# Genetic Data: Machine Learning meets NGS clinical data

São Paulo School of Advanced Science

Learning from data

August 2019

João Carlos Setubal, Murilo Cervato & team

# This is a hands-on course

- The course is provided by the team from Albert Einstein Hospital in São Paulo led by Murilo Cervato
- There will be several members of his team helping you out

ALBERT EINSTEIN
SOCIEDADE BENEFICENTE ISRAELITA BRASILEIRA

# You will be using a proprietary platform called Varstation developed by this group

- The platform was the most convenient way to present this course to you
- This course is not a tutorial on how to use this platform
- I am not part of the team that developed this platform

varstation

# Program

## August 8th

08:00 - 08:30 - Introduction

08:30 - 09:00 - Bioinformatics basics (from raw sequencing data to annotated variants)

09:00 - 09:30 - Bioinformatics live demo

09:30 - 10:00 - Clinical interpretation of hereditary variants

10:00 - 10:30 - Varstation intro: a tool for NGS analyses

10:30 - 12:00 - Exploratory data analysis of whole-exome sequencing samples: 2 case studies

## August 9th

10:00 - 10:30 - Machine learning basics and classification problem definition

10:30 - 11:00 - Modeling and predictions for whole-exome sequencing samples: 2 case studies

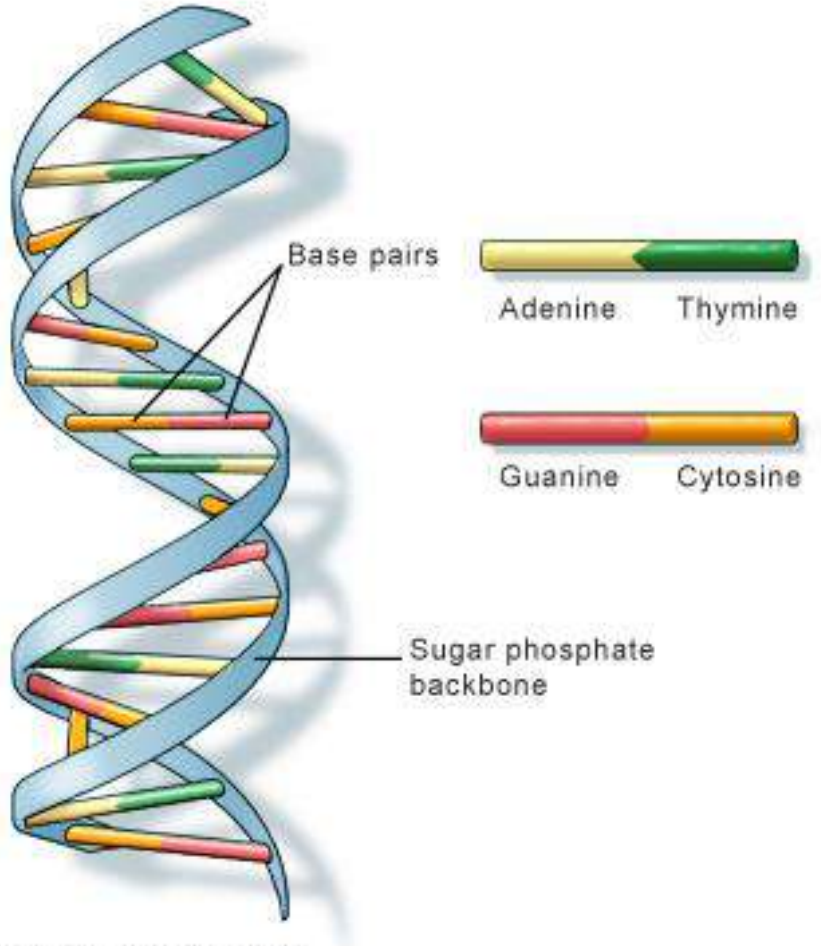
11:00 - 11:30 - Filtering and analyzing on Varstation

11:30 - 11:45 - Can NGS analyses be automatized?

