Data Science and Genomics

Ilan Shomorony

ECE 365

Data Science and Genomics

Ilan Shomorony

ECE 365

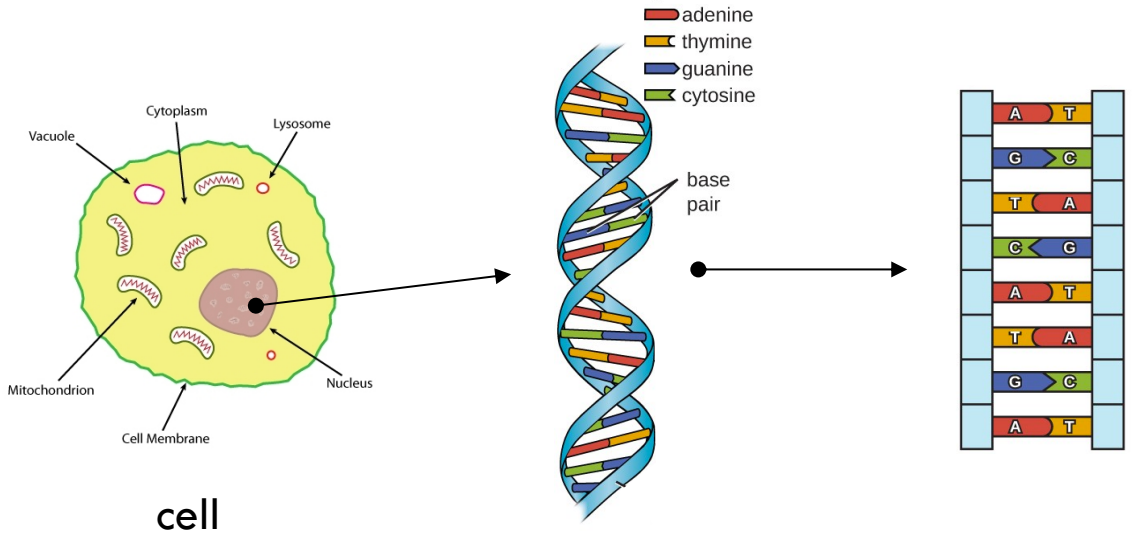
About the module

- Introduction to the analysis of genomic data
- Instructor: Ilan Shomorony (ilans@illinois.edu)
- Lectures: Tuesday/Thursday 3:30 - 4:20 PM via Zoom (10 lectures)
- TA: Shubham Gangil (sgangil2@illinois.edu)
- 4 labs
  - Assigned on Wednesday, due on Thursday of following week
  - First lab released tomorrow, due March 11
  - No lab meeting tomorrow
  - We will use a single submission system (unlike in the first module)
- 2 quizzes (CBTF)
  - 03/23 and 04/01

Lecture 1: A brief introduction to genomics

ECE 365

Deoxyribonucleic Acid (DNA)



cell

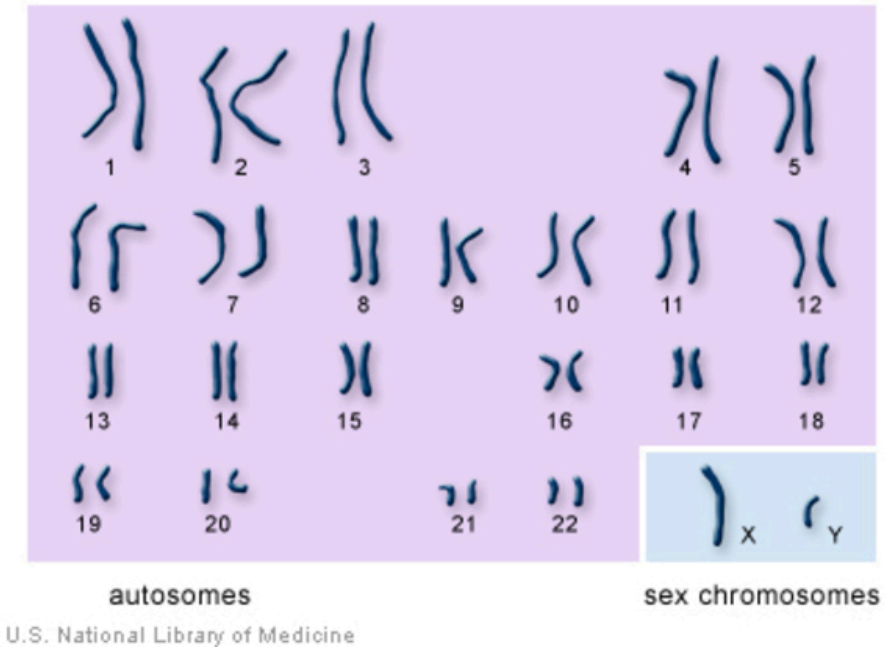
DNA double helix

- Chemical compound that carries genetic information
- Located in the nucleus of every cell of every living being
- Can be described as a sequence of nucleotides, e.g.,

. . . **AGTCATGA** . . .

