Genomic Data Science with Galaxy Assignment

Peer-graded Assignment: Course Project

**Instructions**

The zip file [fastq\_bundle.zip](https://d396qusza40orc.cloudfront.net/gengalaxy/data/fastq_bundle.zip) contains six fastq files. These files contain targetted re-sequencing data for a father mother and daughter trio (identified as NA12877, NA12878, and NA12880 respectively). The data consists of raw reads from an Illumina MiSeq sequencer seuenced as paired ends (R1/R2) to 125bp in length.

Create a Galaxy workflow to identify polymorphic sites in all three individuals. Your workflow will need to map the three sets of paired reads to the appropriate reference genome. You will then need to use a variant caller to identify sites that appear to have strong support for the presence of a polymorphism, and call the genotype at that site for each sample.

You should report your results in VCF (variant call format). You should only include sites where the chance of a false positive call is 1 in 10,000 or better according to the VCF qual field.

Using your resulting VCF determine 1) the number of single nucleotide variants, 2) the number of insertion/deletion variants, 3) the number of multi-necleotide variants, 4) the number of variants with multiple alternate alleles, and 5) the names of the 5 genes with the largest number of polymorphic sites.

Submit the following

1. A short write-up (maximum 300 words) describing your results including the information requested above
2. The exported Galaxy workflow (a .ga file)
3. Your VCF file of filtered variants (a single .vcf file)

**Review criteria**

When you and your classmates evaluate one another's course projects, you will use the following rubric.

Write Up

Please provide feedback about the submitted write-up. Point out strengths and areas for improvement.

Has the student submitted a write-up? 0 = No 1 = Yes

Does the write-up accurately report the number of single nucleotide variants? 0 = No 1 = Yes

Does the write-up accurately report the number of insertion/deletion variants? 0 = No 1 = Yes

Does the write-up accurately report the number of multi-nucleotide variants? 0 = No 1 = Yes

Does the write-up accurately report the number of variants with multiple alternate alleles? 0 = No 1 = Yes

Does the write-up accurately report the names of the 5 genes with the largest number of polymorphic sites? 0 = No 1 = Yes

Workflow file

Please provide feedback about the submitted workflow file. Point out strengths and areas for improvement.

Has the student submitted a workflow file? 0 = No 1 = Yes

Does the workflow file identify polymorphic sites in all three individuals? 0 = No 1 = Yes

Does the workflow file map the three sets of paired reads to the appropriate reference genome? 0 = No 1 = Yes

Did the student use a variant caller to identify sites that appear to have strong support for the presence of a polymorphism and call the genotype at that site for each sample? 0 = No 1 = Yes

VCF file

Please provide feedback about the submitted VCF file. Point out strengths and areas for improvement.

Has the student submitted a VCF file? 0 = No 1 = Yes

Does the VCF file include only sites where the chance of a false positive call is 1 in 10,000 or better according to the VCF qual field? 0 = No 1 = Yes

**PLEASE NOTE: Four additional 1-point criteria will also be included to assess the accuracy of your VCF. To maintain the integrity and rigor of the assignment, we are keeping those criteria hidden until after the submission phase is over.**