COMPILER
C | R
RESEARCH

GENEROOT PROJECT

Progress So far

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

PROJECT GOALS

- Comprehensive benchmarking of the RAM (ROOT Alignment Map) format: file sizes against BAM, CRAM and SAM, query speed for region access, selective column reads, and flagstat-style operations
- New format support: direct BAM read/write for interoperability with existing genomics pipelines
- Fixing correctness with reference to samtools: lossless round-trip validation (SAM→RAM→SAM), region query parity with samtools view
- A public release of RAMTools with documentation and installation instructions

OUR PROGRESS SO FAR

BAM Support

- Direct BAM-to-RAM conversion using htslib, no SAM intermediate needed
- Multithreaded BGZF decompression via hts_set_threads
- Proper 0-based/1-based coordinate handling, CIGAR unpacking, auxiliary tag parsing

Index & Region Query Fixes

- Fixed sparse index strategy: new chromosome, 10kb gap, and every 100th mapped read as triggers
- Eliminated duplicate index entries at same (chromosome, position)
- Fixed CIGAR-based reference span for accurate region overlap

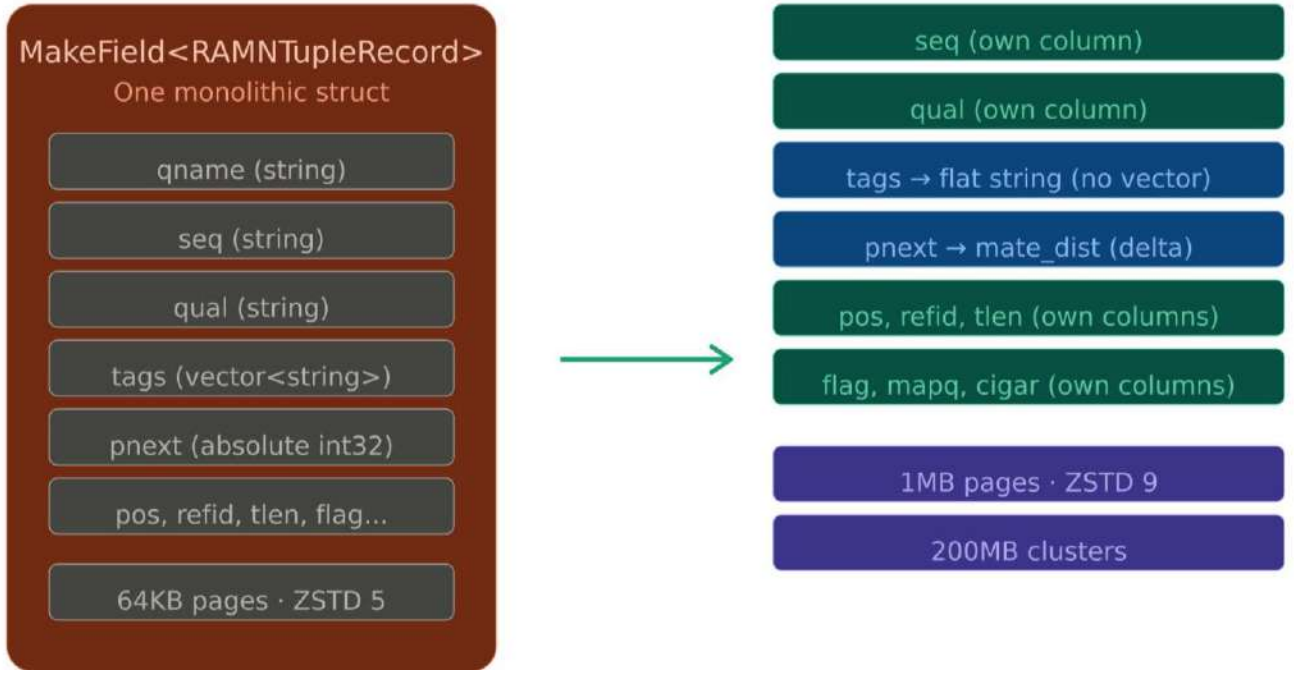
Infrastructure

- Made Codecov and clang-tidy workflow robust

COMPILER C | R RESEARCH

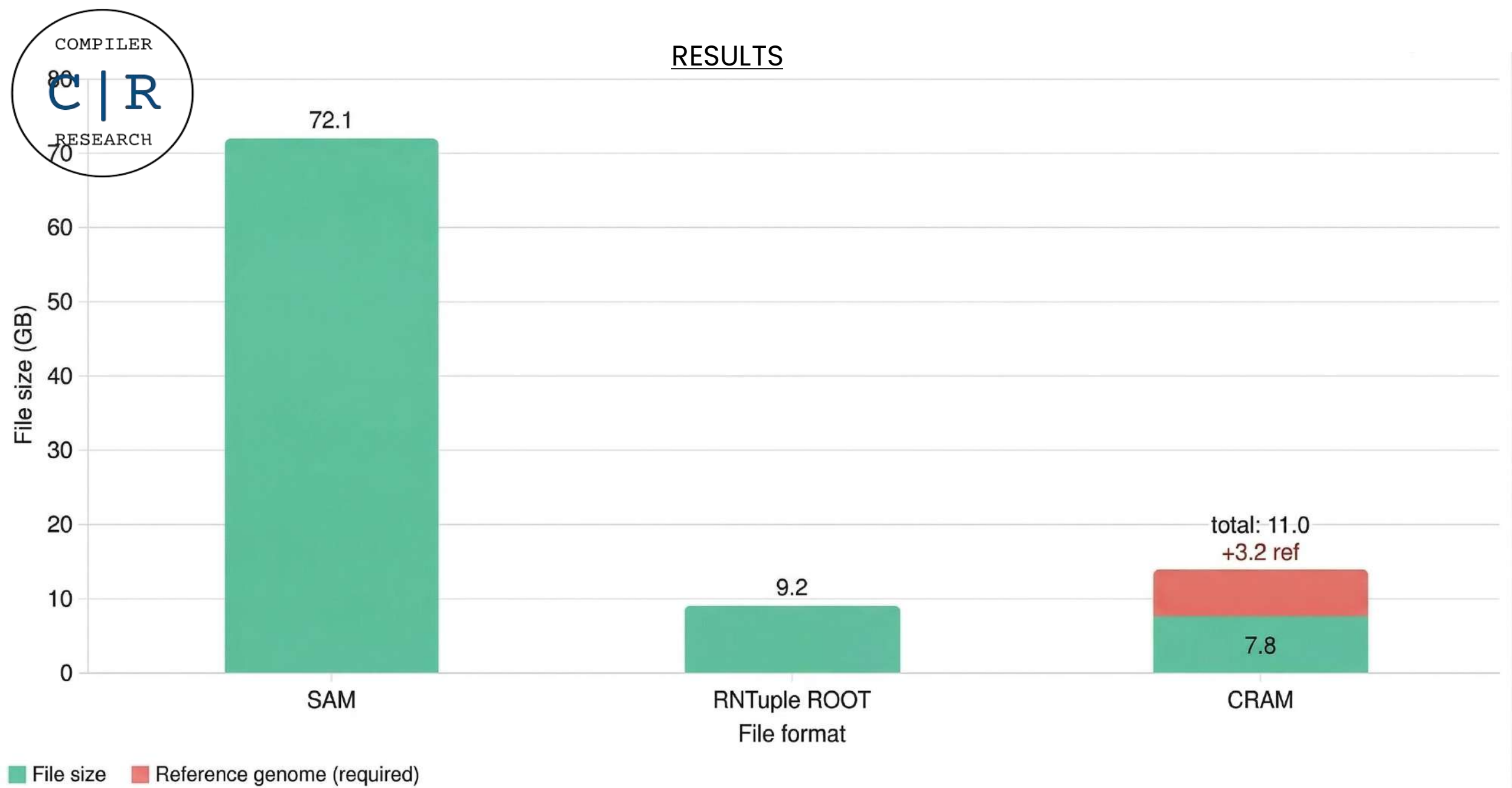
COMPRESSION WORK

- Studied CRAM 3.1 compression techniques and adapted them to RNTuple's columnar architecture
- Decomposed monolithic RAMNTupleRecord into individual top-level columns with independent compression
- Flattened tags to single string, delta-encoded mate position, tuned page/cluster sizes and ZSTD level



COMPILER C | R RESEARCH

RESULTS



COMPILER C | R RESEARCH

REMAINING OBJECTIVES

- Validate correctness against samtools: resolve remaining discrepancies in region query read counts, ensure lossless SAM→RAM→SAM round-trips match samtools output exactly
- Public release of RAMTools: documentation, installation guide.
- Add RAM as a native format in HTSlib: implement basic read/write/index/query through HTSlib's API so that samtools and all HTSlib-based tools gain RAM support without code changes

Thank you for listening