The genome

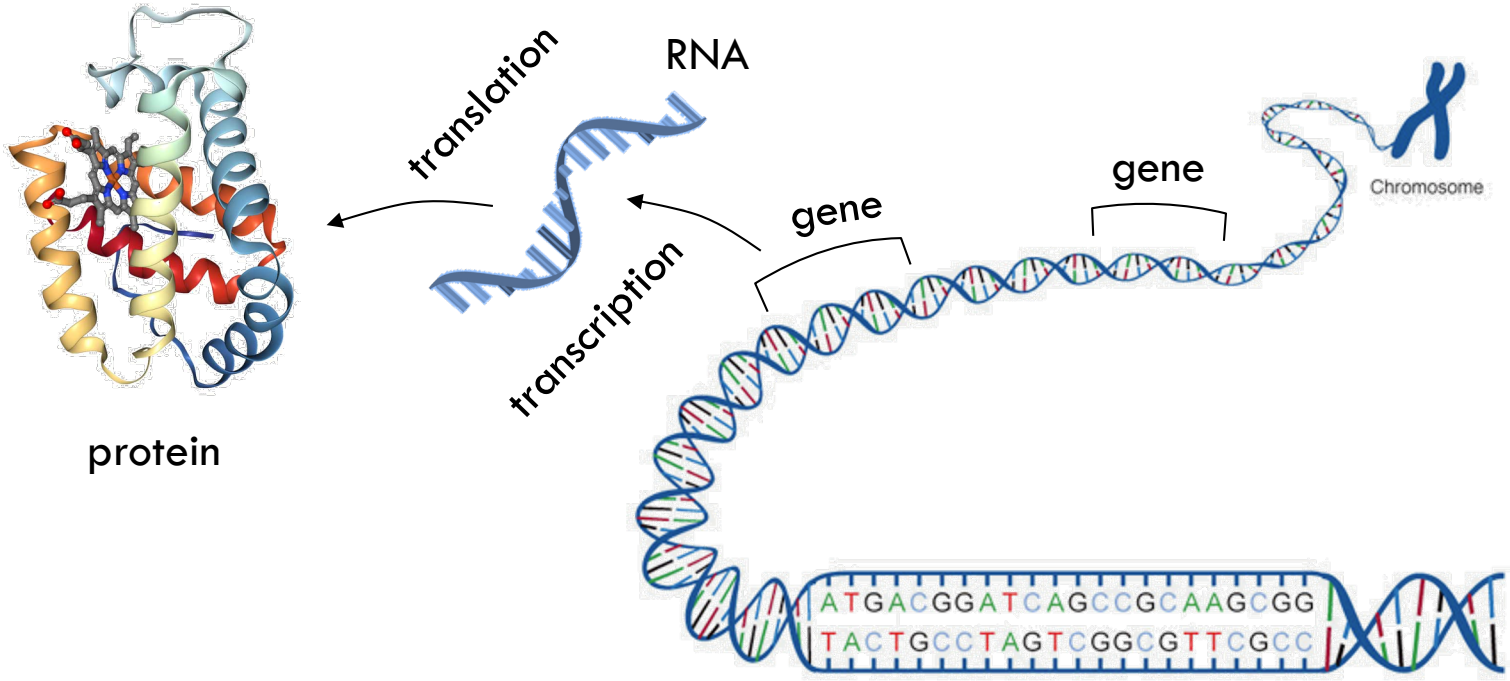
- An organism's complete set of DNA is called its genome
- Every cell contains an entire copy of the genome
- Organized into chromosomes (23 pairs of chromosomes for humans)

