Genomic Data Science with Galaxy - Course Project - Write-up

The project's aim was to identify polymorphic sites in the genomes (more exactly: three sets of paired Illumina reads) of a human father - mother - daughter trio.

After mapping the reads on the hg19 reference genome using BWA-MEM (v0.7.17.1), a quality analysis of the mapped reads, the addition of readgroups for each of the three individuals, filtering according to the map quality and the removal of duplicates, FreeBayes (v1.3.1) was invoked in order to identify sites which show strong support for the presence of a polymorphism.

Having obtained the VCF output from FreeBayes, VCFfilter (v1.0.0) was used to include only those sites where the chance of a false positive call is 1 in 10,000 or better.

The final VCF output yielded the following results:

Total number of polymorphisms: 2,644
Only SNPs: 2,306
Only multi-nucleotide polymorphisms (MNPs): 7
Only insertions: 126
Only deletions: 136
Only complex alleles: 85

Furthermore, 23 sites with multiple alternate alleles were identified.

In order to obtain information as to the genes containing the identified polymorphic sites, Annovar (v0.2) was used to annotate genes for each found variant. The 5 genes with the largest number of polymorphisms were found to be:

RBFOX1: 156 sites
CACNA1H: 50 sites
USP7: 47 sites
ABAT: 46 sites
CLCN7: 39 sites