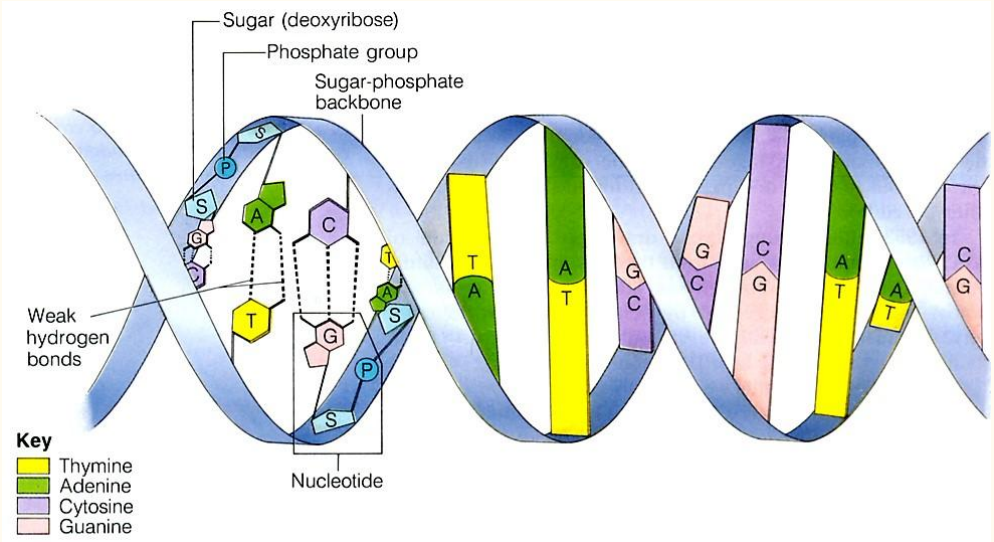
SevenBridges

Latest advances: DNA sequencing, bioinformatics and precision medicine

Vladimir Kovačević

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 pairs (complementary bases)
  - A - T (adenine and thymine)
  - C - G (cytosine and guanine)



DNA code

- 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%)

