Assignment #6-1

Due: 11:59pm, Tuesday, Nov 21, 2017

1. In the previous variant calling lab (check the lecture slides), you obtained the vcf file using bcftools. The VCF file contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. Meta-information is included after the ## string and the header line starts with #CHROM.

   Awk, sed, and bash are command-line tools that provide tremendous power for managing and manipulating text files of arbitrary size, from very small to extremely large.

   Let us suppose that you want to extract and filter the VCF results by the QUAL (phred-scaled quality) field (6th column) where the QAUL value is greater than or equal to 40. Be careful! In the filtered VCF file, all meta-information lines and header like (both start with # character) should be kept. You can work this using the awk with one line. How to do it? The, what's the number of variants compared to the original? Provide the one line command using awk and the short comparison of variants numbers.

   Ref:
   http://williamslab.bscb.cornell.edu/?page_id=235
   http://williamslab.bscb.cornell.edu/?page_id=287
   http://www.thegeekstuff.com/2010/02/awk-conditional-statements
   http://www4.ncsu.edu/~rosswhet/BIT815/Overview/Awk_Sed_Bash_Exercises.html