Genome Informatics 2019

Lesson 1 - An introduction

Lesson overview

- Course info
- Bioinformatics and genomics definitions
- Molecular biology basics
- Genome sequencing technologies
- Introduction to Python

**Why slides in English?*

Course info

- 13 classes (lecture + exercise) - ~2.5h
- Exam will have both theoretical and practical part
  - 40% on the exam
  - 60% during the semester
    - 40% project assignment (with presentation)
    - 20% essay on the certain topic
- Last class - presentation of student assignments
- Exercise will follow lectures - examples in python Jupyter notebook

Communication

github.com/vladimirkovacevic/gi-2019-etf

(All info about the course. Create an issue!)

vladimir.kovacevic@sbgenomics.com

General info (not relevant to all)

Questions about lessons 1-5

vladimir.tomic@sbgenomics.com

Questions about lessons 6-9

marko.zecevic@sbgenomics.com

Questions about lessons 10-12

Course info - syllabus

1	Course info. Bioinformatics and genomics definitions. Molecular biology basics. Genome sequencing technologies. Exercise: Introduction to python and Jupiter environment.
2	Exact string matching: Boyer-Moore, indexing structures, hash tables
3	Exact string matching: Suffix trie, suffix tree, Pigeonhole principle
4	Burrows-Wheeler Transform and FM Index.
5	Bioinformatic workflows, Variant calling, Cancer analysis.
6	Approximate string matching, Edit distance, Dinamičko programiranje, Global alignment
7	Variation on global alignment (end-space-free variant, longest common substring) , local alignment, gaps. Practical: BLAST, Bowtie
8	Shortest common superstring, Overlap graph
9	De-Bruijn graph, scaffolding, error correction
10	Biostatistics and RNA: The central dogma of molecular biology, RNA-Seq motivation and technologies for gene expression measurement; RNA-Seq alignment (high level).
11	Quantification: on gene and transcript levels, between and within-sample normalization procedures; Exploratory analysis.
12	Differential expression: intro to statistical inference, multiple testing corrections; Biological pathways.
13	Presentation of the student projects.

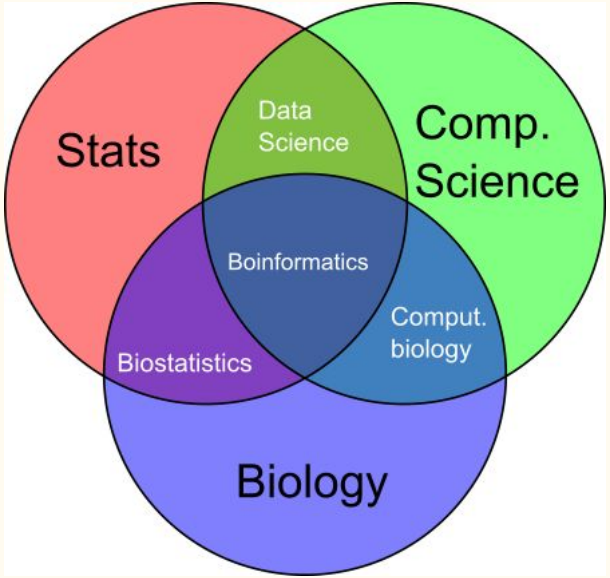
Literature

- 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

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.

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 - 6 billions of base-pairs (A, C, T, G) letters
  - Can be imagined as a sting 6 billion letters long

Genomics: contrast with biology & genetics*

* Everything on this slide is a gross generalization

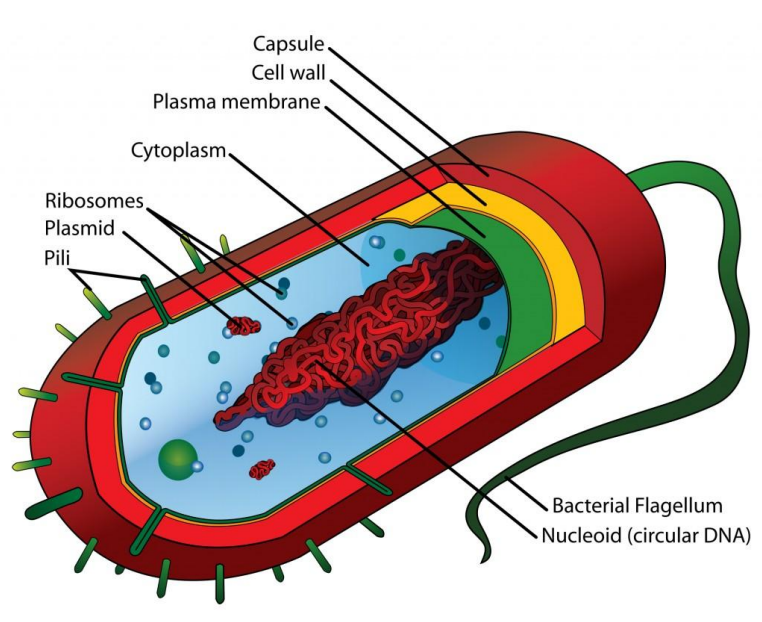
Biology & Genetics		Genomics
Targeted studies of one or a few genes	scope	Studies considering all genes in a genome
Targeted, low-throughput experiments	technology	Global, high-throughput experiments
Clever experimental design, painstaking experimentation	hard part	Tons of data, uncertainty, computation

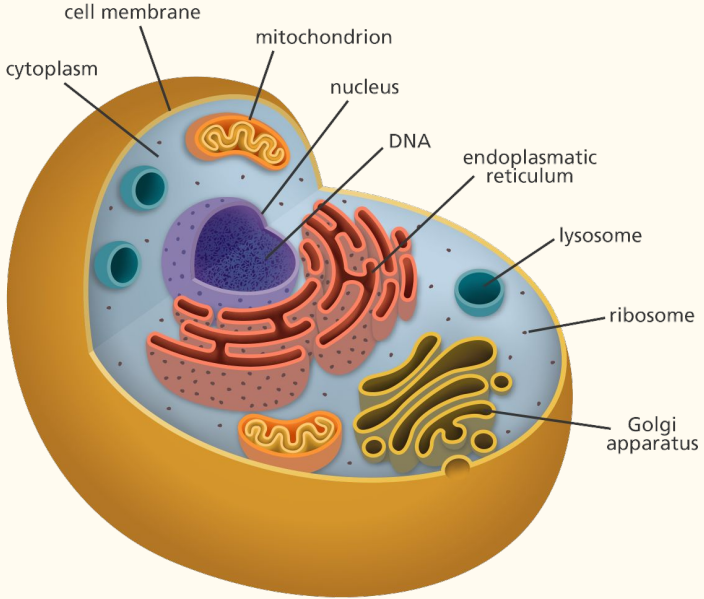
Used with permission of Ben Langmead

The cell

Fundamental working units of every living system.

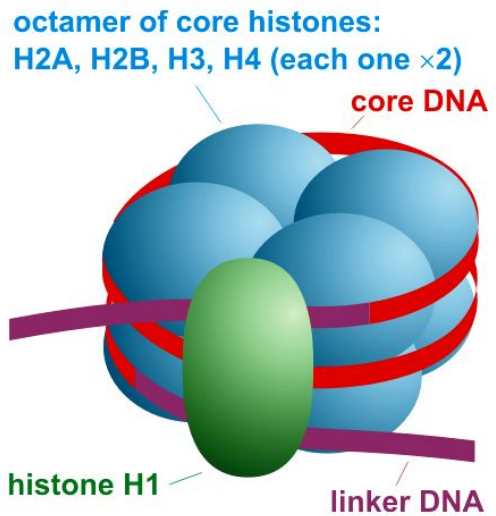
- Prokaryotic (bacteria)
- Eukaryotic (higher organisms - animals, plants)

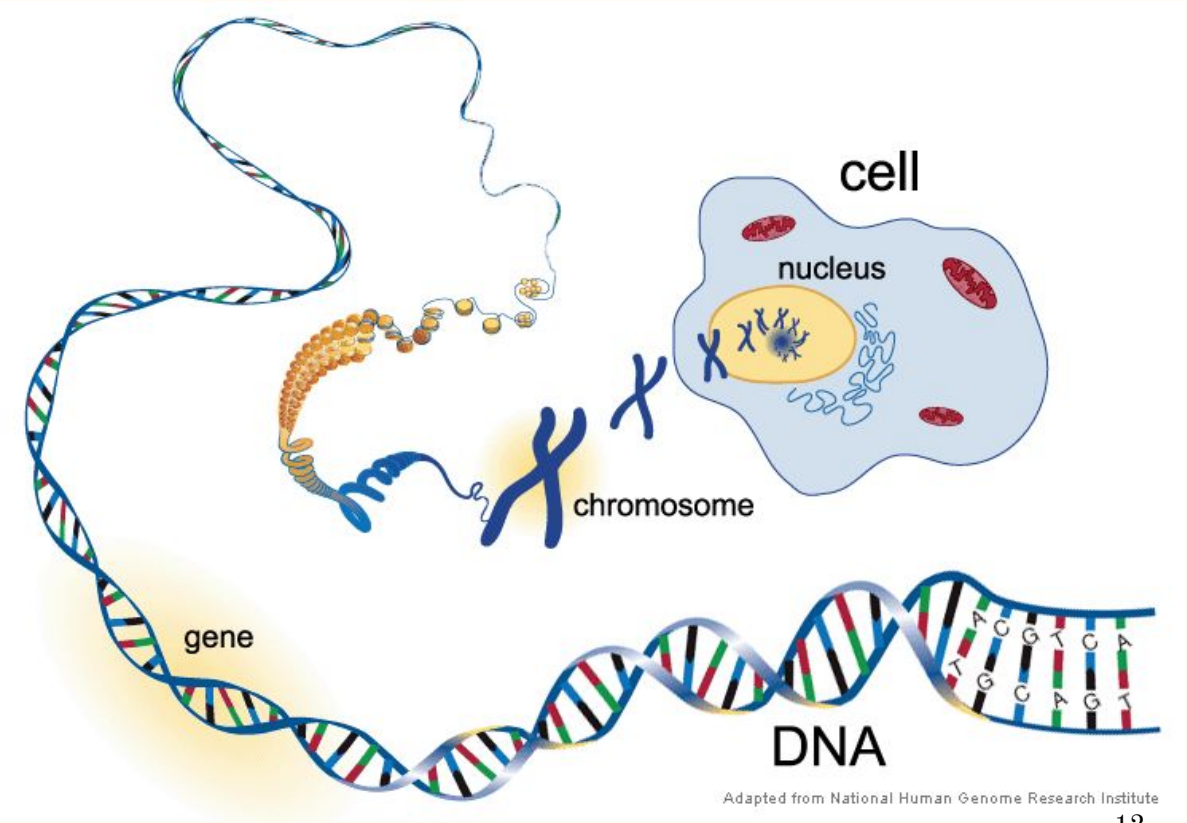




“And inside the nucleus thou lays the mighty DNA”

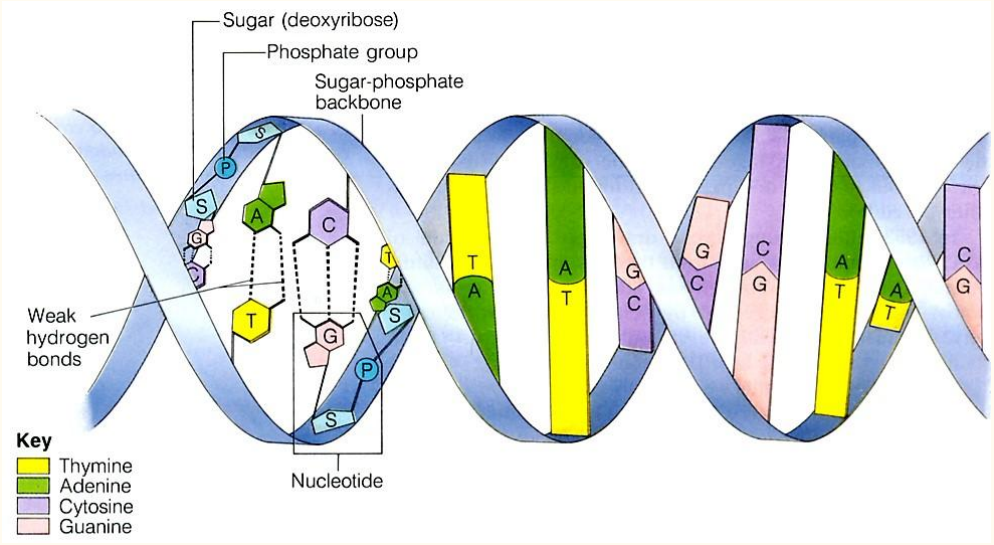
- Chromatin - tightly packed DNA
- A **nucleosome** is a basic unit of DNA packaging in eukaryotes, consisting of a segment of DNA wound in sequence around eight histone protein cores.
- Current model