Karyotype

Central dogma of molecular biology

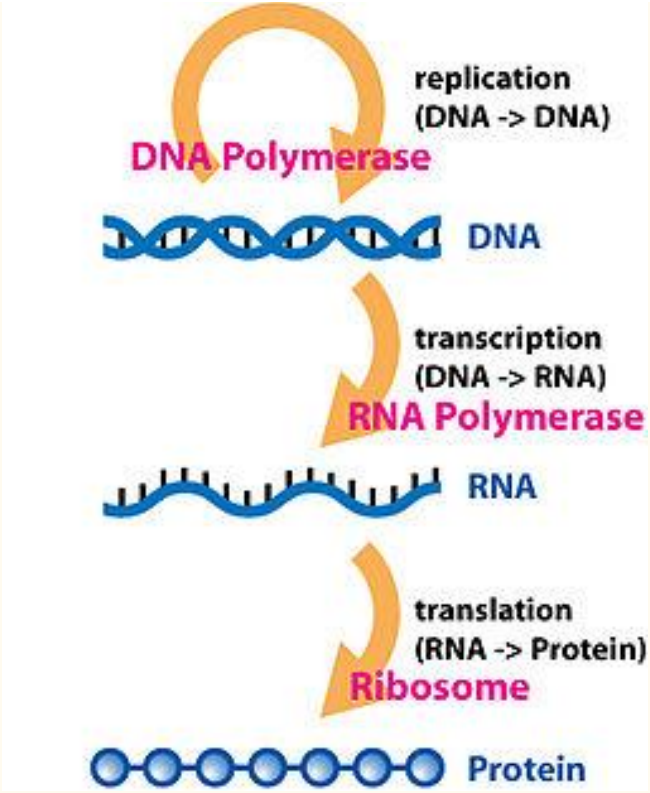
DNA ----> RNA -----> Protein

Transcription: DNA ->RNA

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

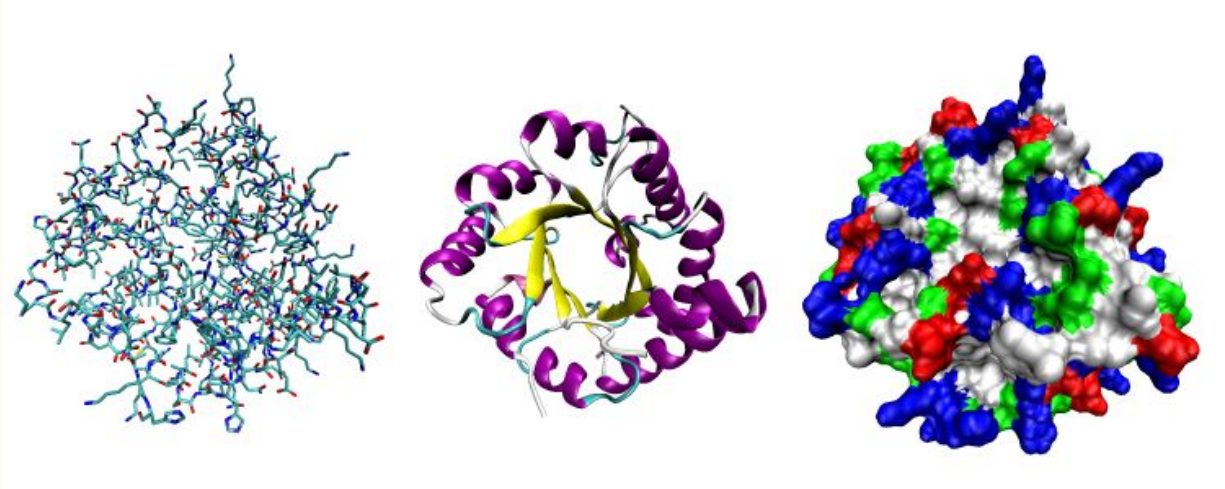
Translation: RNA -> Protein

- ribosomes synthesize proteins using RNA pattern



Proteins

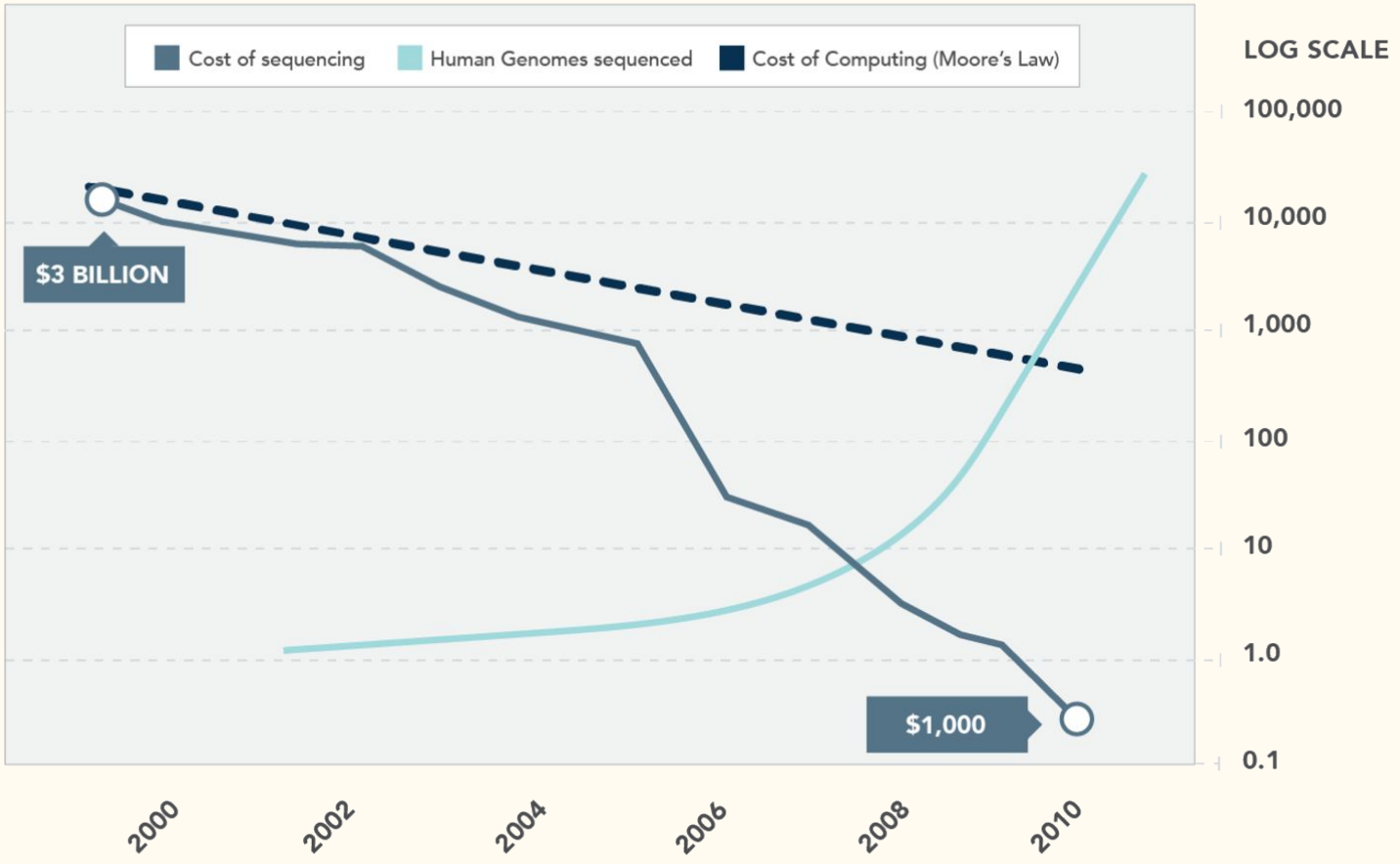
- Building blocks of life
  - Various functions in the organism (transportation, regulation, metabolism, DNA replication)
- Long chains of amino-acids, that also fold into complicated 3D structures
  - We often distinguish protein primary, secondary, tertiary and quaternary structure



Genome sequencing

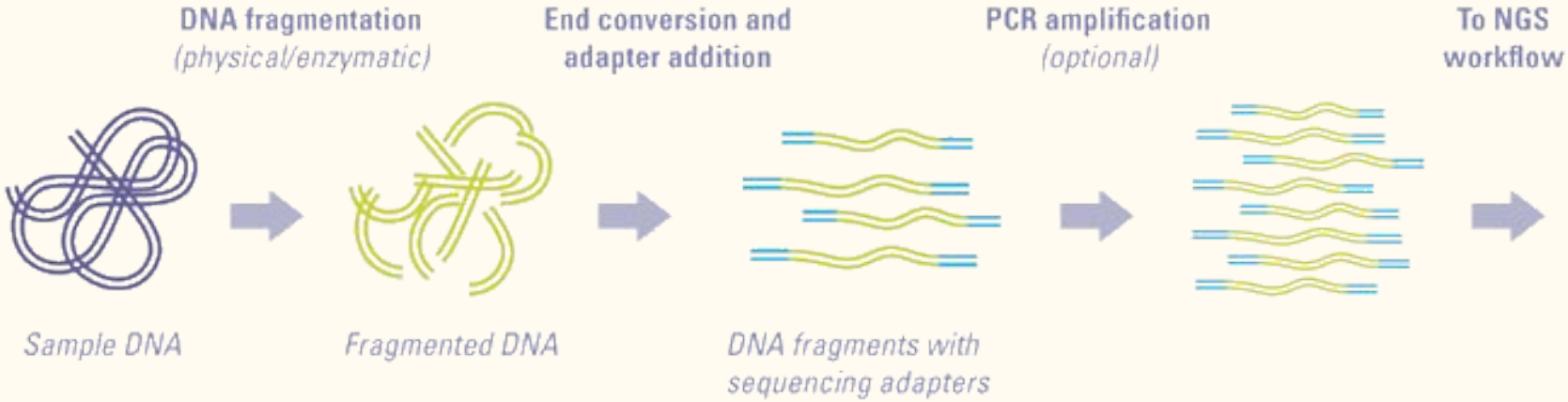
- Digitalization of genome
- **Human Genome Project** (1990-2003), 3B $
- Sanger sequencing (First generation sequencing)
  - Long (took 13 years)
  - Costly (3B$ for one human genome)
- Currently NGS (next generation sequencing)
  - Illumina
  - Around 200$ and 1 day needed to sequence the genome
- 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



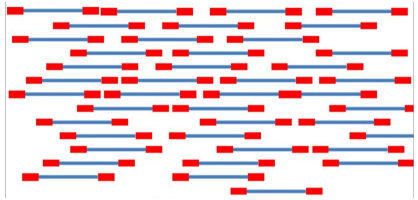
NGS sequencing

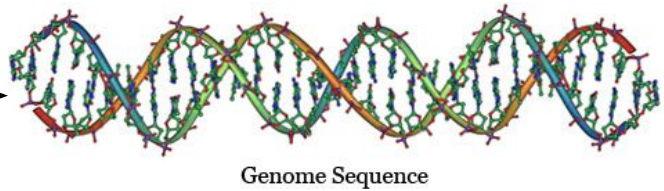
- Read - DNA fragment after reading it in sequencer
- Typical whole genome sequencing experiment:
  - 200-500 million reads
  - 50-150 bases (letters long)



