INSTITUTE FOR APPLIED
COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

CS205 - FINAL PROJECT
SPRING 2018

Genomic Sequencing Parallelization

Kar-Tong Tan
Nripsuta Saxena
Divyam Misra
Andrew Lund

Uses of genomic sequencing

- Quicker diagnosis of mysterious diseases
- Find patients with the same disease
  - Very important for extremely rare disorders
- Testing for hereditary disorders
  - In utero and carrier testing
- Predictive (presymptomatic) testing
- Faster pharmacogenetic testing
  - Testing how someone will respond to a certain medication
  - Used for certain kinds of cancers

Problem Description

The cost of sequencing has dramatically decreased in the last decade.

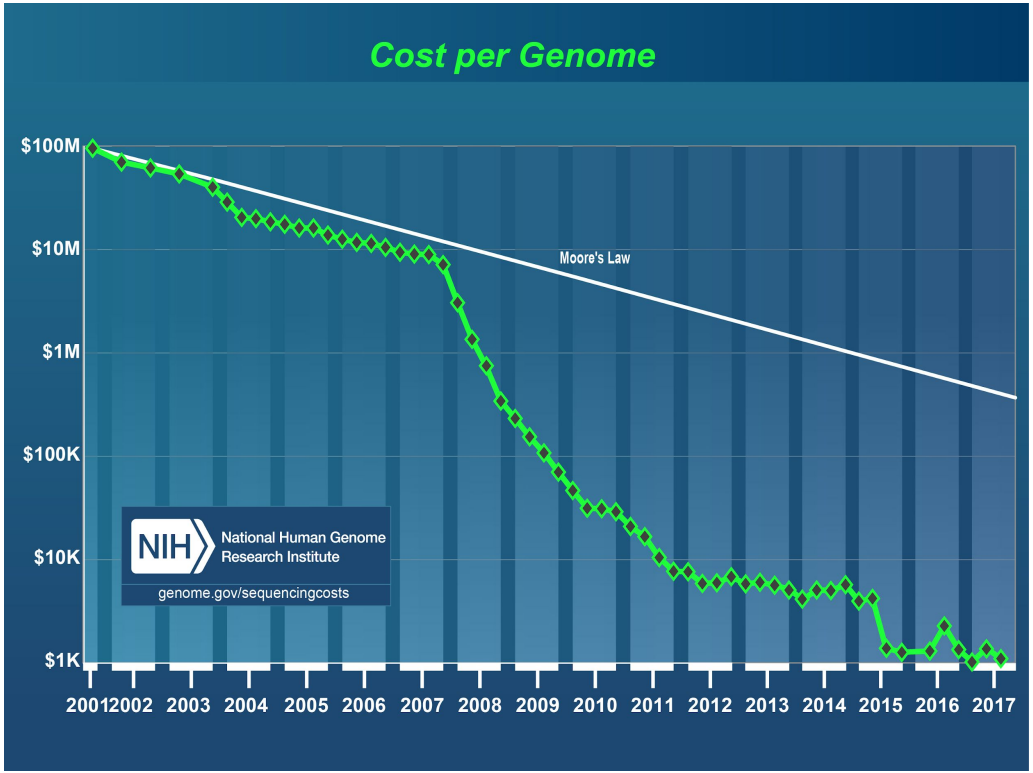
- First Genome: >$2.7B 15 years
- Today: $1,000, ~1 week

The primary overhead is now cost of computation.

- Not easy to parallelize algorithms
- Algorithms do not scale linearly

Results need to be returned in timely fashion for clinical applications

- >1 week algorithmic runtime in some applications
- Too slow for timely diagnostics



Models & Data

Application of interest:

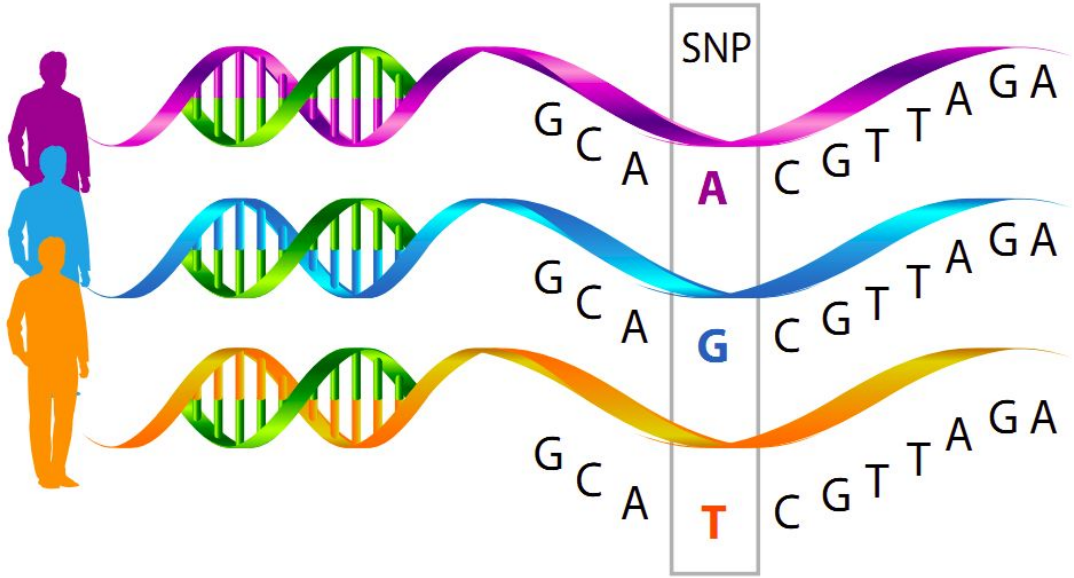
- Single Nucleotide Polymorphisms (SNPs)
- Drives differences between individuals

Algorithms:

- SNP calling
- Most are open-source

Data:

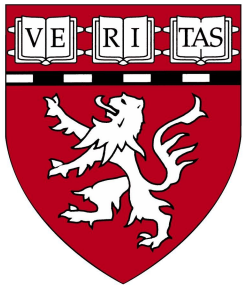
- DNA, RNA alignment files
    - 2 individuals (public data from the 1000 Genomes Project)
    - .bam (~10GB each)
    - .bai (~5MB each)



Tools & Infrastructure

Infrastructure:

- Compute Cluster at Harvard Medical School
  - https://rc.hms.harvard.edu/
- 8,000 cores with several PB network storage
- Nodes support up to 32 cores but capped at 20 cores
  - Known problem in parallelization of related algorithms



Harvard Medical School
Research Computing

Tools:

- SAMtools, bcftools
  - Calling of single nucleotide variations
- MPI + Spark/MapReduce

SAMtools