DNA - the code of life

- DNA (deoxyribonucleic acid) - double stranded molecule
- Same in every cell - DNA replication during cell division
- More stable, redundant information - complementary double helix chain
- Base (nucleotide) pairs (complementary bases)
  - A - T (adenine and thymine)
  - C - G (cytosine and guanine)



Genome

- Set of all pairs of chromosomes

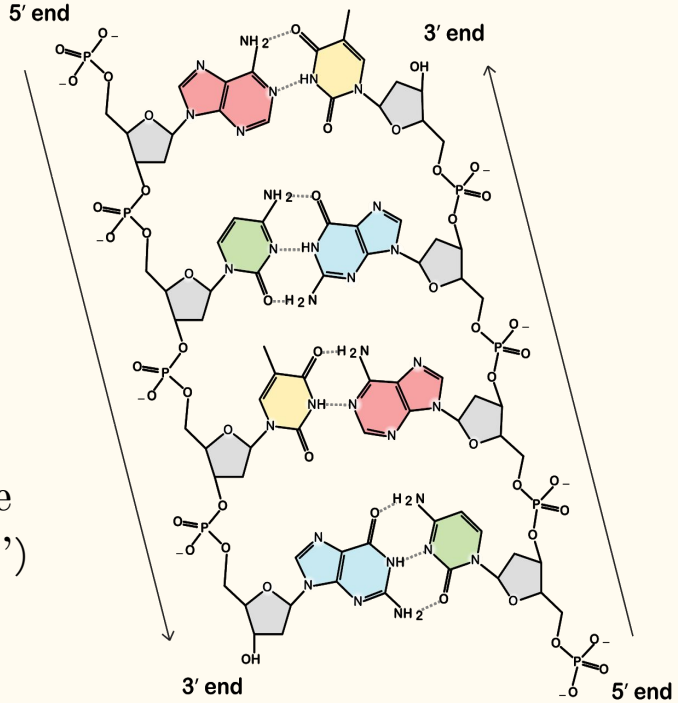
- Human genome:
  - 23 pair of chromosomes (diploid)
  - 22 autosomes
  - 1 sex chromosome (X and/or Y)
  - 3 billion base-pairs x 2
  - Intron and exom (2%)

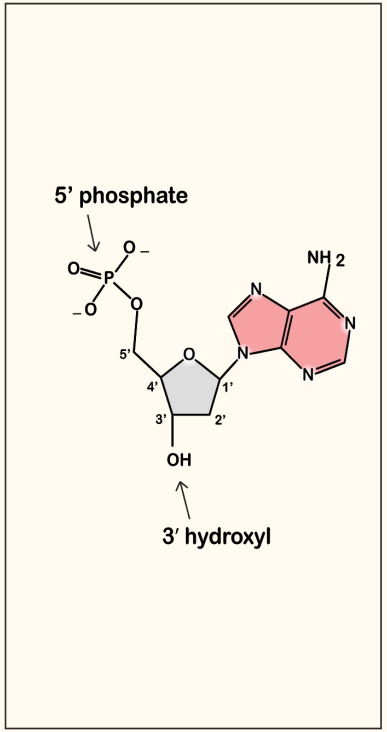
Karyogram

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)





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"

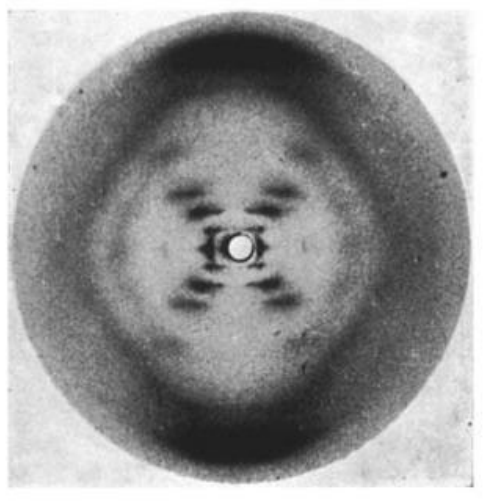


Photo 51

Central dogma of molecular biology

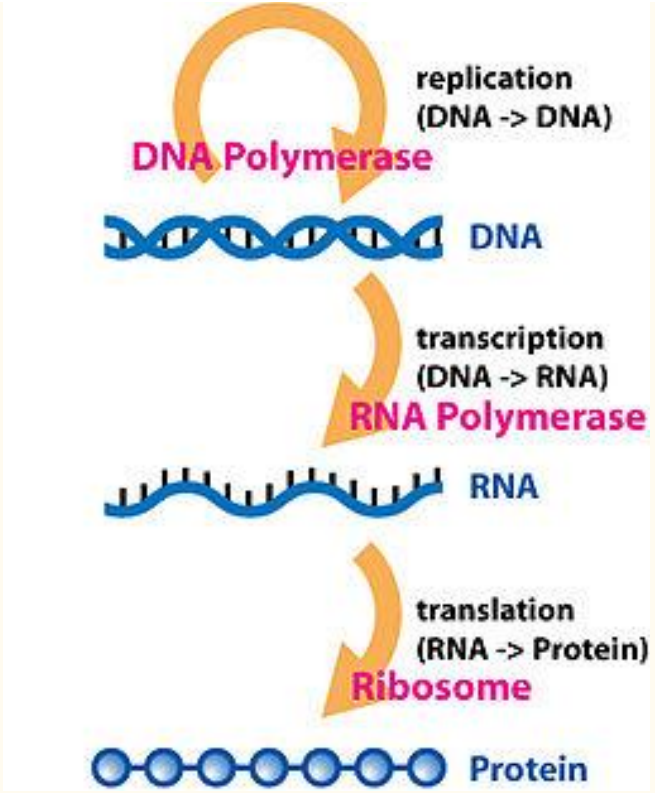
DNA ----> RNA -----> Protein

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.



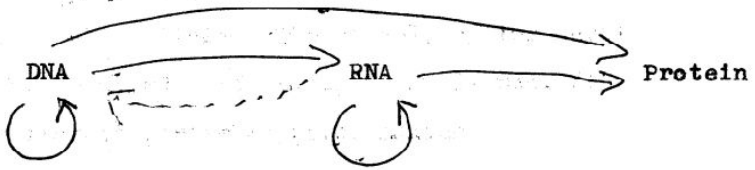
Central dogma of molecular biology

Ideas on Protein Synthesis (Oct. 1956)

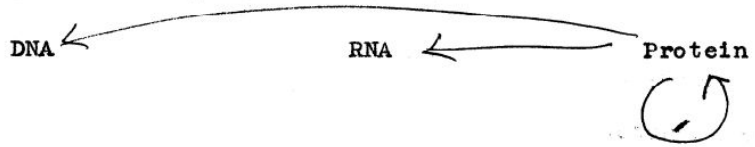
The Doctrine of the Triad.

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



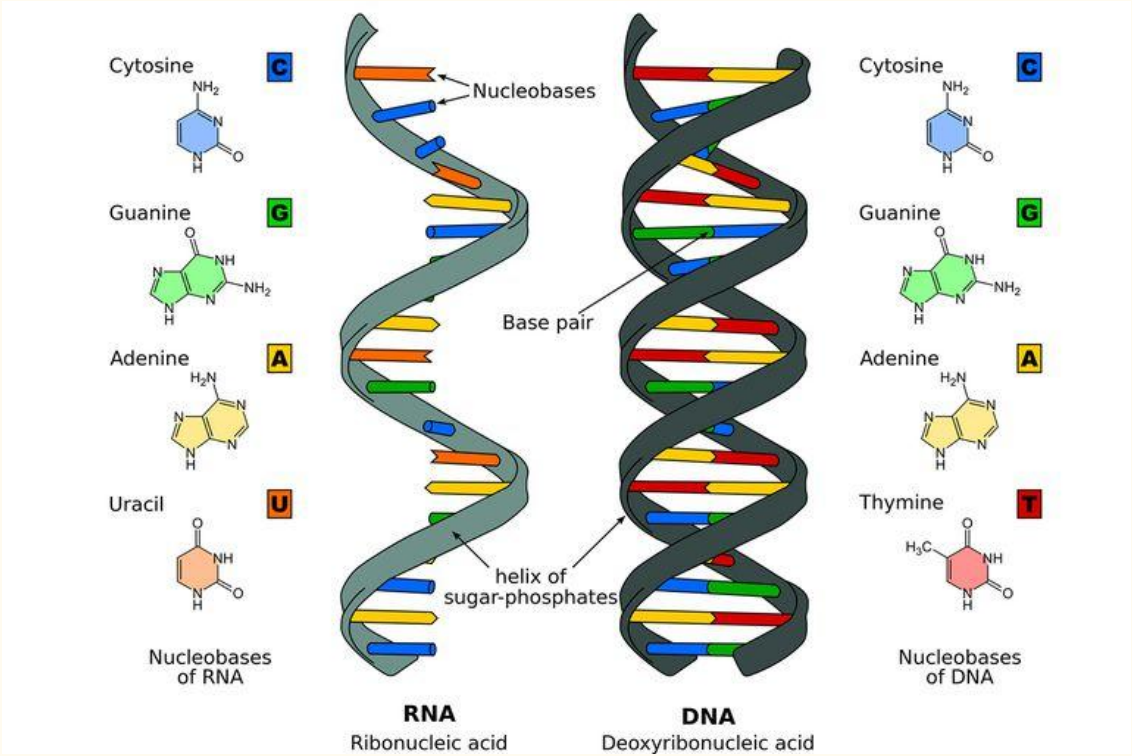
but never



where the arrows show the transfer of information.

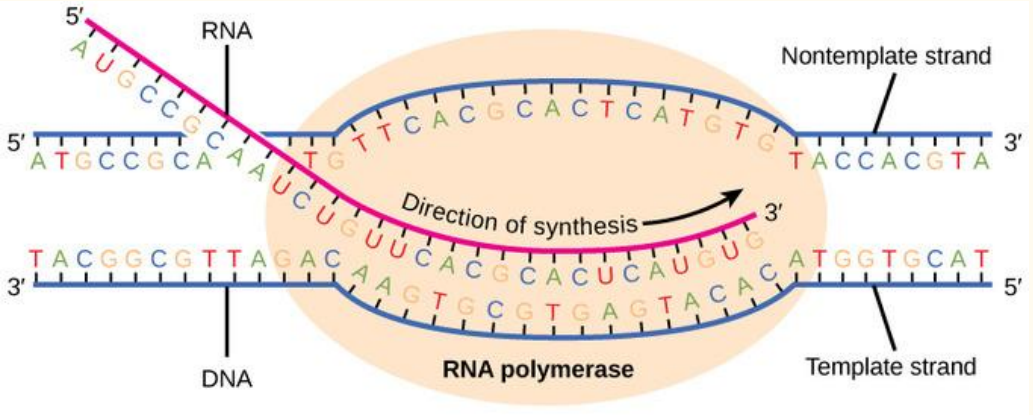
RNA

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

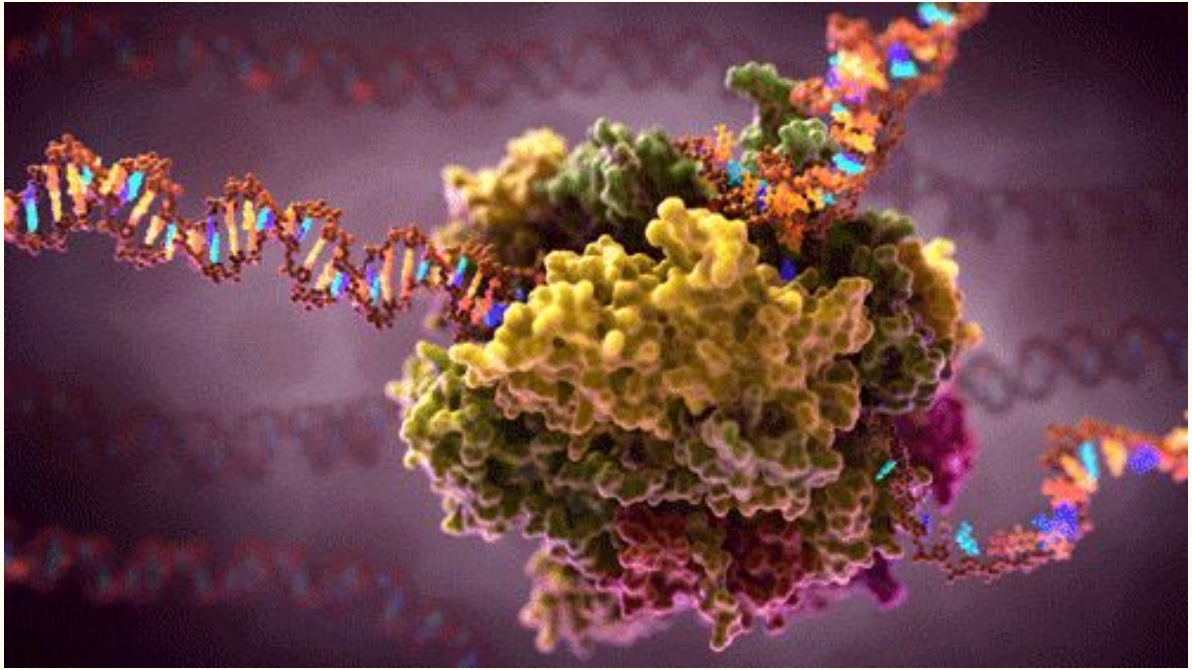


Transcription

- Template (noncoding) strand
  - One which is transcribed by RNAP (RNA polymerase)