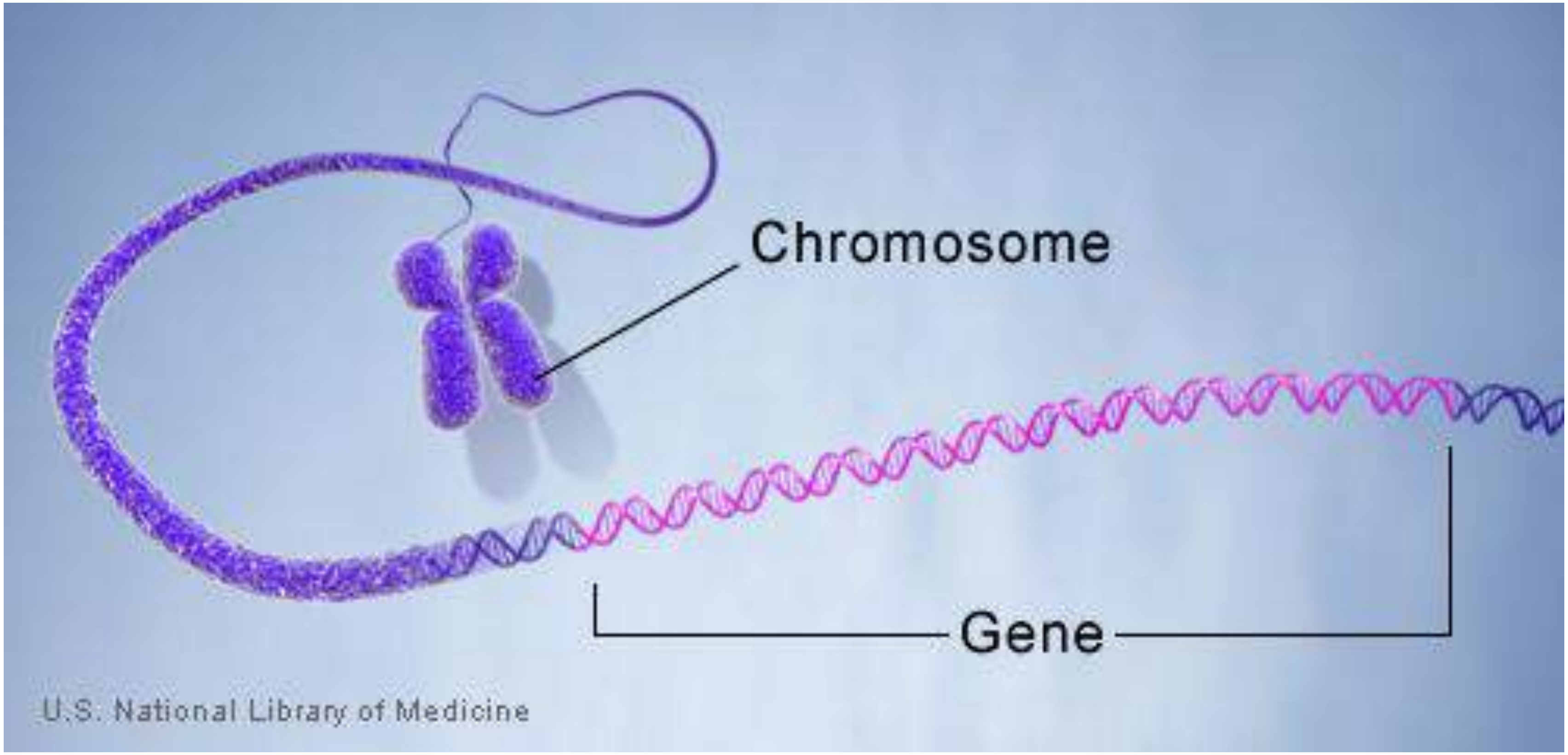
11:45 - 12:00 - Final comments and questions

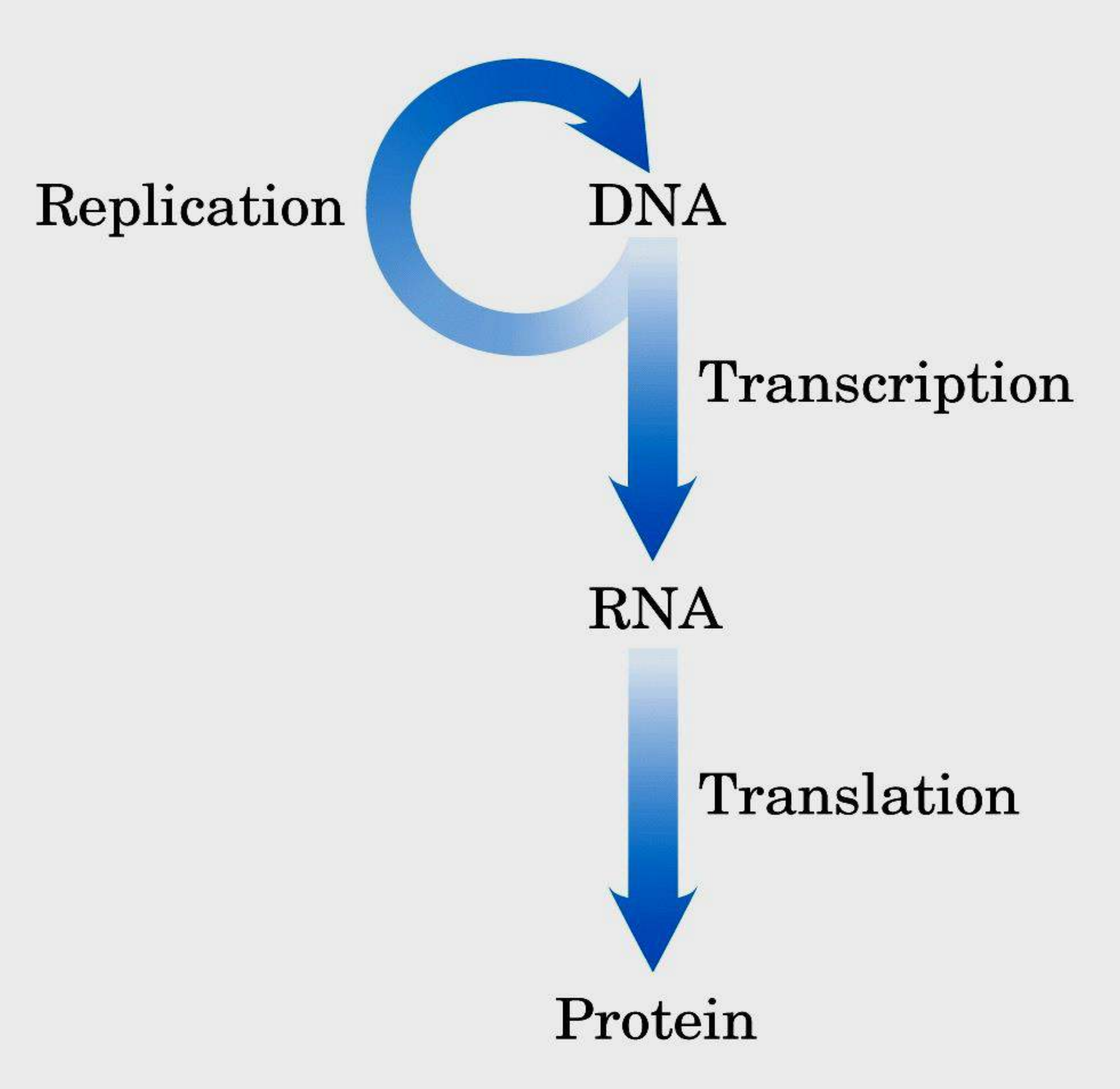
# Quick overview of molecular biology concepts

