAGTCGCCGNCTGAATGCCAGCAATCTCTTTTTGAGTCTCATT
+
HHHHHHHHHHHHHHEHHHHHHHHHHHHHHHHHHHHHHHDHHHHHHGHGHHHHHHHHHH=EF?DHE4#555=;===GGHEGGEGHG@C@<7<3@?F<A9@<
@509.6.47.3027.76579
CCTTTTCGACTAGAGACTGCCAAGTGCCAAAATATCCACTTGCAGATACTACAACAAGAGTGTTTCNAAACTGCTCAATCAAAAGAAATGTTCAACTCTT
+
HHHHHHHHHHHHHHHHHHHHIHHHHHHHHHHHHHHHHHHHGHHHHHHHHGHHHHHHHHHHEH?HH4#554DDADDHHHHHH@GHHFGBBFFHFHFHEHHH
@509.2.7.2951.186312
AAAGATACAACATACCACAATCTTTGAGACACACCTAAGACAATAAGGCAGTGTTAAGAGGAAAATTAATAGCACTAAATGCCCACATCAAAAAGTTAGA
+
HHHHHHHHHHHHHHHHHHGHDHHHHHFHEHHHGHHGHHHHHHHHGHHHHHHEHEF<?<@=BBFFFGCFFE?<;@AFG=GA;@D@D?FDFFB=B;F=>AA@
@509.6.25.8102.140546
GGACACATTCAAACCATTGCATCCATCCTCTGCATTCAGAAAGATAGTCCAACAGAAAGATCTGGANTCAAGAGACCCAGCTGATTACCAATTCCAGTTT
