Biological foundations for bioinformatics

Lesson 3

Lesson overview

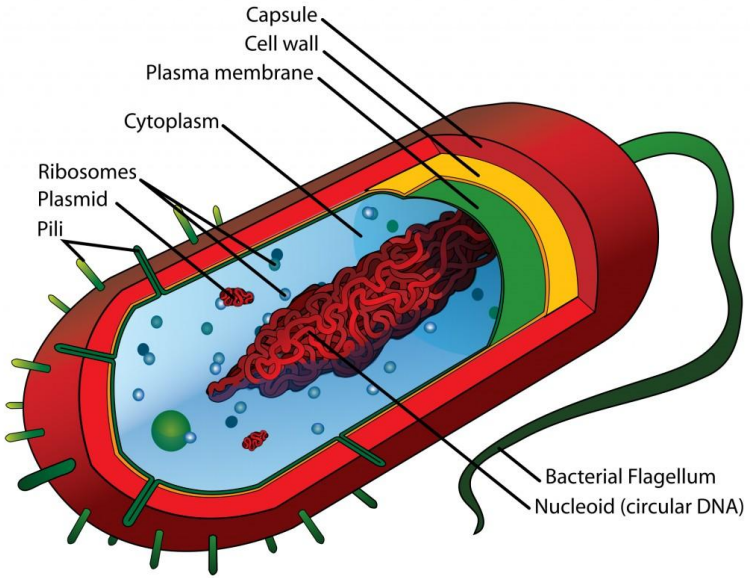
- Molecular biology basics
- Genome sequencing applications and technologies
- Git - Code versioning system

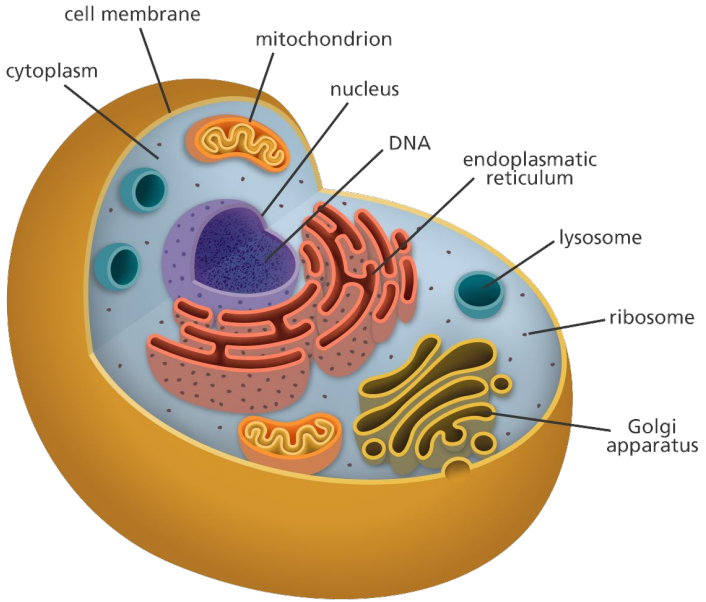
**Why slides in English?*

The cell

Fundamental working units of every living system.

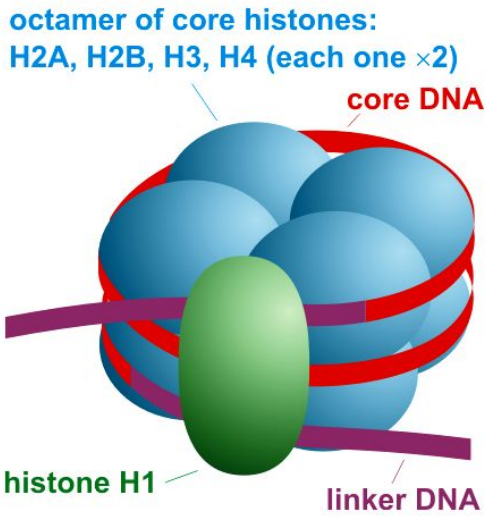
- Prokaryotic (bacteria)
- Eukaryotic (higher organisms - animals, plants)

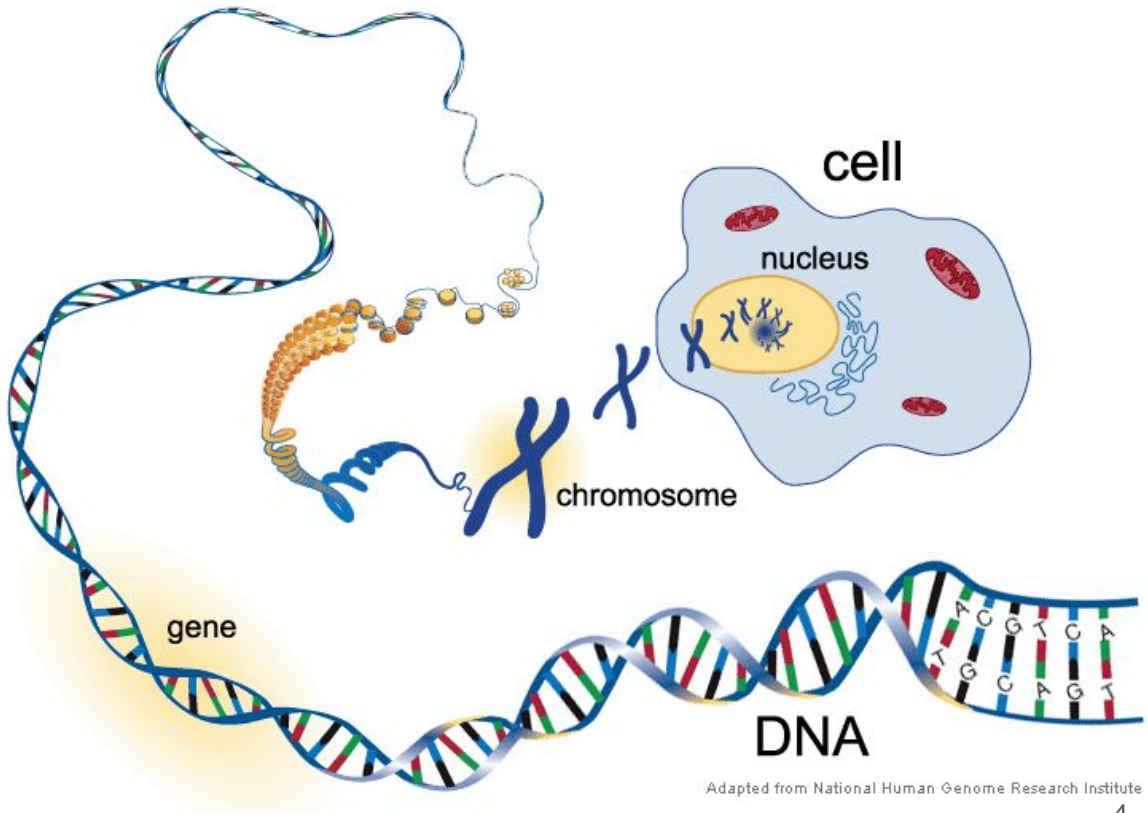




“And inside the nucleus thou lays the mighty DNA”

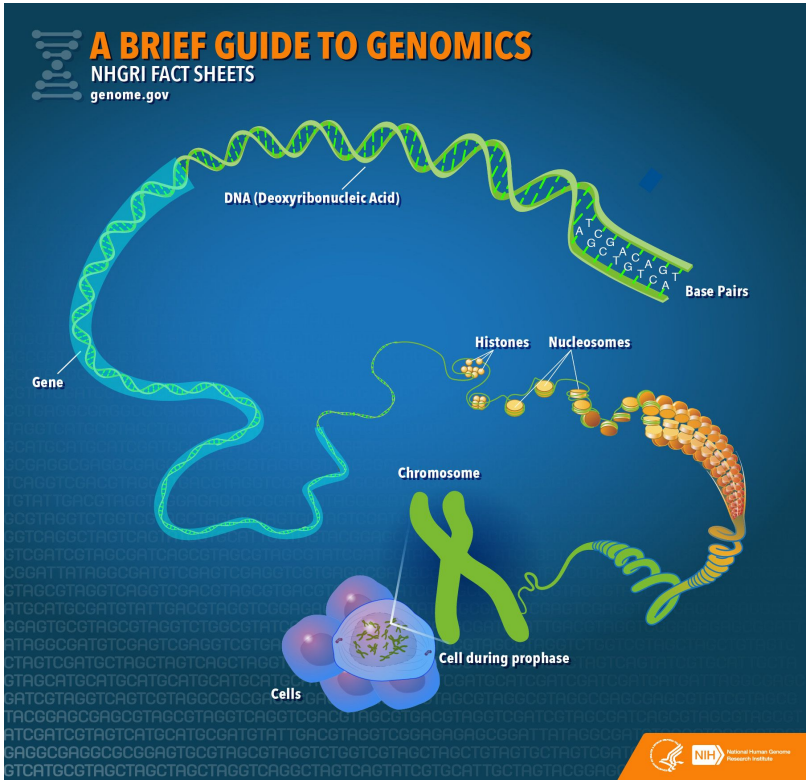
- Chromatin - tightly packed DNA
- A **nucleosome** is a basic unit of DNA packaging in eukaryotes, consisting of a segment of DNA wound in sequence around eight histone protein cores.
- Current model





DNA - the code of life

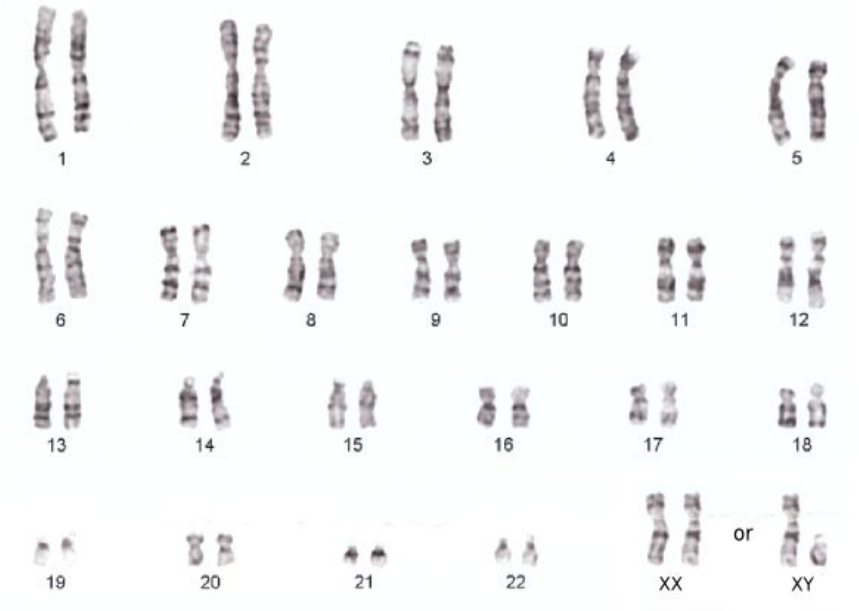
- DNA (deoxyribonucleic acid) - double stranded molecule
- Same in every cell - DNA replication during cell division
- More stable, redundant information - complementary double helix chain
- ~99.6% same between 2 individuals
- Base (nucleotide) pairs (complementary bases)
  - A - T (adenine and thymine)
  - 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
  - CTGGATTATATATCGAAGGGACTAT... etc
  - Intron and exom (2%)

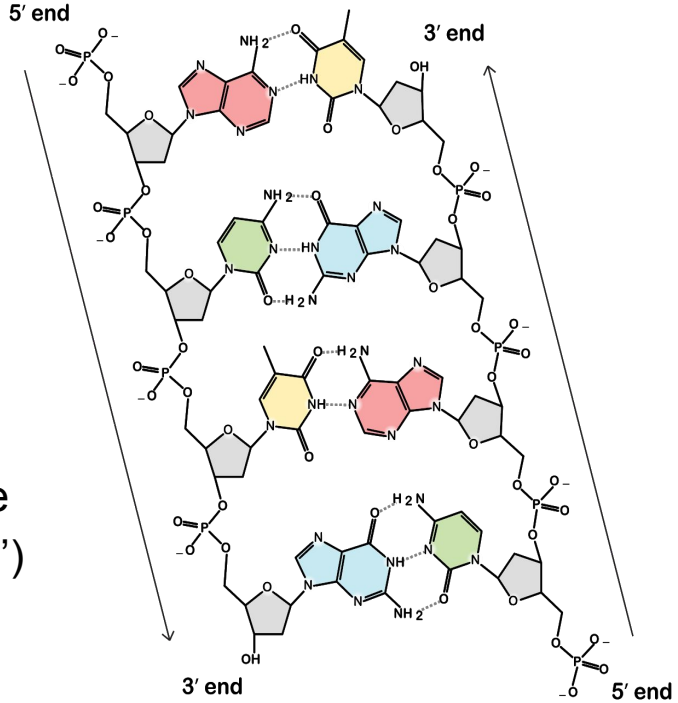


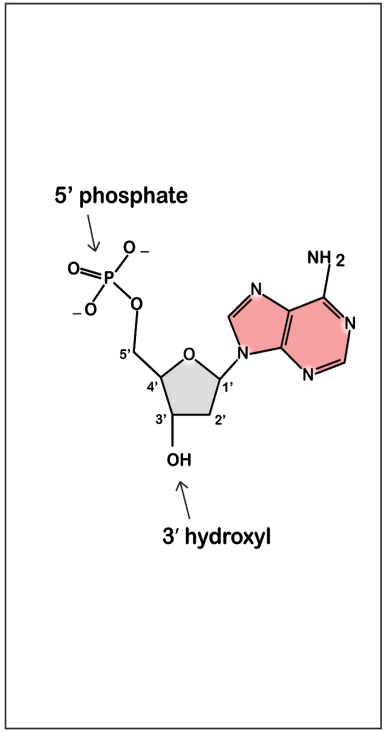
Karyogram

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)