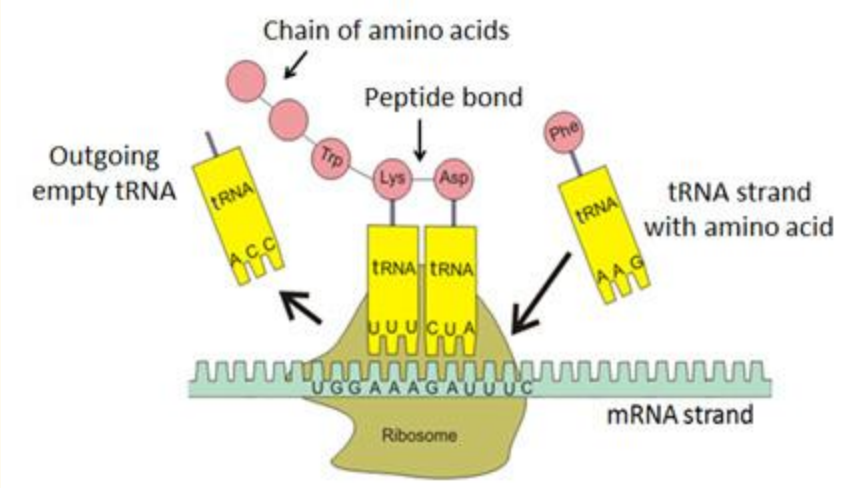


Transcription



Translation

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

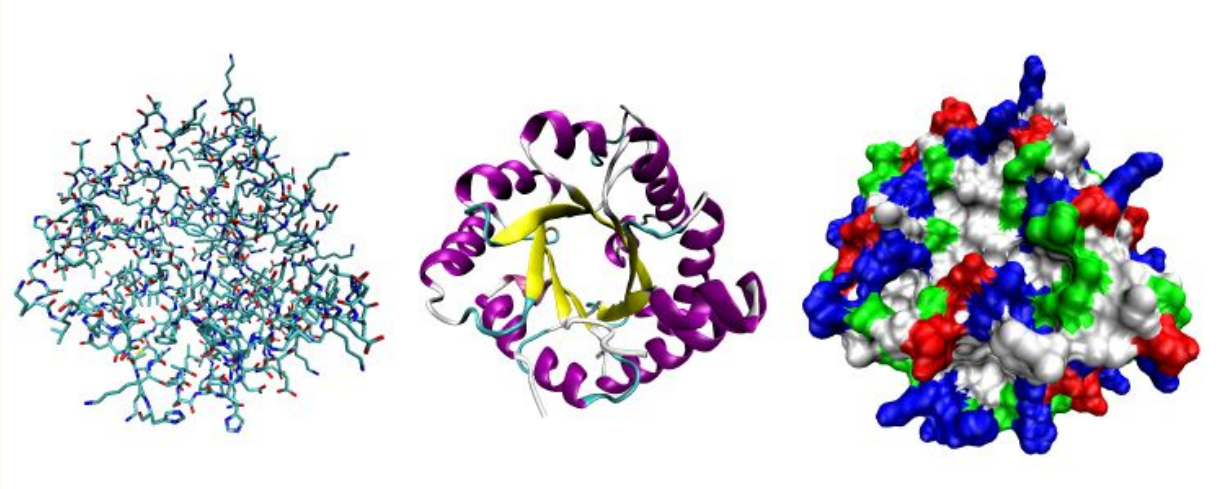


Second letter

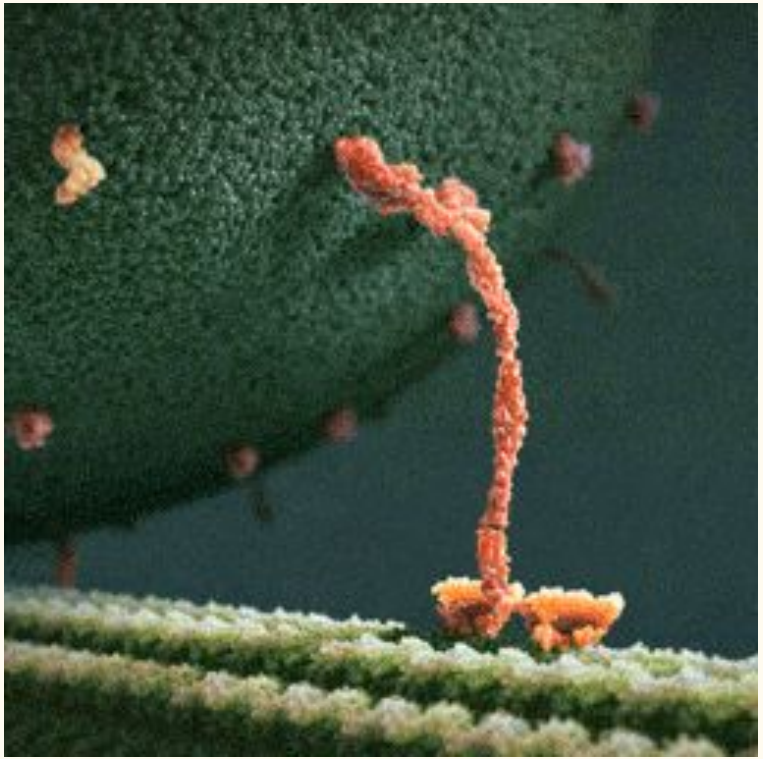
First letter	U	C	A	G	Third letter
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr **UAA Stop** **UAG Stop**	UGU Cys UGC Cys **UGA Stop** UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg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

Proteins

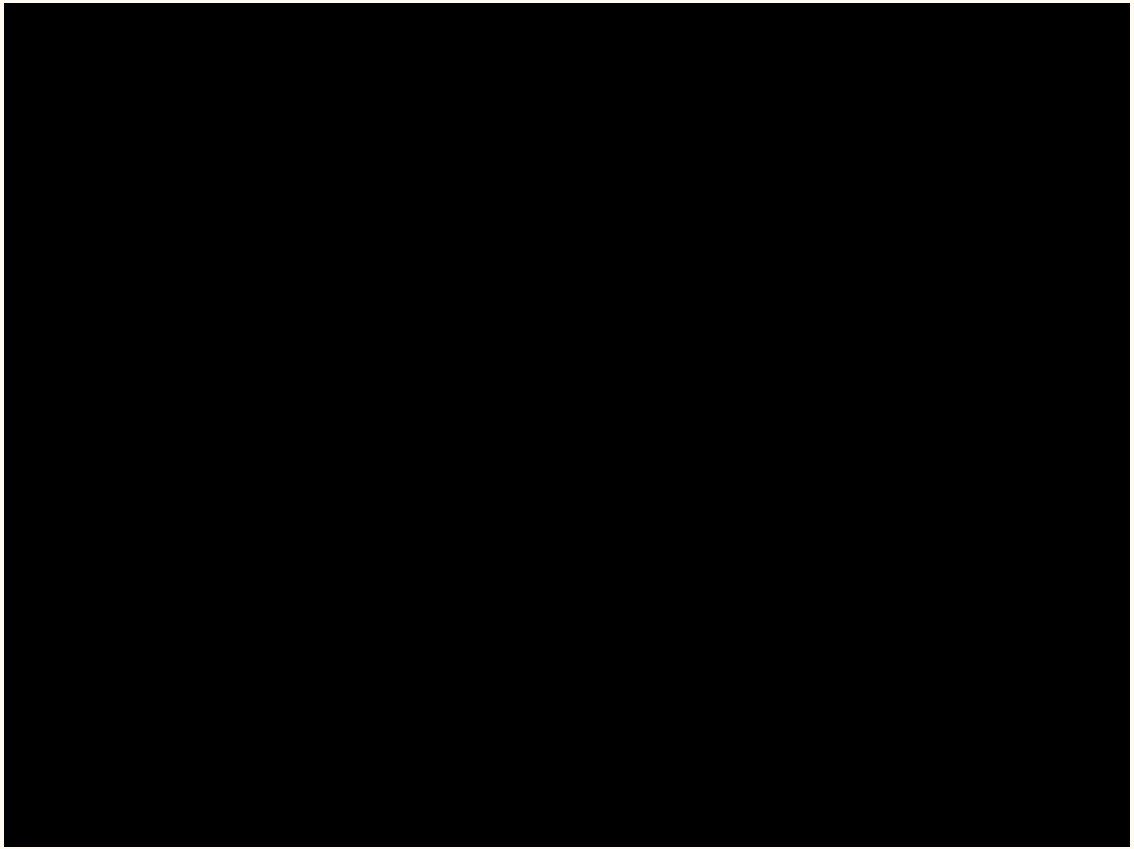
- Building blocks of life
  - Various functions in the organism (transportation, regulation, metabolism, DNA replication)
- Long chains of amino-acids, that also fold into complicated 3D structures
  - We often distinguish protein primary, secondary, tertiary and quaternary structure



Proteins



Proteins



Transcription & Translation: The Central Dogma of Biology - DNA Learning Center

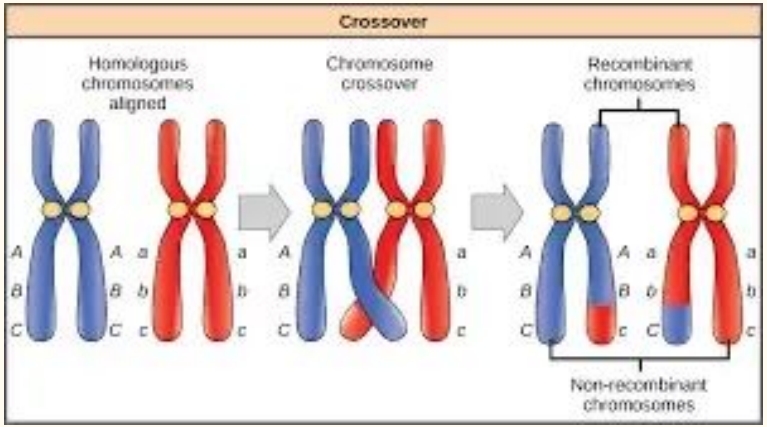
Biology 101 - genotype vs phenotype

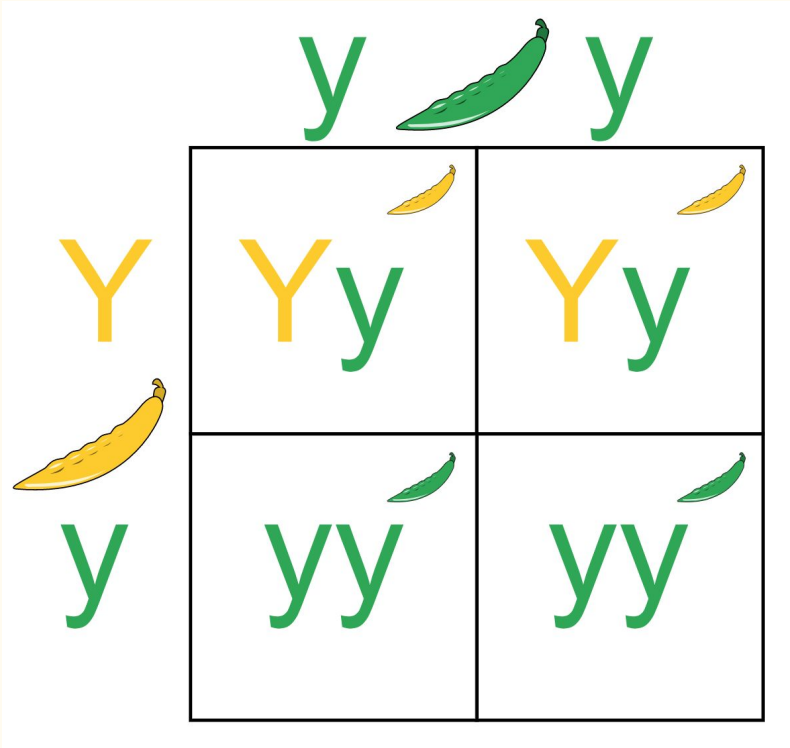
The **genotype** is the part of the genetic makeup of a cell, and therefore of an organism or individual, which determines one of its characteristics (phenotype).