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

Genome reconstruction

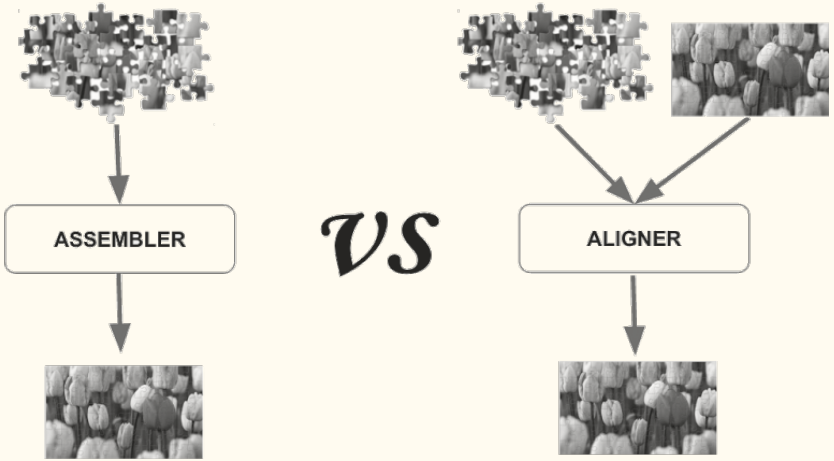
Result of sequencing experiment

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

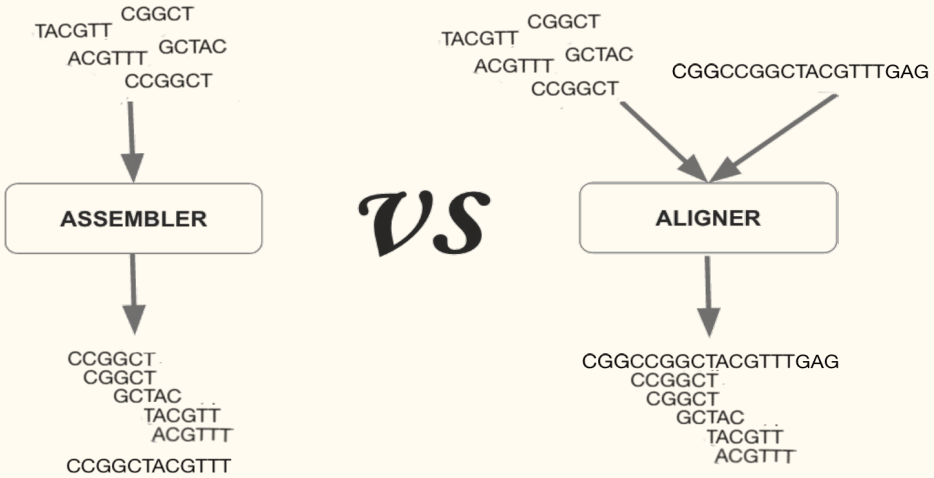
Genome reconstruction

How do we reconstruct genome from reads?

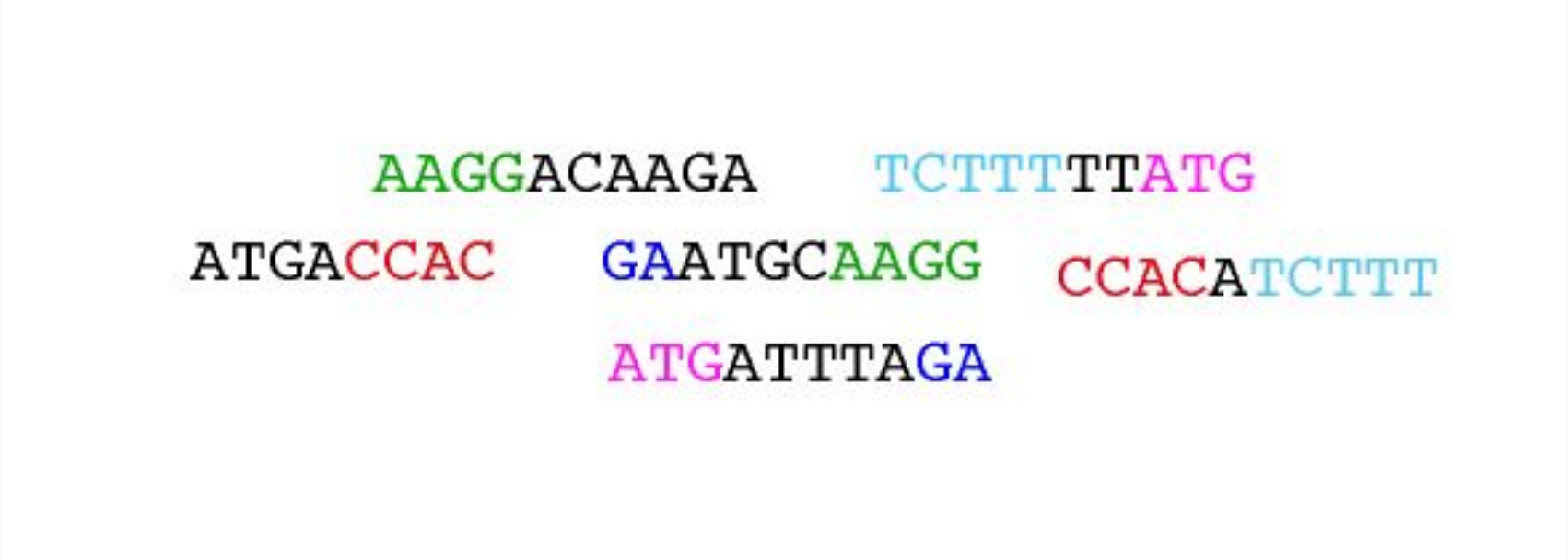
1. Alignment
    - Using reference genome to map the position of the reads (we share 99.9% of DNA)
2. Assembly
    - Reconstructing the genome by finding the links between the reads