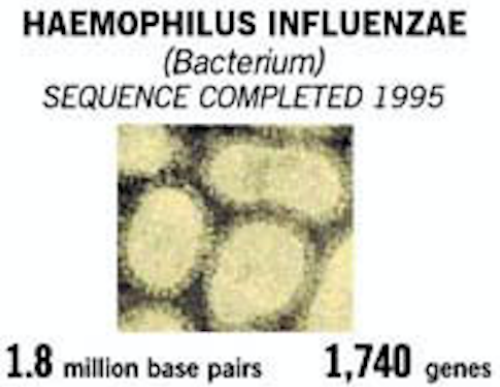


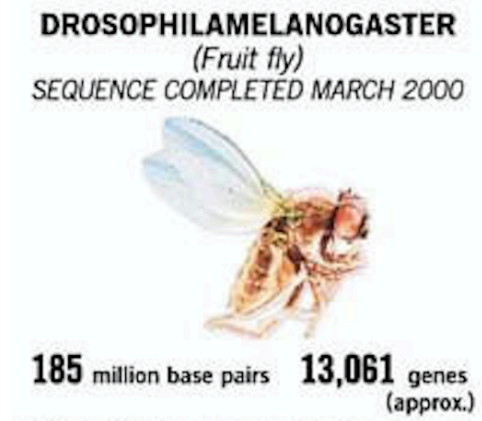
Source: US National Library of Medicine

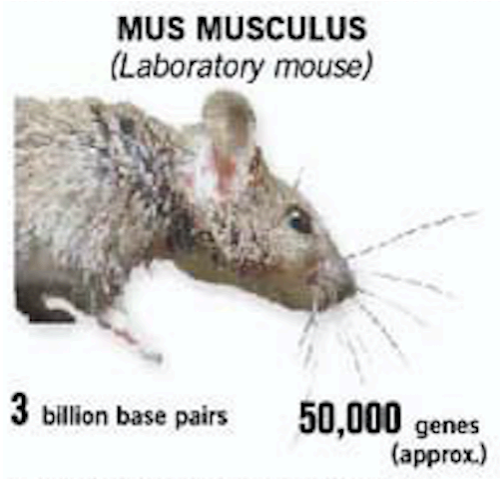
What is Genomics?

- The study of genomes:
  - The genome sequence of different organisms
  - Genes and how they encode proteins
  - **Technology to analyze genome sequences**









(Source: SFGate.com)

How do we read the genome?

- Practical technologies are based on *shotgun sequencing*
- Short reads from random and unknown locations

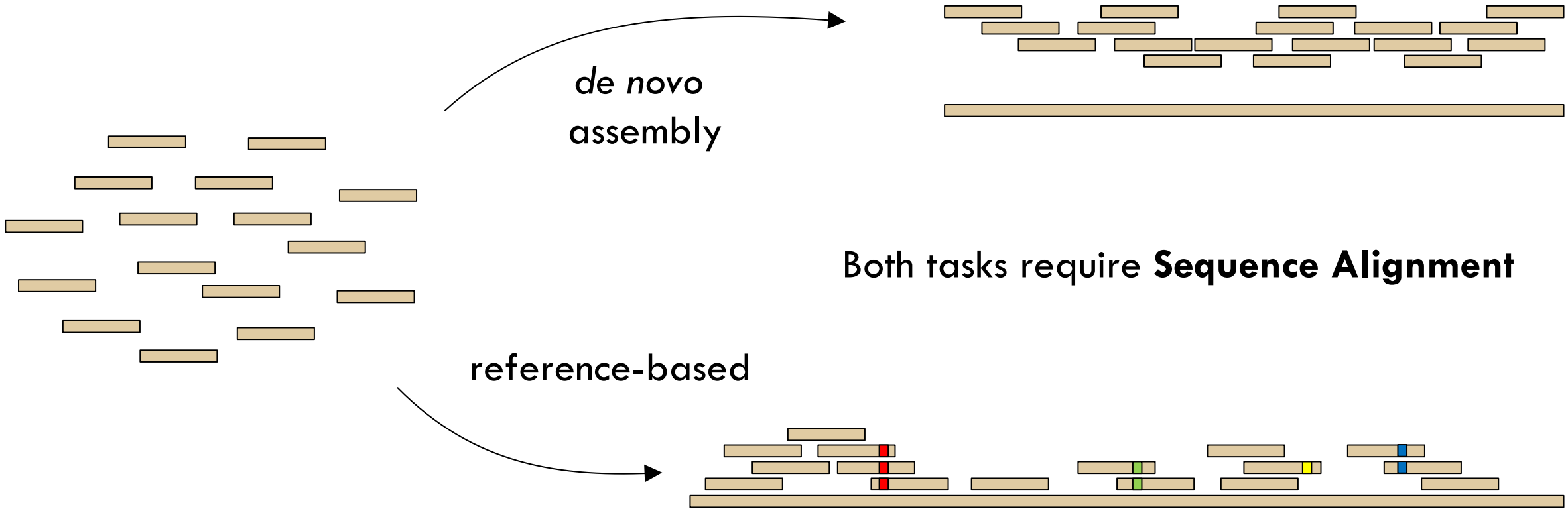
Bacteria:	$\sim 10^6$
Human:	3×10^9
Paris japonica:	150×10^9

$\hat{s}$ = ACGCATTCGCGATT

- Practical problems: short reads, errors, genomic repeats
- Let's look at some real data

How do we process shotgun sequencing data?