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

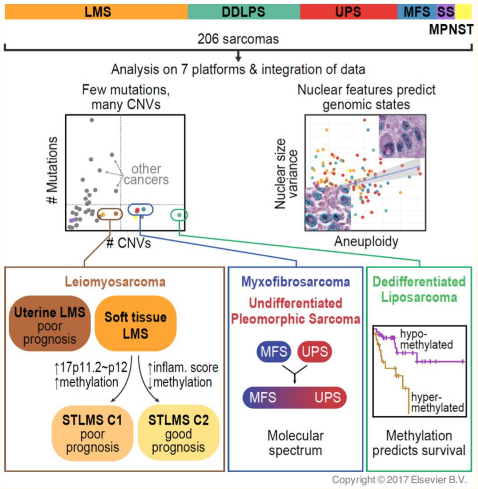


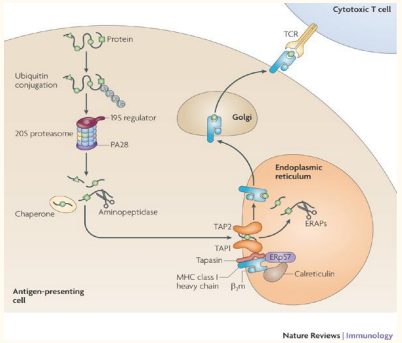


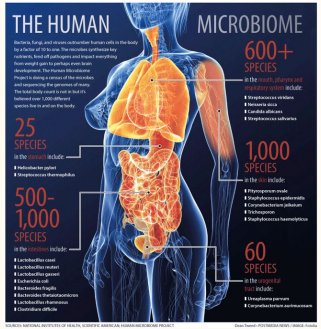
Example of inheritance combinations

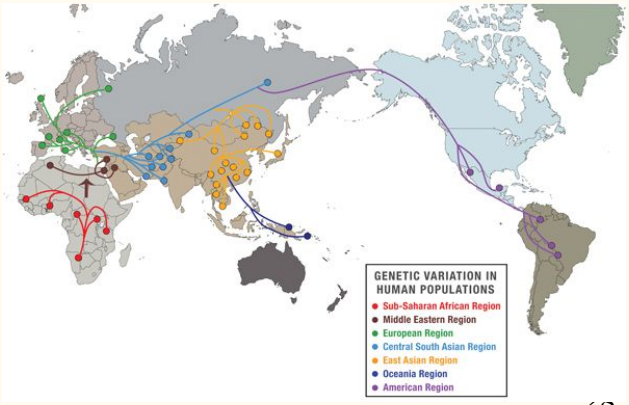
Why perform DNA sequencing?

- Rare genetic diseases
- Origins of humans
- Precision medicine- Cancer treatment (immunotherapy)
- Microbes that live inside us (microbiome)
- Study ways that genomes work









MEDICAL DISPATCH JULY 21, 2014 ISSUE

ONE OF A KIND

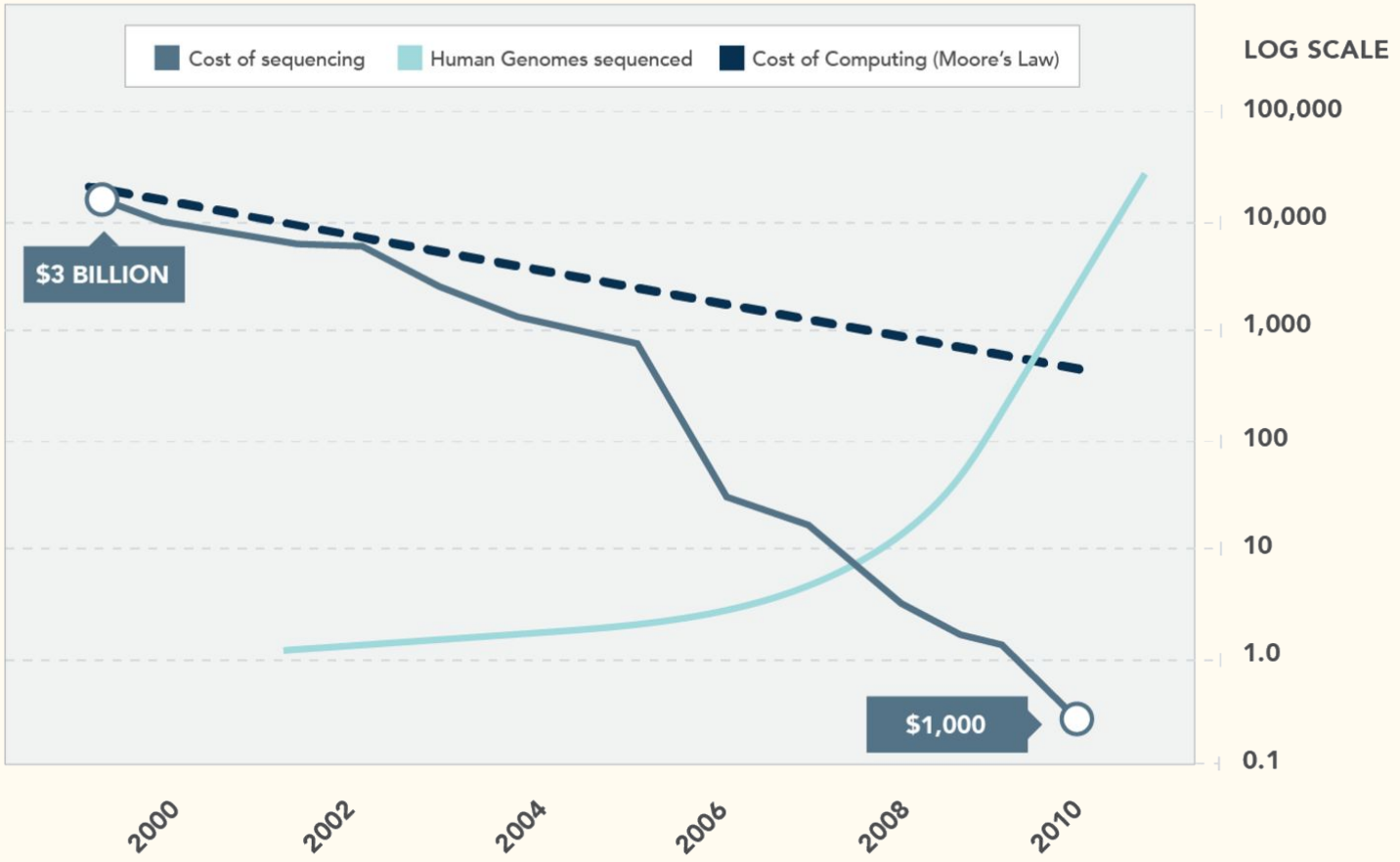
What do you do if your child has a condition that is new to science?

By Seth Mnookin

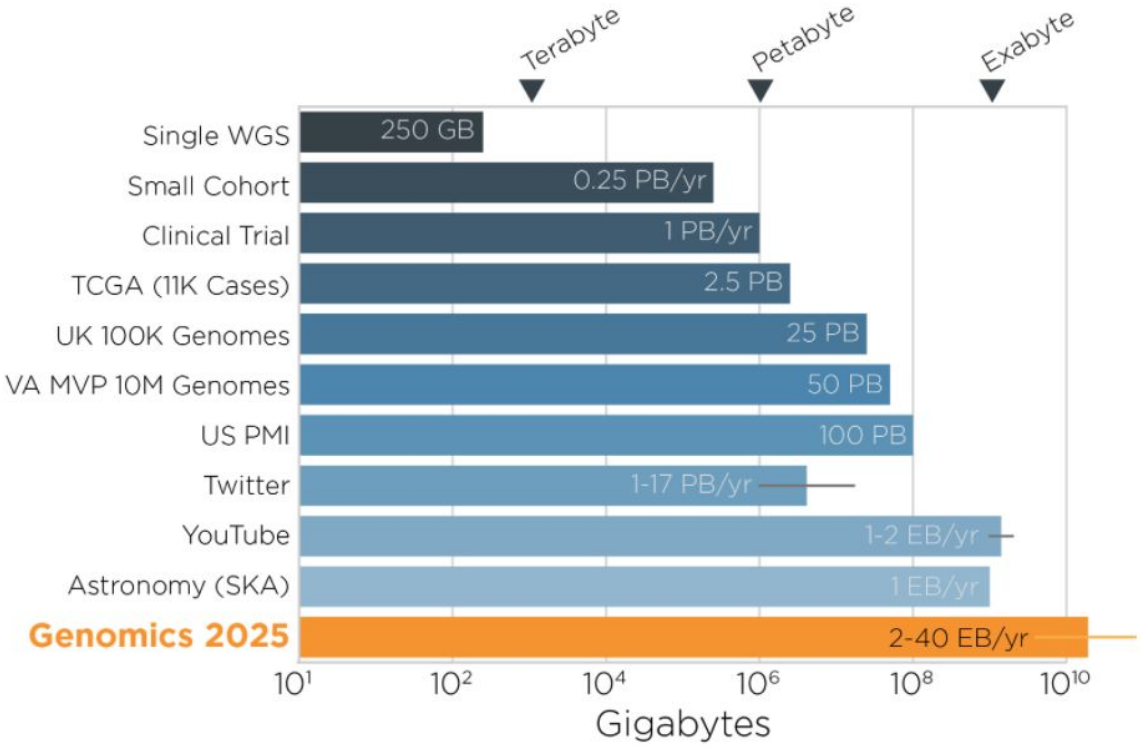
Genome sequencing

- Digitalization of genome
- **Human Genome Project** (1990-2003), 3B $
- Birth of bioinformatics
- Sanger sequencing (First generation sequencing)
  - Long (took 13 years)
  - Costly (3B$ for one human genome)
- Currently NGS (next generation sequencing)
  - Illumina
  - Around 200$ 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
  - Smaller in size
  - Sequencing in space

GROWTH OF DNA SEQUENCING



Genomics is Big Data



Source: "Big Data: Astronomical or Genomical?" *PLoS Biology* (2015).

Sequences:
1 zetta-bases/yr

Storage needs:
2-40 exabytes

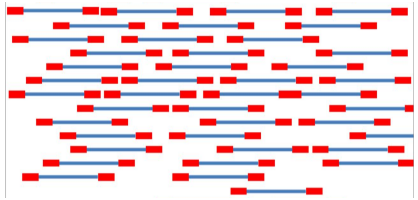
Compute for Alignment:
10,000 trillion CPU hrs
= 83x time since Big Bang

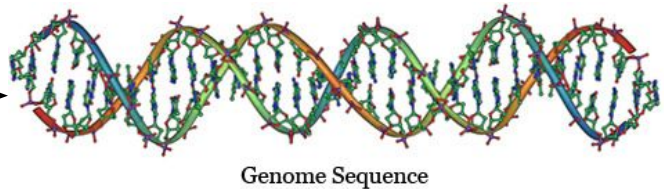
Variant Calling:
~**2 trillion** CPU hrs

Tertiary Analysis:
~**4 trillion** CPU hrs
= time since land-breathing mammals evolved

Bioinformatics to the rescue!

- Genomes of the all species are arrays of nucleotides (A, T, C, G) - strings
- The process of DNA sequencing returns only fragments of it
- Our mission: RECONSTRUCT IT!

