USING ROOT IN THE FIELD OF GENOME SEQUENCING

Google Summer of Code 2025
Mid-Term Presentation

Mentee:- Aditya Pandey
Mentors:-Vassil Vassilev, Martin Vassilev

PROJECT INTRODUCTION

- Modern DNA sequencing generates 200GB per human genome, creating petabytes of data globally that current formats struggle to handle efficiently. Existing solutions force users to choose between compression (CRAM format) or speed (BAM with fast access), but never both.

- ROOT, CERN's framework that manages petabytes of data, offers RNTuple a next generation columnar storage system with advanced compression. This project adapts ROOT's 25 years of big data expertise to create RNTuple RAM (ROOT Alignment Maps) format extending GeneROOT project specifically for genomic data.

PROJECT OBJECTIVES

- Reproduce benchmark results based on previous work done through the GeneROOT project.
- Investigate and compare the latest compression strategies used by Samtools for BAM conversion with ROOT Alignment Maps (RAM)
- Explore ROOT's RNTuple format for efficient storage of RAM data, replacing the previously used TTree
- Investigate and evaluate various ROOT file splitting techniques
- Produce a comprehensive comparison report summarizing findings and recommendations

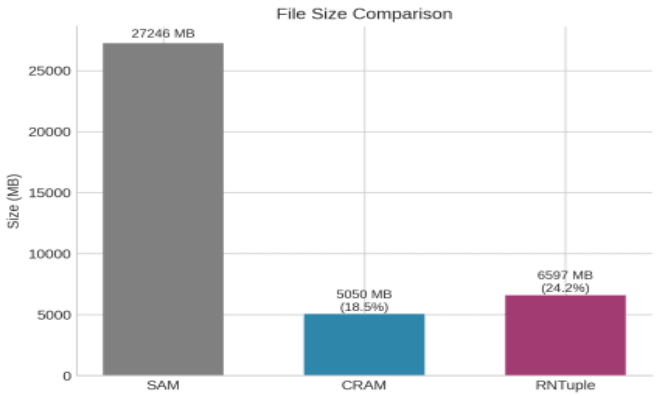
OUR PROGRESS SO FAR

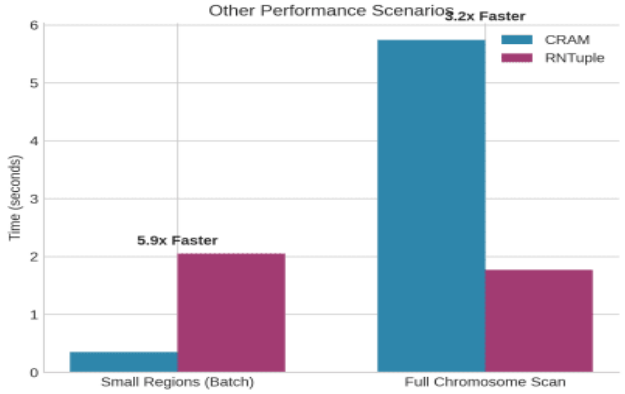
1. Reproduced Previous TTree Comparisons
- Validated baseline TTree performance against BAM/SAM formats
- Established benchmark metrics for file size, compression ratios, and read/write speeds

2. Implemented initial RNTuple-based RAM Format
- RAMNTupleRecord: Full SAM/BAM field support
- Smart reference management with caching
- Fast position-based indexing for region queries

3. Analysis of CRAM format and benchmarking with RNTuple implementation

CRAM vs. RNTuple Performance Benchmark





Key Takeaways

Metric	Winner	Detail
File Size	CRAM	23.5% Smaller
Large Queries	RNTuple	1.9x Faster (Avg)
Full Scan	RNTuple	3.2x Faster
Small Queries	CRAM	Faster Access Times

Notes:- 1)We are using a 1.1 GB FASTA reference file for CRAM file querying.
2) We are using ZSTD compression algorithm for RNTuple file conversion


FASTQ File Size Comparison
14.2 GB
4.5 GB
(31.7%)
6.8 GB
(47.9%)
File Size (MB)
0
2000
4000
6000
8000
10000
12000
14000
FASTQ
(Original)
Gzip
Compressed
RNTuple
Compressed

REMAINING OBJECTIVES

1. Making our RNTuple implementation more concrete and robust.(Adding more features like RAM splitting and RAM merge).

2. Understanding CRAM compression algorithms and exploring if they can help in our work.

3. Investigate ROOT's File Splitting Techniques

4. Produce final comparison report.

Thank you for listening