Nucleotide

DNA - discovery

- In 1953, James Watson and Francis Crick, with significant input from Rosalind Franklin's X-ray diffraction images of DNA, proposed the double helix structure of DNA
- This model elucidated how genetic information is stored and replicated

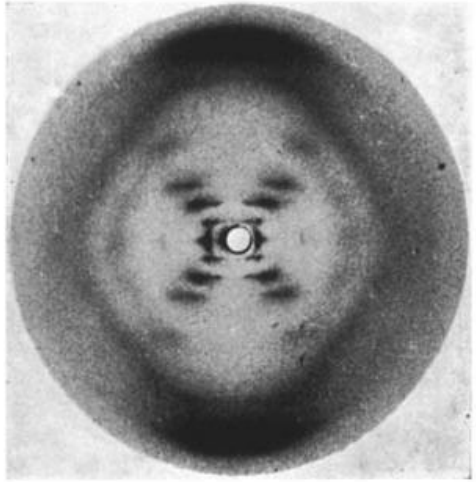


Photo 51

Central dogma of molecular biology

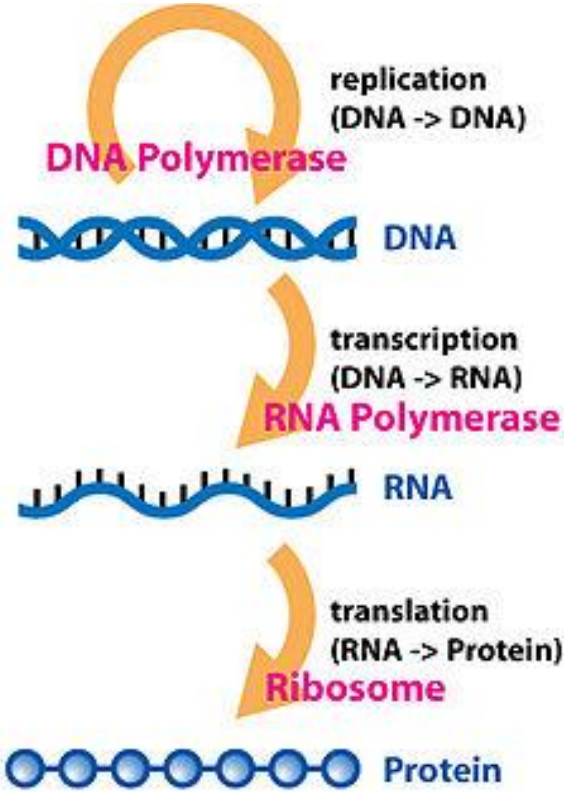
DNA ----> RNA -----> Protein

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.



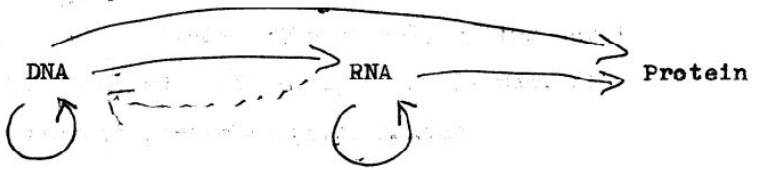
Central dogma of molecular biology

Ideas on Protein Synthesis (Oct. 1956)

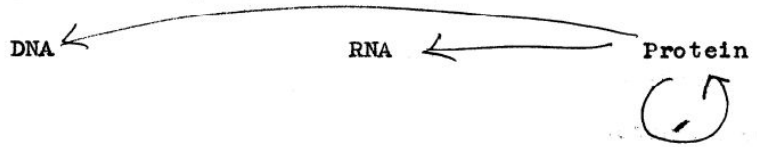
The Doctrine of the Triad.

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



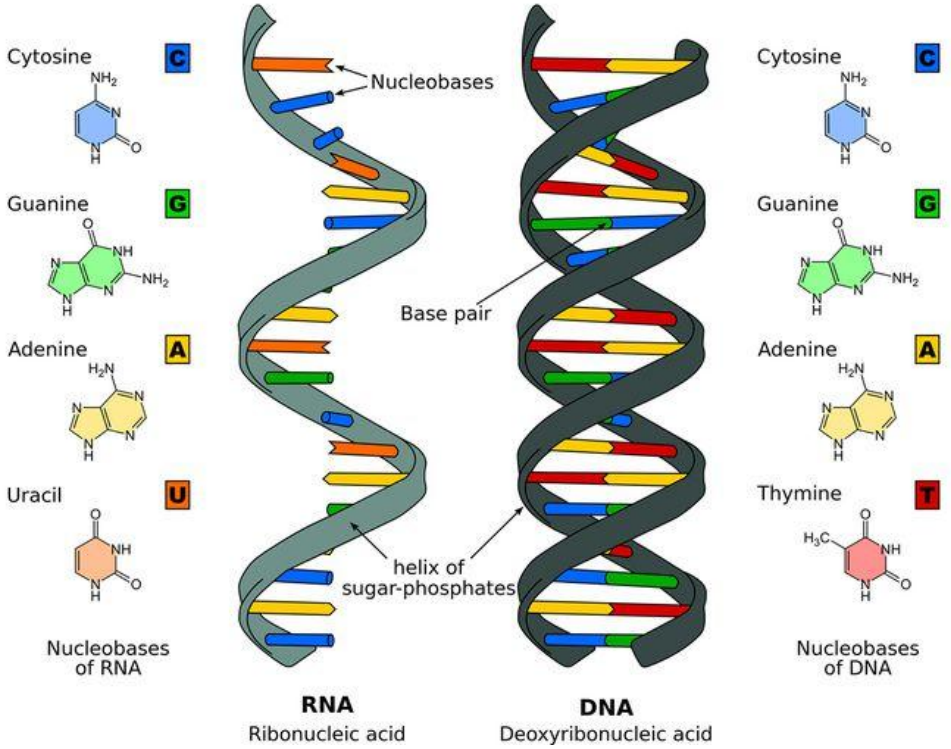
but never



where the arrows show the transfer of information.

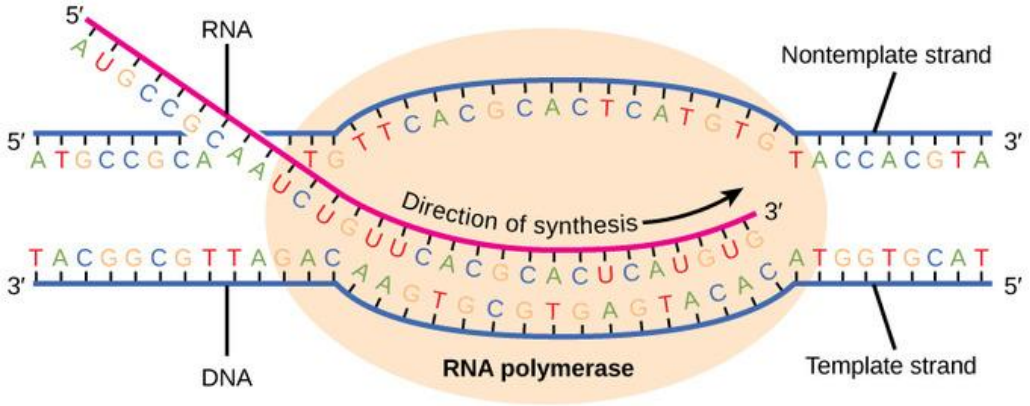
RNA

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

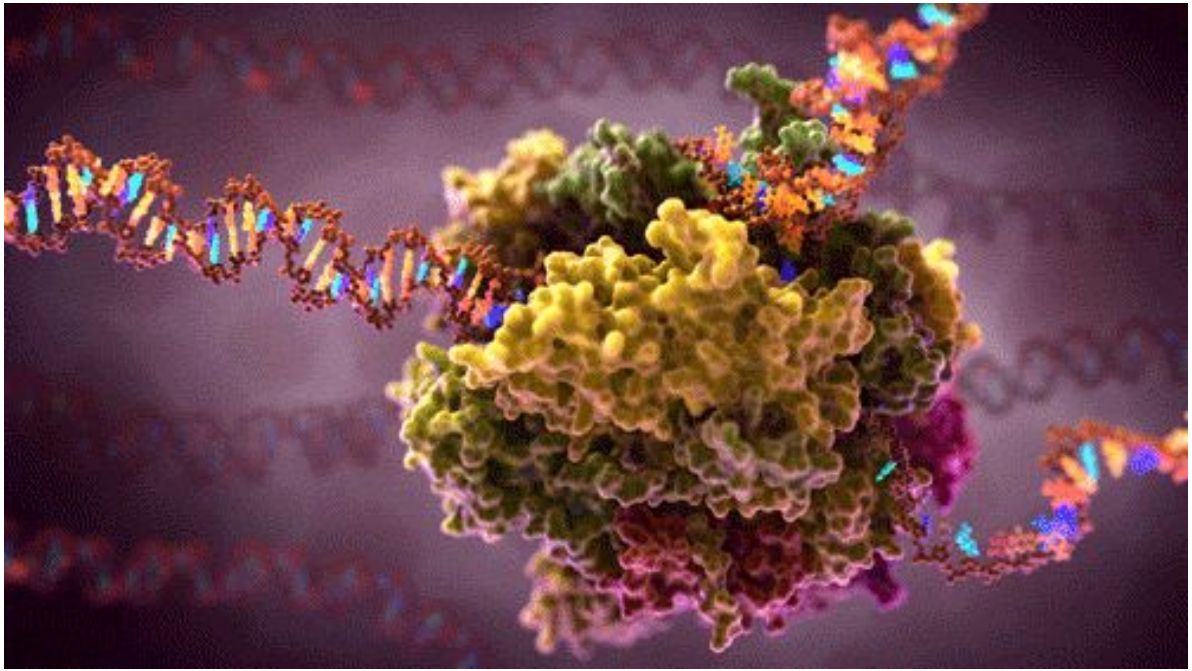


Transcription

- Template (noncoding) strand
  - One which is transcribed by RNAP (RNA polymerase)
- Pre-mRNA is synthesized from a DNA template in the cell nucleus

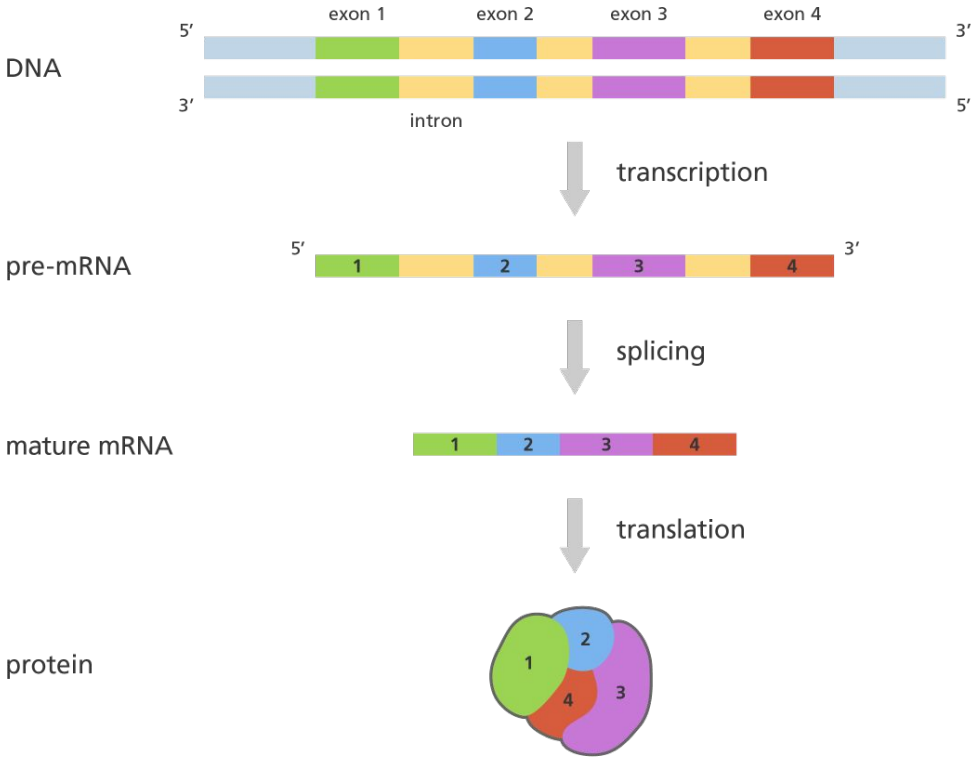


Transcription



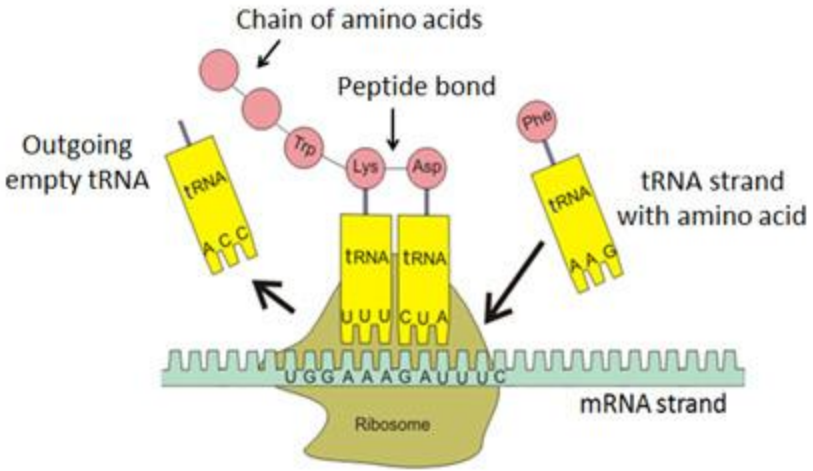
Transcription and splicing

- Splicing: removing all the introns (non-coding regions of RNA) and splicing back together exons (coding regions)
- Alternative splicing
- Transcript (form of the gene, coding sequences for protein synthesis)