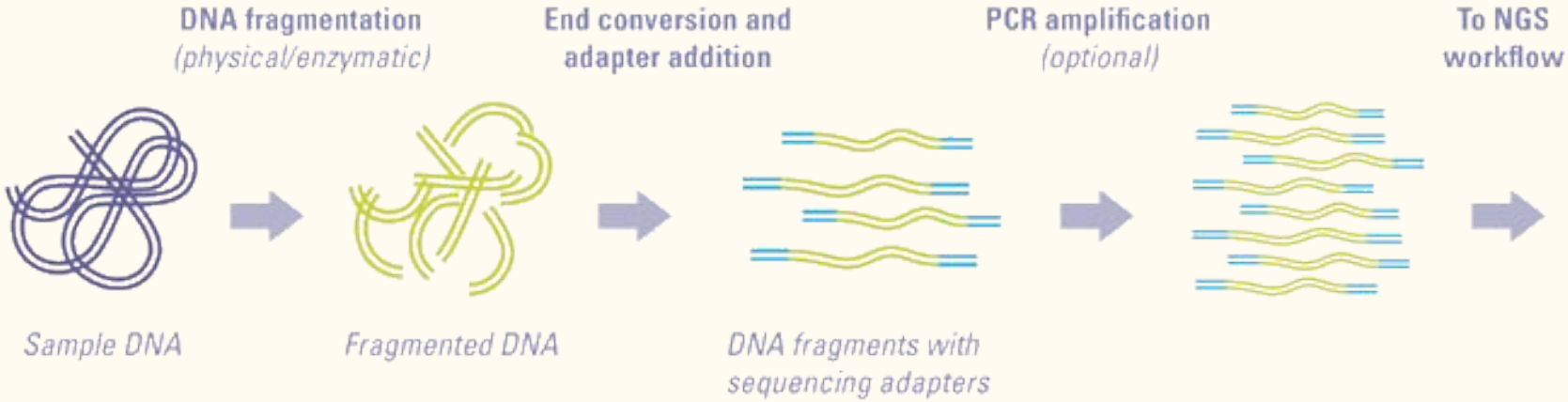




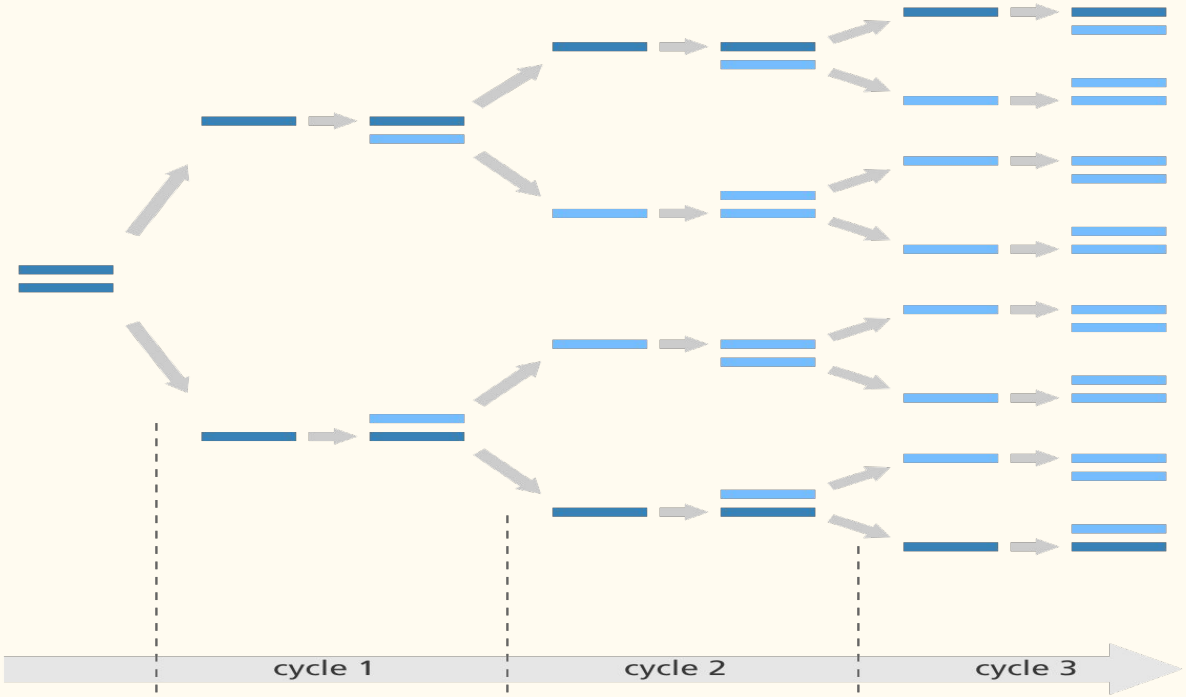
Genome Sequence

Illumina sequencing

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

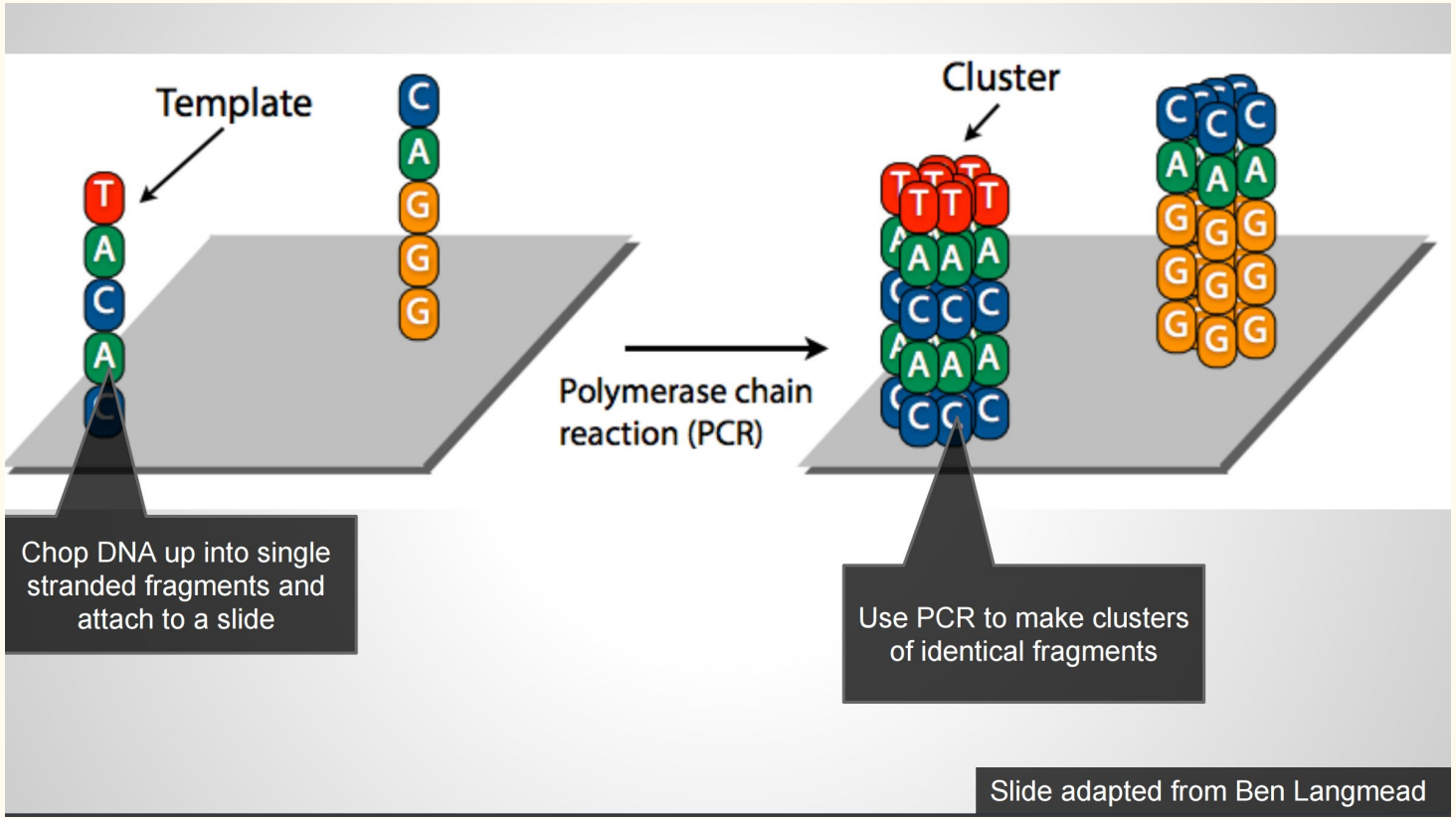


Sequencing - PCR (polymerase chain reaction)