# DNA double helix



U.S. National Library of Medicine


Chromosome
Gene
U.S. National Library of Medicine



Replication

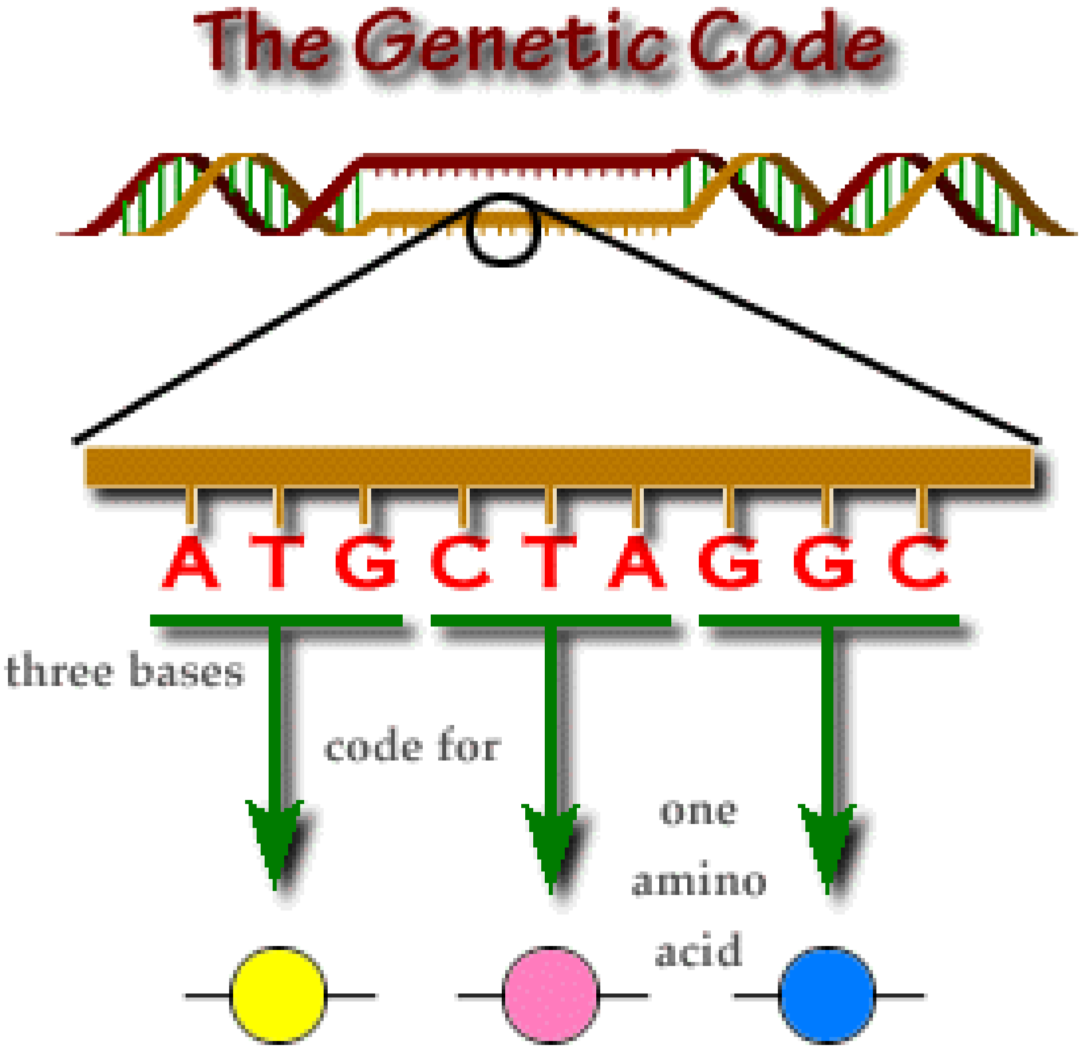
DNA

Transcription

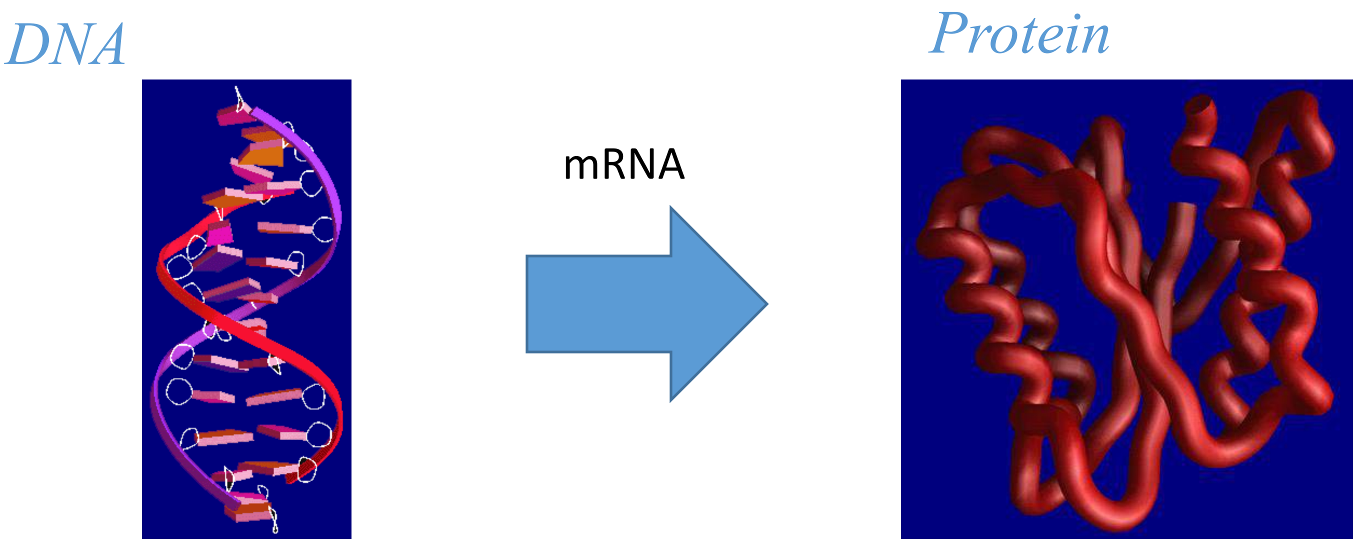
RNA

Translation

Protein


The Genetic Code
A T G C T A G G C
three bases
code for
one
amino
acid

# Genes and proteins



```
> DNA sequence
ATGCATAAAATCGTATACTGGTCTGGTACCGGCAACAC
TGAGAAAACGGCAGAGCTCATCGCTAAAGGTATCATCGAA
TCTGGTAAAGACGTCAACACCATCAACGTGTCTGACGTTA
ACATCGATGAACTGCTGAACGAAGATATCCTGATCCTGGG