Translation

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

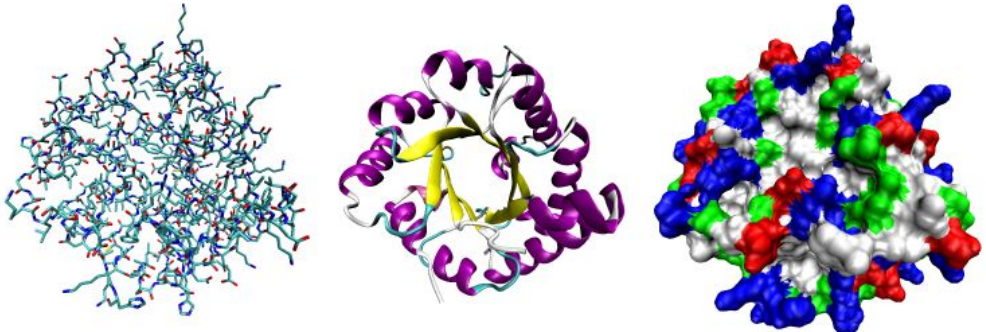


Second letter

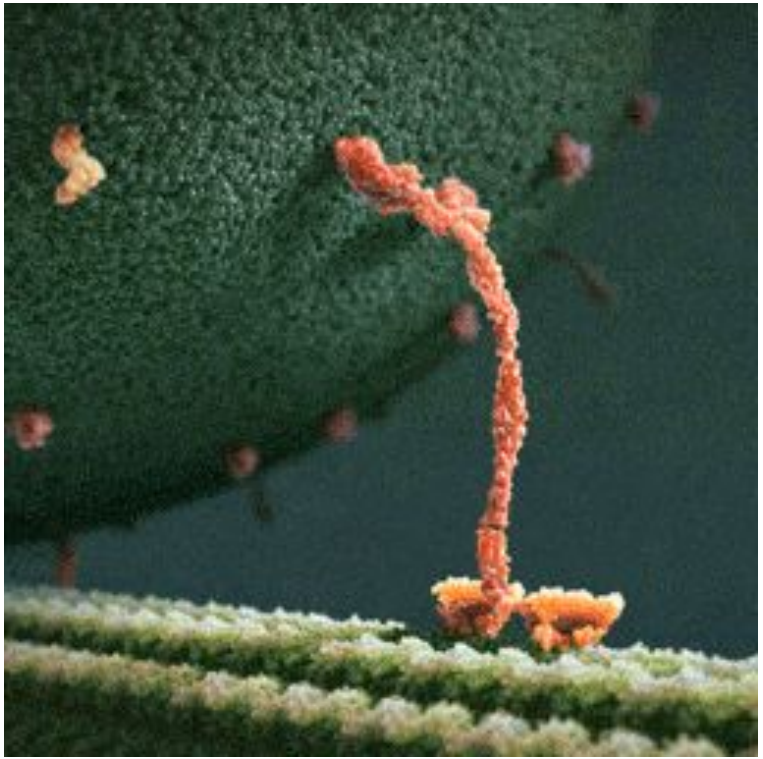
First letter	U	C	A	G	Third letter
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr **UAA Stop** **UAG Stop**	UGU Cys UGC Cys **UGA Stop** UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg 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 (linear sequence of amino acids linked together by peptide bonds)
  - Secondary (local spatial arrangements of the polypeptide chain - alpha helices, beta sheets)
  - Tertiary (three-dimensional conformation of the entire polypeptide chain)
  - Quaternary (multiple polypeptide subunits forming a functional protein complex structure)



Proteins



Proteins

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

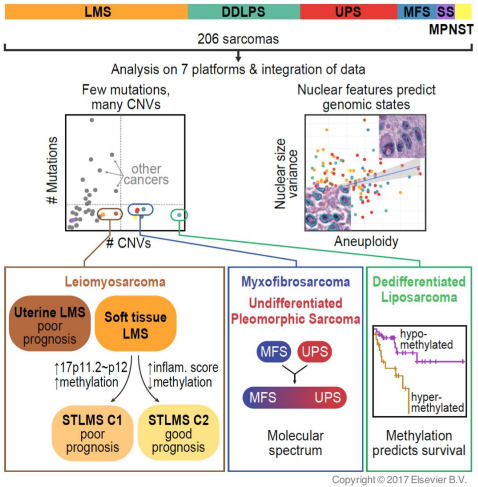
Biology 101 - genotype vs phenotype

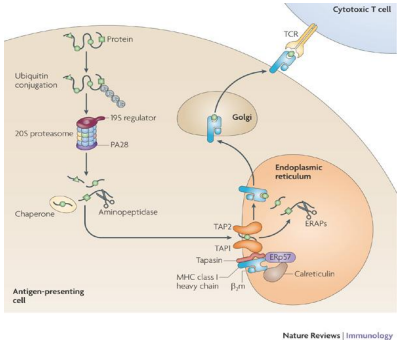
The **genotype** is the part of the genetic makeup of a cell, and therefore of an organism or individual, which determines one of its characteristics (phenotype).

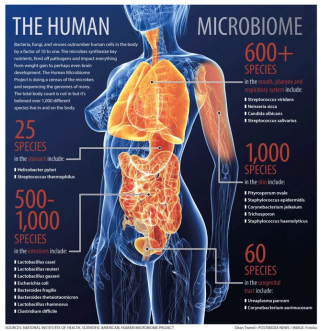
A **phenotype** (from Greek *phainein* , meaning 'to show ', and *typos* , meaning 'type') is the composite of an organism's **observable characteristics** or traits, such as its morphology, development, biochemical or physiological properties, behavior, and products of behavior (such as a bird's nest).

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
- Gene editing
- Forensics







MEDICAL DISPATCH JULY 21, 2014 ISSUE

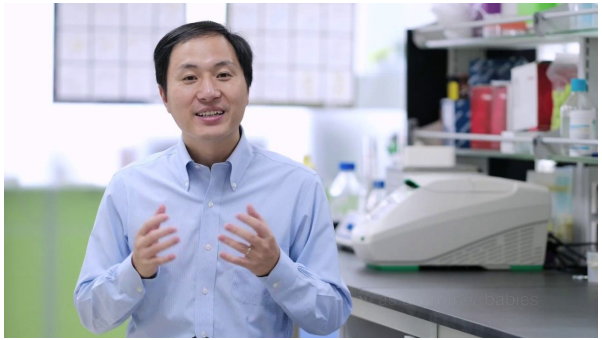
ONE OF A KIND

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

By Seth Mnookin

The world's first germline genetically edited babies

- Clinical project: standard in vitro fertilization
- + CRISPR-Cas9 (technology that can modify DNA)
- Mother was HIV positive
- Modify the CCR5 gene on single egg cell before fertilization to confer genetic resistance to the HIV virus
- CCR5 codes for a protein that HIV uses to enter cells
- Clinical project was conducted secretly until November 2018
- Lulu and Nana are born healthy crying babies



He Jiankui

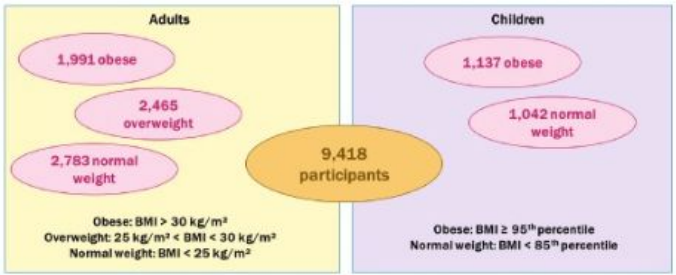
Exome sequencing of 25000 schizophrenia cases and 100 000 controls

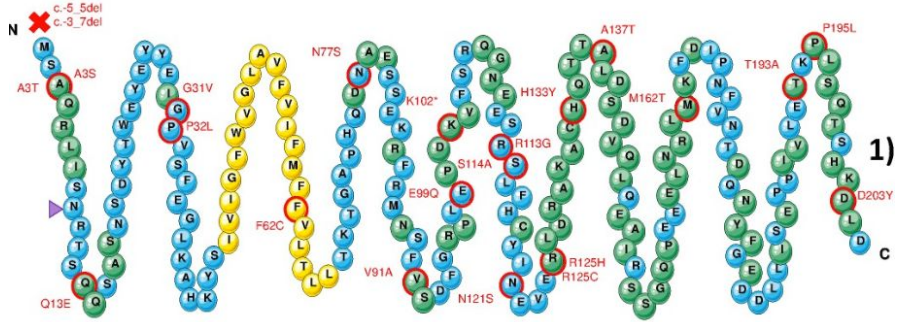
Tarjinder Singh

- The Schizophrenia Exome Sequencing Meta-analysis (SCHEMA) consortium (2017) - aggregating and analyzing high-throughput sequencing data
- Understand the genetic causes of schizophrenia
- Motivate the development of new therapeutics
- Life expectancy - 12-15 years shorter life expectancy
- **10 genes** that when disrupted, dramatically increase risk for schizophrenia
- Odds ratios 4 - 50, $P < 2e\text{-}6$
- 2 genes code glutamate receptors - crucial in brain cells communication
- 10 genes have no protein-truncating variant signal

Pathogenics loss-of-function in MRAP2 cause metabolic syndrome - A. Bonnefond

- Melanocortin receptor accessory protein 2 (MRAP2) is a transmembrane accessory protein predominantly expressed in the brain
- Deletion of Mrap2 results obesity in both mice and human
- 23 rare mutations in MRAP2
  - 2 frameshift
  - 1 non-sense
  - 20 missense





Insight into genetic architecture of autism

Adam Rocke

- Autism sequencing consortium (SPARK)
- 18,381 autism spectrum disorder (ASD) cases and 27,969 controls
- ASD affects 1–1.5% of individuals and is highly heritable
- identifies 5 risk loci