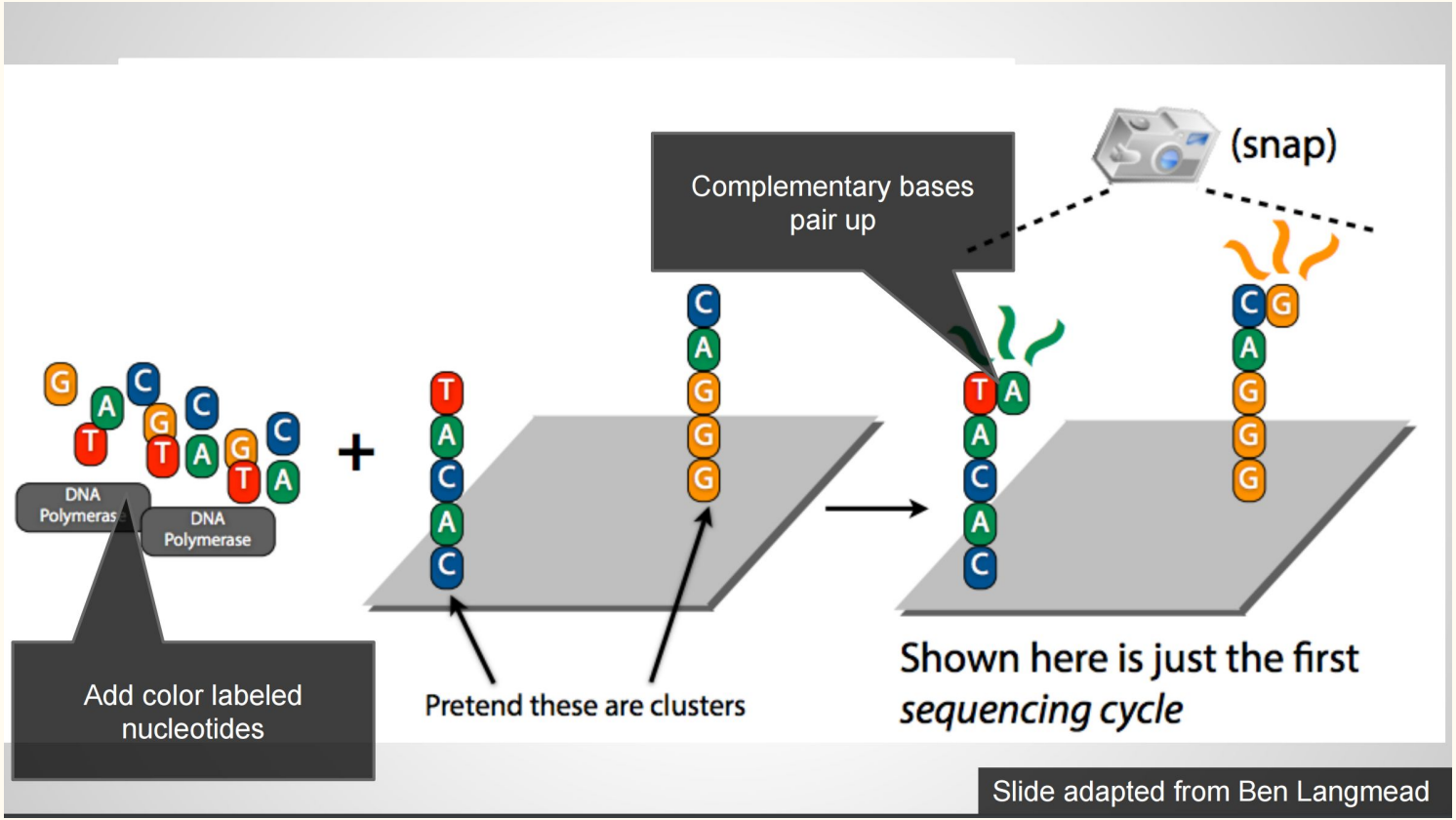


Bridge amplification

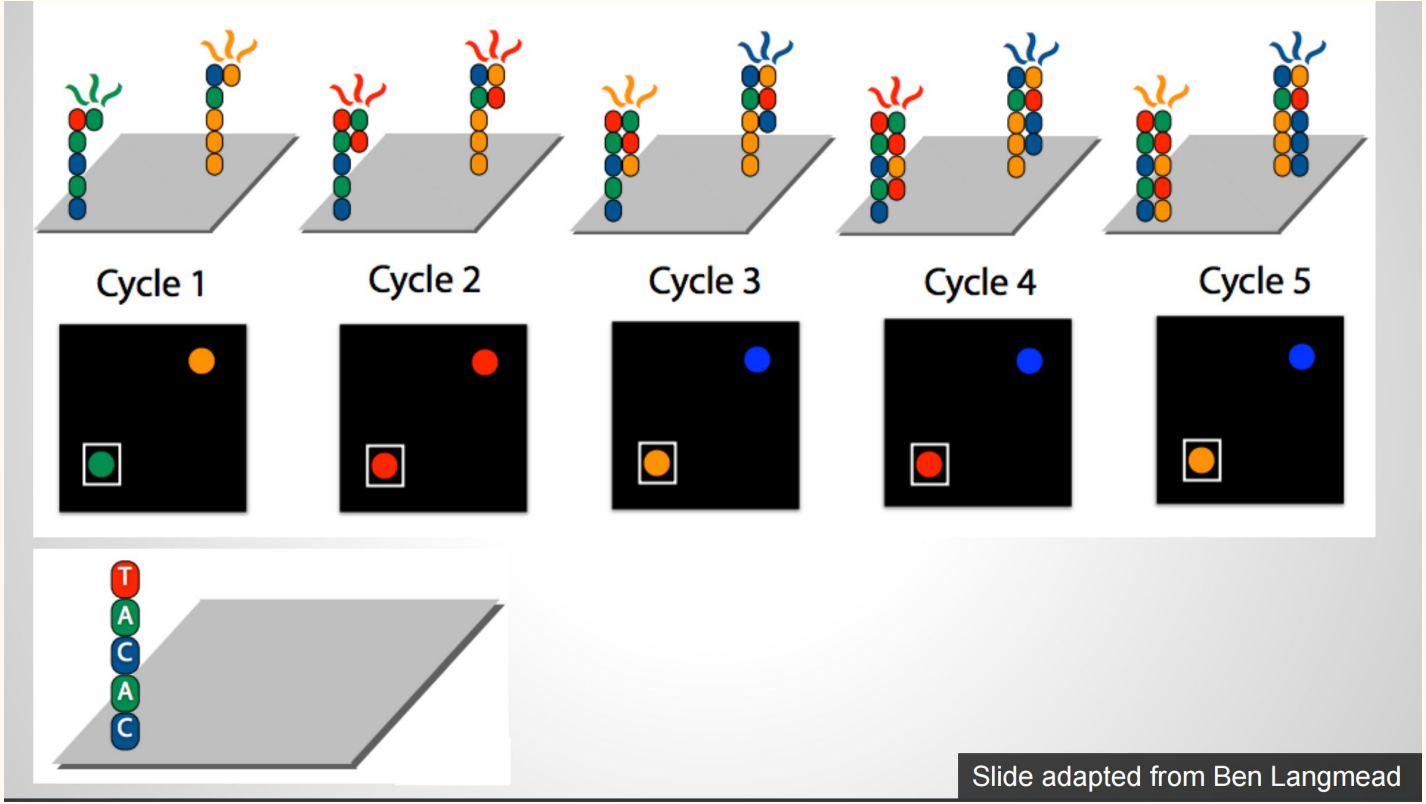
Sequencing (Illumina)



Sequencing (Illumina)

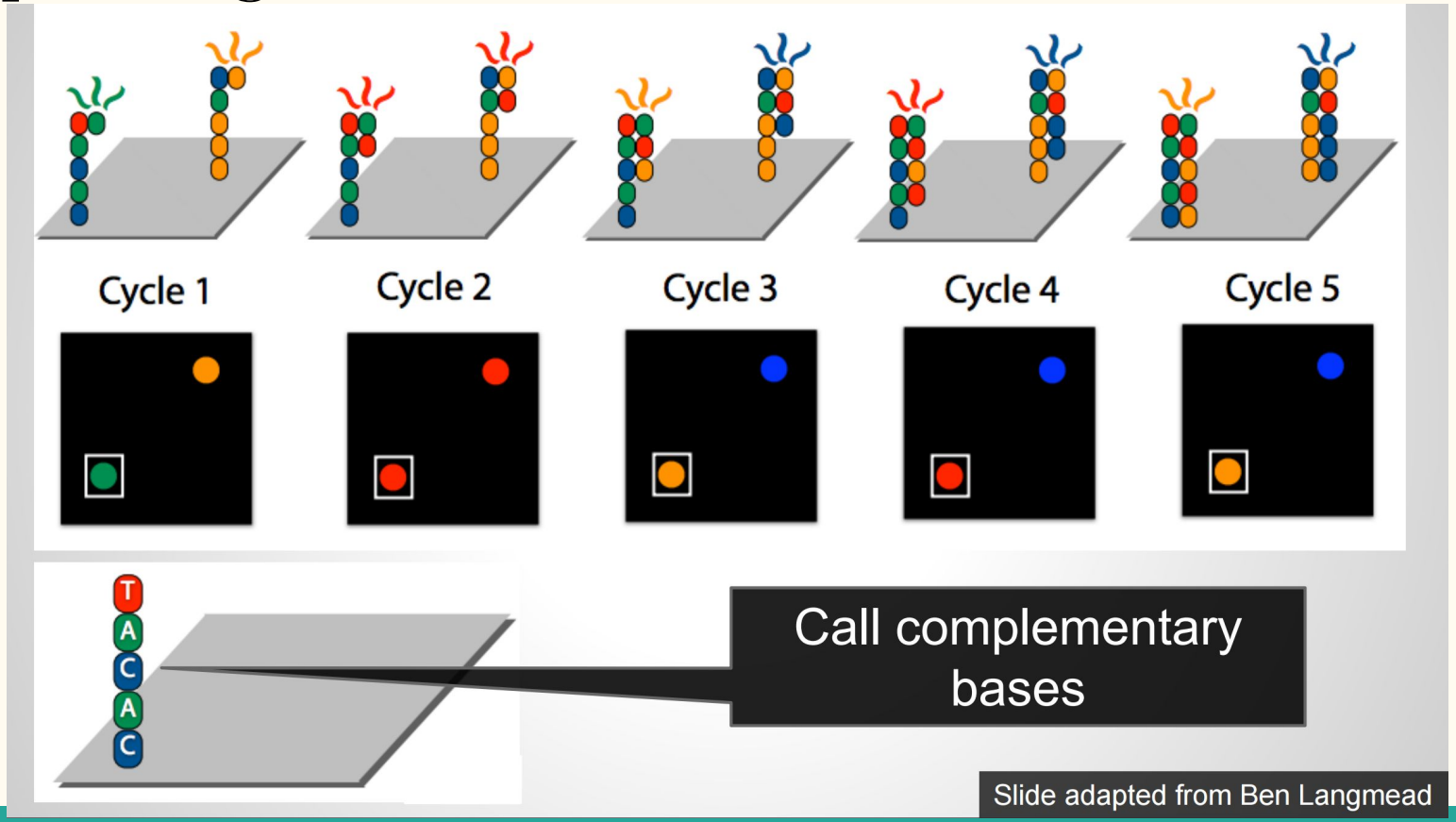


Sequencing (Illumina)



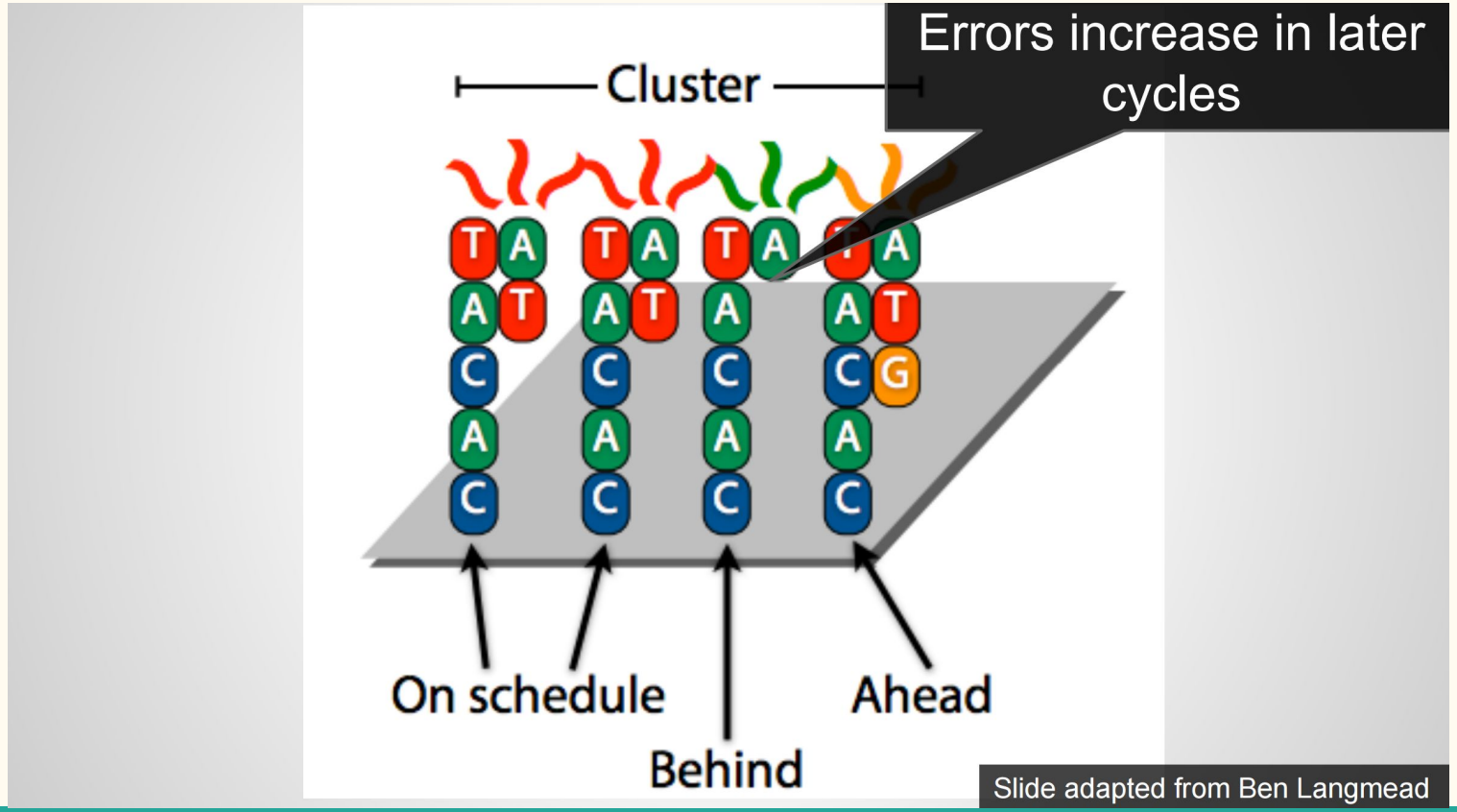
Slide adapted from Ben Langmead

Sequencing (Illumina)