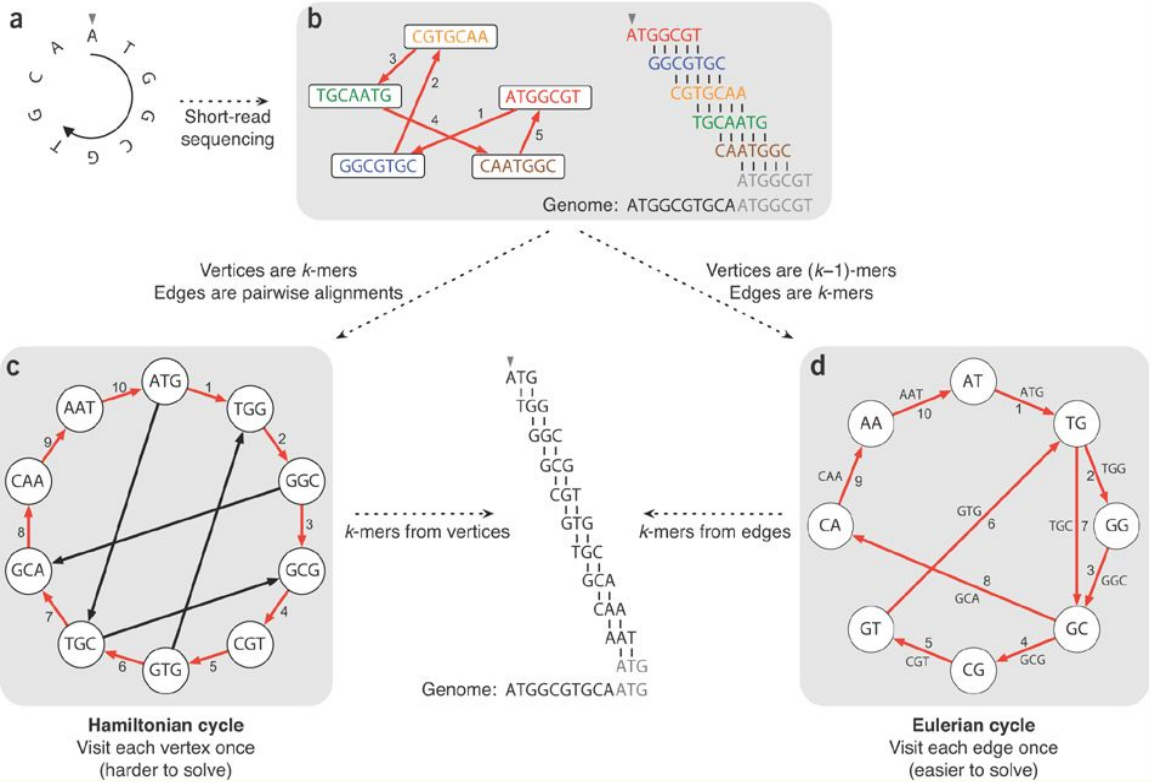
Genome reconstruction



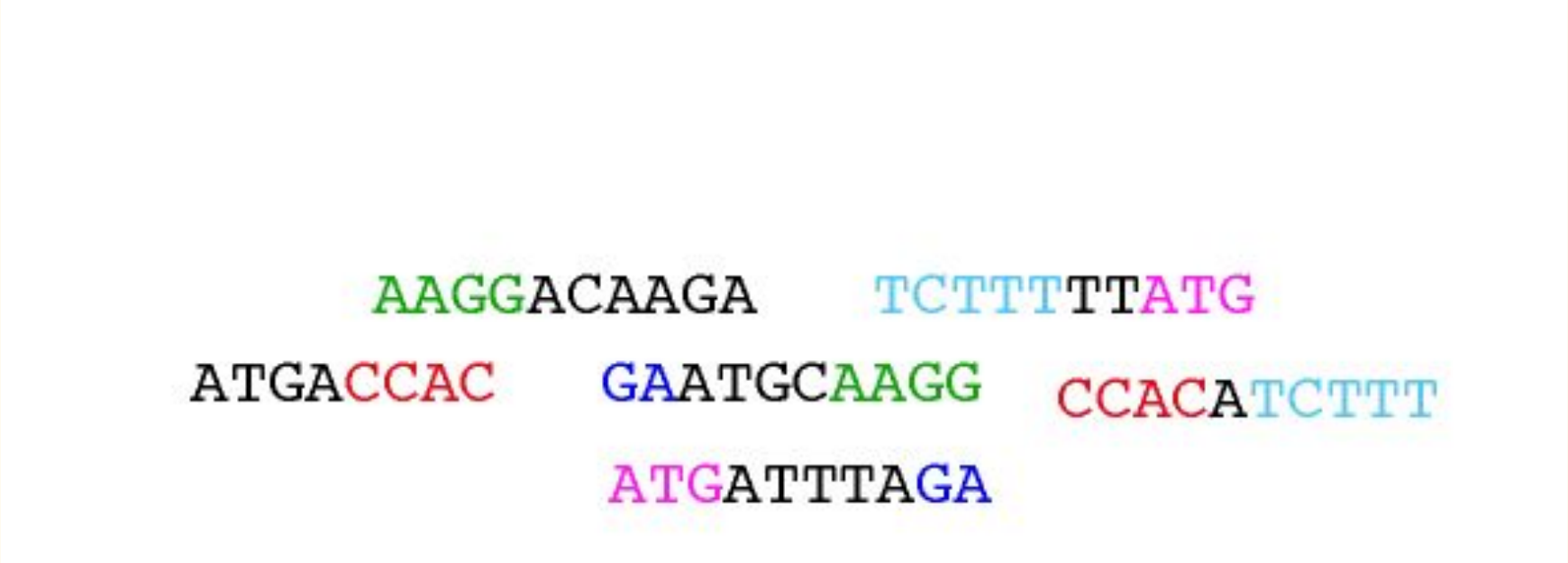
Assembly


AAGGACAAGA
TCTTTTTATG
ATGACCAC
GAATGCAAGG
CCACATCTTT
ATGATTTAGA

Assembly


a
Short-read sequencing
b
CGTGCAA
TGCAATG
ATGGCGT
GGCGTGC
CAATGGC
Genome: ATGGCGTGCAATGGCGT
Vertices are k-mers
Edges are pairwise alignments
Vertices are (k–1)-mers
Edges are k-mers
c
ATG
TGG
GGC
GCG
CGT
GTG
TGC
GCA
CAA
AAT
k-mers from vertices
k-mers from edges
Genome: ATGGCGTGCAATG
d
AT
TG
GG
GC
CG
GT
CA
AA
Hamiltonian cycle
Visit each vertex once
(harder to solve)
Eulerian cycle
Visit each edge once
(easier to solve)