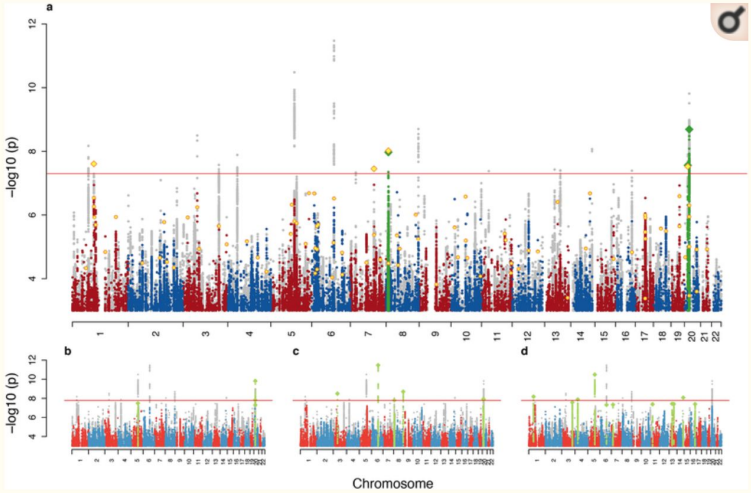
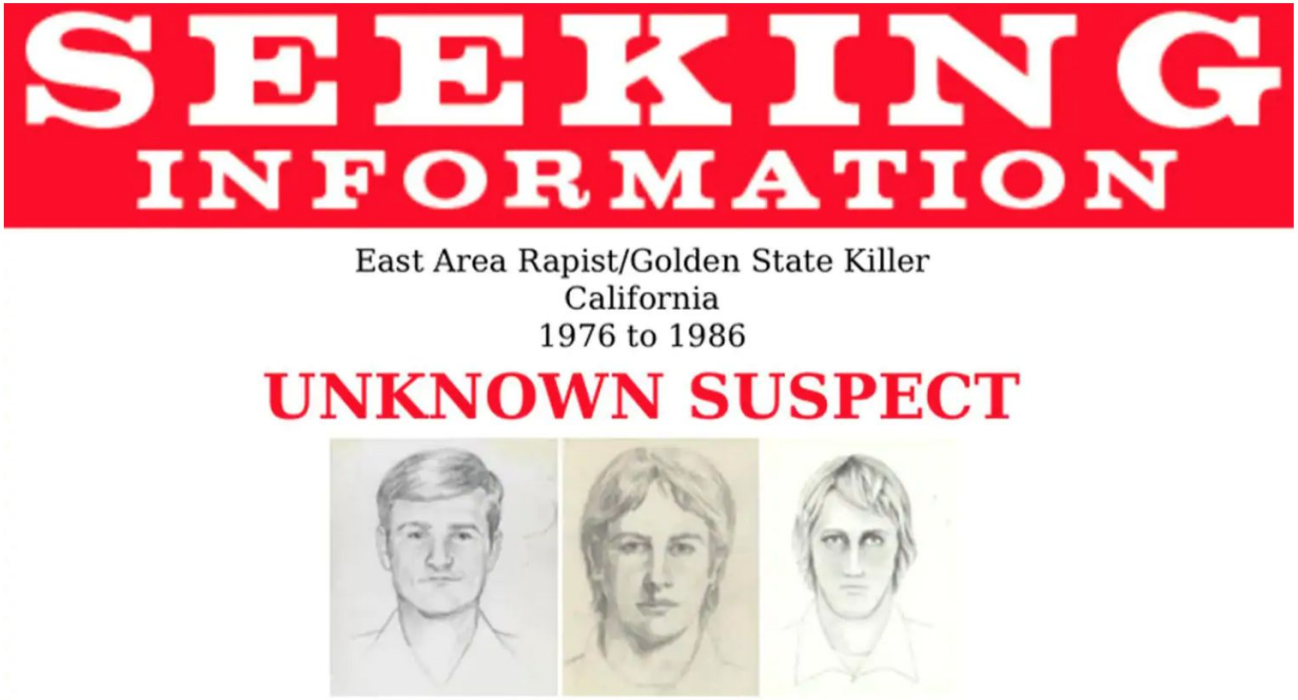


Figure 1.

Manhattans plots:

with the x axis showing genomic position (chromosomes 1–22) and the y axis showing statistical significance as −log10 (*P*) of z statistics. **a:** The main ASD scan (18,381 cases and 27,969 controls) with the results of the combined analysis with the follow-up sample (2,119 cases and 142,379 controls) in yellow in the foreground. Genome-wide significant clumps are painted green with index SNPs as diamonds. **b-d**: Manhattan plots for three MTAG scans of ASD together with, respectively, schizophrenia[12] (34,129 cases and 45,512 controls), educational attainment[29] (*N* = 328,917) and major depression[29] (111,902 case and 312,113 controls). See Supplementary Figures 45–48 for full size plots. In all panels the results of the composite of the five analyses (consisting for each marker of the minimal p-value of the five) is shown in grey in the background.

Golden state killer



Golden state killer

- https://www.gedmatch.com, tutorial
- Applications for comparing your DNA test results with other people
- Genealogical Data Communication
- Software developed by the Church of the Latter Day Saints

- Site found 10-20 distant relatives of the killer, roughly, equivalent of third cousins
- “When you go that far back in time, you have trees that grow huge,” Holes said.
- Census data, old newspaper and a gravesite locator relatives, websites such as LexisNexis.

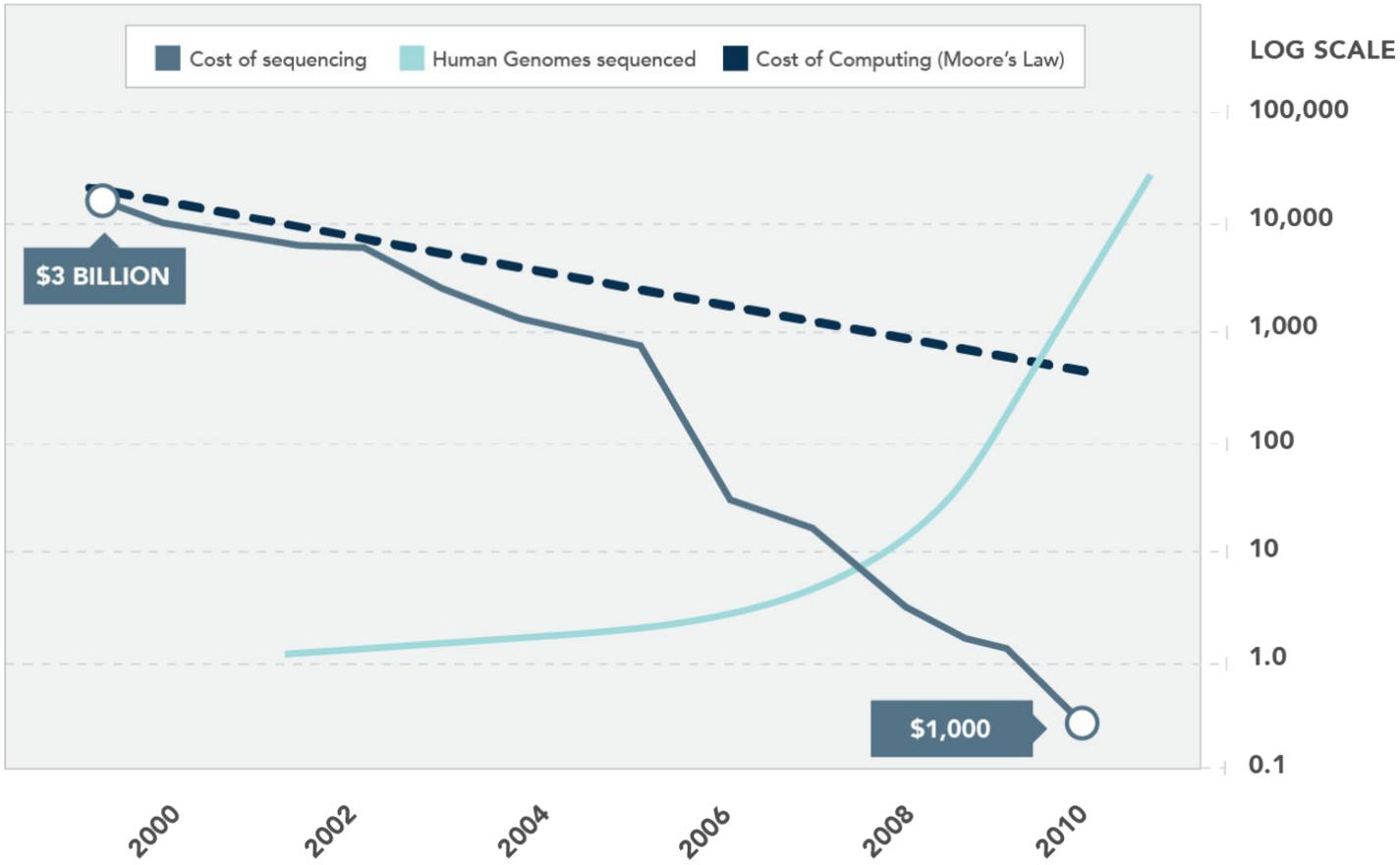
Golden state killer

Joseph James DeAngelo

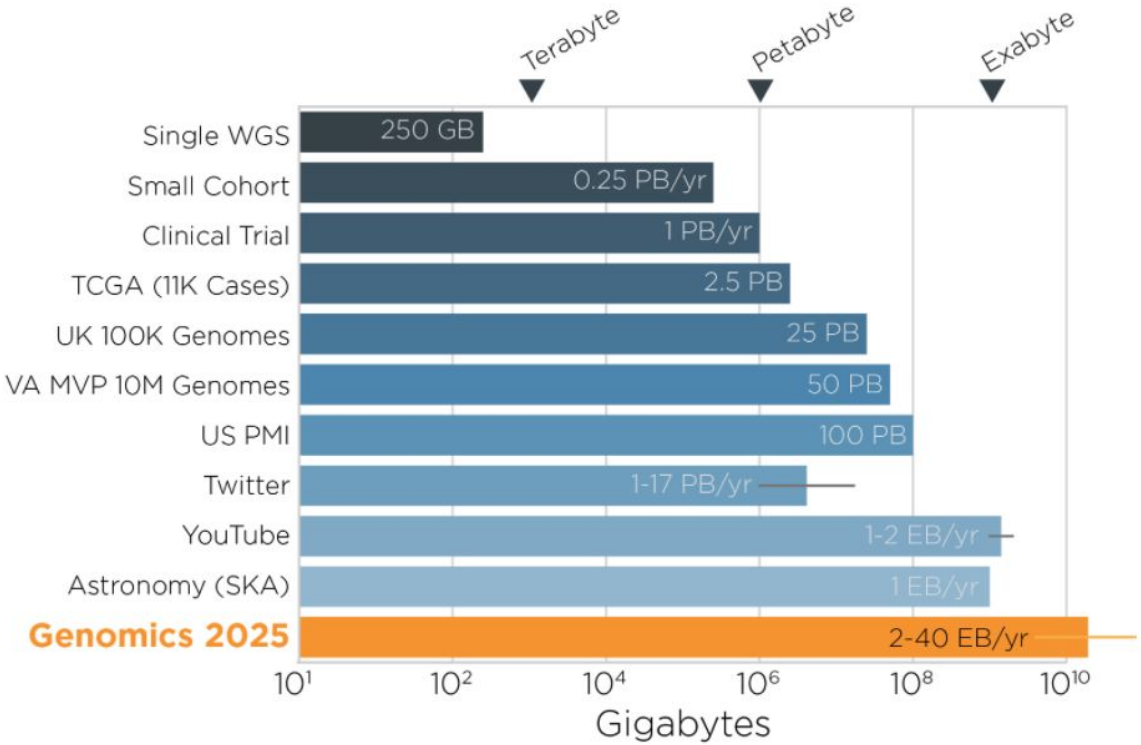
Genome sequencing

- 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

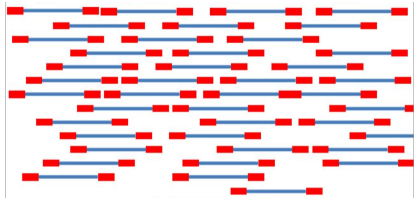
Compute for Alignment:
10,000 trillion CPU hrs
= 83x time since Big Bang

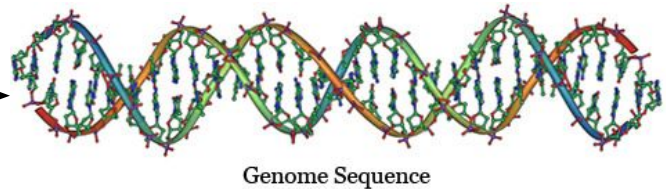
Variant Calling:
~**2 trillion** CPU hrs

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!