Slide adapted from Ben Langmead

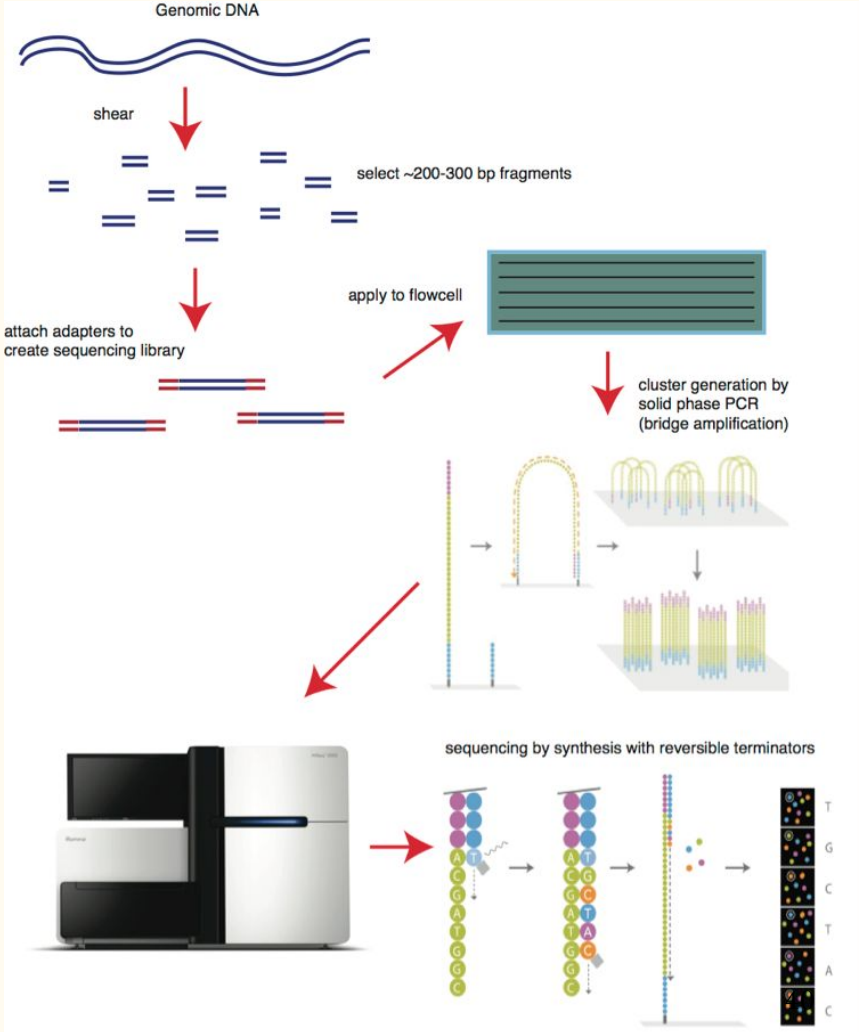
Sequencing error



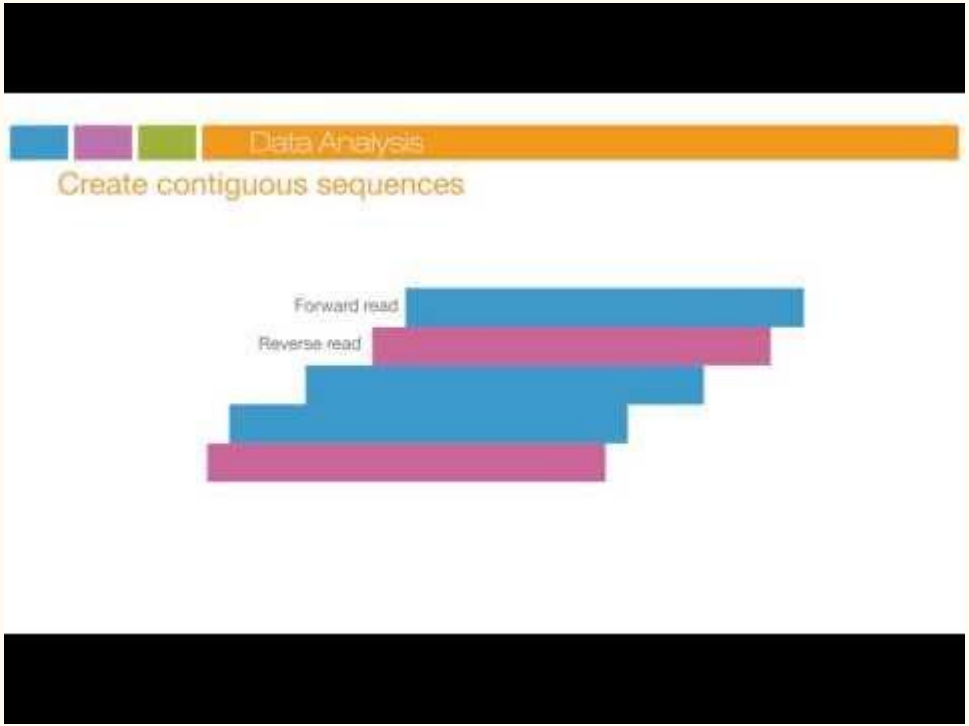
Slide adapted from Ben Langmead

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



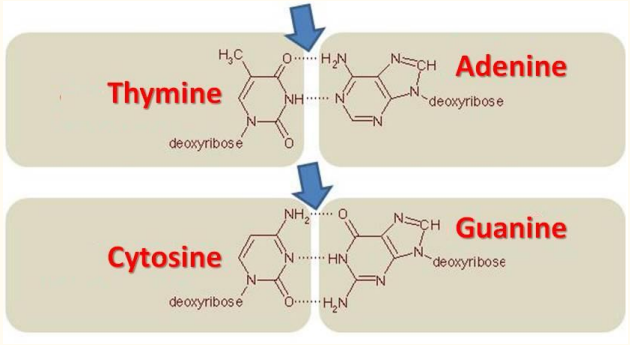
https://www.youtube.com/watch?v=womKfikWlxM

