CDER Scientific Computing
Genome Analysis Toolkit (GATK) Pipeline
On
CDRH HPC Cluster

SCD20
September 30, 2020
Xiaoqing He
CDER GATK Pipeline Programming Platform is specifically designed to promote Next Generation Sequencing (NGS) analytics on High Performance Computing (HPC) cluster.

### Overview of GATK Pipeline Programming Platform

- **User submits workflows written in WDL script to Cromwell, which parses the script and generates individual tasks and input parameters for execution via the scheduler.**
- **WDL scripts call GATK tools and non-GATK tools.**
- **Input file provides workflow parameters for execution.**

### WDL Script for GATK Workflow

- **WDL is a user-friendly scripting language.**
- **WDL scripts define tasks and call GATK tools to perform workflows.**
- **A sample WDL script for Sequence data format validation workflow: (validatebarn.Wdl)***

### Input JSON File

- Provides input variables to the commands in WDL script, may include the following:
  - file names,
  - path to the input data,
  - the parameters for the GATK tools,
  - any other resources that are needed for the workflows.
- Allows you to be able to customize input values from run to run.
- If input variables were hard-coded in WDL scripts, you will have to make a new copy and edit the inputs each time you run your script on a new batch of data—which undermines the advantages of setting up a pipeline.

### Cromwell SGE Backend

- **Cromwell provides a way to configure SGE backend.**
- **It launches jobs using qsub command.**
- **It supports runtime attributes configurations, such as queue, cpu and memory-unit.**

### Pre-Data Processing for Variant Discovery Workflow

- **Purpose:** To read unmapped sequencing data and process the data for variant discovery analysis.
- **Workflow Name:** processing-for-variant-discovery-gatk4
- **Workflow Requirements:**
  - Pair-end sequencing data in unmapped BAM (uBAM) format.
  - One or more SAM or BAM files to validate.
  - An explicit request of either SUMMARY or VERBOSE mode.
- **Outputs:**
  - Set of .txt files containing the validation reports, one per input file.

### GATK Best Practice Workflows

- The GATK Best Practices provide step-by-step recommendations for performing variant discovery analyses.
- There are several different GATK Best Practices workflows tailored to particular applications depending on the type of variation of interest and the technology employed.
- These are dependent on many factors including sequencing technology and the hardware infrastructure, so one may need to adapt their recommendations to one specific situation.
- The following GATK best practice workflows have been uploaded to CDER GATK pipeline project repository:
  - gatk-exome-analysis-pipeline
  - gatk-germline-snp-indels
  - gatk-somatic-snp-indels
  - gatk-somatic-cnv
  - seq-format-conversion
  - gatk-mapp-gemline-snp-indels
  - gatk-germline-variant-calling
  - gatk-exome-calling
  - gatk-germline-calling
  - gatk-germline-variant-calling
  - gatk-somatic-calling
  - gatk-exome-calling

### Pipelineing GATK With Cromwell/WDL

- **Cromwell** is an open sourced Workflow Management System geared towards scientific workflows.
- **Used by Broad Institute GATK Best Practice Pipelines.**
- **Executes tasks defined in WDL scripts.**
- **Uses stdin JSON file to fill in the values of inputs to commands in WDL script where appropriate.**
- **Cromwell has a way to configure Sun Grid Engine (SGE) backend relying on CDIR Betsy High Performance Computing (HPC) frameworks, and with access to General Parallel File System (GPFS).**
- **Cromwell tracks the execution of workflows.**

**Command to run Cromwell:**
```
java -jar cromwell.jar run workflow.wdl --inputs.json
```