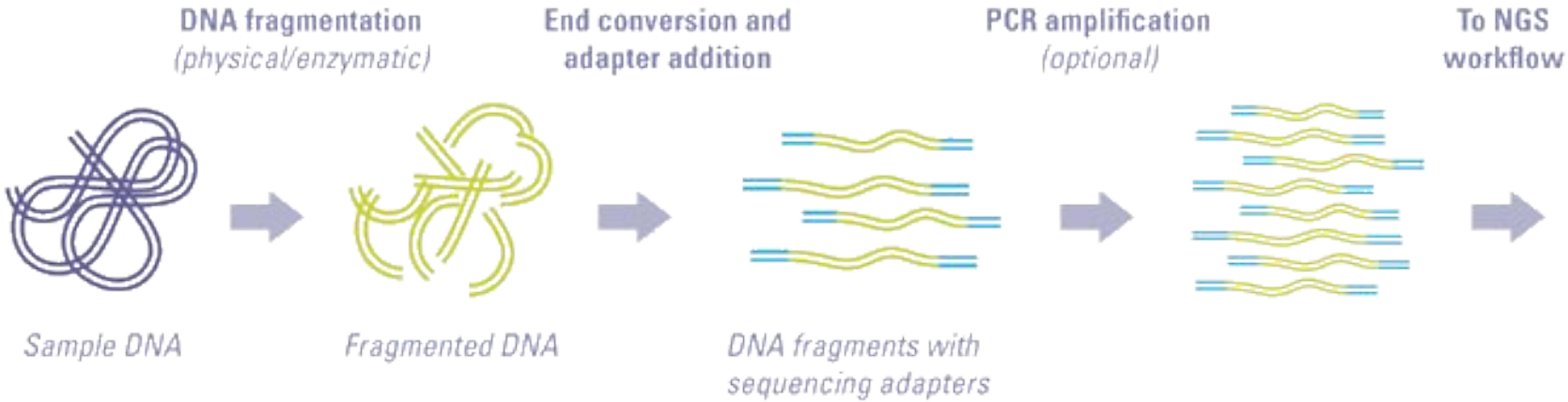




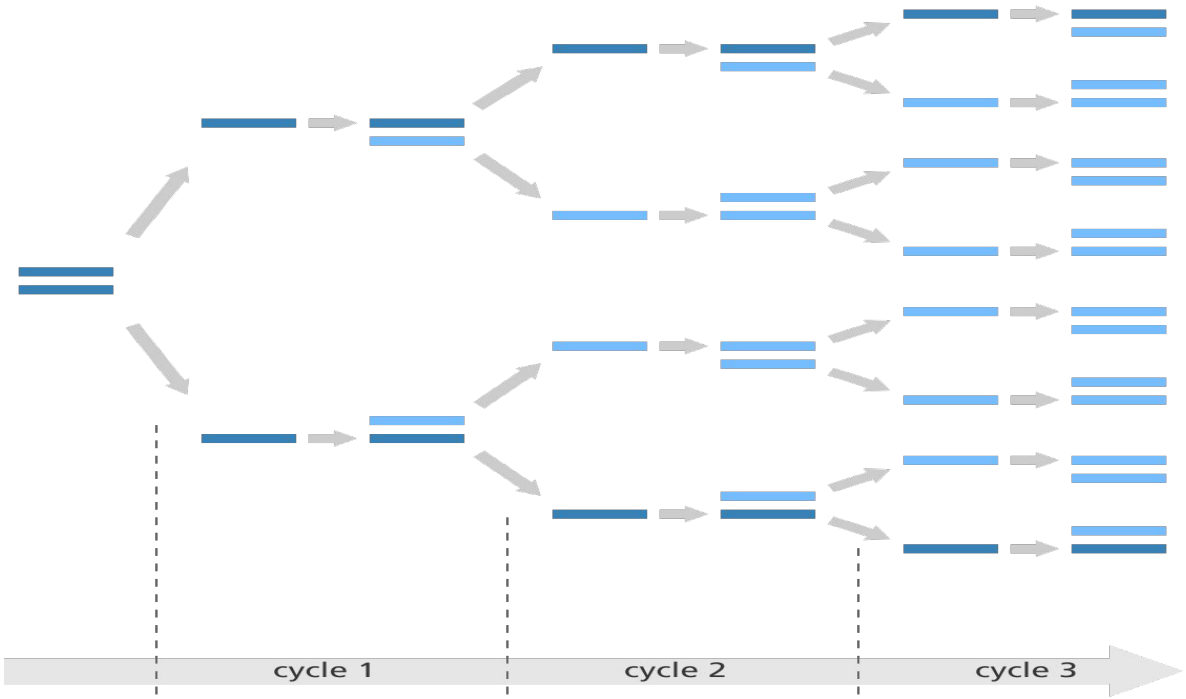
Genome Sequence

Illumina next generation sequencing

- Read - DNA fragment after reading it in sequencer
- Typical whole genome sequencing experiment:
  - 200-500 million reads
  - 150-250 bases (letters long)

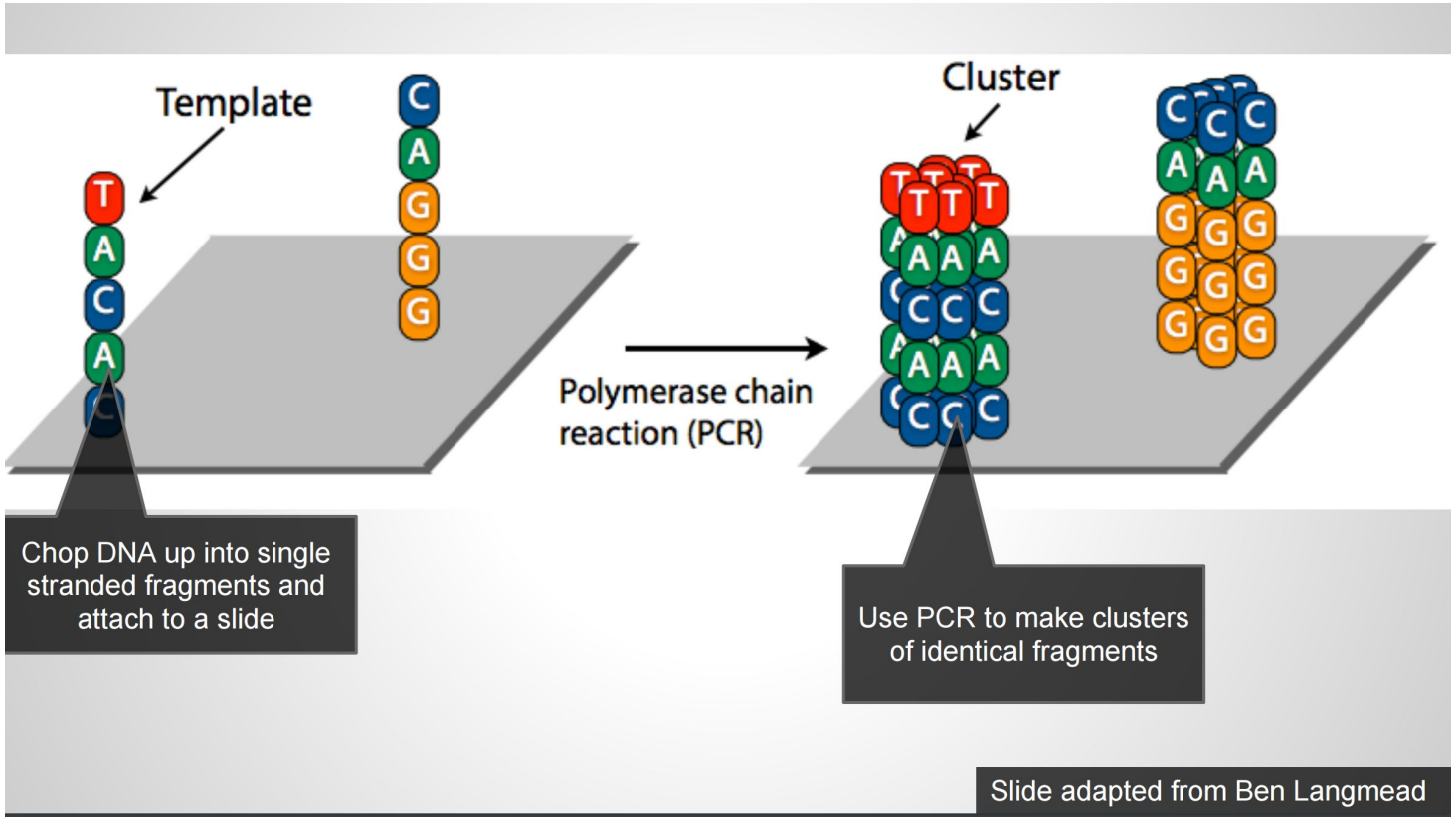


Sequencing - PCR (polymerase chain reaction)

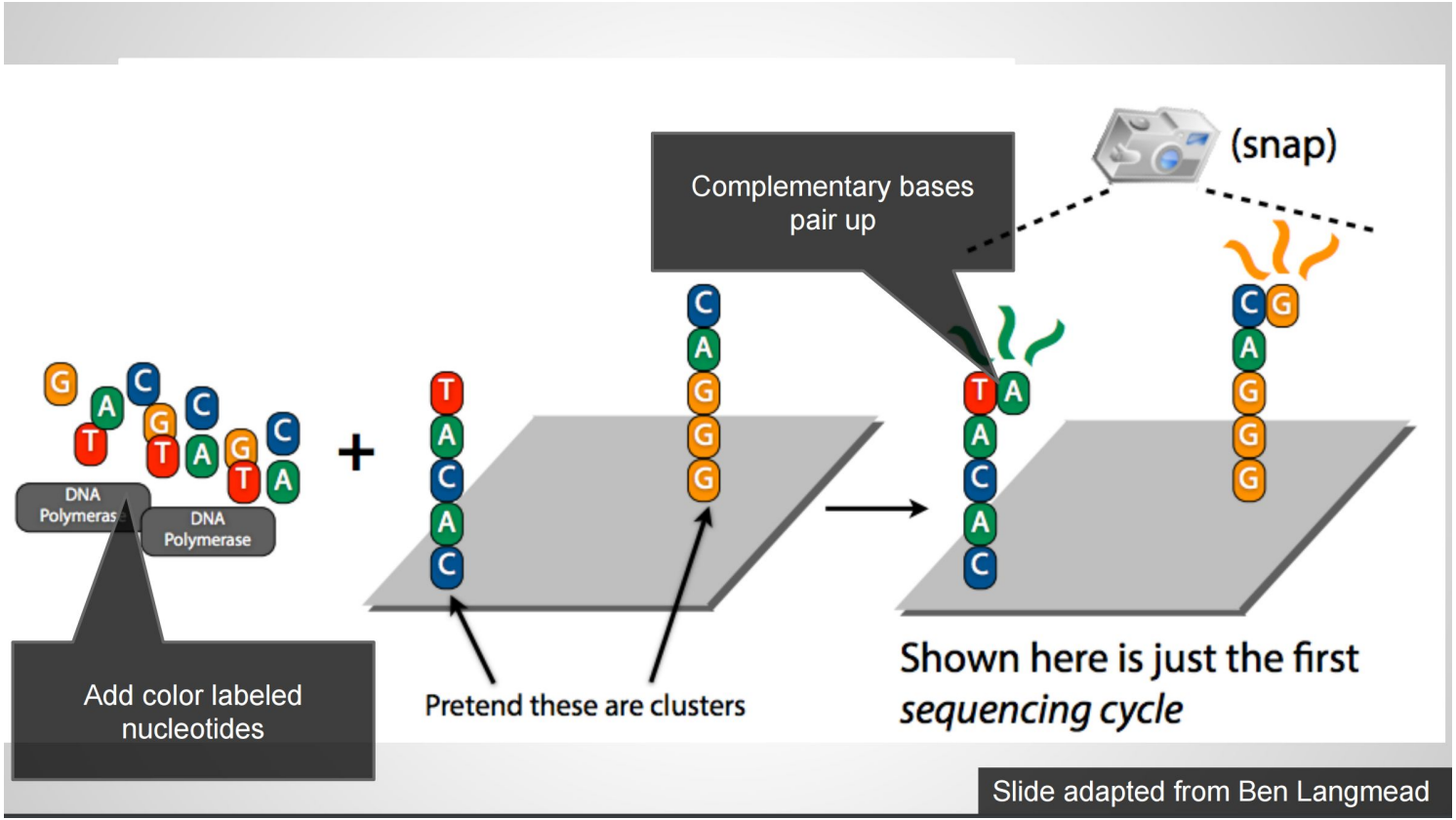


Bridge amplification

Sequencing (Illumina)

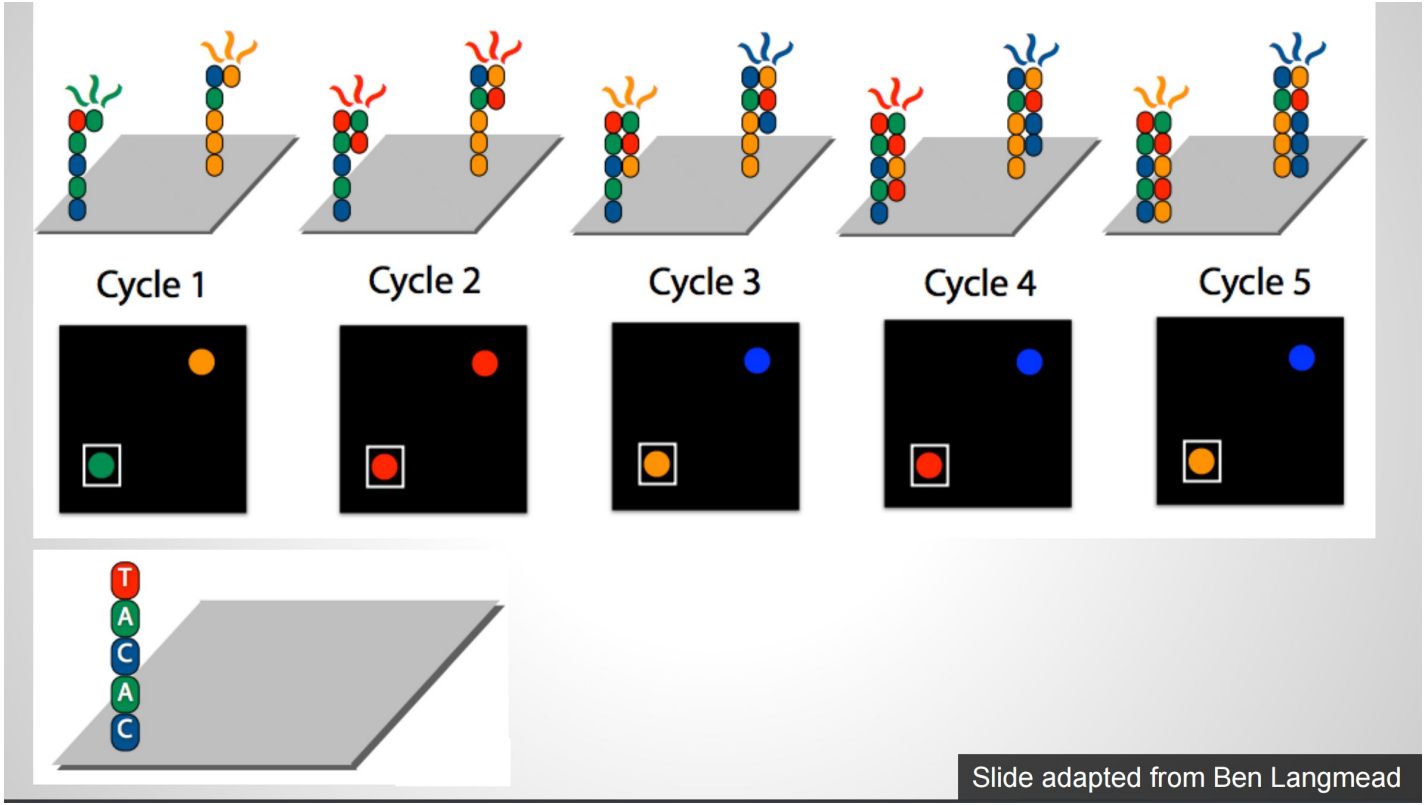


Sequencing (Illumina)



