IACS
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

Speedup analysis of single nucleotide polymorphisms in human genomes

Primary Focus: load balancing & sorting

Program Suite: SAMtools

Application Type & Parallelism

Application Type:

- Big Data: I/O-bound reading large .bam files, up to 200TB
- HTC: High-frequency sequence reads

Levels of Parallelism:

- Inherently partially parallelized with pthreads.h for sorting bam files (bam_sort.c)

Parallel Execution Model:

- Embarrassingly parallel
- MPI
- Potentially OpenMP

Performance Analysis and Optimization of SAMtools Sorting

Nathan T. Weeks[1,2](✉) and Glenn R. Luecke[2]

[1] Department of Computer Science, Iowa State University, Ames, USA
[2] Department of Mathematics, Iowa State University, Ames, USA
weeks@iastate.edu

Published Code Profiling (.bam sorting)

Single thread performance

- Similar for Samtools and optimized Samtools

Two thread performance

- Slight regression for optimized Samtools, probably due to OpenMP and communication overheads

More than four threads

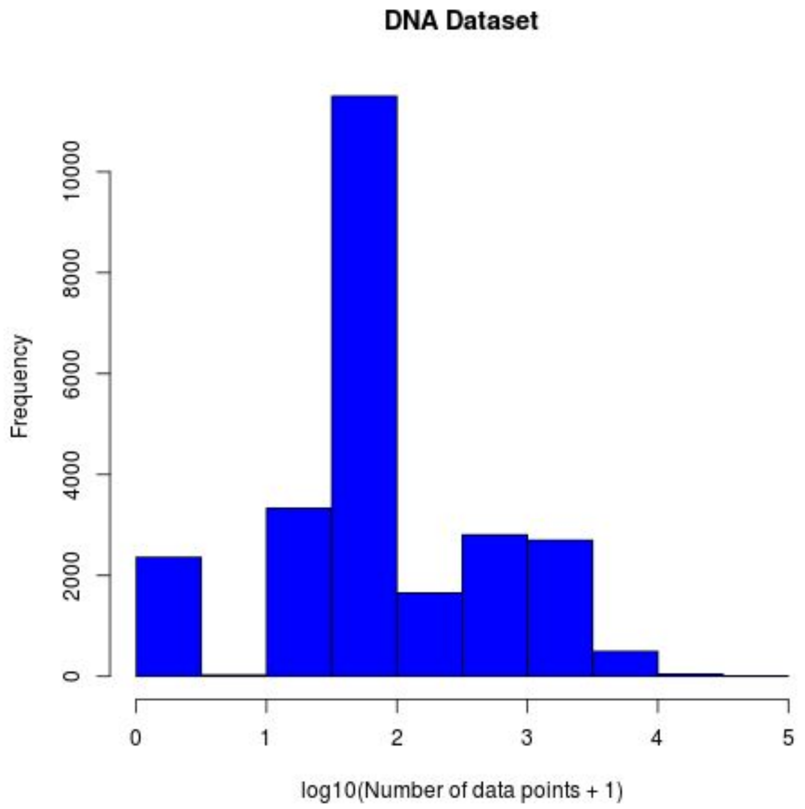
- Optimized Samtools 29% and 73% faster

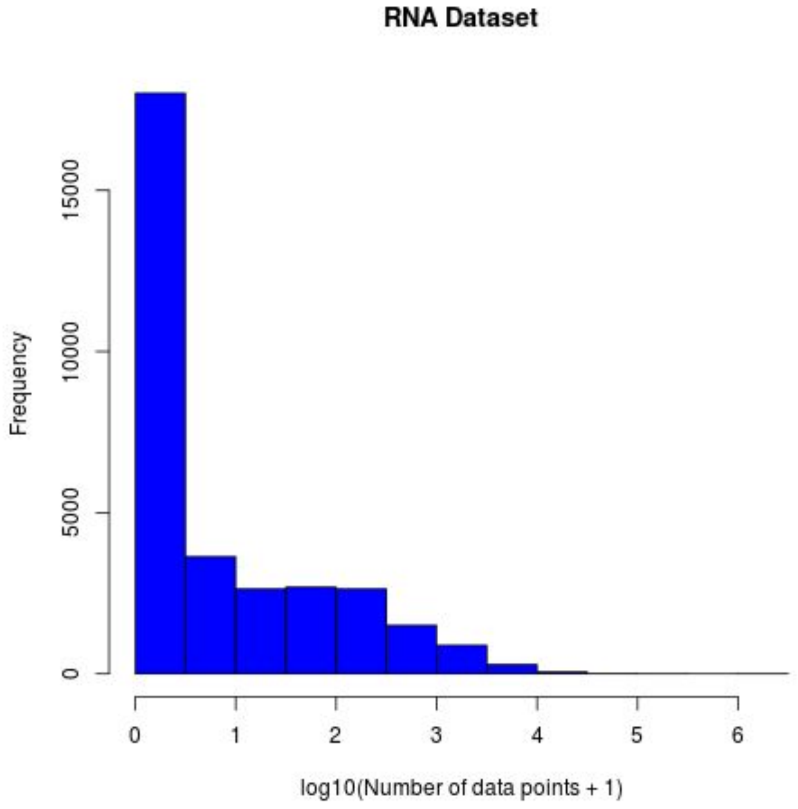
Main Overheads

Load balancing: .bai & .bam data distribution heterogeneity

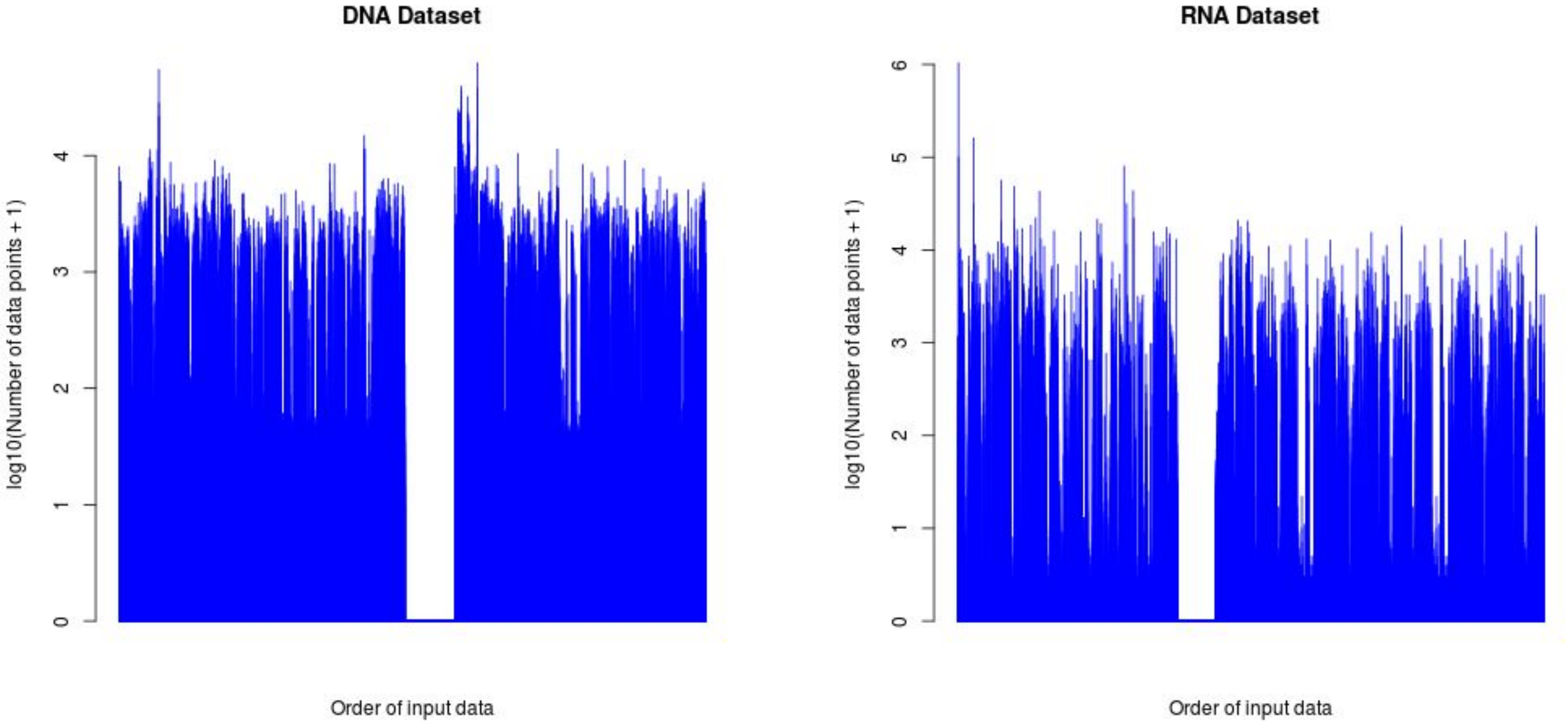
Sorting: .bai & .bam

Load Balancing: Uneven input data chunk sizes





Unpredictable sizing in index file order