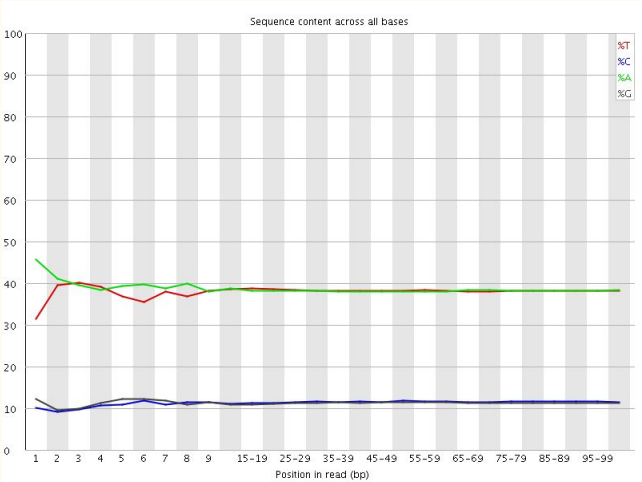
Sequencing errors

1. GC bias

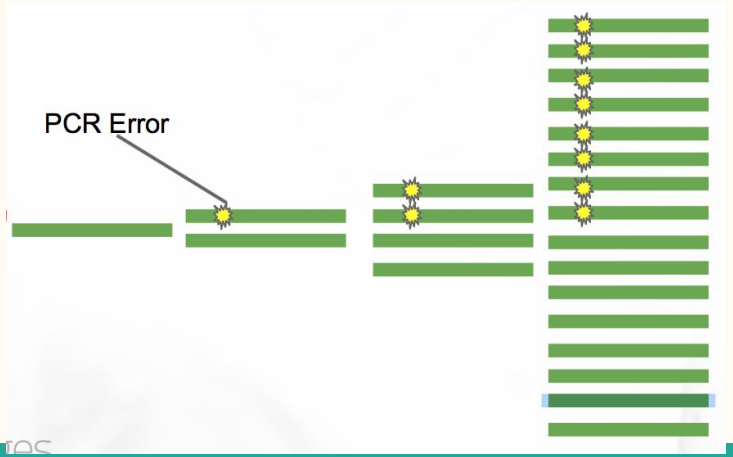
35% to 60% - human

~20% - Plasmodium falciparum

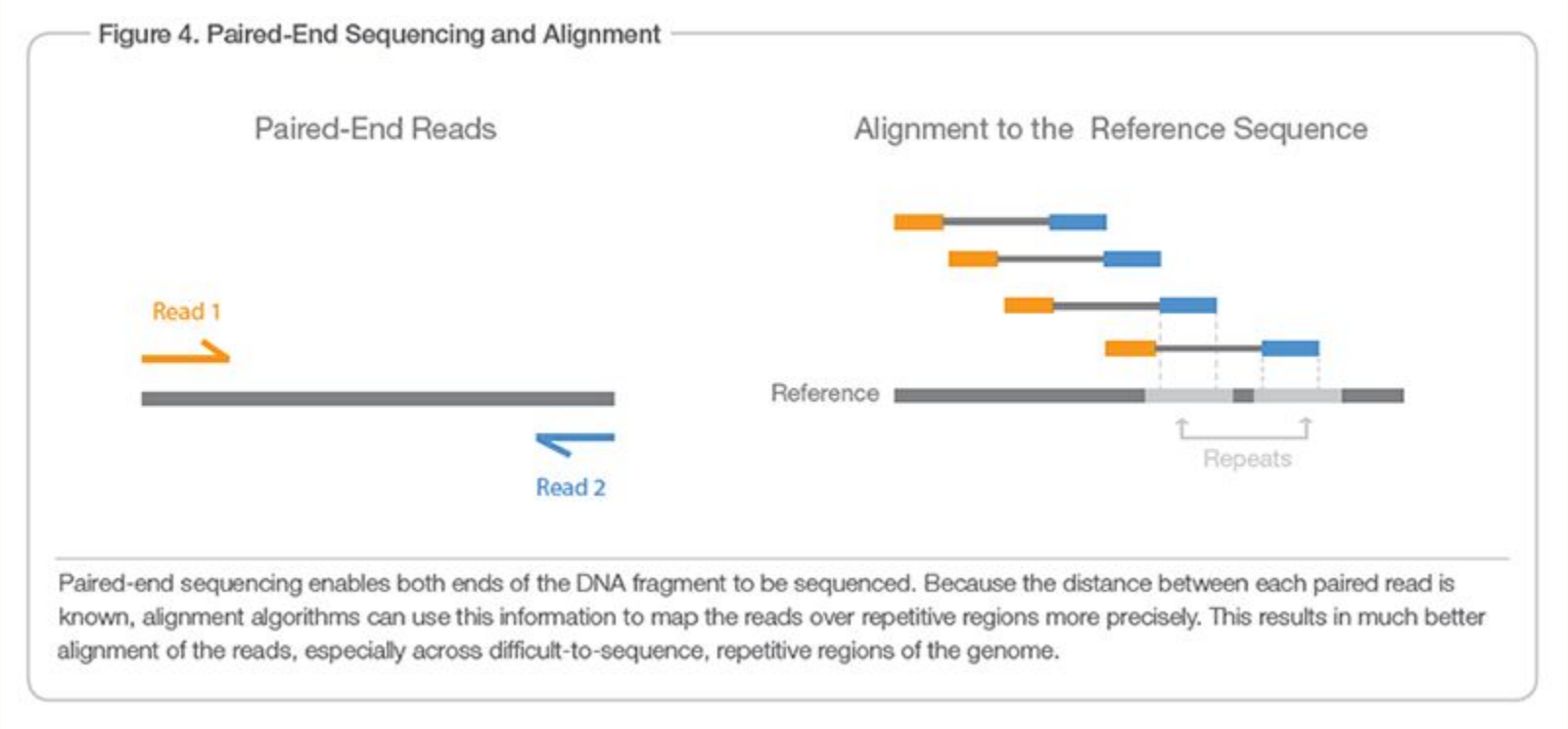




2. Error propagation (1 in 10.000 error rate)



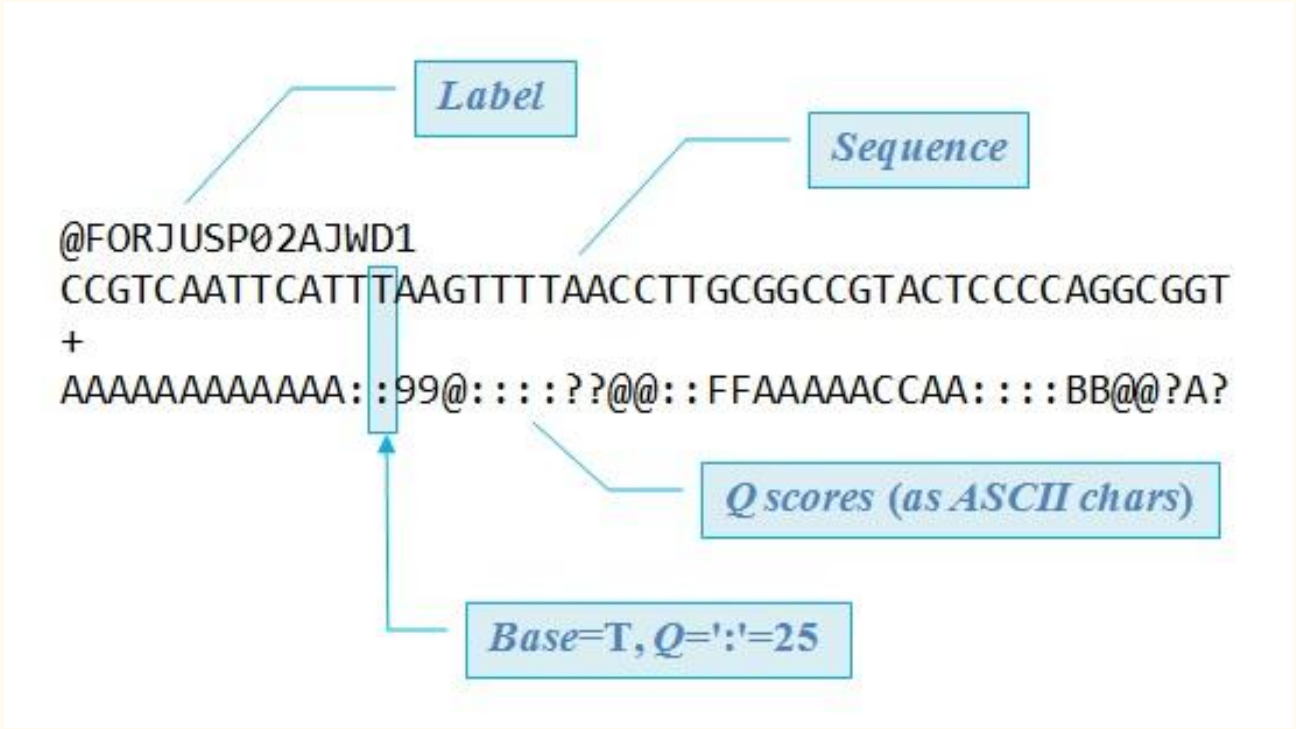
Paired-end sequencing



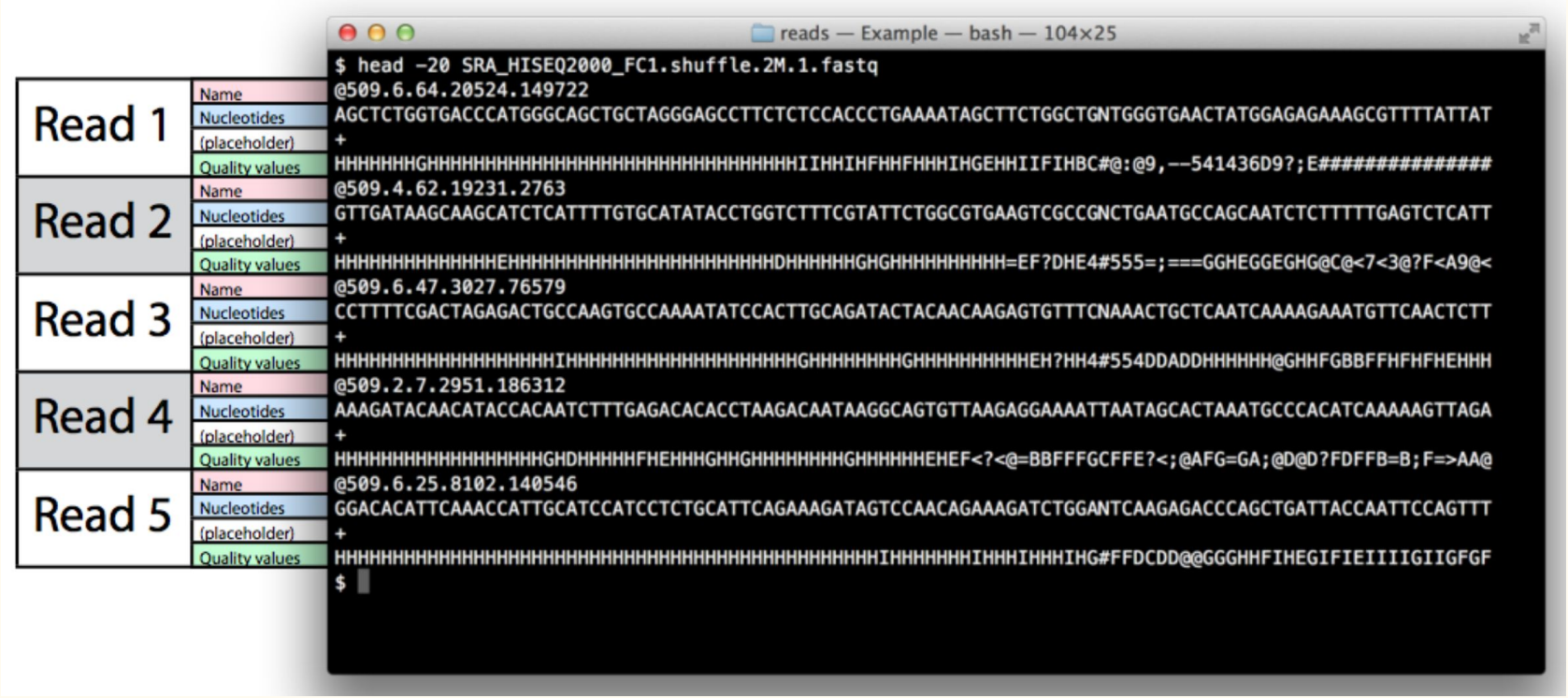
Sequencing data - FASTQ file

4 lines for each read

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



Sequencing data - FASTQ file



Genome reconstruction

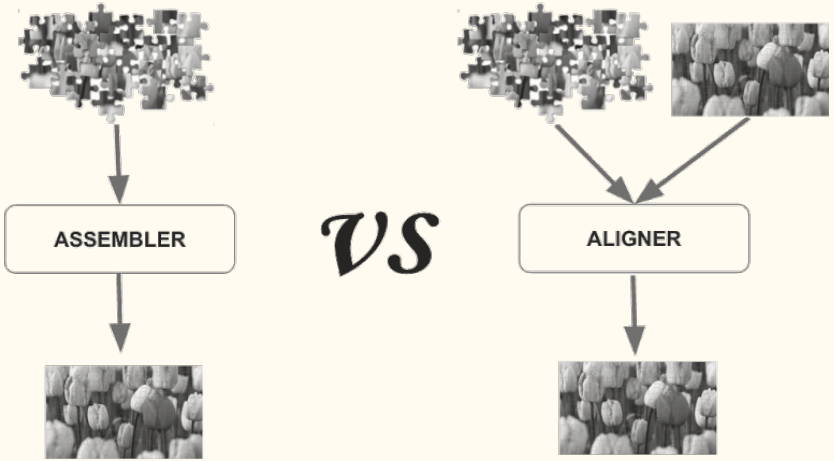
Result of sequencing experiment

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

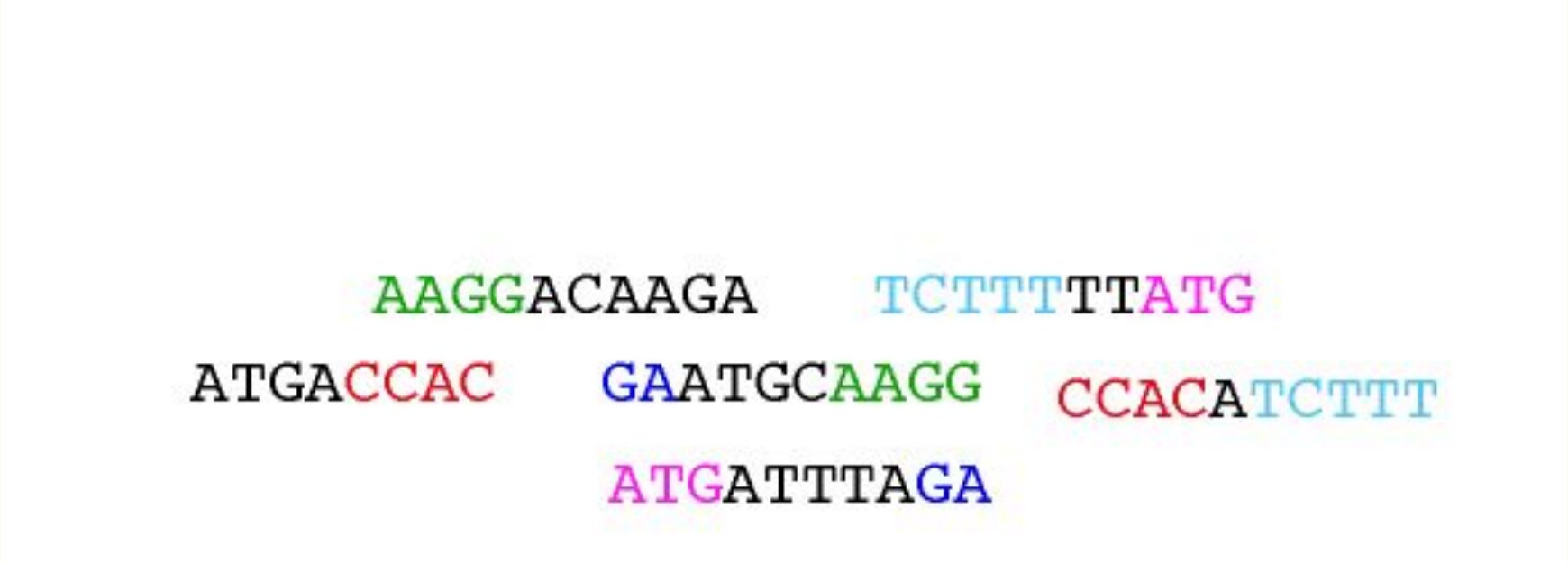
Genome reconstruction

How do we reconstruct genome from reads?

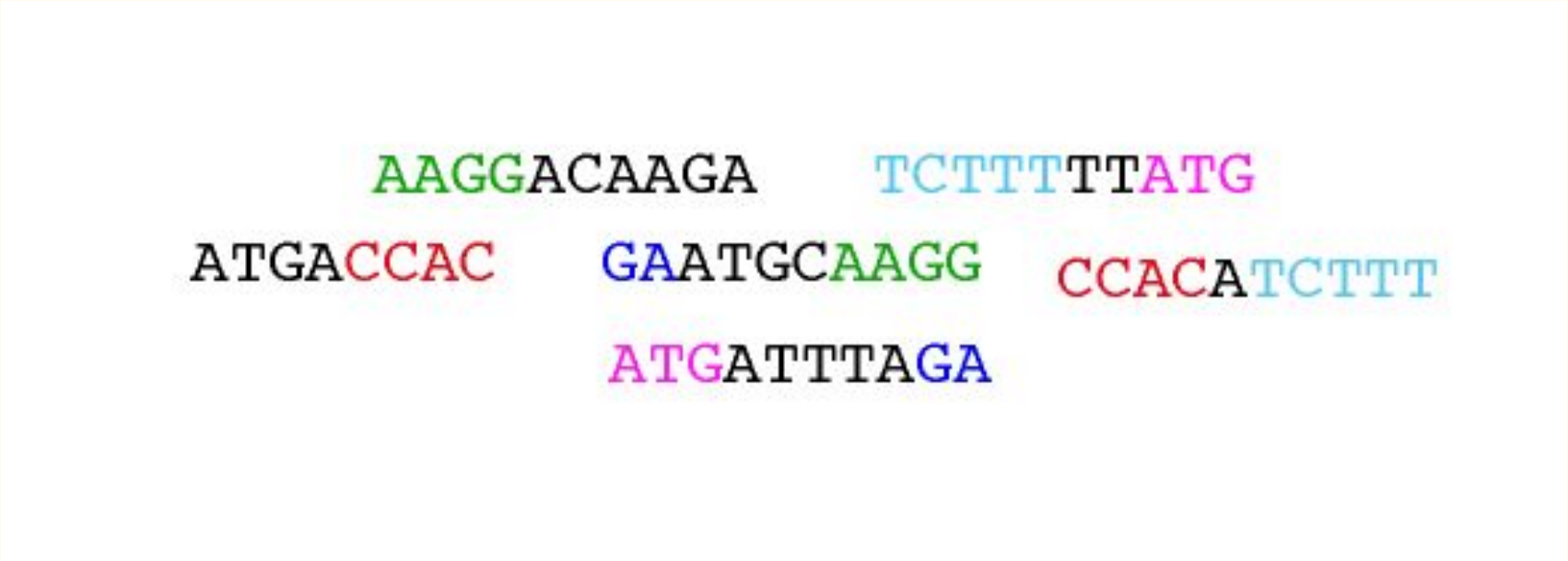
1. Alignment
    - Using reference genome to map the position of the reads
2. Assembly
    - Reconstructing the genome by finding the links between the reads



Alignment


AAGGACAAGA
TCTTTTTTATG
ATGACCAC
GAATGCAAGG
CCACATCTTT
ATGATTTAGA

Assembly


AAGGACAAGA
TCTTTTTATG
ATGACCAC
GAATGCAAGG
CCACATCTTT
ATGATTTAGA

Resources and additional reads

Presentation available at: github.com/vladimirkovacevic/gi-2019-etf

A Computer Scientist's Guide to Cell Biology, A Travelogue from a Stranger in a Strange Land

Genomics 101, Edition 2016

Bioinformatics at COMAV - SNP Calling

Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM

High-Throughput Sequencing Technologies - Review paper