Homework Exercise 10

1)

Go to ClinVar's FTP site and download their entire dataset in tab-delimited format. The file (***variant_summary.txt.gz***) should contains all known genetic variants in human with their clinical significance annotations (pathogenic, benign, or uncertain).

The ClinVar variants contain HGVS terms that describe the effects of the variants (in both gene and protein level, where relevant). For example, according to the HGVS standard, ***p.Gly1046Arg*** expresses a variant in a protein coding region that changes the 1046th residue in the protein from Glycine to Arginine.

A)

According to ClinVar's annotations, do amino-acid substitutions of pathogenic missense variants have different BLOSUM62 scores than benign variants? Explain the results.

B)

According to ClinVar, is there a connection between the number of variants recorded in a gene and the percentage of those variants that are pathogenic? How could such a connection have arisen? Suggest several potential explanations.