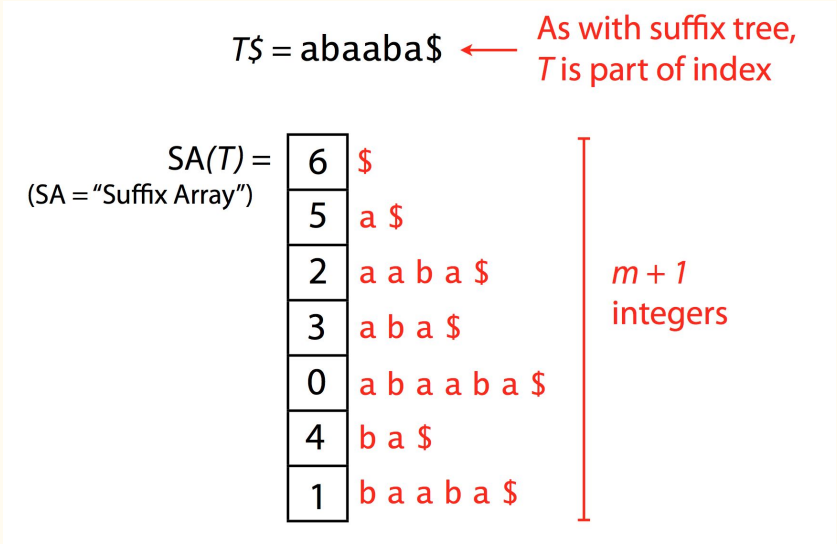
Alignment


AAGGACAAGA
TCTTTTTTATG
ATGACCAC
GAATGCAAGG
CCACATCTTT
ATGATTTAGA

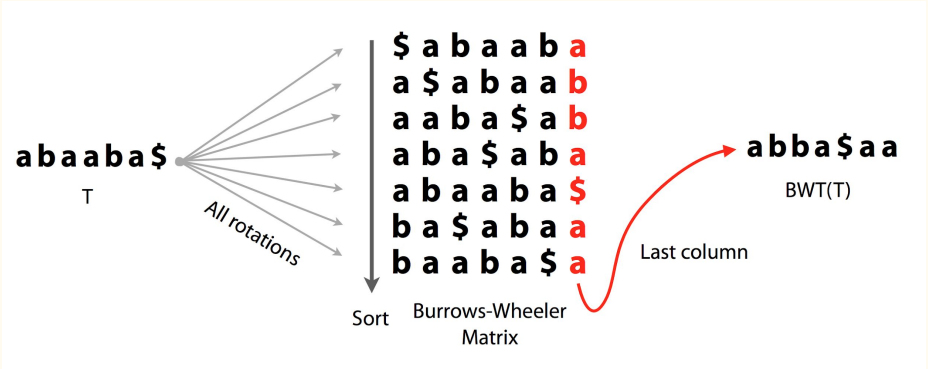
Alignment

- Use indexing structures - fast sequence search
- Able to align whole genome sample with 880 million reads against 3-billion long reference genome in 3 hours (36 CPUs, 60Gb memory)

