Genomic Sequencing Analysis Parallelization

CS205 - Computing Foundations for Computational Science - Spring 2018 Final Project

Andrew Lund, Divyam Misra, Nripsuta Saxena, Kar-Tong Tan

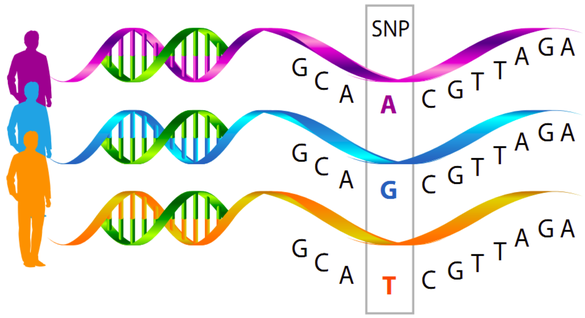
HARVARD School of Engineering and Applied Sciences

IACS INSTITUTE FOR APPLIED COMPUTATIONAL SCIENCE AT HARVARD UNIVERSITY

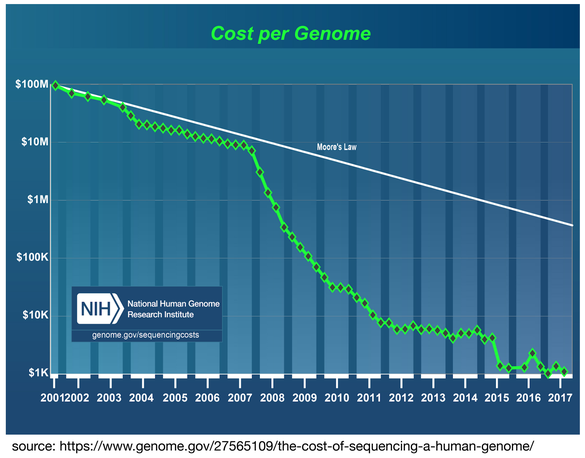
Introduction

Project Goal

To speed up the identification of single nucleotide polymorphisms (SNP) in DNA and RNA through big data and high throughput parallelization techniques.



- SNPs are the differences between people and can identify diseases, disorders, or medication responses
- The cost of sequencing has dramatically decreased in the last decade
- The primary overhead is now computation



source: https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/

- SNP algorithms are not easy to parallelize
- Some analyses take more than 1 week
- Too slow for timely clinical diagnostics

Tools & Infrastructure

SAMtools

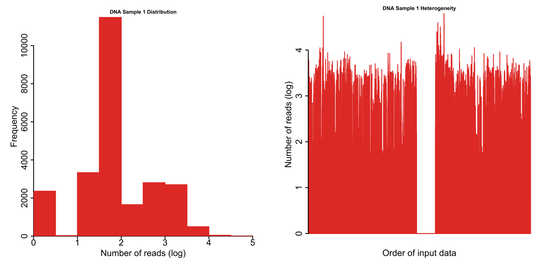
- Suite of programs for interacting with high-throughput sequencing data
- Open-source

HARVARD MEDICAL SCHOOL

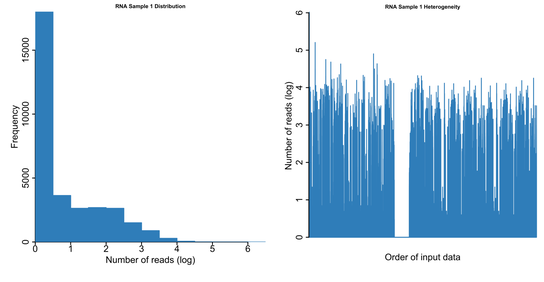
- HMS Research Computing cluster
- 8000 cores with several PB network storage

Data

- DNA, RNA alignments for 2 people (1000 Genomes Project)
- .bam (~10GB each), .bai (~5MB each)