1. If a genome from that species has never been sequenced before

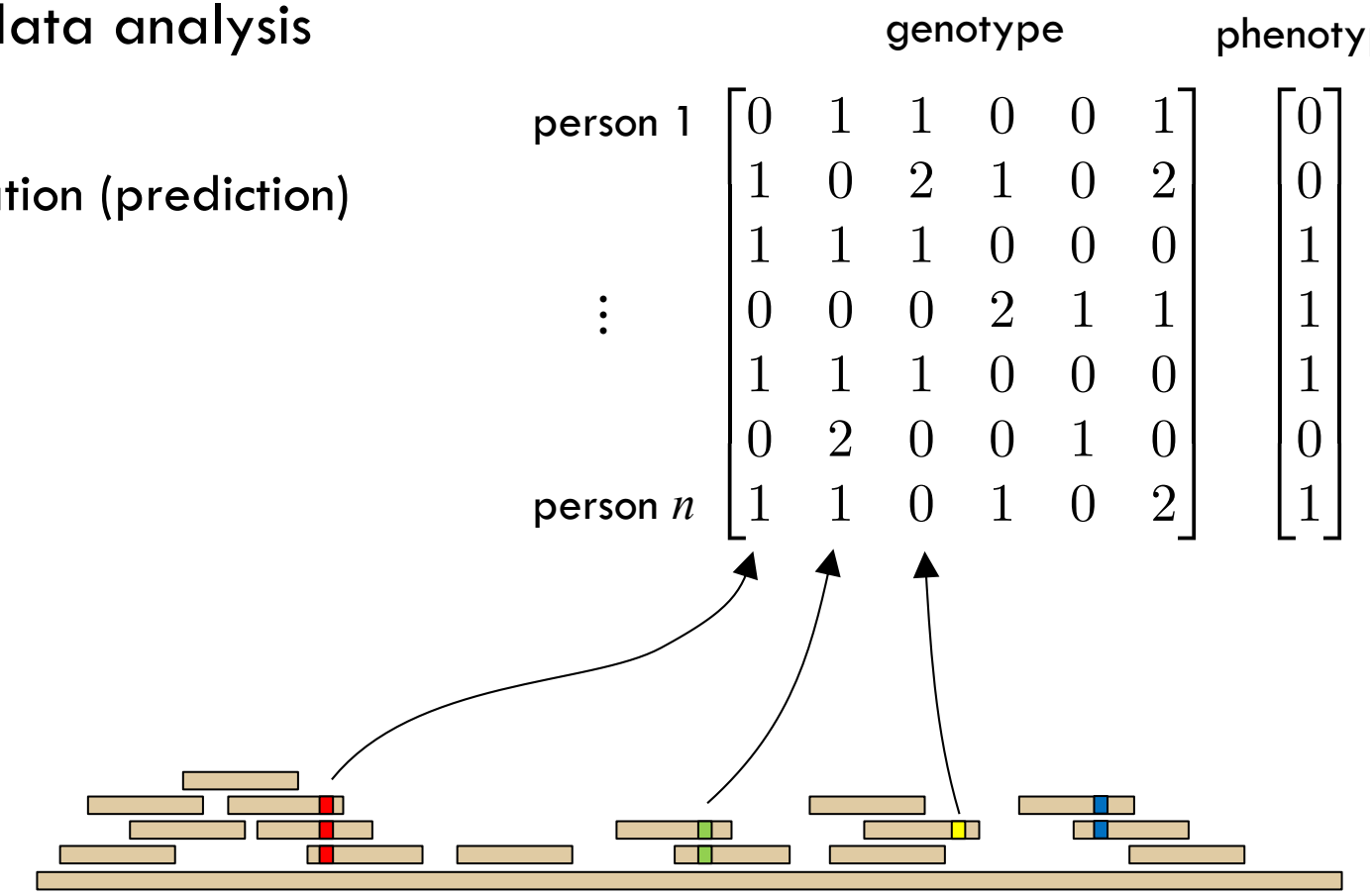


Both tasks require **Sequence Alignment**

2. If a genome from that species has already been sequenced

Course outline

- First part: DNA sequencing data analysis
  - Sequence alignment
  - Genotype-phenotype association (prediction)



Course outline

- First part: DNA sequencing data analysis
  - Sequence alignment
  - Genotype-phenotype association (prediction)
- Second part: RNA sequencing data analysis
  - RNA quantification (maximum likelihood estimation)
  - Clustering cells by RNA profile