TTGCTCTGCCATGGGCGATGAAGTTCTCGAGGAAAGCGAA
TTTGAACCGTTCATCGAAGAGATCTCTACCAAAATCTCTG
GTAAGAAGGTTGCGCTGTTCGGTTCTTACGGTTGGGGCGA
CGGTAAGTGGATGCGTGACTTCGAAGAACGTATGAACGGC
TACGGTTGCGTTGTTGTTGAGACCCCGCTGATCGTTCAGA
ACGAGCCGGACGAAGCTGAGCAGGACTGCATCGAATTTGG
TAAGAAGATCGCGAACATCTAGTAGA
```

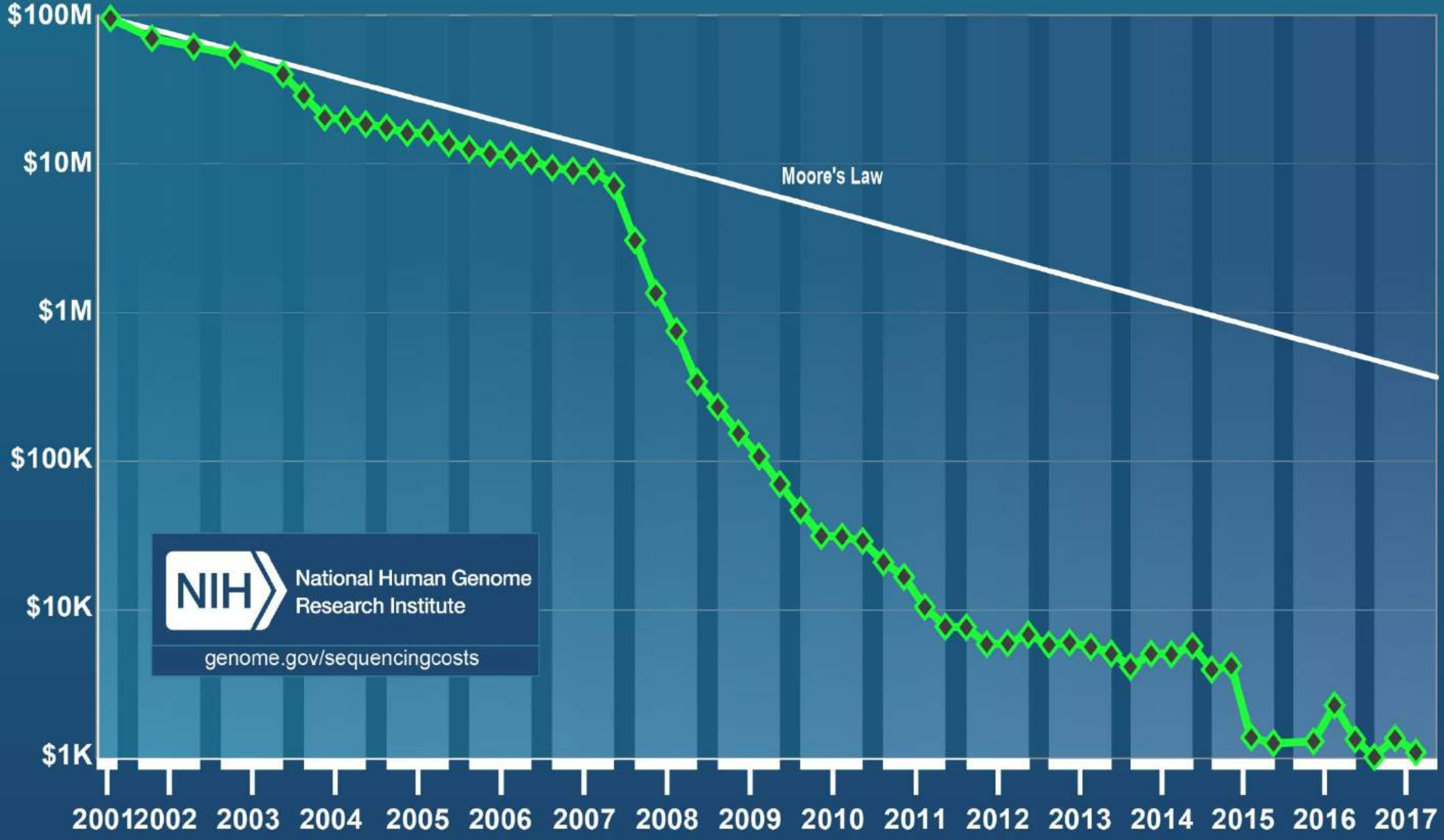
```
> Protein sequence
MHKIVYWSGTGNTEKTAELIAKGIIESGKDVNT
INVSDVNIDELLNEDILILGCSAMGDEVLEESE
FEPFIEEISTKISGKKVALFGSYGWGDGKWMRD
FEERMNGYGCVVVETPLIVQNEPDEAEQDCIEF
GKKIANI
```