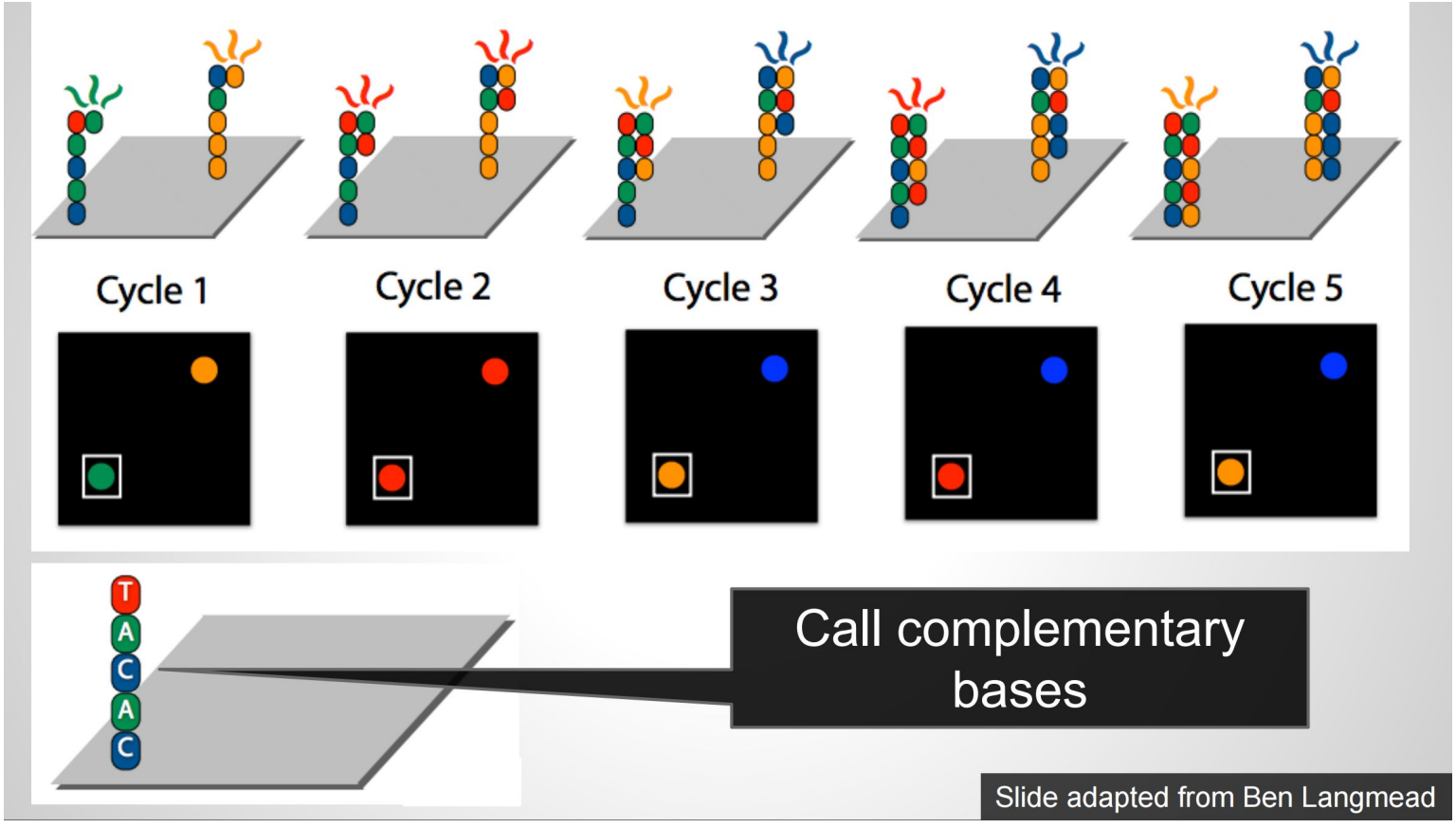
Slide adapted from Ben Langmead

Sequencing (Illumina)



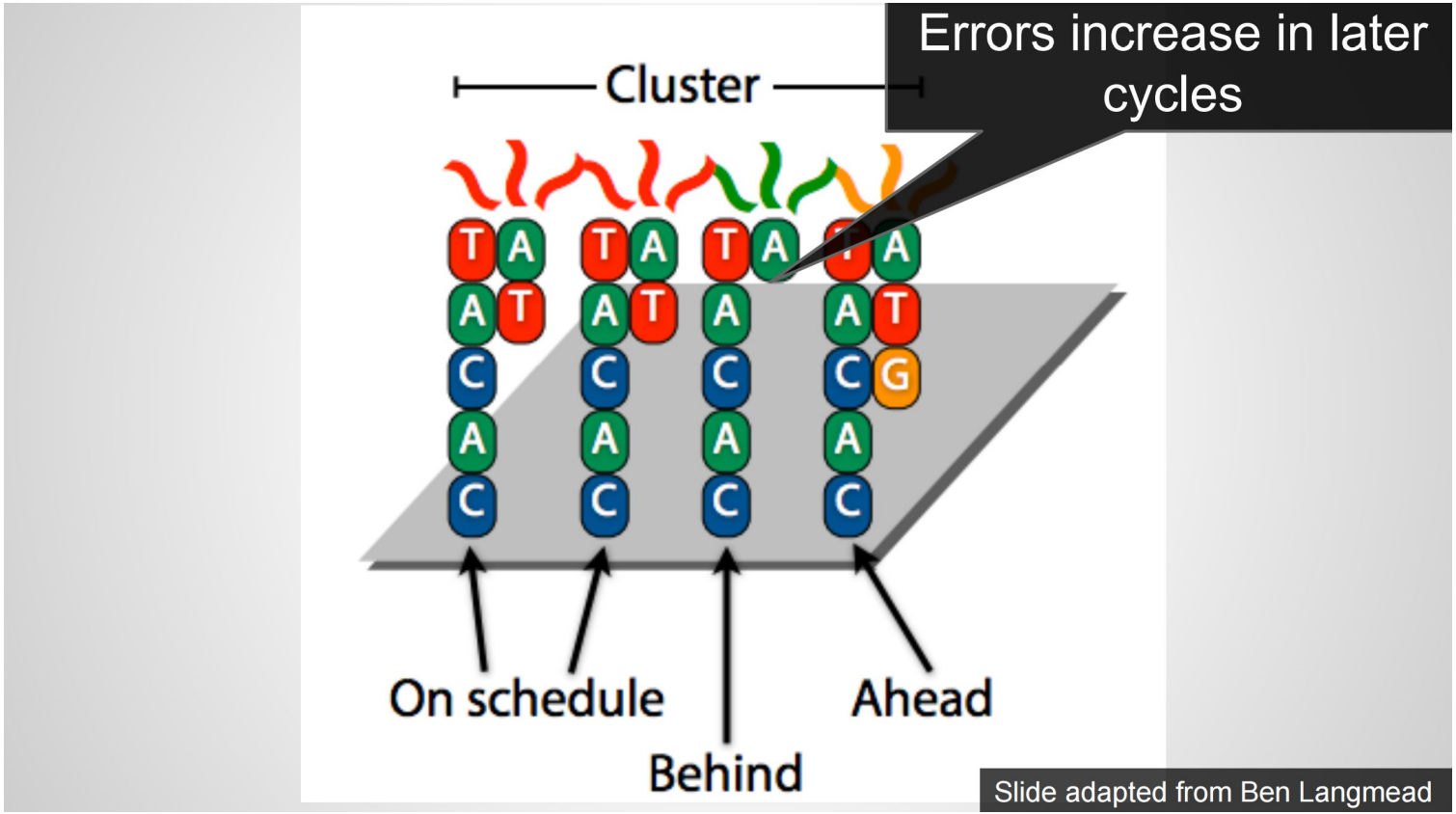
Slide adapted from Ben Langmead

Sequencing (Illumina)



Slide adapted from Ben Langmead

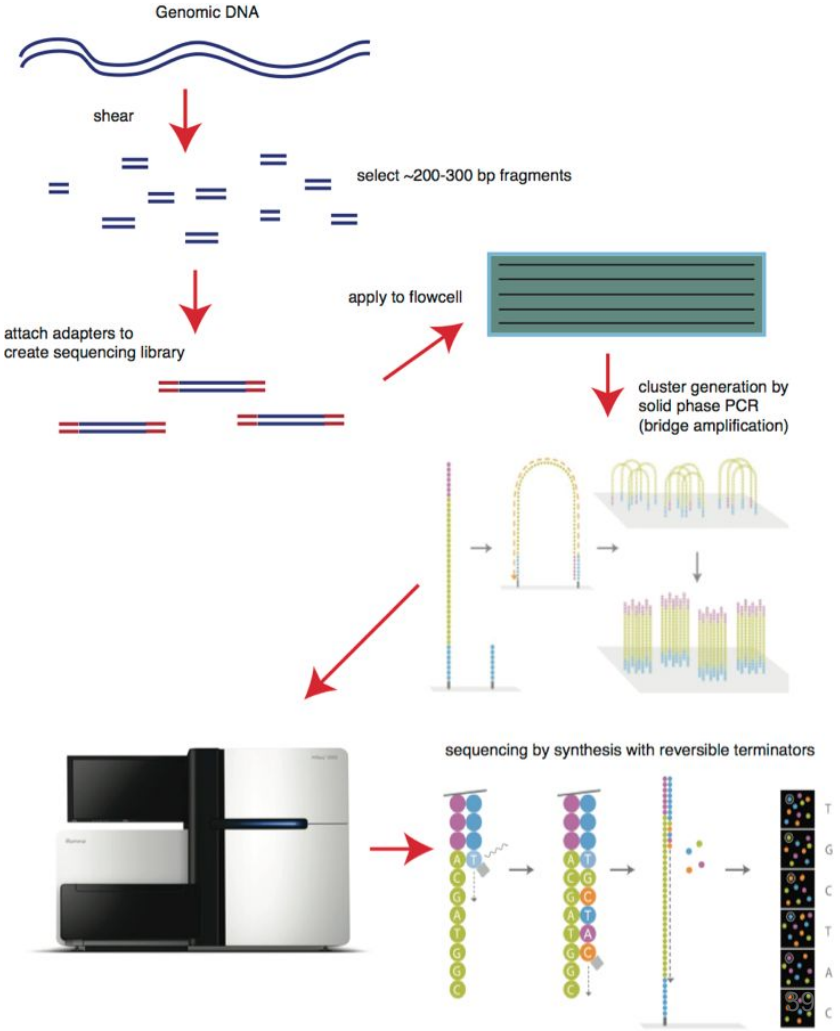
Sequencing error



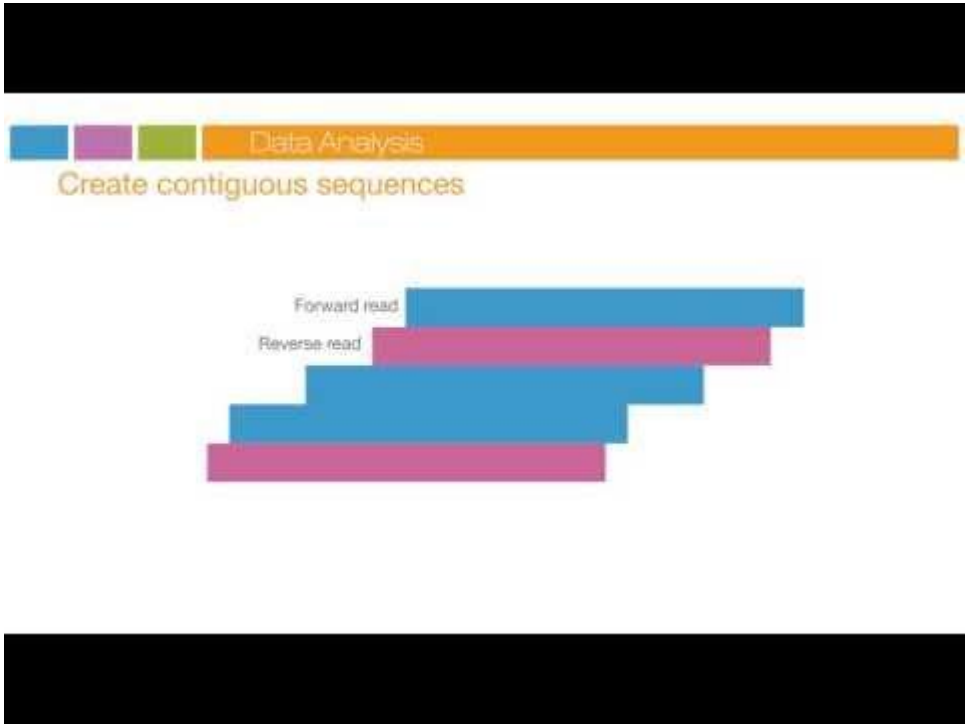
Slide adapted from Ben Langmead

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



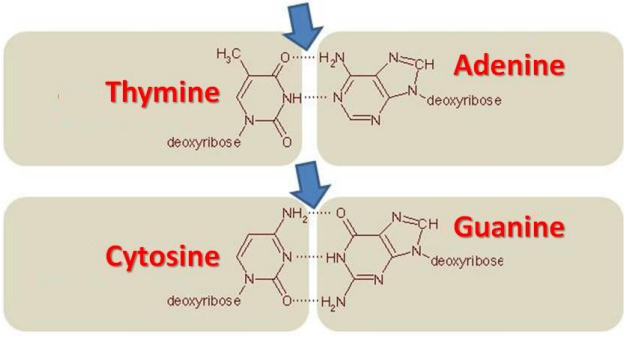
https://www.youtube.com/watch?v=womKfikWlxM

