IN-BIOS(9/5)000 Home exam (Autumn 2018): Variant calling module

Overview
##

TITLE: Variant calling of sample NA12877 (restricted to the first 3 MBp of chromosome 5) with two variant callers

GOAL: The overall goal is for you to perform variant calling and interpretation of a whole genome sample. This whole genome sample was sequenced on Illumina with 100 bp single reads. The variant calling should be done with both the UnifiedGenotyper (UG) and the HaplotypeCaller (HC).

You are provided with the required data in exactly the same environment that you had during the course. You are also provided with a “draft” file containing a definition of where key files are located (050_exam.bash). You do not have to use the variables if you do not wish to. What you need to do is:
* PART 1: Make use of the exercises in the course to generate the files you need to perform the analysis. Note that you will need to make some adjustments with regards to the names of input and output files (and perhaps variables) when you combine commands from different exercise pipelines (see below for extra advice).
* PART 2: Prepare a report that you will present orally. You should provide an overview of the analysis that you performed in the script and the answers to a set of specific questions (see further down). You must be able to present the essence of the report in 10-15 minutes.

With regards to BAM file metrics, you only need to compute the coverage metrics (do not perform alignment, insert size, or duplication metrics). Compute this statistic both for the entire region first 3 MBp of chromosome 5 and for the exons in this region (the necessary files are defined in the "draft" file).

Part 1: Performing the analysis
##

1. Log into the VM and open a terminal.

2. Make sure you have all the files you need by running:
source /share/inf-biox121/data/vc/exerDefinitions/setupEnv.bash
/share/inf-biox121/data/vc/exerDefinitions/copyFiles.bash

3. There are two files in /share/inf-biox121/data/exam/vc:
050_exam.txt that are the exam questions (i.e. this file)
050_exam.bash that contains a script with variables defining the files you need. It has an identical format to the exercises we did on the course.

Advice:
* When re-using code from the exercises, be very careful to make the appropriate modifications so that you carry out the correct analysis in relation to the tasks set in the exam.
* You will also need to be careful with file naming as you are asked to use both the UnifiedGenotyper and HaplotypeCaller variant callers. If you are not careful, you might overwrite the output of one caller with the output of the other.
* Sometimes a command produces more than one output file: make sure that you adjust both of the output file names.
* When doing the variant calling, the code will run a lot faster if you limit the calling to the region for which we have data. You can do this by using "-L 5:1-3000000"
* Remember that you can use ls –lrt to get a listing of files in the order they were produced which is useful for identifying the files that were most recently produced.
* If you do not feel comfortable manipulating VCF files in the VM, you can export them (for example using google docs) and work with them in a spreadsheet.
* You should contact me (timothy.hughes@medisin.uio.no), if you get stuck with technicalities of command execution (after you have had a decent go at solving the problem yourself).

IMPORTANT: In the process of using the files generated by the commands, make sure you understand the commands. This means understanding:
* What the command does
* Why the command is necessary
* How to interpret the output.
* Many of the tools that are used in the script belong to the GATK suite that has excellent documentation:
https://www.broadinstitute.org/gatk/guide/tooldocs/
* For extra details on hard filtering of variants:
https://gatkforums.broadinstitute.org/gatk/discussion/11069/hard-filtering-germline-short-variants
* At the presentation, you may be asked to give an explanation for specific parts of the code that you used. If you are unsure of what a piece of code does, you should return to the exercises that we did during the course where you will find explanations in the exercise comments. You will find a list of the exercise files here: /share/inf-biox121/data/vc/exerDefinitions

Part 2: Producing a report on variant calling
##################################

Using the data that you generated in PART 1 and other further manipulation of the output files you may deem necessary, you should produce a **short** report on the variant calling of this sample. You should present an overview of the computations you executed. You can also include screenshots from IGV if you wish to illustrate some of the concepts.

SPECIFIC QUESTIONS (that need to be addressed in addition to the general report)

1. HC and UG comparison:
* Produce two tables (one for SNPs and one for indels) detailing the numbers found with the different variant callers both before and after variant filtering.
  Remember that:
  1. The VCF file has a header that needs to be removed (on the command line to return lines that do not start with "#": grep -v "^#" yourFile.vcf)
  2. The filtering step does not actually remove the variants: all input variants are in the output, but the FILTER column is changed to indicate either PASS or the name of the filter that is failed (hint: grep "PASS")
* Load into IGV the BAM file and the HC and UG files (both the SNP files and the indel files). Variants that do not pass the filters will appear in a lighter shade (make sure that when you right click on the variant track name, "Suppress filtered sites" is not selected). Use a viewing window of 100 kbp and scroll through the first 3000 kbp of chr5. Find 5 locations where HC detects an indel but UG does not AND where coverage is above 10X AND where there are differences in the SNPs called in this region by HC and UG. Zoom in on each of the regions (window of a few hundred bps) and take a screenshot.
* Use the tables you produced above and your visual inspections to provide an explanation of why you see these different results with the different callers.

2. Finding and interpreting specific variants:
* Give an example of a UG variant that did not pass filtering.
* Find two SNPs one with high quality and one with low quality, and include the full VCF record for each SNP in the report.
* Of the indels called with HC that PASS filtering, what is the indel with the highest quality and what is the individual's genotype at this position?
* What is the biggest indel detected by UG. Does it pass the variant filters? Is it an insertion or a deletion? What is the genotype at this site?