Hugues Sicotte, NCBI

# DNA sequencing as applied to the human genome

- The human genome has about 3 x $10^9$ bp
    - and about 20,000 genes
- First version was made available in 2001
- Now there are thousands of human genomes available
    - and very soon there will be millions

# Cost per Genome



NIH National Human Genome Research Institute

genome.gov/sequencingcosts

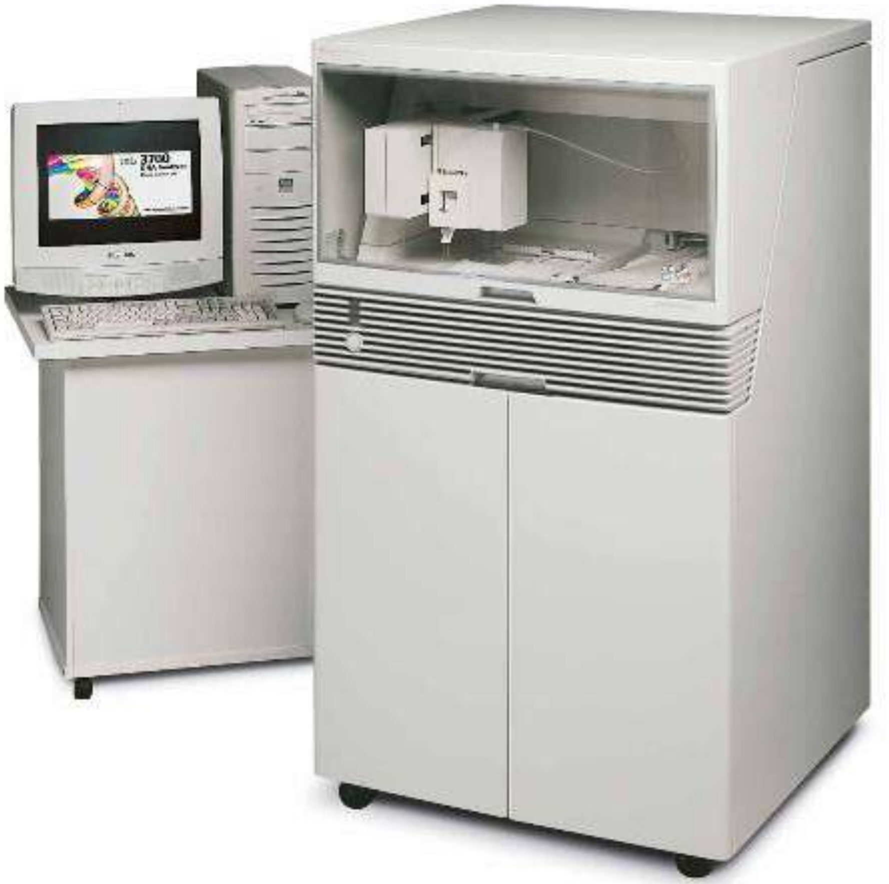
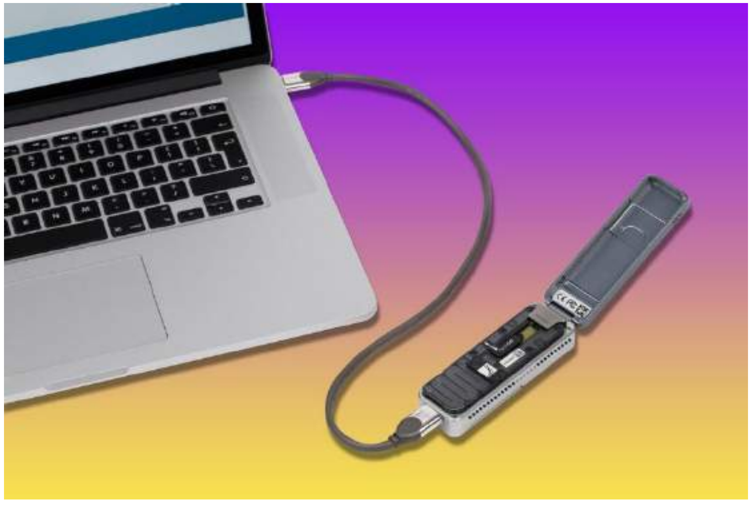
