BMEG3105 Fall 2025

Data analytics for personalized genomics and precision medicine

Data & Python & Sequence

Lecturer: YuLI(李煜)fromCSE

Liyu95.com, liyu@cse.cuhk.edu.hk **Friday, 5th September 2025**

Today's agenda

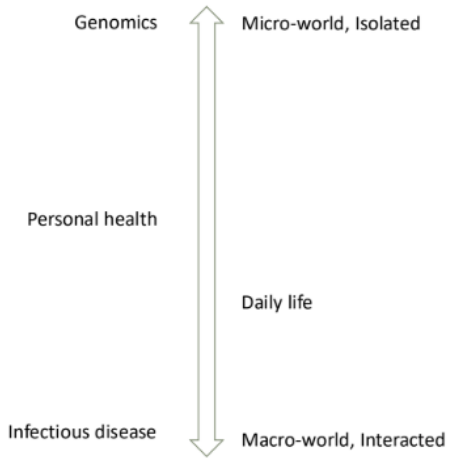
- Different data types
- Introduction to Python programming
- Sequence data
- Sequence comparison and alignment score (Not finished)

Post-course survey results

Positive responses:	Requests:
"fun"," clear," "engaging," "humorous and enthusiastic", "Pretty good"	More rest time, increased student participation, simpler definitions of complex terms

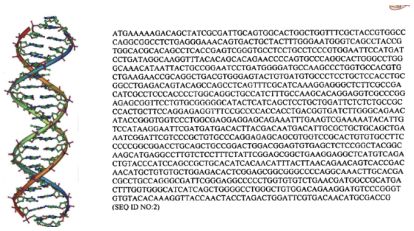
Recap from Last Lecture

- Overview of multi-scale health data:
  - Molecular: Genome, transcriptome, proteome, metabolome
  - Clinical: Medical imaging, blood tests, ECG
  - Lifestyle: Diet, exercise, social media, environment
  - Other: Drug history, family history, travel history
- Emphasis on genomics in personal health and infectious disease



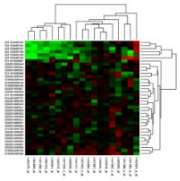
Data Types Encountered in the Course

1. **Sequential Data**: e.g., DNA, RN and protein sequences

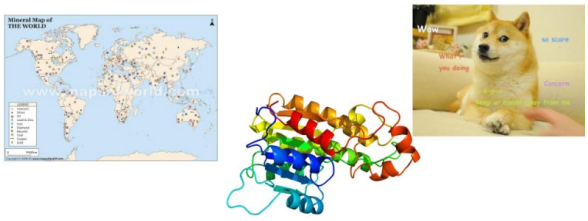


2. **Data Matrix**: it is a collection of records, each of which consists of a fixed set of attributes. Which the tows=objects and columns = attributes.

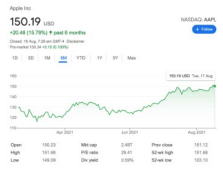
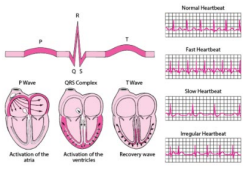
Person	Height (m)	Weight (kg)
P1	1.79	75
P2	1.64	54
P3	1.70	63
P4	1.88	78



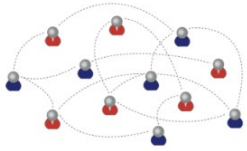
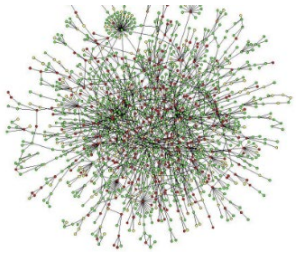
3. **Spatial Data**: Geographically or spatially located information



4. **Temporal Data**: data which involves time (e.g. ECG, stock p

5. **Graph/Networks**: Objects and connections (e.g., social networks, PPI)

6. **Text**: Short or long documents
7. **Multi-modality Data**: Combined types (e.g., video, EHRs, spatial transcriptomics)
8. **Unknown Data**: Diet and Exercise.

