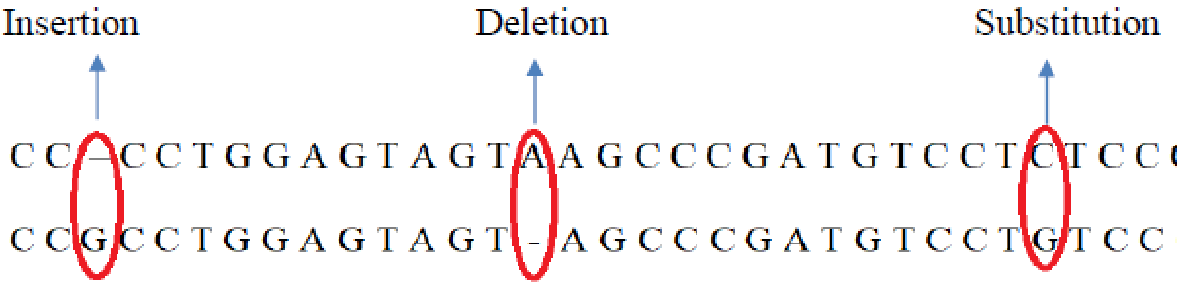
ALIGNMENT STRATEGY

Alignment: It is a genomic feature that maps between the letters of the two sequences, with some spacers (indels). We perform alignment to inhibit polymorphism, or the induced error introduced while sequencing. The alignment will take into account differences such as polymorphism and sequencing errors, and introns (for genes).


Insertion
Deletion
Substitution
CC—CCTGGAGTAGTAAGCCCGATGTCCTCTCC
CCGCCTGGAGTAGT-AGCCCGATGTCCTGTCC

In above sequences, sequence read is matching with its corresponding nitrogenous base but some places are empty known as insertions or deletions and some are wrongly filled known as substitution. These errors are corrected in the alignment phase.

- For alignment, we can use spliced aligner including Tophat2, HISAT, STAR tools/algorithms But in this study, we have used **Bowtie2** alignment tool. Because Bowtie2 is, time consuming and memory-efficient tool for aligning reads to long sequences or high-sequencing throughput. It is predominantly better at aligning sequences of about 50s to 100s of characters to relatively long mammalian genomes and it uses 3.2 gigabytes of RAM that makes more feasible to work it on GVL.
- Bowtie2 takes a bowtie2 index and a set of sequencing read files and outputs a set of alignments in SAM/BAM format.
- For short, alignment is the process by which we determine how much genetic fragment is similar to reference genome.

After performing alignment, the results are following:

1. For dataset sample one, SRR1554535

91185969 reads; of these: 91185969 (100.00%) were unpaired; of these: 8945201 (9.81%) aligned 0 times 56816398 (62.31%) aligned exactly 1 time 25424370 (27.88%) aligned >1 times **90.19% overall alignment rate**

2. For dataset sample second, SRR1554536

46971267 reads; of these: 46971267 (100.00%) were unpaired; of these: 2130239 (4.54%) aligned 0 times 17434738 (37.12%) aligned exactly 1 time 27406290 (58.35%) aligned >1 times **95.46% overall alignment rate**

3. For dataset sample third, SRR1554561

89061863 reads; of these: 89061863 (100.00%) were unpaired; of these: 12590443 (14.14%) aligned 0 times 57693153 (64.78%) aligned exactly 1 time 18778267 (21.08%) aligned >1 times **85.86% overall alignment rate**

4. For dataset sample fourth, SRR1554541

162403466 reads; of these: 162403466 (100.00%) were unpaired; of these: 17489969 (10.77%) aligned 0 times 112129877 (69.04%) aligned exactly 1 time 32783620 (20.19%) aligned >1 times **89.23% overall alignment rate**

5. For dataset sample fifth, SRR1554537

121992326 reads; of these: 121992326 (100.00%) were unpaired; of these: 13500649 (11.07%) aligned 0 times 83156647 (68.17%) aligned exactly 1 time 25335030 (20.77%) aligned >1 times **88.93% overall alignment rate Page 40**

6. For dataset sample sixth, SRR1554567

136312421 reads; of these: 136312421 (100.00%) were unpaired; of these: 14107633 (10.35%) aligned 0 times 94795954 (69.54%) aligned exactly 1 time 27408834 (20.11%) aligned >1 times **89.65% overall alignment rate**

NOTE: We have used Homo sapiens (hg19) genome for referencing the sample datasets.
The alignment is performed on GVL, to support the reusability or reproducibility we are providing the link of workflow for aligned sequence.

[Link: https://usegalaxy.org/u/ajay.ducs/h/genomic-data-science-capstone]

The output files are in BAM file format could be downloaded from above link to perform downstream analysis.
*Phenotypic information is given in phenotypic table.

[**Please Rate above workflow**]