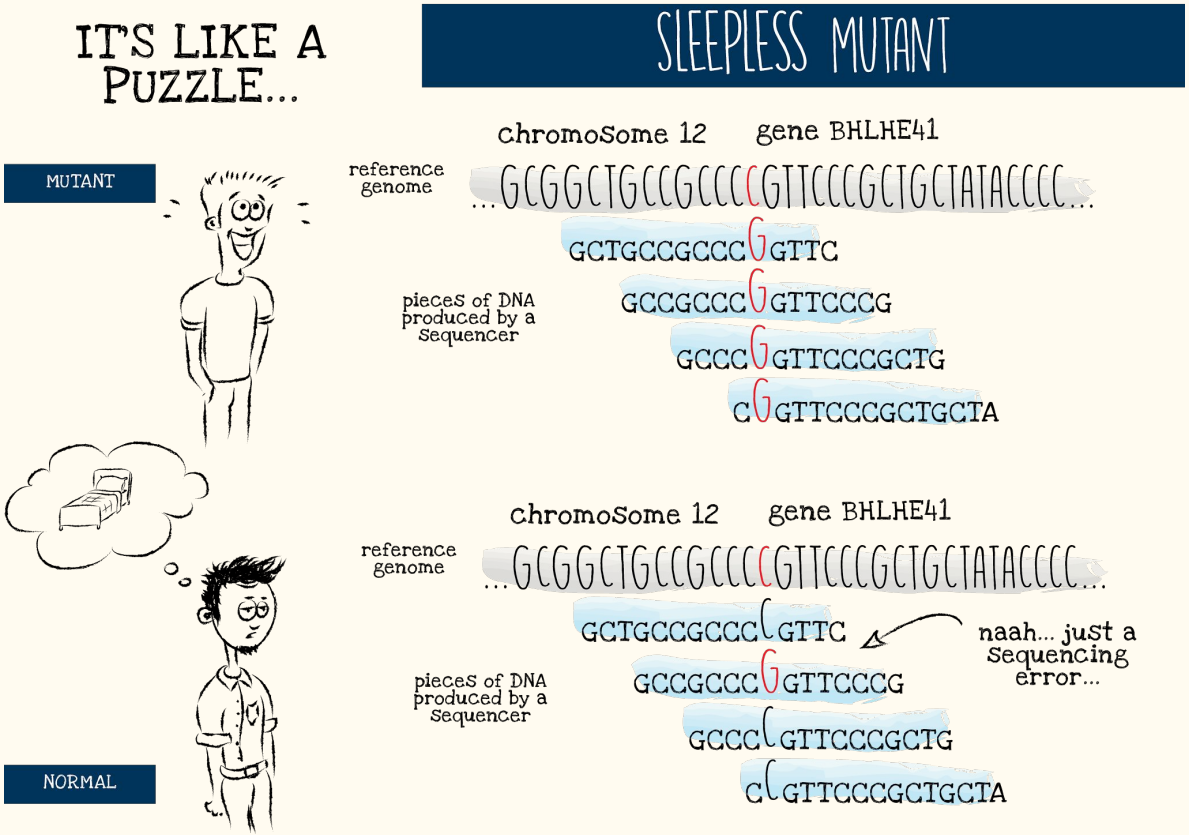
Suffix Array



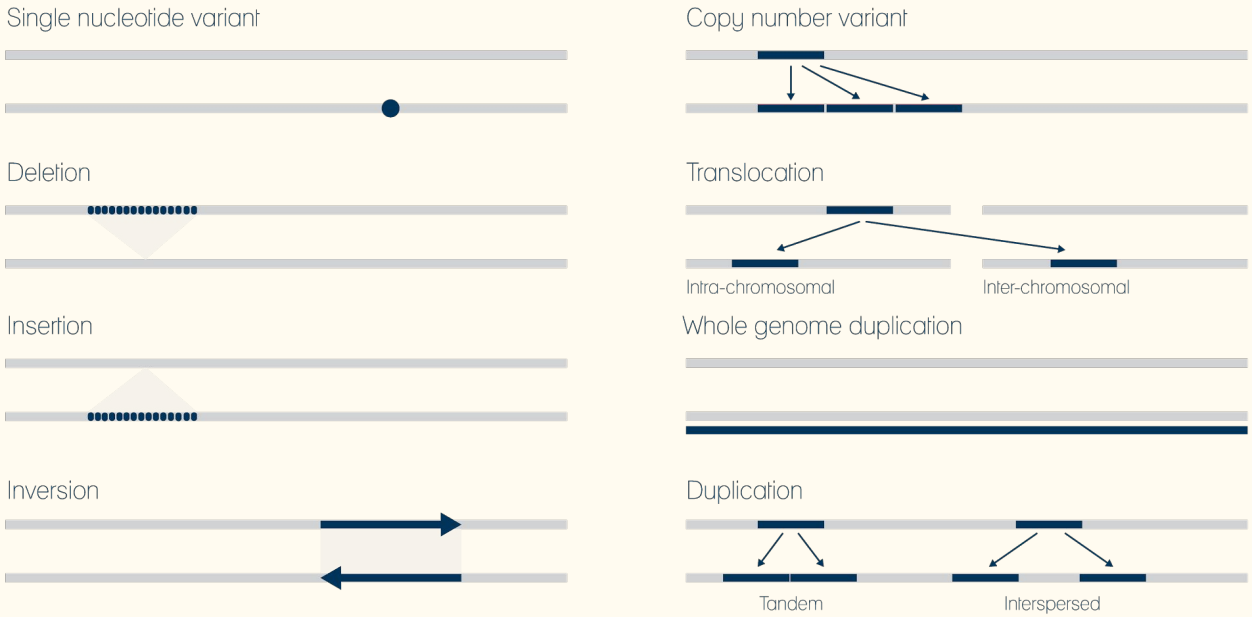
Burrows Wheeler transformation



Variants (mutations)



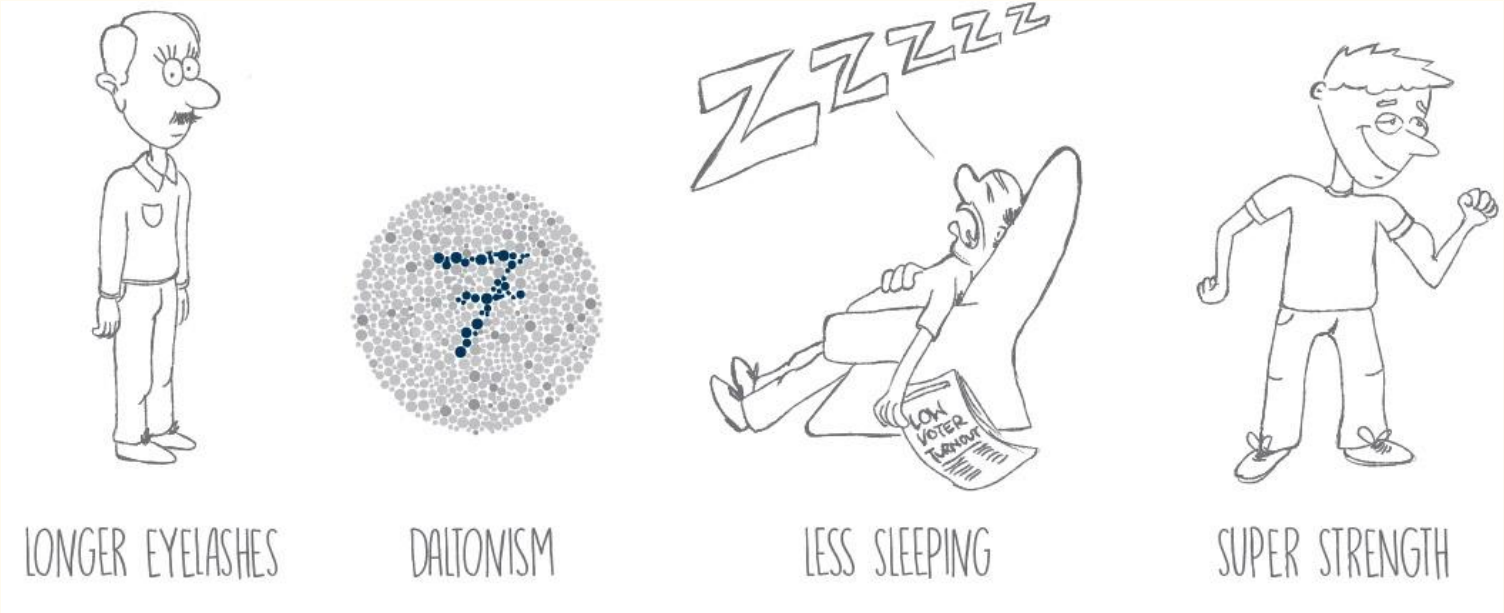
Genomic Variants (mutations)



We share ~99.9% of DNA, 99% with chimps, 80% with mouse, 50% with banana

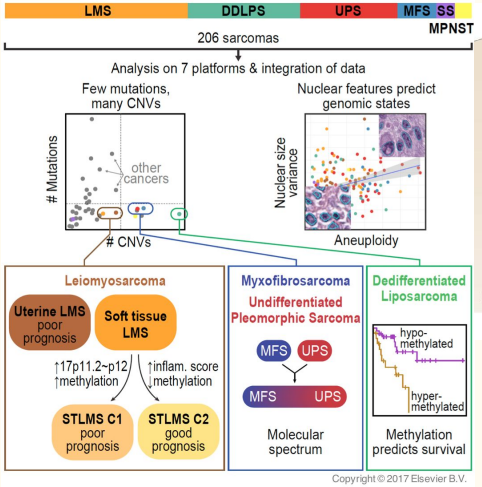
Genomic Variants

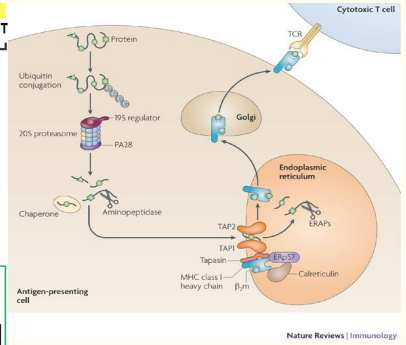
Each of those characteristics is caused by one Single Nucleotide Variant

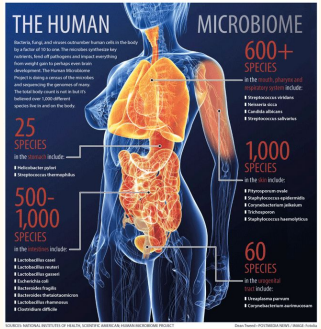


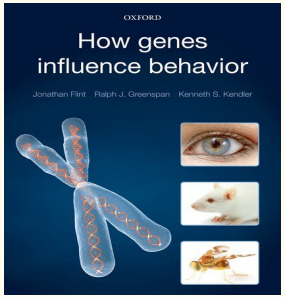
Why perform DNA sequencing?

