Somatic and Germline variant identification in Tumor and Normal sample using Galaxy platform

Suchitra Thapa

Team: Rosalind

OBJECTIVE:

To study the provided tutorial on variant identification and reproduce it by using galaxy platform

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:

login to the Galaxy platform using the id and password is needed to complete the tutorial.
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

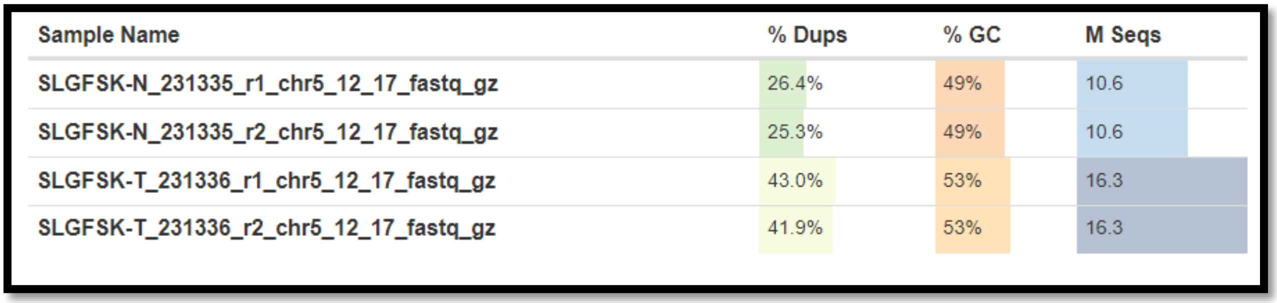
The mapped reads were filtered with a BAM tools filter. Duplicates were removed by RmDup. Thereafter, BAM Left Align tool was used to left-align reads around indels . Then CalMD was used to recalibrate the read quality and finally BAM tools Filter was used for BAM dataset to ensure high quality mapping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

RESULT:

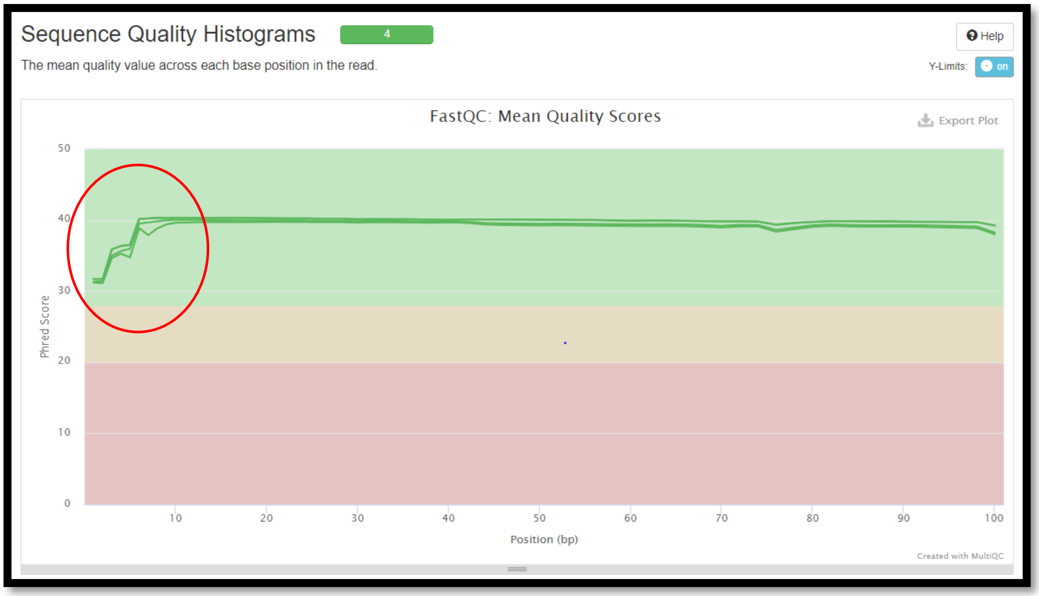
The dataset was exome sequence from Normal and Tumor cell. The data was a paired end sequence and contains a forward and reverse read file for each sample. The quality checking of each file showed that all the files are almost a good quality reads.



