Identification of somatic and germline variants from tumor and normal sample pairs using linux terminal

Amarachukwu Okechukwu

Team: Rosalind

Background

The objective of the stage 2 task on HackBio internship is study the provided tutorial on somatic and germline variant identification and reproduce it by using linux terminal. Variant detection of somatic mutations is best performed when both tumor and normal (germline) samples are available. Sequence may be obtained from whole genome, exome capture, or small gene panels, where only the exons of clinically relevant genes are sequenced. Adequate sequencing depth is paramount to detecting somatic variants, especially when tumor heterogeneity is concerned. Selection of sample is crucial in the comparison as we want to understand the spectrum of mutations in both tissues. These mutations can be germline (inhrited) or somatic (acquired after birth), and a common kind of genetic mutation as a result of either is Loss of Heterozygosity (LOH). LOH usually leads to loss of one normal copy or a group of genes, which is a common even in cancer development. Germline mutations can easily be identified by comparing a sample genome to a reference while for somatic mutations, comparison of tumor cell with a reference genome is not enough to achieve this hence there is need to use the same person's normal cell. Therefore, this tutorial is tries to identify both the somatic and germline variants in the exome sequence of normal and tumor cell from the same person.

Workflow

The script is written in bash language and can be run on any linux terminal. The script can be found at https://github.com/AmarachukwuOkechukwu/Rosalind-team/blob/main/Stage_two.sh

- **Data Acquisition:** The sequencing reads we are going to analyze are from real-world data from a cancer patient's tumor and normal tissue samples. The samples (paired end) were two in number. Two folders were created one for the samples and the other the reference. All four sequence file was downloaded where the first two files represent the forward and reverse reads sequence data from a patient's normal tissue, and the last two represent the data of the same patient's tumor tissue. The reference genome hg19 was also downloaded in similar manner but unzipped since it is a compressed file.

- **Pre-processing and trimming:**

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. But before mapping the reference genome hg19 was indexed using bwa index command.

- **Post processing after mapping:**

After mapping, the ouput sam file was converted to bam file then the bam file was sorted and indexed using samtools sort and samtools index command respectively. The sorted reads file was filtered with samtools view. The bam files were viewed using samtoosl flagstat. Duplicates were checked using a combination of commands i.e. samtools collate, samtools fixmate, samtools sort and samtools markdup command. After confirming the duplicates, they were removed by samtools rmdup. Thereafter, bam left-align command was used to left-align reads around indels . Then samtools calmd command was used to recalibrate the read quality and finally bamtools filter (using <= 254 map quality parameter) was used for bam dataset to ensure high quality mapping reads before variant calling.

- **Variant calling and classification:**

The variants file was downloaded using. Then a pileup file was created with the reference file and the refiltered file. The variant calling was done using the varscan which created a vcf output file for each dataset. These vcf files were compressed and then indexed using tabix command. Further, they were merged using bcftools merge command.

- **Creating database for annotation/Variant annotation and reporting:**

snpEff command was used to create the database. For that snpEff zip file was downloaded and unzipped. The called variants were annotated with snpEff using the Homo sapiens: hg19 as a reference genome. Only functional annotation was done using SnpEff tool

Result

Mapping result

For 335:

```
Input Read Pairs: 10602766 Both Surviving: 10526288 (99.28%) Forward Only Surviving: 76478 (0.72%) Reverse Only Surviving: 0 (0.00%) Dropped: 0 (0.00%)

TrimmomaticPE: Completed successfully
```

For 336:

```
Input Read Pairs: 16293448 Both Surviving: 16093238 (98.77%) Forward Only Surviving: 200210 (1.23%) Reverse Only Surviving: 0 (0.00%) Dropped: 0 (0.00%)
```

Samtool filter result

```
samtools flagstat SLGFSK-T_231336.filtered1.bam

[W::bam_hdr_read] EOF marker is absent. The input is probably truncated 31383006 + 0 in total (QC-passed reads + QC-failed reads)

0 + 0 secondary

30349 + 0 supplementary

 0 + 0 duplicates

31383006 + 0 mapped (100.00% : N/A)

31352657 + 0 paired in sequencing

15676922 + 0 read1

15675735 + 0 read2

31352657 + 0 properly paired (100.00% : N/A)

31352657 + 0 with itself and mate mapped

0 + 0 singletons (0.00% : N/A)

0 + 0 with mate mapped to a different chr

0 + 0 with mate mapped to a different chr (mapQ>=5)

samtools flagstat SLGFSK-N_231335.filtered1.bam 9053504 + 0 in total (QC-passed reads + QC-failed reads)

0 + 0 secondary

1484 + 0 supplementary
```

```
0 + 0 duplicates

9053504 + 0 mapped (100.00% : N/A)

9052020 + 0 paired in sequencing

4526163 + 0 read1

4525857 + 0 read2

9052020 + 0 properly paired (100.00% : N/A)

9052020 + 0 with itself and mate mapped

0 + 0 singletons (0.00% : N/A)

0 + 0 with mate mapped to a different chr

0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Conclusion

To conclude, *somatic variant calling* tries to distinguish *somatic mutations*, which are private to tumor tissue, from *germline* mutations. The need for identification of somatic and germline variants was fully understood and the tutorial was reproduced successfully.

Challenges

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