Current Work

Hummingbird

- Performance prediction for genomics applications on Google cloud
- Downsamples FASTQ and BAM files
- Executes each stage of pipeline on downsampled file
- Best configuration for downsampled file = Best configuration for whole file
- Aim: Extend Hummingbird to include other frameworks

Pipelines

- The Genome Analysis Toolkit(GATK)
    - MuTect
- High-performance genomic analysis framework with in-memory computing(GPF)
- Persona: A High-Performance Bioinformatics Framework
- ADAM: Genomics Formats and Processing Patterns for Cloud Scale Computing
- Deep Learning Toolkit
- OAI Pipeline
- TPC-DS: Comparison with Ernest and CherryPick

Method

- Use cluster instead of Google cloud for training(save money)
- Downsample files
- Replicate Google cloud environment on cluster
- Execute pipelines using downsampled files
- Predict best configuration