Distributed execution of bioinformatics tools on Apache Spark with ADAM and Cannoli

Michael L Heuer, UC Berkeley

Bioinformatics Open Source Conference 2018 (BOSC2018)
Portland, OR USA
28 June 2018
Getting started with ADAM

$ brew install brewsci/bio/adam

Reproducible research.
Install the software from bioconda:
$ conda install adam
See bioconda.github.io

$ docker run -it quay.io/biocontainers/adam
Getting started with ADAM, resources

- https://github.com/bigdatagenomics/adam
- https://gitter.im/bigdatagenomics/adam
- https://twitter.com/bigdatagenomics
- http://adam.readthedocs.io
- http://bdgenomics.org
AMPLab/RISE Lab Big Data Genomics organization stack

**ADAM**
Transforms: Read ETL

**avocado:**
Scalable variant calling

**gnocchi:**
Parallel variant analysis

**deca:**
Exome copy number variant calling

**cannoli:**
Stream to external tools

**lime:**
Set theoretic primitives for genomics

**mango:**
Fast, multi-sample visualization

**IntervalRDD**

---

**Core ADAM APIs**

**Pipe API**

**SortedRDD/RegionJoin**

---

**Apache Spark**

Workflows (Nextflow, Toil, CWL)

Clouds (EMR, Google, Azure, Databricks)

On-prem (CDH, Slurm/LSF, PSB)

Flat file formats (BAM/VCF/GFF3)

Apache Parquet

GA4GH Schemas
ADAM transforms data into efficient distributed representations

**Sequences**
EMBL, FASTA, Genbank, INSDC

**Reads**
FASTQ

**References**
FASTA, 2bit

**Aligned reads**
SAM, BAM, CRAM

**Features**
BED, GFF2/GTF, GFF3, NarrowPeak, IntervalList

**Variants, genotypes**
VCF

**Graph assembly**
GFA

**Generic formats**
TXT, CSV, SQL, JSON

Avro objects in Parquet for Spark RDDs

Typed products in Parquet for Spark Datasets

Untyped rows in Parquet for Spark DataFrames

Optionally partitioned by genomic region for efficient range queries
ADAM GenomicDataset pipe API

- No need to scatter/gather within a workflow, data are already partitioned by Spark
- Stream in common file formats (SAM/BAM/CRAM, VCF, BED/GTF/GFF3/NarrowPeak, FASTQ)
- Pipe API scales with the number of executor nodes, reducing single sample alignment time with BWA to < 20 minutes
import org.bdgenomics.adam.models.{ RecordGroup, RecordGroupDictionary }
import org.bdgenomics.adam.rdd.ADAMContext._
import org.bdgenomics.adam.rdd.fragment.InterleavedFASTQInFormatter
import org.bdgenomics.adam.rdd.read.{ AlignmentRecordRDD, AnySAMOutFormatter }
import org.bdgenomics.adam.sql.{ AlignmentRecord => AlignmentRecordProduct }
import org.bdgenomicsformats.avro.AlignmentRecord

val reads = sc.loadPairedFastqAsFragments("sample_1.fq", "sample_2.fq")

implicit val tFormatter = InterleavedFASTQInFormatter
implicit val uFormatter = new AnySAMOutFormatter

val alignedReads = reads.pipe[
  AlignmentRecord, AlignmentRecordProduct, AlignmentRecordRDD, InterleavedFASTQInFormatter] (Seq("bwa", "mem", "-t", "1", "-p", "/data/hs38DH.fa", "-"))
  .replaceRecordGroups(RecordGroupDictionary(Seq(RecordGroup("sample", "sample"))))
Cannoli builds on ADAM GenomicDataset pipe API

Cannoli supports:
- Aligning reads with Bowtie, Bowtie2, BWA, Minimap2, and SNAP
- Calling variants with FreeBayes and Samtools mpileup
- Normalizing variants with BCFtools norm and vt normalize
- Annotating variant effects with SnpEff and Ensembl Variant Effect Predictor (VEP)

Cannoli provides:
- Distribution of files across the cluster using SparkFiles mechanism (useful for e.g. BWA indexes)
- Calling external process installed locally, or via Docker, or Singularity (mounting files to the container as necessary)
- Functional API for interactive analysis in cannoli-shell or workbooks such as Jupyter, Zeppelin, or Spark Notebook
import org.bdgenomics.adam.rdd.ADAMContext._
import org.bdgenomics.cannoli.cli._

val reads = sc.loadPairedFastqAsFragments("sample_1.fq", "sample_2.fq")

val bwaArgs = new BwaArgs()
bwaArgs.sample = "sample"
bwaArgs.indexPath = "/data/hs38DH.fa"
bwaArgs.useDocker = true
//bwaArgs.useSingularity = true

val alignedReads = new Bwa(bwaArgs, sc).apply(reads)
#!/bin/bash

cannoli-submit \        
  interleavesFastq \    
  sample_1.fq \       
  sample_2.fq \       
  sample.ifq

cannoli-submit \        
  bwa \              
  --index /data/hs38DH.fa \  
  --use_docker \        
  # --useSingularity \   
  sample.ifq \         
  sample.alignments.adam \  
  sample
Big Data Genomics Contributors

- **UC Berkeley**: Matt Massie, Timothy Danford, André Schumacher, Jey Kottalam, Karen Feng, Eric Tu, Alyssa Morrow, Niranjan Kumar, Ananth Pallaseni, Michael L. Heuer, Frank Nothaft, Justin Paschall, Taner Dagdelen, Devin Petersohn, Anthony D. Joseph, Dave Patterson
- **Mt. Sinai**: Arun Ahuja, Neal Sidhwaney, Ryan Williams, Michael Linderman, Jeff Hammerbacher
- **GenomeBridge**: Carl Yeksigian
- **Cloudera**: Uri Laserson, Tom White
- **Microsoft Research**: Ravi Pandya, Bill Bolosky
- **UC Santa Cruz**: Benedict Paten, David Haussler, Hannes Schmidt, Beau Norgeot, Audrey Musselman-Brown, John Vivian
- And many other open source contributors, especially Neil Ferguson, Andy Petrella, Xavier Tordior, Deborah Siegel, Denny Lee
- Total of >75 contributors to ADAM/BDG from >12 institutions
Thank you!

- [https://github.com/bigdatagenomics/cannoli](https://github.com/bigdatagenomics/cannoli)

Demo D06 at 3 pm