Introduction to Python Programming

What is programming

- Communicating with the computer

What is Python

- Tool for communication with the function like a translator to translate our language to the way computer can understand(like WhatsApp and WeChat)

What is Numpy

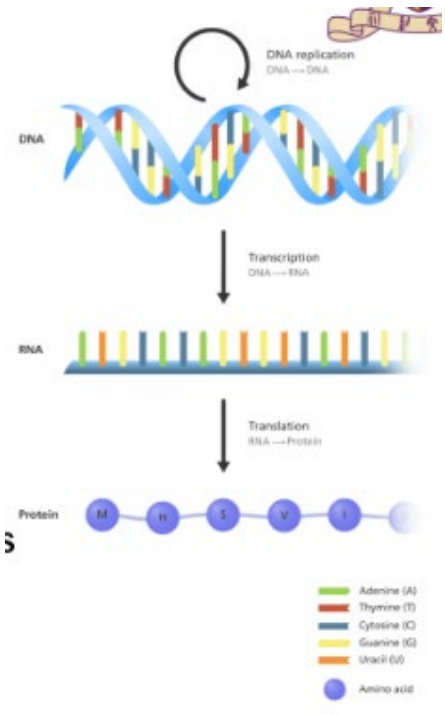
- Additional plug-in library to make Python more powerful usually helpful for math calculation. Usually use for calculate mean, variance, median, max, min and lots of other things.

Homework 0: don't need submit.

Why Sequence Data

Central dogma

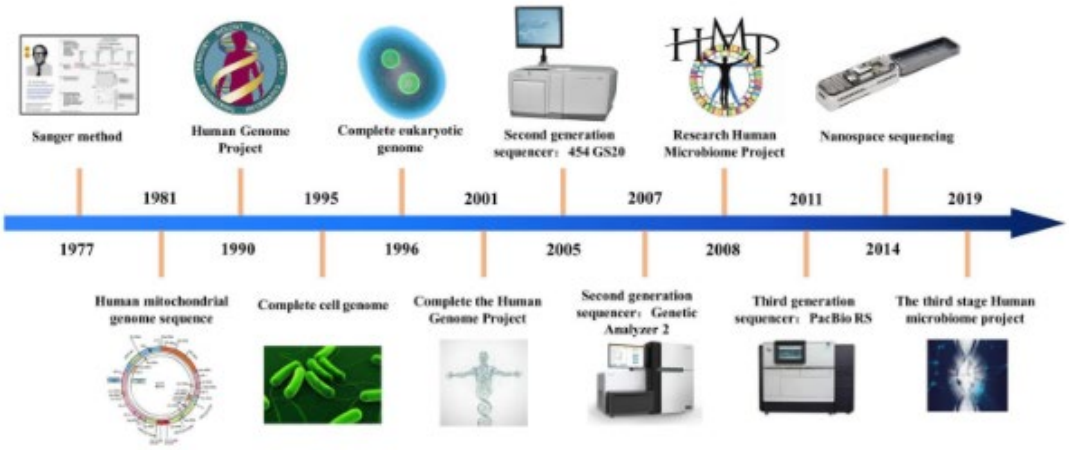
- The genetic information hidden in DNA sequences
- Biologists believe genotypes are determined by the sequences.



DNA, RNA and protein sequences

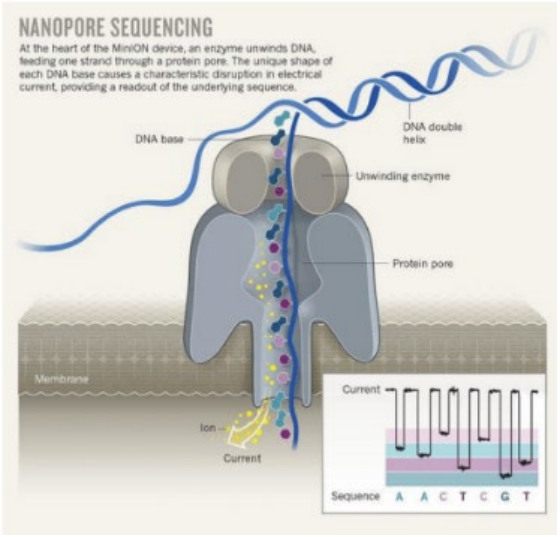
- DNA: composed of ATCG, Complementary double strand
- RNA: composed of AUCG
- ProteinL usually composed of 20 amino acids

