Somatic and Germline variant identification in Tumor and Normal sample

Suchitra Thapa

Team: Rosalind

Objective:

To study the provided tutorial on variant identification and reproduce it either using galaxy platform or google cloud shell.

Background:

Tumor cells are variant of the normal cell and therefore comparative genomic analysis of these two cells can provide a detail spectrum of mutation in these cells. Since there are different types of mutation, selection of sample is crucial in the comparison. Normally, for germline mutation (inherited), comparison of tumor cell with a reference genome is enough but in case of somatic mutation (after birth) , comparison of tumor cell with the same person normal cell is mandatory. Therefore, this tutorial is identifying both the somatic and germline variants in the exome sequence of normal and tumor cell from the same person.

Workflow:

The user can choose either Linux terminal or Galaxy platform to complete the tutorial. This report is for the Galaxy platform tutorial and the entire history of the task can be found here.

- *Data Acquisition:*
  The dataset was paired end data from Normal and Tumor tissue. All four sequence file was uploaded via upload data in the Galaxy homepage following the paste /fetch instruction for web link.

- *Quality control and reads mapping*
  The read quality was determined by using FastQC tools and MultiQC for combined quality report. Since the dataset seems quite of a good quality but had some adapters. Trimmomatic tool was used to trim out the adapters. Further FastQC and MultiQC were run for the trimmed dataset to confirm the final quality. The trimmed reads were proceeded to alignment with the reference genome using BWA MEM tool.

- Post processing after mapping

The mapped reads were filtered with a BAM tools filter. Duplicates were removed by RmDup. Thereafter, BAM Left Align tool was used to left-align reads around indels . Then CalMD was used to recalibrate the read quality and finally BAM tools Filter was used for BAM dataset to ensure high quality mapping reads before variant calling.

- Variant calling and classification
  The variants in the high quality mapped reads were detected using the VarScan Somatic tools. The VCF output was used to classify the variants as germline, somatic and LOH.

- Variant annotation and reporting
  The called variants were annotated using SnpEff tools using the Homo sapiens: hg19 as a reference genome. Functional annotation by using SnpEff tool and genetic and clinical evidence based annotation by using GEMINI was done in the end.

Conclusion:
The tutorial was successfully reproduced using Galaxy platform.

Problem faced:
Except for some time delay for running the tools, all the instructions could be replicated as written.