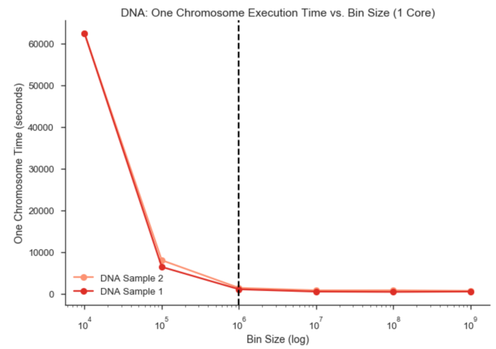


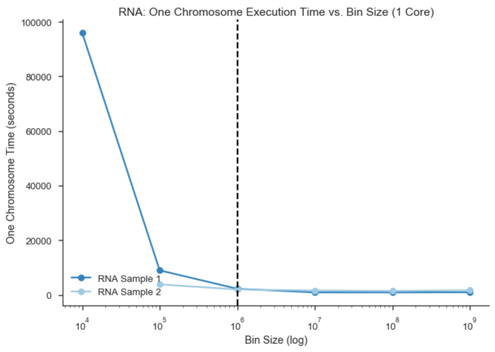
- Uneven input data chunk sizes
- Unpredictable sizing in index file



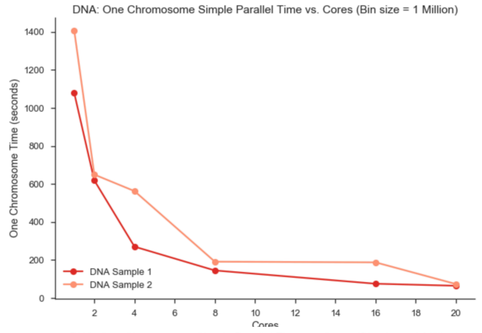
Binning

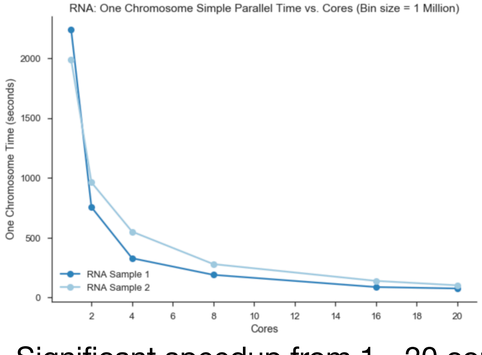
- Distribute DNA & RNA reads (sequence strings) into “bins” amongst cores





- Best bin size: at least 1,000,000 chromosome positions





- Significant speedup from 1 - 20 cores

Profiling

- **gprof** profiler on 10,000,000 read DNAsample

```
Each sample counts as 0.01 seconds.
%           cumulative       self              self     total
time        seconds  seconds calls  s/call  s/call  name
51.10       1.16     1.16    1     1.16    2.27  mpileup
18.06       1.57     0.41    4519010  0.00    0.00  bam_plp_next
12.11       1.85     0.28    26195998 0.00    0.00  pileup_seq
9.69        2.07     0.22    26896170 0.00    0.00  resolve_cigar2
5.73        2.20     0.13    4213250  0.00    0.00  bam_mplp_auto
0.44        2.21     0.01    353719   0.00    0.00  kh_get_olap_hash
0.44        2.22     0.01    330015   0.00    0.00  bam_read1
0.44        2.23     0.01    305760   0.00    0.00  mp_free
0.44        2.24     0.01    305760   0.00    0.00  mplp_func
0.44        2.25     0.01    305759   0.00    0.00  bam_copy1
0.44        2.26     0.01    17474    0.00    0.00  tweak_overlap_quality
0.44        2.27     0.01    2670     0.00    0.00  printw
0.22        2.27     0.01    4213249  0.00    0.00  mplp_get_ref
```

OpenMP

- Shared-memory technique to reduce execution
- **mpileup** identified in profiling (above) as primary function to focus parallelization
- 457 lines, 22 for/while loops

MPI

- Distributed-memory technique to reduce execution
- Requires wrapper around SAMtools to scale up to 640 nodes on HMSRC cluster
- Currently limited to 20 cores with binning and OpenMP

Load Balancing

- Read the index (.bai) file
- Apply sorting algorithm to bin by size rather than index in current binning method

Desired Outcome

- Ultimate goal to combine binning, OpenMP, MPI to achieve maximum possible speedup
- Scale to ~200GB (whole genome)