1. The data was loaded into Galaxy (cloud), set to .fastqsanger format and hg19 database.
2. FASTQC was run on all datasets (optional).
3. Mapped the datasets using Map with BWA for Illumina (version: 0.5.9-r16).
4. Filtered using SAMTools (Filter SAM version: 1.0.0) for properly mapped pairs only.
5. Converted the SAM format to BAM using SAM-to-BAM (version: 2.1).
6. Added the read groups to the BAM files using Picard – AddorReplacereadgroups (version: 1.136.0).
7. Merged the 3 BAM files (Merge BAM File version: 1.2.0).
8. Used FreeBayes tool (version: 0.4.1) to generate a vcf file based on hg19 genome.
9. Filtered this file to select sites where the chance of false positive call is 1 in 10,000 or -f “QUAL > 40” (VCFfilter version: 0.0.3) and downloaded this vcf file for analysis.
10. Extracted the workflow.

**Analysis of the VCF file**

1. VCFfilter was again used on the vcf file to identify the different types of polymorphisms using the following filtering expressions:
   * 1. -f “TYPE = snp” **Single Nucleotide Variants = 1,977**
     2. -f “TYPE = ins” -o -f “TYPE = del” **Insertion/Deletion Variants = 202**
     3. Tried the ins and del separately **Insertion = 114 and Deletion = 93**
     4. -f “TYPE = mnp” **Multi Nucleotide Variants = 22**
     5. -f “TYPE = complex” **Multiple Alternate Alleles = 56**
2. The genes were identified by using the ANNOVAR Annotate VCF tool (version: 0.2) on the vcf file. Grouped on the Gene.refGene column and counted the polymorphic sites using the Group tool (version: 2.1.1). And finally, the grouped results were sorted in descending order using the Sort tool (version: 1.0.3). The 5 names of the genes with the largest number of polymorphic sites:

**GENE** **COUNT**

* + 1. **RBFOX1 146**
    2. **CACNA1H 84**
    3. **ABAT 50**
    4. **PKD1 44**
    5. **TPSB2 41**