Genome Analysis Toolkit (GATK) 2017

Overview

The Genome Analysis Toolkit (GATK) is a set of programs developed by the broad institute with an extensive website. As mentioned in the final presentation, it has the ability to perform much of the analysis required for calling genomic variants as well as many many other things. Why you may ask yourself did this magical tool only appear on the final day of the class? GATK uses read mappers, read aligners, variant callers, and all the other things (or similar things) that you have been introduced to throughout the course so we actually have been teaching you what you needed to know in smaller more digestible chunks.

This tutorial is quite small and does not showcase but the smallest drop in a bucket of what GATK is capable of doing. This is because the broad itself has developed many many tutorials for all the different things GATK does and extensive forums are available if the tutorials are not enough to get you through what you are trying to do. Finally, as the makers of the software they have put out and maintain what they regard as the best way to use their product in the form of 'best practices'. If you are going to use GATK, its a real real real good idea to make sure you are following their best practices because that is a situation where people will raise a big eyebrow if you say you are going against the flow.

While GATK is great, one stop shops often are often not the best at everything they do, don't be afraid to use other programs.

Objectives

1. Load GATK on lonestar
2. Use the sample data provided by the broad (through TACC) to verify that TACC is working
3. Explore a little of what is under the hood.

Tutorial: Loading GATK

While you may correctly think based on the overview that GATK is an obvious choice for a module on TACC, you may be surprised to learn that last year this was not the case, and that installation on TACC could be quite difficult. Luckily it is now a module and will remain one. In several of the previous tutorials we had to load additional programing languages (such as perl) in order to use specific scripts or programs. Like the velvet assembly tutorial, GATK requires **java.** Load the required modules.

Click here for the answer. If you didn't have the answer before reading it, please raise your hand by this point in the course you should have gotten these 2 lines.

```
module load java
module load gatk
```

Getting sample data

One of the module commands that we have not previously emphasized other than in passing is the module spider command. In addition to giving you a general description of what the module is and does, it provides you information of where the files are stored within TACC. Use the module spider command to see if you can copy the test files to a new folder on scratch with our existing naming structure and check what is there.

TACC (and bash) store things as variables in all capital letters, and bash requires a **$** in front of a variable name to know that what is next is a variable.

Click here to check your answer

```
cds
mkdir BDIB_gatk
cp $TACC_GATK_RESOURCES/* BDIB_gatk
cd BDIB_gatk
ls
```

Assuming you copied everything over correctly you should see the following 5 files.

```
exampleBAM.bam
exampleBAM.bam.bai
exampleFASTA.dict
exampleFASTA.fasta
exampleFASTA.fasta.fai
```

Tutorial: Testing GATK is working

As mentioned, gatk is a java program. As such it must be invoked with the java -jar command. From **module spider gatk** you may have noticed a directory containing not only the resources you just copied but also a .jar file. Unfortunately the entire path to that .jar file must be supplied due to the nature of java, but after that the standard --help option can tell you how many different things gatk can do.

Command to display gatk help

```
java -jar /opt/apps/gatk/3.5.0/GenomeAnalysisTK.jar --help

# alternatively as a bit of a shortcut (but not much of one)
java -jar $TACC_GATK_DIR/GenomeAnalysisTK.jar --help
```

If you see 316 lines of a long scrolling output detailing some copyright information and a bunch of different commands everything is correctly loaded. While individual tools will require different options and the program itself takes many different options only 3 things are **ALWAYS** required:

flag	Description
-T	Tool name, what tool are you trying to use
-R	Reference sequence file
-I	Input bam file

Stealing a nice mnemonic devices from a GATK toturial (which is condensed below), these 3 arguments don't have to be in this order, but if you learn them in this order, you will be able to remember them if you **TRI.** Remember, specific tools will require additional arguments.

Tutorial: Use GATK to count the number of reads in a bam file

Using the above information we will use the **CountReads** tool to count the number of reads in the **exampleBAM.bam** file which was from the **exampleFASTA.fasta** reference file. pay attention to the the words in bold and the table/discussion in the previous tutorial section and see if you can figure out how to do this on your own.

Don't forget you will still need to start your command with **java -jar /opt/apps/gatk/3.5.0/GenomeAnalysisTK.jar** to envoke java and gatk before specifying your arguments.

