due in 2day 18h

Submission Phase

1. Do assignment ☑ (/gengalaxy-003/human_grading/view/courses/975240/assessments/3/submissions)

Evaluation Phase

2. Evaluate peers 🔒 (/gengalaxy-003/human_grading/view/courses/975240/assessments/3/peerGradingSets)

Results Phase

3. See results 🔒 (/gengalaxy-003/human_grading/view/courses/975240/assessments/3/results/mine)

Your work was submitted. Review your work (https://class.coursera.org/gengalaxy-003/human_grading/view/courses/975240/assessments/3/submissions/45) to make sure everything looks OK. X

✔ Submitted. You can still make changes and re-submit before the deadline.
☑ In accordance with the Honor Code, I certify that my answers here are my own work, and that I have appropriately acknowledged all external sources (if any) that were used in this work.

Re-submit for grading

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: A short write-up (maximum 300 words) describing your results including the information requested above, along with two additional files:

- The exported Galaxy workflow (a `.ga` file)
- Your VCF file of filtered variants (a single `.vcf` file)

Use the text box below to submit a short write-up describing your results including the information requested above.

B | *I* | Link | <code> | Math | Edit | Preview

In this analysis, the sequences are mapped to reference genome hg19 using BWA-MEM tool (version: 0.1) and the mapping is merged with MergeSamFiles (version: 1.126.0) and cleaned with Filter (version: 1.1.0, remove low quality mapping), MarkDuplicates (version: 1.126.0, filter out duplicated mapping), CleanSam (version: 1.126.0)(see workflow for details). There are 2666 polymorphic sites are detected with a false positive call is 1 in 10,000 or better according to the VCF qual field.
Based on the VCF file and VCFfilter (version: 0.0.3) and Filter (version: 1.1.0) tools (see workflow for details), the following results were obtained:
1) the number of single nucleotide variants: 2328
2) the number of insertion/deletion variants: 277
3) the number of multi-necleotide variants: 23
4) the number of variants with multiple alternate alleles: 62

Then, the ANNOVAR Annotate VCF tool (version: 0.1) together with Group (version: 2.1.0) and Sort (version: 1.0.3)tools were used to identify the top 5 genes with the largest number of polymorphic sites. The genes and the number of polymorphic sites were listed below:

Rank	Gene	number of sites
1	RBFOX1	161
2	CACNA1H	55
3	USP7	48
4	ABAT	44
5	CLCN7	39
5	UNKL	39

Words: 194 / 300

Upload your workflow file (.ga) by using the button below.

B | I | Link | <code> | Math | Edit | Preview

Galaxy-Workflow-Workflow_constructed_from_history__Genomic_Data_Science_with_Galaxy_Project_-_completed_

Attach a file *(supports: txt, png, jpg, gif, pdf, ga, vcf)*

Upload your variation data file (.vcf) by using the button below.

B | I | Link | <code> | Math | Edit | Preview

Genomic Data Science with Galaxy Project - identify polymorphic sites

Attach a file *(supports: txt, png, jpg, gif, pdf, ga, vcf)*

✔ Submitted. You can still make changes and re-submit before the deadline.

☑ In accordance with the Honor Code, I certify that my answers here are my own work, and that I have appropriately acknowledged all external sources (if any) that were used in this work.

Re-submit for grading