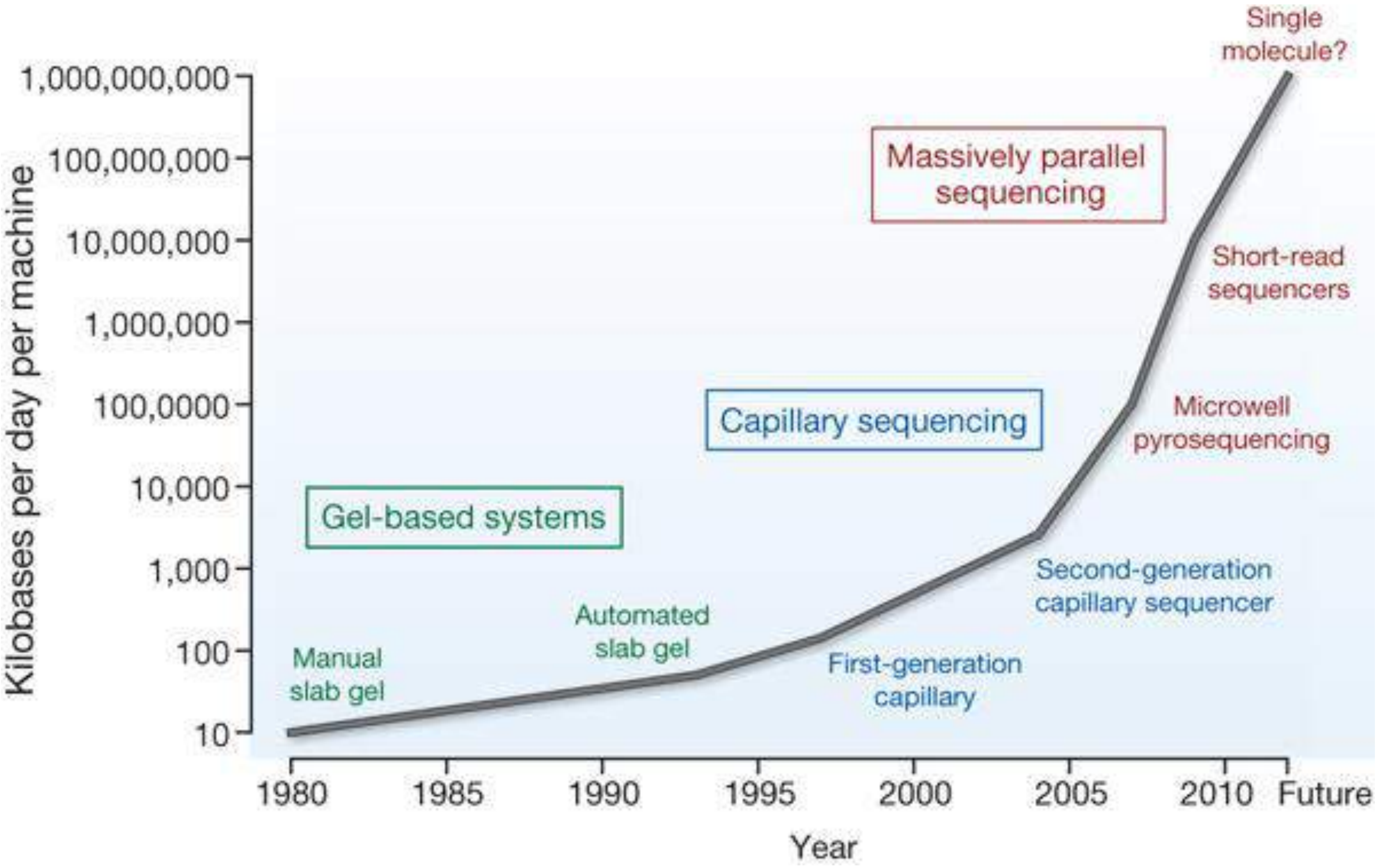


*Figure 3* ABI PRISM 3700 DNA *analyzer.*

Oxford Nanopore minION



# Improvements in the rate of DNA sequencing over the past 30 years and into the future



MR Stratton *et al.* *Nature* **458**, 719-724 (2009) doi:10.1038/nature07943

The UK's Department of Health and Social Care has announced its plans to sequence five million [human] genomes in the UK over the next five years