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







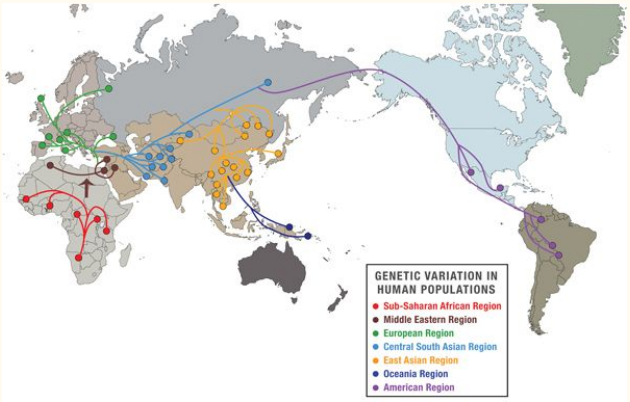



MEDICAL DISPATCH JULY 21, 2014 ISSUE

ONE OF A KIND

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

By Seth Mnookin



Precision medicine

“Precision medicine is a medical model that proposes the customization of healthcare, with medical decisions, treatments, practices, or products being tailored to the individual patient.”

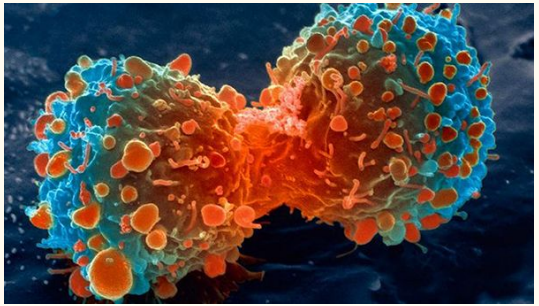
What is cancer?

Mutation during DNA replication can fall to:

1. Intron (no change)
2. Important gene (cell dies, organism lives)
3. Gene that stops cell division (cell lives, organism...)

What causes cancer (increases probability of mutation)?

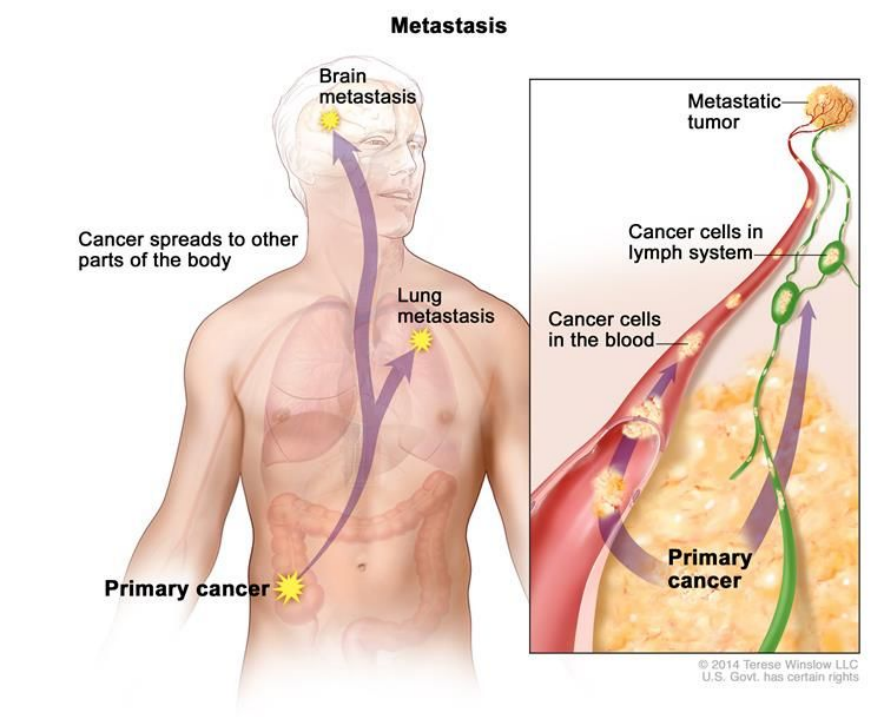
1. EM radiation
2. Chemical agents
3. Free radicals
4. Genetic factors
5. Infections (viruses)



A dividing lung cancer cell.
Credit: National Institutes of Health

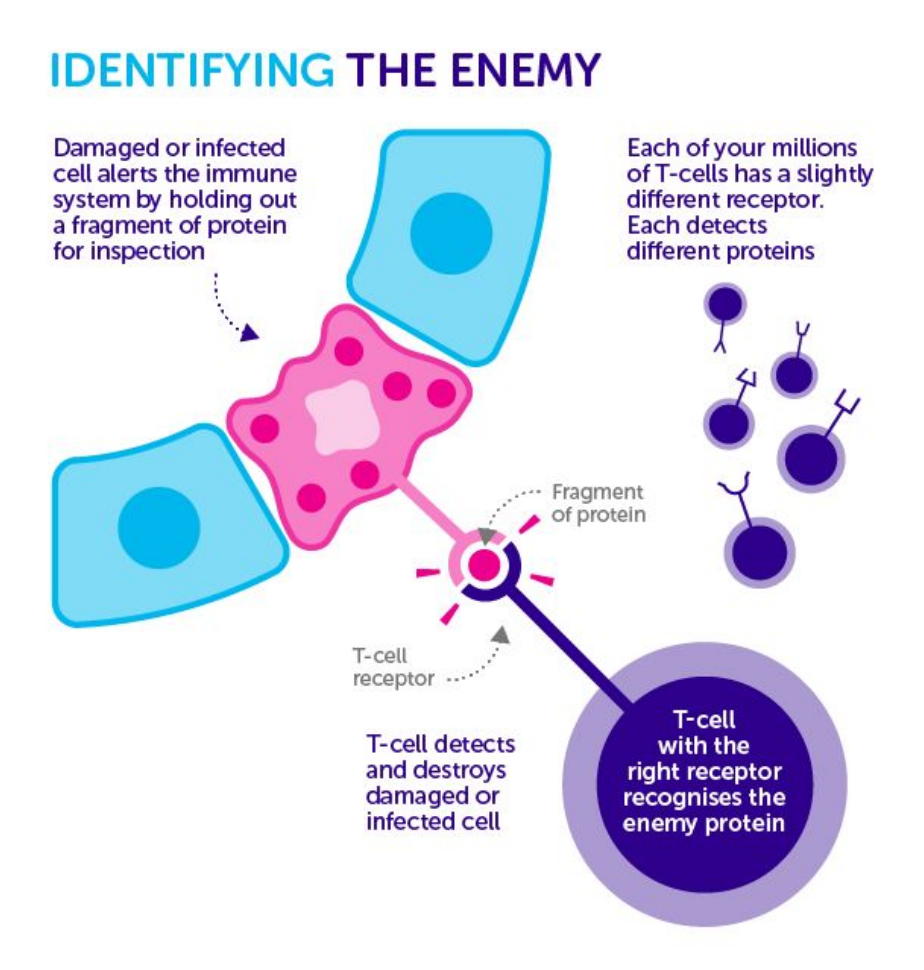
What is metastasis?

