Project: Identify DNA polymorphic sites

In the analysis, quality of sequencing data is checked by FASTQC at first. Second, the sequences are mapped to reference genome hg 19 using botwie2 (Version 2.3.4.2) and read groups are renamed by AddOrReplaceReadGroup (version: 1.126.0). Third, the mapping is merged with MergeSamFiles (version: 1.126.0), cleaned with Filter (version: 1.126.0) to remove low quality mapping, MarkDuplicates (version: 1.126.0) to filter out duplicated mapping and CleanSam (version: 1.126.0) to perform BAM grooming. Afterward, **FreeBayes** tool (version: 0.4) is used to identify polymorphic sites based on hg19 genome and VCFfilter (version: 0.0.3) is used to select sites where the chance of a false positive call is 1 in 10,000 or better.

According to the vcf file and the VCFfilter tool (version:0.0.3), the following results are the number of snp, mnp, del, ins or complex counted by excel:

1) The number of single nucleotide variants: 1917

2) The number of insertion/deletion variants: 204

3) The number of multi-nucleotide variants: 3

4) The number of variants with multiple alternate alleles: 48

Then, the ANNOVAR Annotate VCF tool (version: 0.1) together with Group (version: 2.1.0) and Sort tools (version: 1.0.3) are used to identify the top 5 genes with the largest number of polymorphic sites. The list is the genes and the number of polymorphic sites.

Rank	Gene	Number of sites
1	RBFOX1	146
2	CACNA1H	43
3	USP7	42
4	ABAT	37
5	CLCN7	35