Click here for the solution

```
java -jar /opt/apps/gatk/3.5.0/GenomeAnalysisTK.jar -T CountLoci -R exampleFASTA.fasta -I exampleBAM.bam
```

INFO 13:05:20,542 HelpFormatter - ---

INFO 13:05:20,543 HelpFormatter - The Genome Analysis Toolkit (GATK) v3.5-0-g36282e4, Compiled 2015/11/25 04:03:56

INFO 13:05:20,543 HelpFormatter - Copyright (c) 2010 The Broad Institute

INFO 13:05:20,543 HelpFormatter - For support and documentation go to http://www.broadinstitute.org/gatk

INFO 13:05:20,546 HelpFormatter - Program Args: -T CountLoci -R exampleFASTA.fasta -I exampleBAM.bam

INFO 13:05:20,552 HelpFormatter - Executing as ded@nid00315 on Linux 3.0.101-0.35.1_1.0502.8640-cray_ari_c amd64; Java HotSpot(TM) 64-Bit Server VM 1.8.0_77-b03.

INFO 13:05:20,552 HelpFormatter - Date/Time: 2017/05/24 13:05:20

INFO 13:05:20,552 HelpFormatter - ---

INFO 13:05:20,552 HelpFormatter - ---

INFO 13:05:20,892 GenomeAnalysisEngine - Strictness is SILENT

INFO 13:05:20,947 GenomeAnalysisEngine - Downsampling Settings: Method: BY_SAMPLE, Target Coverage: 1000

INFO 13:05:20,952 SAMDataSource$SAMReaders - Initializing SAMRecords in serial

INFO 13:05:20,967 SAMDataSource$SAMReaders - Done initializing BAM readers: total time 0.01

INFO 13:05:21,026 GenomeAnalysisEngine - Preparing for traversal over 1 BAM files

INFO 13:05:21,035 GenomeAnalysisEngine - Done preparing for traversal

INFO 13:05:21,035 ProgressMeter - [INITIALIZATION COMPLETE; STARTING PROCESSING]

INFO 13:05:21,036 ProgressMeter - | processed | time | per 1M | | total | remaining

INFO 13:05:21,036 ProgressMeter - Location | sites | elapsed | sites | completed | runtime | runtime

2052

INFO 13:05:21,090 ProgressMeter - done 2052.0 0.0 s 26.0 s 97.3% 0.0 s 0.0 s

INFO 13:05:21,091 ProgressMeter - Total runtime 0.06 secs, 0.00 min, 0.00 hours

INFO 13:05:21,093 MicroScheduler - 0 reads were filtered out during the traversal out of approximately 33 total reads (0.00%)

INFO 13:05:21,093 MicroScheduler - -> 0 reads (0.00% of total) failing BadCigarFilter

INFO 13:05:21,093 MicroScheduler - -> 0 reads (0.00% of total) failing DuplicateReadFilter

INFO 13:05:21,093 MicroScheduler - -> 0 reads (0.00% of total) failing FailsVendorQualityCheckFilter

INFO 13:05:21,094 MicroScheduler - -> 0 reads (0.00% of total) failing MalformedReadFilter

INFO 13:05:21,094 MicroScheduler - -> 0 reads (0.00% of total) failing NotPrimaryAlignmentFilter

INFO 13:05:21,094 MicroScheduler - -> 0 reads (0.00% of total) failing UnmappedReadFilter

INFO 13:05:21,698 GATKRunReport - Uploaded run statistics report to AWS S3

What in all that are we actually looking for you might ask?

```
INFO  13:05:21,093 MicroScheduler - 0 reads were filtered out during the traversal out of approximately 33 total reads (0.00%)
```

This tells us that the bam file contains 33 total reads and that none were removed by any filtering options. The lack of anything being removed should make sense since we didn't try to filter anything out. As mentioned this is a very small introduction to GATK adapted from one of the broad's tutorials which can be found here http://gatkforums.broadinstitute.org/gatk/discussion/1209/howto-run-the-gatk-for-the-first-time. Feel free to explore that link and the other tutorial links for taking GATK further.

Return to GVA2017 home page.