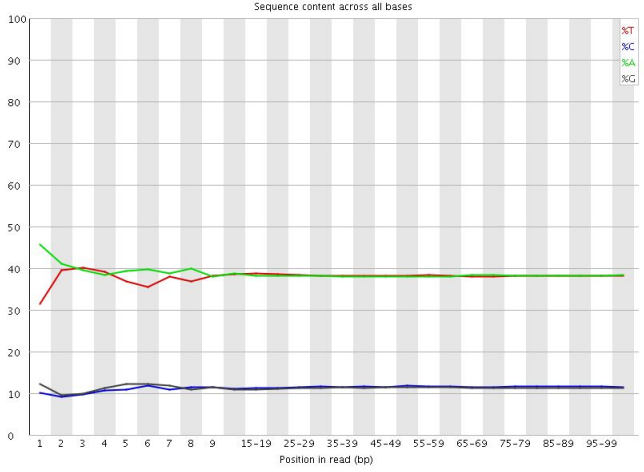
Sequencing errors

1. GC bias

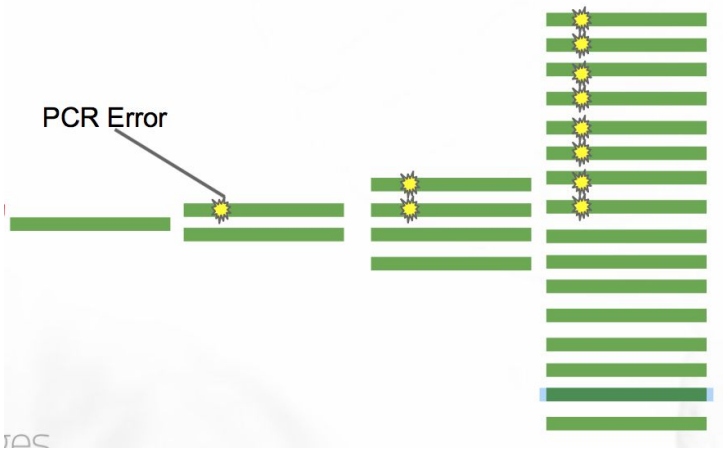
35% to 60% - human

~20% - Plasmodium falciparum



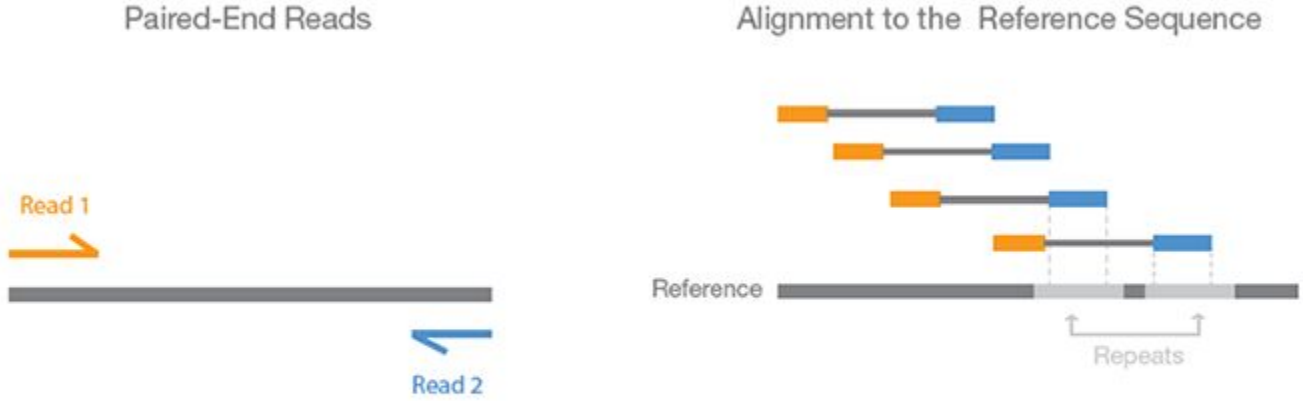


2. Error propagation (1 in 10.000 error rate)



Paired-end sequencing

Figure 4. Paired-End Sequencing and Alignment

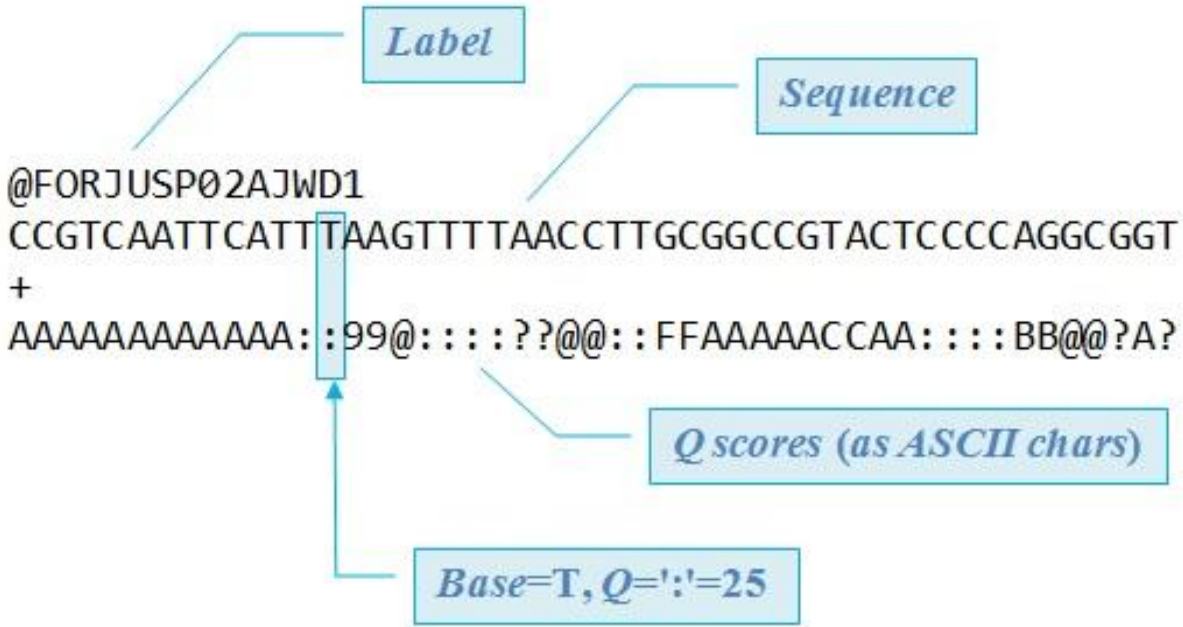


Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across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

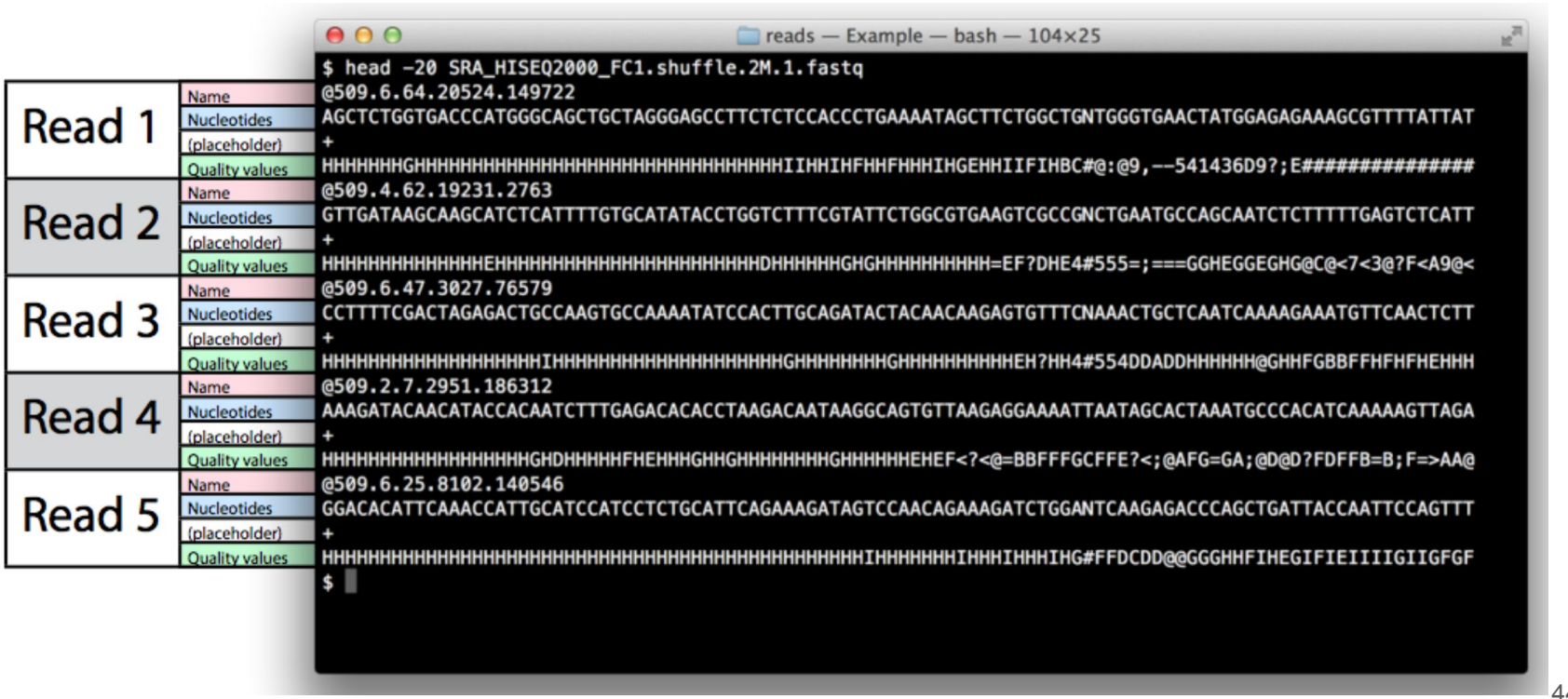
Sequencing data - FASTQ file

4 lines for each read

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



Sequencing data - FASTQ file



Genome reconstruction

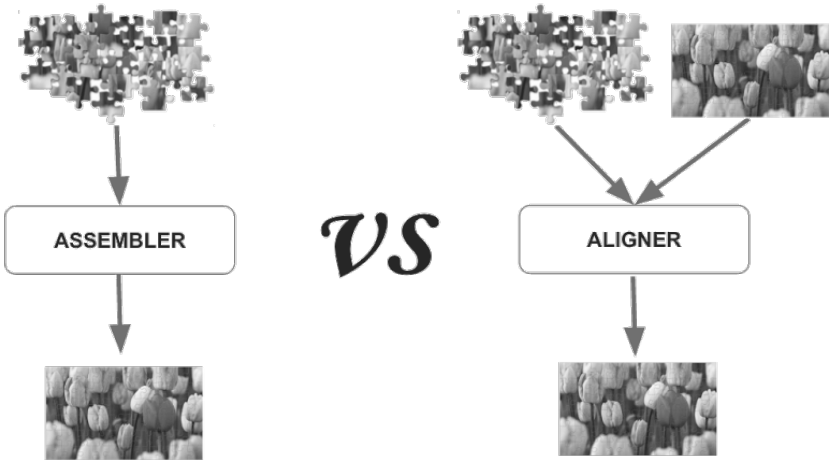
Result of sequencing experiment

- 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



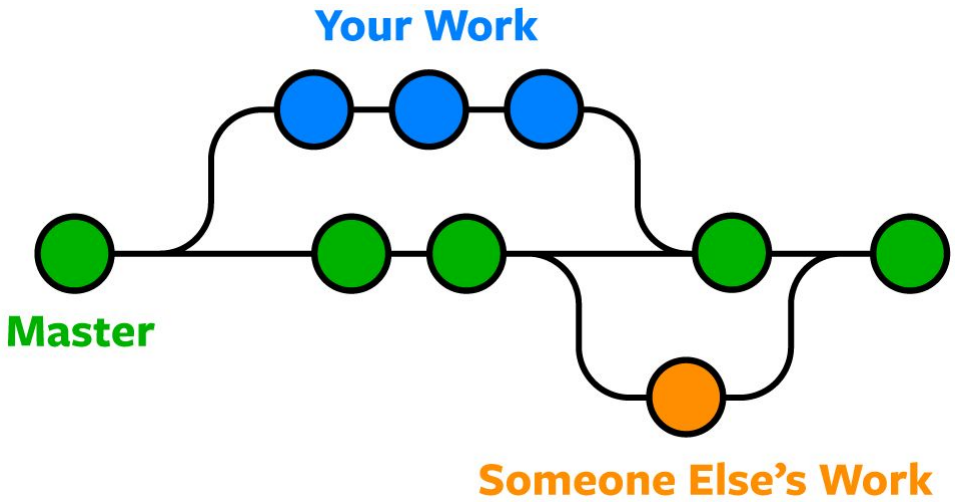
Alignment


AAGGACAAGA
TCTTTTTTATG
ATGACCAC
GAATGCAAGG
CCACATCTTT
ATGATTTAGA

Assembly


AAGGACAAGA
TCTTTTTATG
ATGACCAC
GAATGCAAGG
CCACATCTTT
ATGATTTAGA

Why Git?