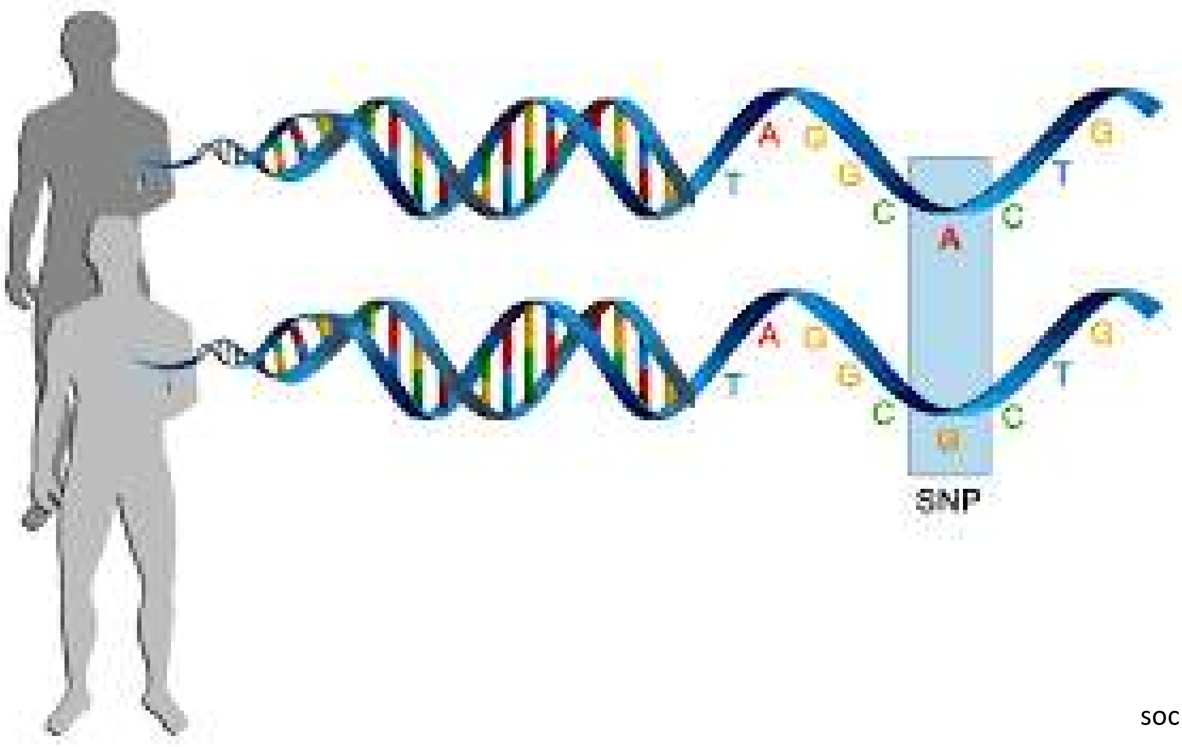
October 2018

# Except for identical twins, any pair of genomes will be different

- On average, any two random genomes will have 1 bp difference every ~1,000 bp
- our genomes are about 99.9% similar

# Genome differences between individuals or populations

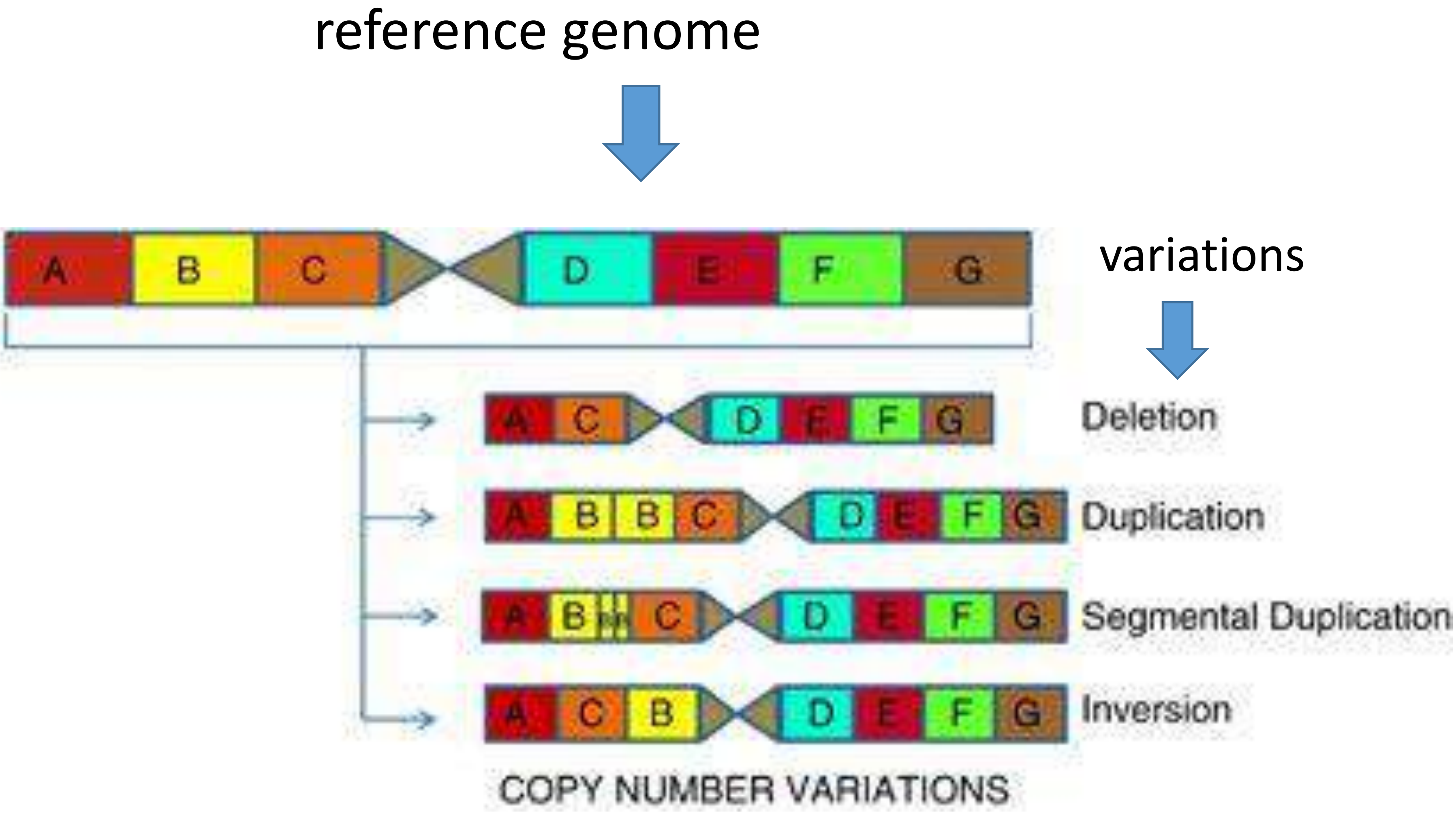
- Polymorphisms
  - single nucleotide polymorphisms (SNPs)