Overhead Mitigation Techniques

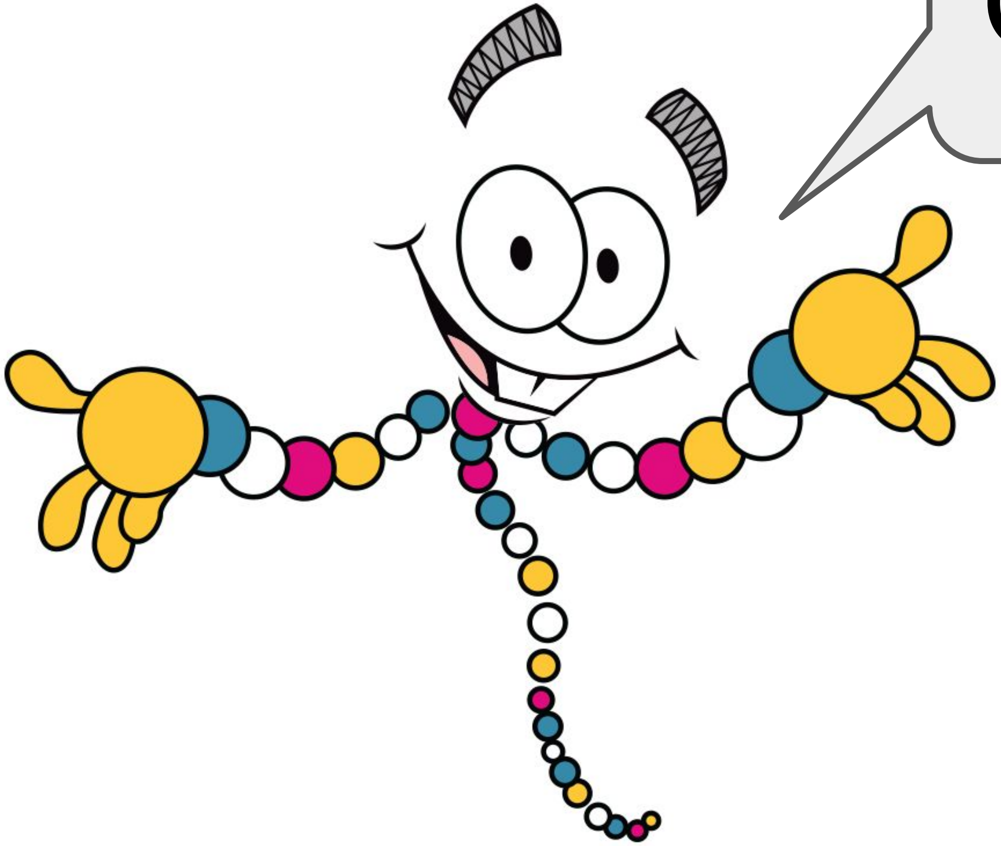
Mitigating Heterogeneity

- Files are 20+ GB. Difficult to directly assess data distribution and density quickly.
- Analyze data distribution, determine ideal way to shard data. Format of index file is an r-tree.
- Index tells where to start and stop
- Develop module which splits a chromosome into equal bins before running samtools analysis.

Speedup & Scalability

Theoretical Speedup & Scalability:

- Hard to estimate speedup.
- Something we will be determining through our project analysis.
- Naively hoping for 2x speedup.
- Can scale up to 600 cores on the HMS cluster which we hope to exploit using MPI.



Questions?