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIHHHHHHHIHHHIHHHIHG#FFDCDD@@GGGHHFIHEGIFIEIIIIGIIGFGF
$
```

Genome reconstruction

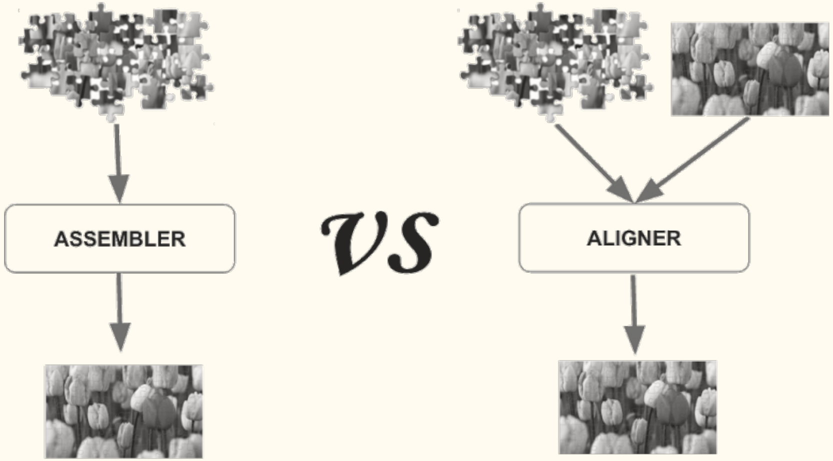
Result of sequencing experiment

- FASTQ file
- 100-500 GB
- Each read(line) containing a genome sequence 50-250 bp long

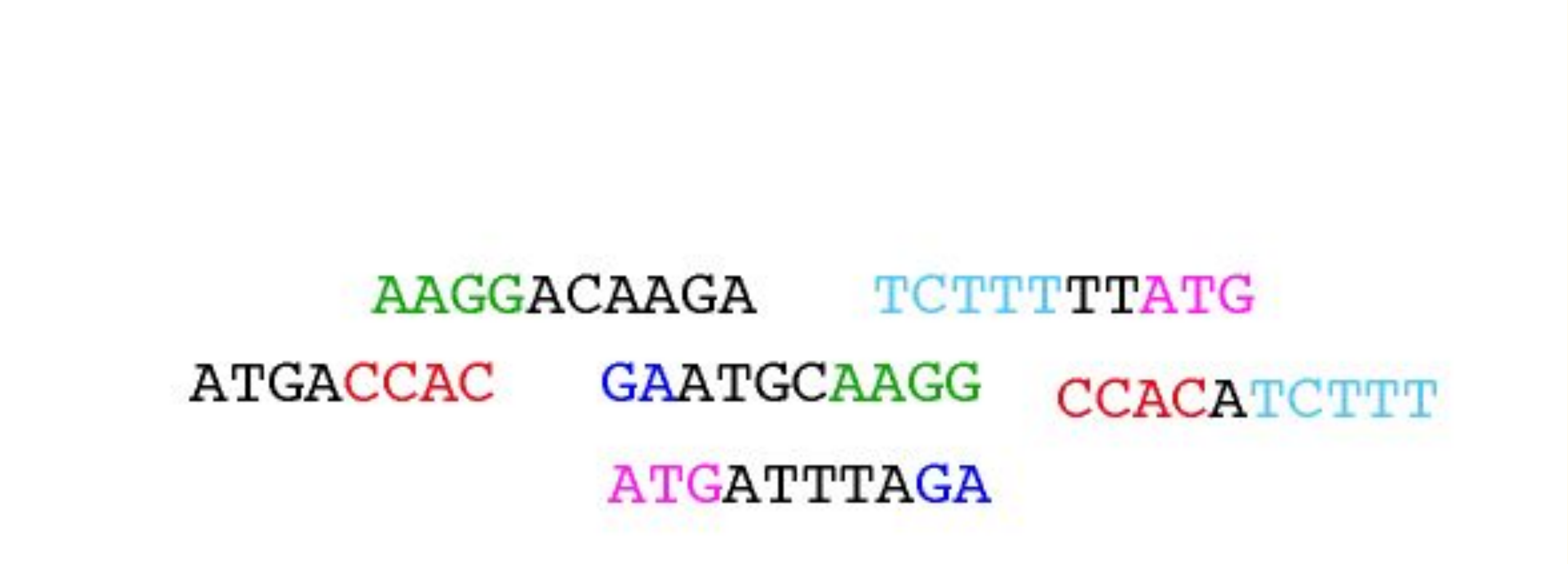
Genome reconstruction

How do we reconstruct genome from reads?

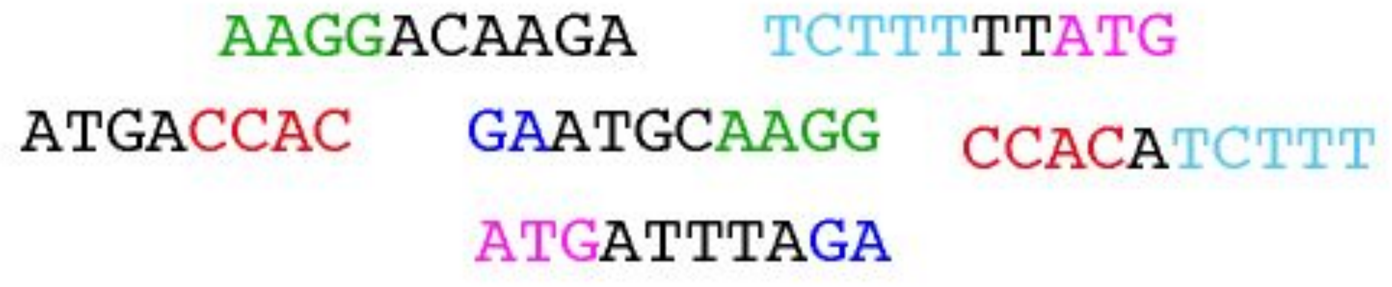
1. Alignment
   - ○ Using reference genome to map the position of the reads
2. Assembly
   - ○ Reconstructing the genome by finding the links between the reads



Alignment


AAGGACAAGA
TCTTTTTATG
ATGACCAC
GAATGCAAGG
CCACATCTTT
ATGATTTAGA

Assembly


AAGGACAAGA
TCTTTTTATG
ATGACCAC
GAATGCAAGG
CCACATCTTT
ATGATTTAGA