We use Git!

- Created by Linus Torvalds, creator of Linux, in 2005
- Came out of Linux development community
- Designed to do version control on Linux kernel
- Goals of Git:
  - Speed
  - Support for non-linear development (thousands of parallel branches)
  - Fully distributed
  - Able to handle large projects efficiently

(A "git" is a cranky old man. Linus meant himself.)

- Instructions to install Git: https://git-scm.com/book/en/v2/Getting-Started-Installing-Git

Installing/learning Git!

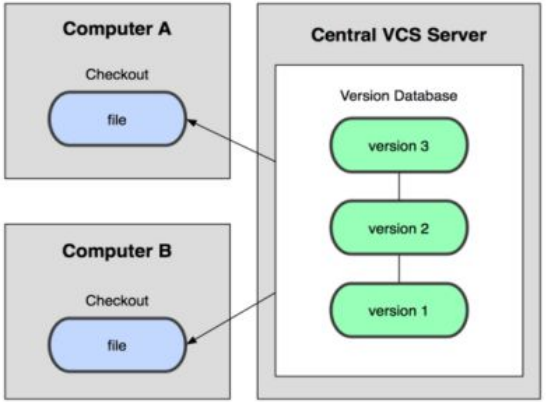
- Git website: http://git-scm.com/
- Free online book: http://git-scm.com/book
- Reference page for Git: http://gitref.org/index.html
- Git tutorial: http://schacon.github.com/git/gittutorial.html
- Git slides: https://courses.cs.washington.edu/courses/cse403/13au/lectures/git.ppt.pdf
- Git for Computer Scientists: http://eagain.net/articles/git-for-computer-scientists
- At command line: (where verb = config, add, commit, etc.)

```
git help verb
```

- Instructions to install Git: https://git-scm.com/book/en/v2/Getting-Started-Installing-Git

Centralized Versioning Control System

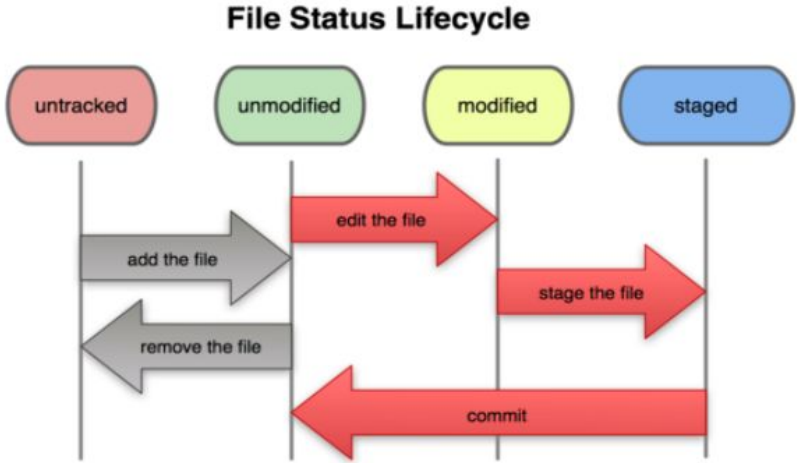
- A central server repository (repo)
  holds the "official copy" of the code
- The server maintains the sole version history of the repo
- You make "checkouts" of it to your local copy
- You make local modifications
- Your changes are not versioned
- When you're done, you "check in" back to the server
- your check in increments the repo's version



Basic Git flow

1. Modify files in your working directory
2. Stage files, adding snapshots of them to your staging area
3. Commit, which takes the files in the staging area
4. Store that snapshot permanently to your Git directory

```
git add file.py
git commit -m "Description of change."
git push origin master
```



Initial Git configuration

Set the name and email for Git to use when you commit:

- git config --global user.name "Bugs Bunny"
- git config --global user.email bugs@gmail.com

You can call git config –list to verify these are set.

Git commands

command	description
`git clone url [dir]`	copy a Git repository so you can add to it
`git add file`	adds file contents to the staging area
`git commit`	records a snapshot of the staging area
`git status`	view the status of your files in the working directory and staging area
`git diff`	shows diff of what is staged and what is modified but unstaged
`git help [command]`	get help info about a particular command
`git pull`	fetch from a remote repo and try to merge into the current branch
`git push`	push your new branches and data to a remote repository
`git checkout filename`	undoes your changes
Others: init, reset, branch, checkout, merge,log, tag	

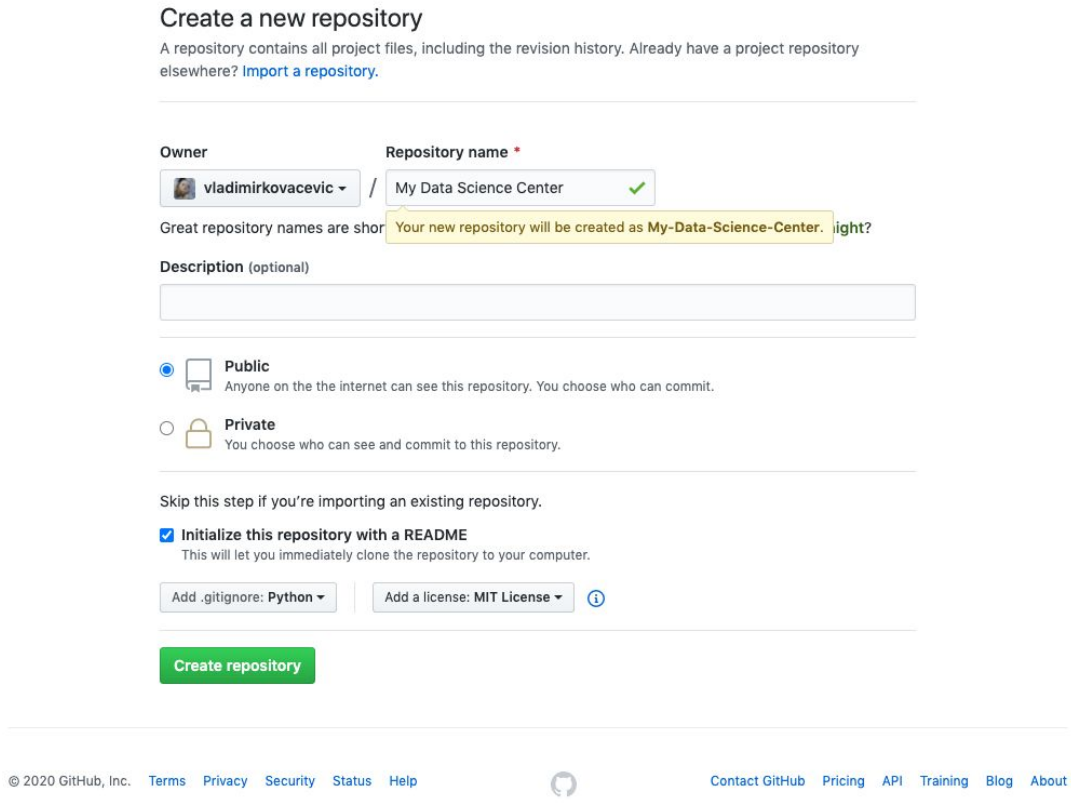
We use Github!

- GitHub.com is a site for online storage of Git repositories.
- You can create a remote repo there and push code to it.
- Many open source projects use it, such as the Linux kernel.
- You can get free space for open source projects, or you can pay for private projects.
- Free private repos for educational use: github.com/edu
- Question: Do I always have to use GitHub to use Git?
  - Answer: No! You can use Git locally for your own purposes.
  - Or you or someone else could set up a server to share files.
  - Or you could share a repo with users on the same file system, as long everyone has the needed file permissions).

Setup Github repo

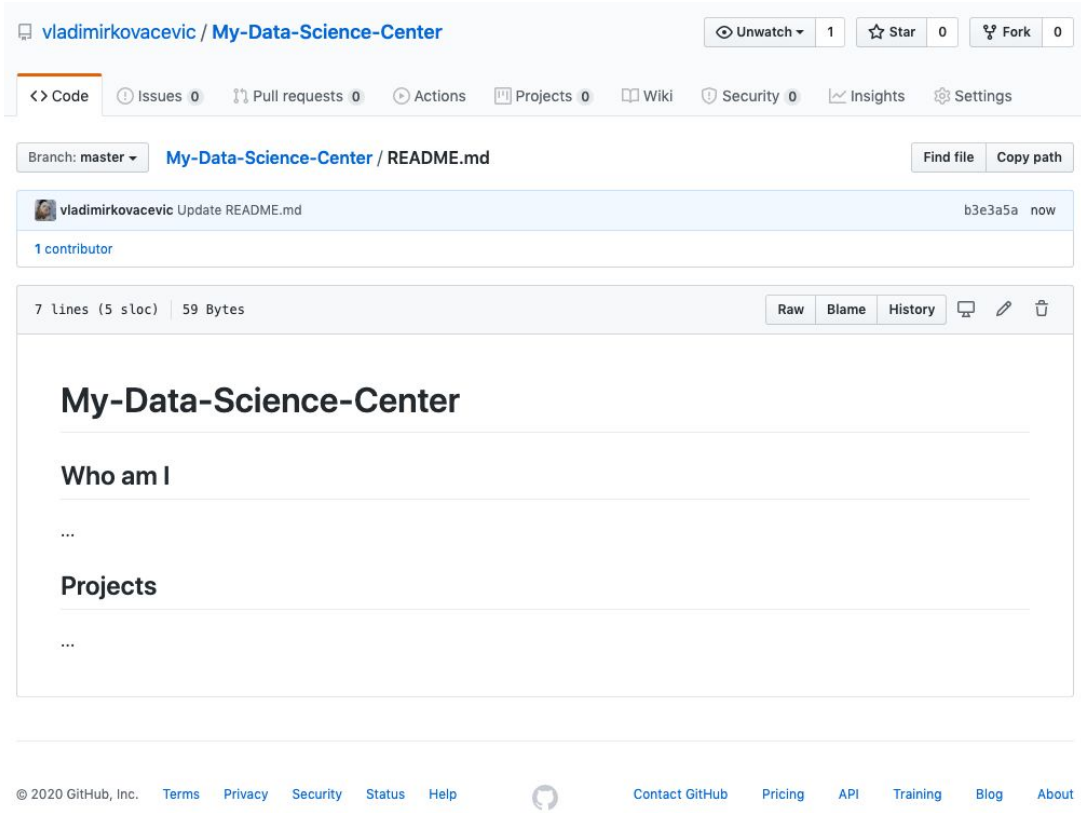
- Create account on www.github.com
- Set an image :)
- Create repository

  My Data Science Center
  - Initialize with README
  - .gitignore Python
  - MIT License



Setup Github repo

- Add short biography
- Projects will come on the way



Resources and additional reads

Presentation available at: github.com/vladimirkovacevic/gi-2023-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
- Vince Buffalo: Bioinformatics Data Skills
- Dan Gusfield: Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology, Cambridge
- Pavel Pevzner, Neils Jones: An Introduction to Bioinformatics Algorithms (Computational Molecular Biology), MIT
- 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
- Molekularna biologija 1; Dušanka Savić-Pavićević, Gordana Matić; NNK International, 2020
- RNA-seqlopedia; Cresko Lab, University of Oregon