Body's cells begin to divide without stopping and spread into surrounding tissues
Cancer cells - ignore signals that normally tell cells to stop dividing or that begin a process known as programmed cell death, or **apoptosis**, which the body uses to get rid of unneeded cells



Cancer cells

Our body develops thousands cancer cells every day. OMG! OMG! OMG!

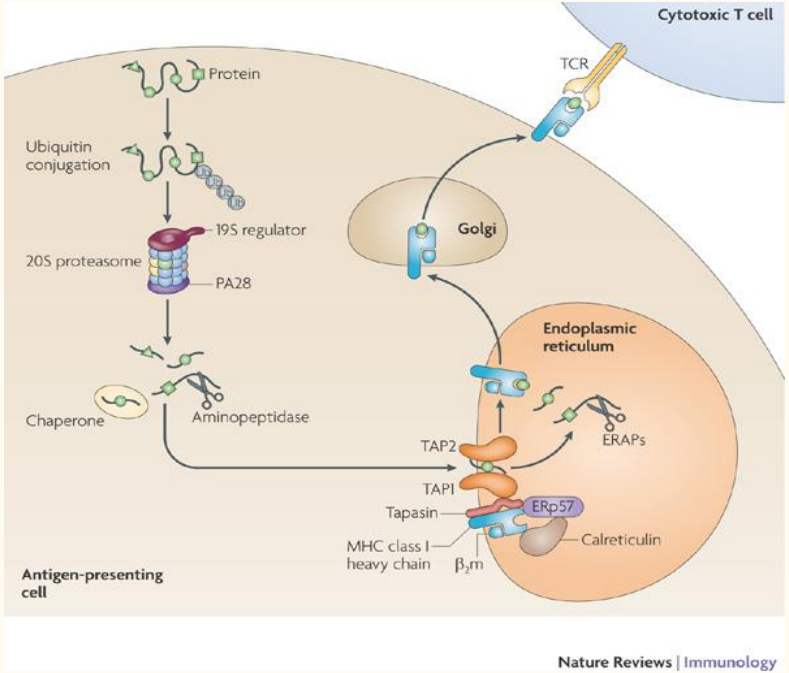


MHC Complex

MHC is a set of cell surface proteins essential for the acquired immune system to recognize foreign molecules (translated from HLA regions from the genome for humans)

MHC molecules bind to **protein fragments available in the cell**

MHC molecule with **antigen** (MHC complex) is "presented" outside of the cell to cytotoxic T cells and helper T cells

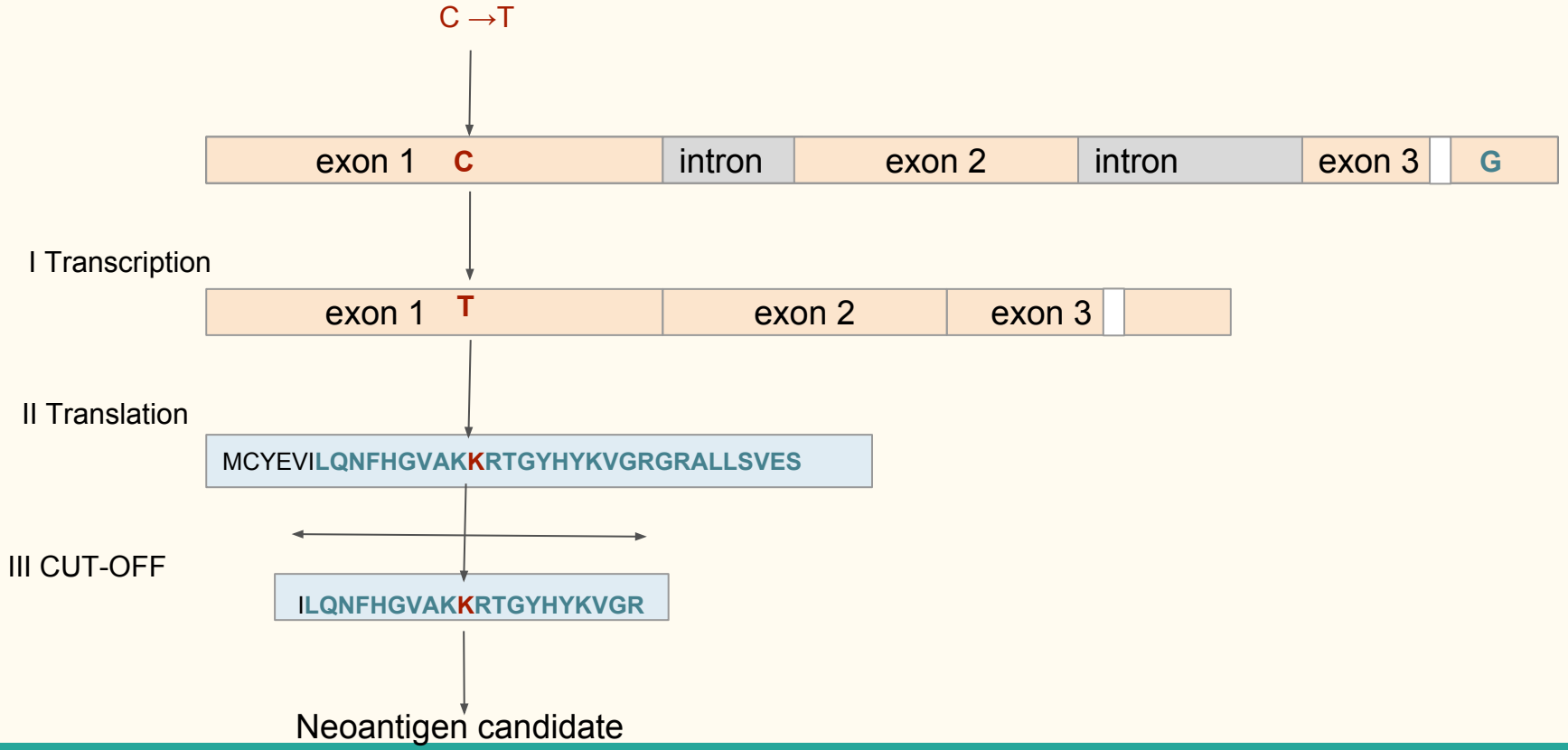


So, what can be done there?

1. Identify NEOANTIGENS - proteins presented only by cancer cells - precision medicine
2. “Program” T-cells to recognize neoantigens

Compare DNA from Tumor and Normal tissue
Mutations present in tumor - somatic mutations

From DNA somatic mutation to neoantigen



Neoantigen cancer vaccine

Two gene therapy drugs obtained FDA approval:

- Novartis - 83% of patients complete or partial remission
- Advaxis - multiple neoantigens presented to immune system

Cons of immunotherapy

- Autoimmune disease
- Very expensive

Questions?