Improving ramtools query speeds

Internship project proposal by Georgi Haralanov

About me

Education: High School of Mathematics “Akademik Kiril Popov” Plovdiv, Bulgaria (2023-Present)

Technical skills:

- Proficient: C, C++
- Familiar: Python, C#, JavaScript

Experience:

- Implemented an assembler for the Hack assembly language
- Working on a compiler for the Jack programming language

Reasons for the changes

- The amount of data used in genomics and bioinformatics is increasing rapidly
- A single large scale query can take more than 10 minutes to process
- In order to facilitate a smooth workflow query times must be lowered
- The current implementation fails to utilize the full potential of modern multi-core CPUs

Project overview

Increasing query speeds using ROOT's RDataFrame:

- This allows us to use ROOT's built-in Implicit Multithreading to use the full amount of cores found on modern machines
- It features caching capabilities which increase access times and lower RAM usage

Project goals

- Implement a multithreaded query tool
- Benchmark and compare the new implementation with the previous one
- Add any new features introduced for the single threaded mode across the duration of the project
- Clear any and all bugs and errors found throughout the project

Project timeline

1. Benchmark single threaded implementation
2. Implement base features in a multithreaded environment
3. Clear any bugs
4. Implement more advanced features for filtering and indexing
5. Clear any bugs
6. Benchmark and compare with earlier benchmarks
7. Add any missing features which were made in parallel with steps 1-6
8. Last check for bugs
9. Compile findings and publish, land all code changes to main repo

Thank you!