Sample Name	% Dups	% GC	M Seqs
SLGFSK-N_231335_r1_chr5_12_17_fastq_gz	26.4%	49%	10.6
SLGFSK-N_231335_r2_chr5_12_17_fastq_gz	25.3%	49%	10.6
SLGFSK-T_231336_r1_chr5_12_17_fastq_gz	43.0%	53%	16.3
SLGFSK-T_231336_r2_chr5_12_17_fastq_gz	41.9%	53%	16.3

MultiQC report before trimming

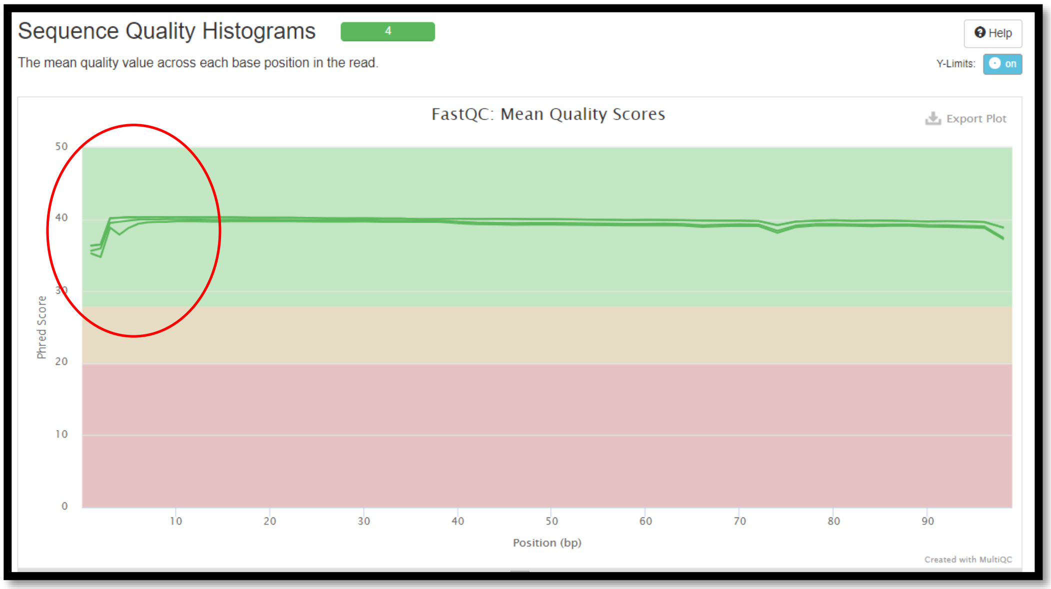
The combined analysis of the Fastqc output files using MultiQC showed the same result. However, trimming and filtering was done to improve the reads quality further. The figure below showed some improvement in the read quality after trimming (seen in red circle). But this trimming step can be skipped as the reads are already a good quality reads.



Sequence Quality Histograms 4

The mean quality value across each base position in the read.

Before trimming

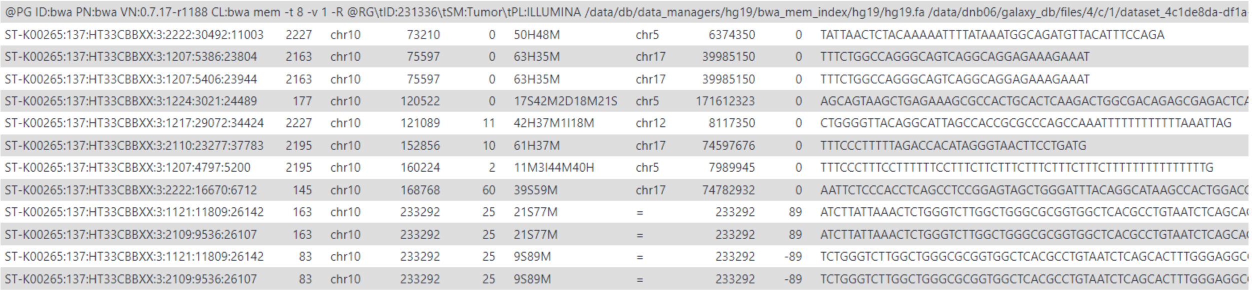


Sequence Quality Histograms 4

The mean quality value across each base position in the read.

