Verily (formerly of Google X)
Baseline project

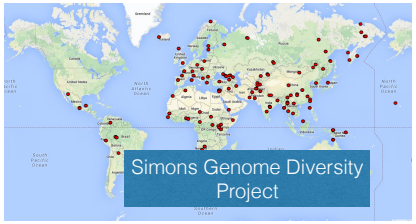
23andMe

ancestryDNA™

VeritasGenetics

1000 Genomes
A Deep Catalog of Human Genetic Variation

Personal Genome Project



- "Understand your genetics and know more about your health." — Veritas Genetics
- "Uncover your ethnic mix, discover distant relatives, and find new details about your unique family history with a simple DNA test." — AncestryDNA
- "We bring the world of genetics to you." — 23andMe

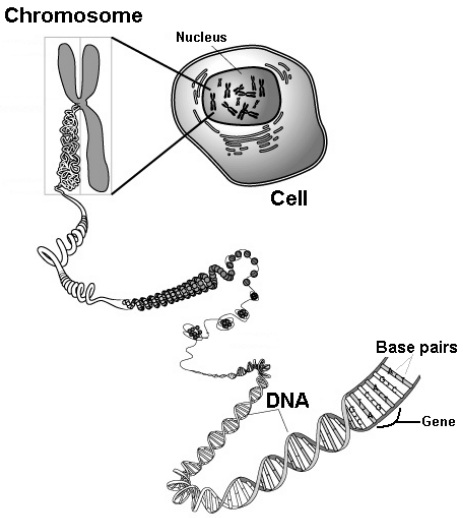
- In this course you will:
  - be exposed to different types of genomic data
  - see examples of how the data is analyzed and used
  - explore the social and ethical issues surrounding genomic data
- Core skills: quantitative analysis, communication (written and spoken)

Let's start at the very beginning

When you read, you begin with ABC

When you study genomic data, you begin with ACGT

- **D**eoxyribo**N**ucleic **A**cids: adenine (A), cytosine (C), guanine (G), thymine (T)
- all you need to know for now: DNA are the rungs of the double helix
  - also referred to as *base pairs*, *bases*, *nucleotides*



Putting DNA into context

We can 'read' DNA

- A, C, G, T are the 'letters' of the DNA alphabet
- a **gene** is a sequence of DNA letters that codes for something important, usually a protein. E.g. *LCT* gene for making lactase, which breaks down lactose in the gut

—> they are the 'words'

- a **genome** contains all the genetic material in an organism', e.g. the human genome, the fruit fly genome

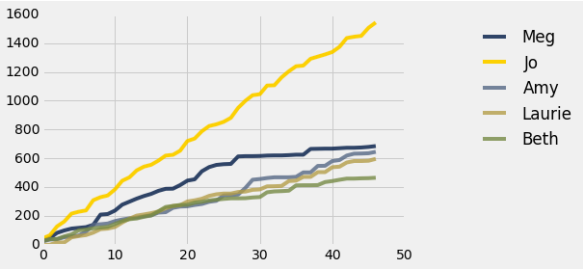
—> they are the 'books'

- strings of DNA are called **sequences**

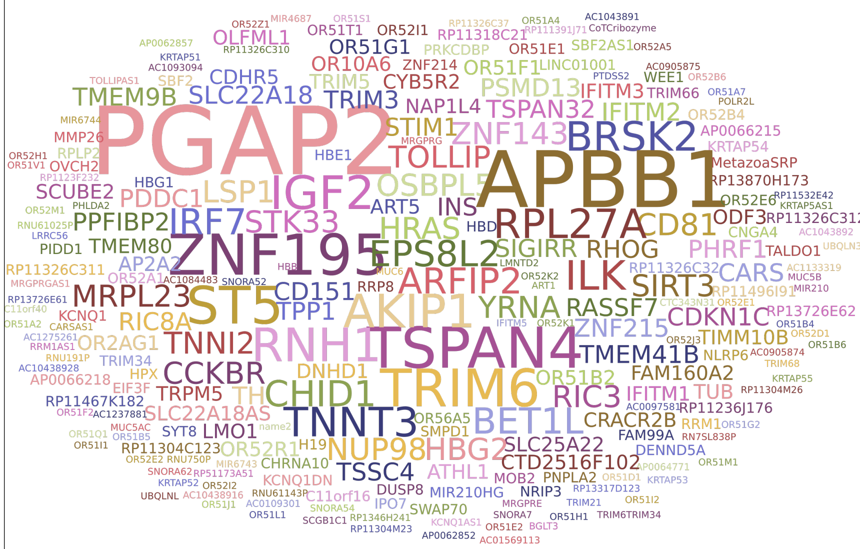
But this analogy is very loose

- let's take a look at the human genome (3 billion base pairs)
- how about the much smaller HIV genome (10,000 base pairs)

Compare with text



Guess the book from the word cloud



DNA data is not comprehensible in its raw form

But because of its biological role, we know it contains useful information

What information is contained in our DNA?

- record of change
- set of instructions (i.e. 'genetic code')
- unique identifier

Course themes

Topics	Data type	Information wanted	Main tool
HIV	nucleotide sequences (DNA)	evolutionary relationship	distances, visualizations
Personal Genomics	single nucleotide polymorhisms (SNPs)	trait association	hypothesis tests
Forensics	short tandem repeats (STRs)	identity	probability

10 minute break

Assessment

- Attendance and participation — 50%
- "Genomics & data science in the news" — 10%
- Group project
  - written proposal — 10%
  - final write-up and presentation — 30%

Other administrivia

- Office hours: Thu 10-11am
- Classes will be mixture of lecture and labs
- https://github.com/shishiluo/Genomics-DataScience

Molecular biology in 10 minutes

- Central dogma video

Just some of the online resources out there

- NIH genetics home reference: https://ghr.nlm.nih.gov/primer/basics/DNA
- NIH National Human Genome Research Institute glossary: https://www.genome.gov/Glossary/index.cfm
- Be careful about sources for which you cannot establish credibility