How do we get the sequences



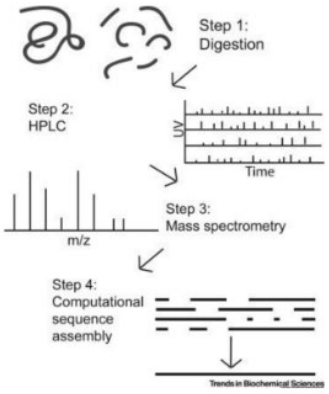
Nanopore sequencing

- DNA goes through a chemical pore
- Able to obtain very long (up to 4Mb VS 1000bp)
- Error rate is high (5% VS 0.001%)
- Using the current change tot detect Sequence



Protein sequencing

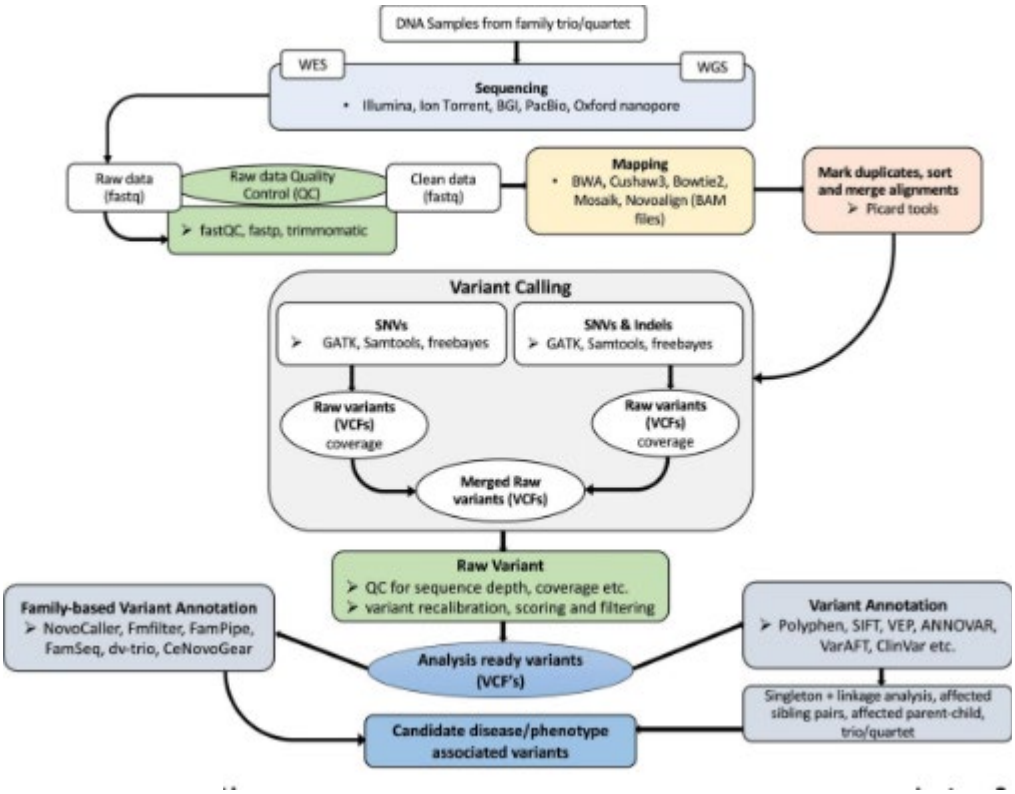
- Based on mass spectrometry
- Break down the long sequence to short pieces
- Each piece will be detected by mass spectrometry
- Assemble the short pieces into the raw sequence



What are the raw data and what do we do them

DNA Sequence:

1. Reading raw sequences
2. Quality control to delete noises
3. Map reads to reference genome
4. Variant calling for mutation
5. Check the phenotype are associated with variants or not



Protein Sequences:

1. Comparison of two or more sequences
2. Multiple sequence alignment
3. Similar sequences imply similar structures which will further imply similar functions.
4. Similar sequences will find out common ancestors.