socratic.org

# Other kinds of polymorphisms

- Indels
  - small insertions or deletions
- Copy number variants (CNVs)
- rearrangements
  - inversions
  - translocations



reference genome

variations

Almal and Padh, 2012

# Why are SNPs and other variations so important?

- many of them are related to health and disease

# Sickle-cell anemia

- Nonsynonymous mutation in one position of the hemoglobin gene (aminoacid #7)
- GAA (glu) → GUA (val)
- GAG (glu) → GUG (val)
- Valine is hydrophobic and Glutamic acid is not

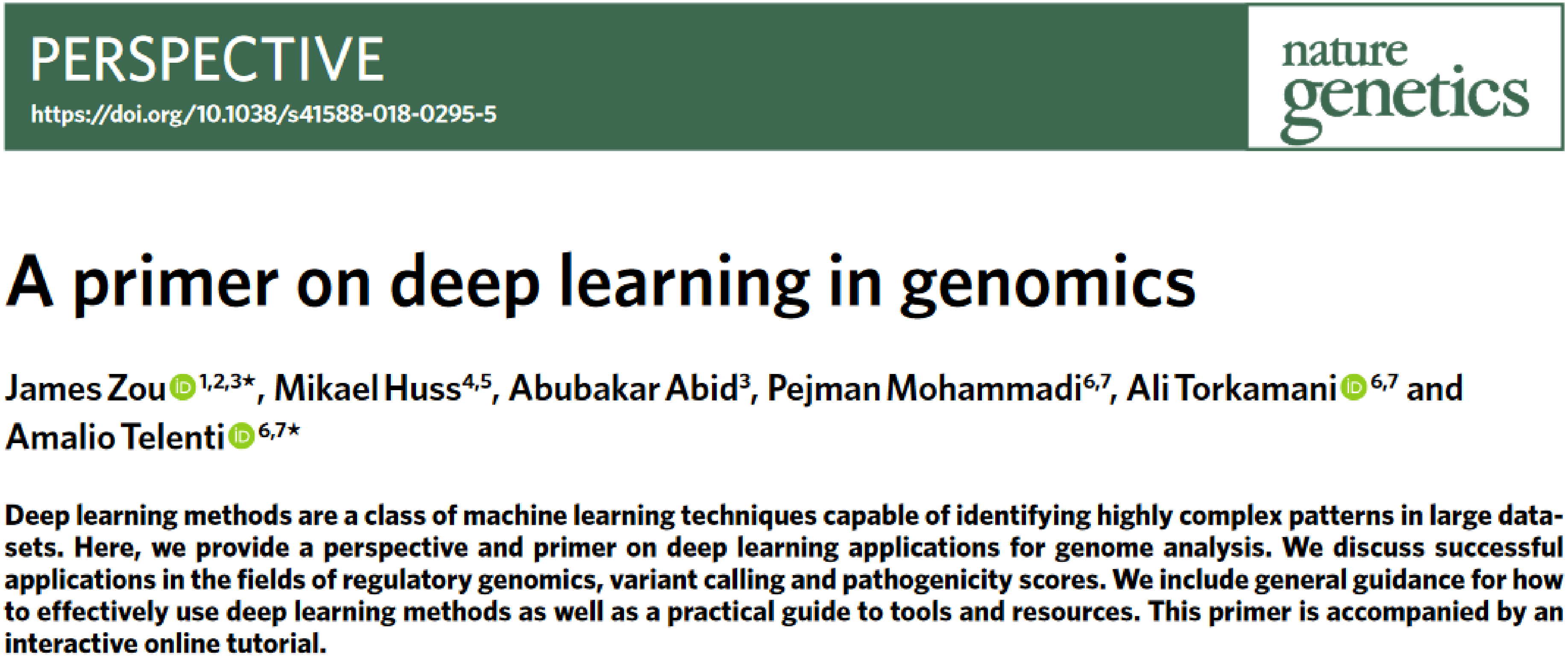
# Bioinformatics
# Computational Biology

- Essential to make sense of the flood of data coming out of the genomics revolution
- Is it any different from
  - Computational Chemistry?
  - Computational Astronomy?
  - etc?
- In a sense, it is not
  - BIG DATA
- But in another sense, it is
  - the genome can be seen as a digital information storage system

# Machine learning and bioinformatics

- This field has exploded in the past few years

November 2018



PERSPECTIVE

https://doi.org/10.1038/s41588-018-0295-5

nature genetics

# A primer on deep learning in genomics

James Zou[1,2,3*], Mikael Huss[4,5], Abubakar Abid[3], Pejman Mohammadi[6,7], Ali Torkamani[6,7] and Amalio Telenti[6,7*]

**Deep learning methods are a class of machine learning techniques capable of identifying highly complex patterns in large datasets. Here, we provide a perspective and primer on deep learning applications for genome analysis. We discuss successful applications in the fields of regulatory genomics, variant calling and pathogenicity scores. We include general guidance for how to effectively use deep learning methods as well as a practical guide to tools and resources. This primer is accompanied by an interactive online tutorial.**

# My research group

- setubal@iq.usp.br
- Bioinformatics for Microbiome Analysis
- ML is one of the tools we use
    - More info: Deyvid Amgarten