EXAM INFBIOX120 (Autumn 2017): Variant caling module

Overview ##

TITLE: Variant calling of sample NA19238 captured with Agilent exome capture kit (restricted to chromosome 5)

GOAL: The overall goal is for you to perform variant calling and interpretation of this exome capture sample.

You are provided with the required data in exactly the same environment that you had during the course for performing exercises. 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 probably need to make some adjustments with regards to the names of input and output files when you combine commands from different exercise pipelines.

* PART 2: Prepare a report that you will present orally. This report should explain:
1. What the commands in the script do
2. Why these computations are necessary
3. What the key outputs are and, in the case of metric calculations, how you use the metrics to determine the quality of the data generated

You are free to include what information you like in the report as long as you respect two constraints:
1. You must be able to present the essence of the report in 10-15 minutes
2. You must include in the report the answers to a set of questions (see further down)

Part 1: Running the commands ##

1. Log into the VM and open a terminal.

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

3. There are two files in ~/vc/exerDefinitions:
050_exam.docx that are the exam questions (ie this file)
050_exam.bash that contains a script with the commands you need. It has an identical format to the exercises we did on the course.

Remember that you can use **ls –rt** to get a listing of files in the order they were produced which is useful for identifying the files that were most recently produced.

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).

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 and what metrics you looked at to evaluate the data. You can also include screenshots from IGV if you wish to illustrate some of the concepts.

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/

Specific questions that need to be addressed in addition to the general report
* What is sample NA19238? Can you find out any information about this sample?
* What percentage of bases in the capture have at least 10X coverage?
* What percentage of PCR duplication do we have in this sample?
* How many indels do you find pre-filtering? Remember that the VCF file has a header that needs to be removed (on the command line: grep -v "^#" yourFile.vcf)
* Give an example of a 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.
* For both SNPs and indels, how many pass filtering? 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.
* Of the indels 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 that passes filtering? Is it an insertion or a deletion?
* OPTIONAL QUESTION: The individual that provided this sample is not known to have any pathological conditions, but can you find any variants that might be considered disease causing? In order to answer this question, you will need to use the relevant section of 040_functionalAnnotation.bash code.

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.

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.
ls /share/inf-biox121/data/vc/exerDefinitions