After trimming

The reads file after the quality checking was mapped and a BAM file was generated which looked like below:



```
@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa mem -t 8 -v 1 -R @RG\tID:231336\tSM:Tumor\tPL:ILLUMINA /data/db/data_managers/hg19/bwa_mem_index/hg19/hg19.fa /data/dnb06/galaxy_db/files/4/c/1/dataset_4c1de8da-df1a
ST-K00265:137:HT33CBBXX:3:2222:30492:11003   2227   chr10      73210    0   50H48M          chr5     6374350    0   TATTAACTCTACAAAAATTTTATAAATGGCAGATGTTACATTTCCAGA
ST-K00265:137:HT33CBBXX:3:1207:5386:23804    2163   chr10      75597    0   63H35M          chr17   39985150    0   TTTCTGGCCAGGGCAGTCAGGCAGGAGAAAGAAAT
ST-K00265:137:HT33CBBXX:3:1207:5406:23944    2163   chr10      75597    0   63H35M          chr17   39985150    0   TTTCTGGCCAGGGCAGTCAGGCAGGAGAAAGAAAT
ST-K00265:137:HT33CBBXX:3:1224:3021:24489     177   chr10     120522    0   17S42M2D18M21S  chr5   171612323    0   AGCAGTAAGCTGAGAAAGCGCCACTGCACTCAAGACTGGCGACAGAGCGAGACTCA
ST-K00265:137:HT33CBBXX:3:1217:29072:34424   2227   chr10     121089   11   42H37M1I18M     chr12    8117350    0   CTGGGGTTACAGGCATTAGCCACCGCGCCCAGCCAAATTTTTTTTTTTAAATTAG
ST-K00265:137:HT33CBBXX:3:2110:23277:37783   2195   chr10     152856   10   61H37M          chr17   74597676    0   TTTCCCTTTTTAGACCACATAGGGTAACTTCCTGATG
ST-K00265:137:HT33CBBXX:3:1207:4797:5200     2195   chr10     160224    2   11M3I44M40H     chr5     7989945    0   TTTCCCTTTCCTTTTTTCCTTTCTTCTTTCTTTCTTTCTTTCTTTTTTTTTTTTTTTG
ST-K00265:137:HT33CBBXX:3:2222:16670:6712     145   chr10     168768   60   39S59M          chr17   74782932    0   AATTCTCCCACCTCAGCCTCCGGAGTAGCTGGGATTTACAGGCATAAGCCACTGGACC
ST-K00265:137:HT33CBBXX:3:1121:11809:26142    163   chr10     233292   25   21S77M          =         233292   89   ATCTTATTAAACTCTGGGTCTTGGCTGGGCGCGGTGGCTCACGCCTGTAATCTCAGCA
ST-K00265:137:HT33CBBXX:3:2109:9536:26107     163   chr10     233292   25   21S77M          =         233292   89   ATCTTATTAAACTCTGGGTCTTGGCTGGGCGCGGTGGCTCACGCCTGTAATCTCAGCA
ST-K00265:137:HT33CBBXX:3:1121:11809:26142     83   chr10     233292   25   9S89M           =         233292  -89   TCTGGGTCTTGGCTGGGCGCGGTGGCTCACGCCTGTAATCTCAGCACTTTGGGAGGC
ST-K00265:137:HT33CBBXX:3:2109:9536:26107      83   chr10     233292   25   9S89M           =         233292  -89   TCTGGGTCTTGGCTGGGCGCGGTGGCTCACGCCTGTAATCTCAGCACTTTGGGAGGC
```

Mapping doesnot create all mapped reads but has unmapped reads as well so the mapped reads were filtered for any such reads and only the retained mapped reads were further processed. Duplicates were removed.

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.