Projektni zadaci

Prijava projekta se može obaviti slanjem e-maila na vladimir.kovacevic@sbgenomics.com. Obavezno je navesti ime, prezime i broj indeksa. Rok za izradu projekta i seminarskog rada je **mesec dana** od dana prijave. U tom roku je potrebno poslati sve tražene rezultate na gore navedeni e-mail. Projektni zadaci su predviđeni za maksimalno dva studenta. Moguće je prijaviti samo zadatke koji nemaju ime uz sebe (one koji nisu rezervisani).

Za sve što nije egzaktno definisano tekstom projekta uvesti logičnu pretpostavku. Slajdovi, prezentacija i komentari u programskom kodu poželjno je da budu na engleskom, ali je i srpski jezik u potpunosti prihvatljiv. **Poželjno je da se u toku razvoja projekta sav programski kod čuva na Github-u ili drugom sistemu za verzionisanje koda.** Budite kreativni!

Zadatak 4

Project task is consisted of wrapping GATK4 CNNScoreVariants deep-learning based tool for variant filtering in CWL and running it with test samples. This project will be done on Cancer Genomics Cloud platform where anyone with academic email address can register and receive $300 free credit. The process of adaptation (wrapping) and running should be done in a similar way as described in quickstart and in the tutorials (pages: YOUR TOOL, TOOL WRAPPING TIPS AND TRICKS, RUN AN ANALYSIS). The Docker image for the tool is available: images.sbgenomics.com/vladimirk/gatk:4.1.0.0 and tool can be run from it with the command: `/gatk/gatk CNNScoreVariants`
All parameters of the tool need to be supported (added as inputs) to be adjustable in the task. Successful end of this part of project considers completing filtering task on CGC platform with input VCF and BAM file from Public reference files. (10 points)
Second part of the project assignment is benchmark of the filtering results. The tool will be run on the set of provided samples with available truth set (HG001-HG007 + CHM11-CHM13). For all of these samples precision, recall and f-score should be calculated using all provided VCF Benchmark workflow, truth-set VCFs and confidence BED regions for each of the samples. The precision, recall and f-score results for all samples, for filtered and non-filtered variants should be added to the comparison table and bar diagrams should to be created. (15 points).
Create slides (Google Slides or Power point presentation) with summarized work being done and record a 5-10 minute (upload to Youtube - optional) video which will present it to potential audience. (15 points)
The project will be done in a cooperation with Vladimir Kovacevic, who will be informed on the progress of the project and answer to all relevant questions or concerns.