1. (2 points) Provided the following set of sequences, draw the EHH plots for each allele (a and A) (given that core SNP is at position 0) [by hand/in Paint/in R]. When drawing the plot, calculate the allele-specific EHH - that is, treat sets of sequences for each allele independently.

$$\mathrm{EHH}_{s,t} = \frac{1}{n_{a_s}(n_{a_s} - 1)} \sum_{k=1}^{K_{a_s,t}} n_k(n_k - 1)$$

a	A
012345	012345
AGCTAA	GAGTGA
AGCTCC	GTCCAA
AGGTAG	GACTGA
AGCTCT	GTGCAA
AACTAC	GAACAA

Which allele appears to be under positive selection?

2. (1 point) Let's mine some data in public resources. Using gnomAD browser (https://gnomad.broadinstitute.org/) find some data for your favorite protein-coding human gene. gnomAD provides summary gene constraint statistics at the top of gene page. Look at the observed and expected counts of synonymous, missense, and loss-of-function (pLoF) SNPs. Is your gene evolutionary conserved? What is the evidence for your conclusion?.

3. (5 points) dN/dS value is usually calculated in a rigorous fashion, comparing the observed and expected frequencies of synonymous and non-synonymous substitutions. Nevertheless, we can use the dN/dS logic to calculate a simpler value which I call f(N)/f(S) = sum($\mathrm{AF}_{\text{nonsynonymous variants}}$)/sum($\mathrm{AF}_{\text{synonymous variants}}$). This value can not be interpreted as dN/dS, but can still be used to rank genes with the highest selective pressure. You are provided with a file **ESP_SNP_data.tsv**, which contains filtered data about coding variant sites from the Exome Sequencing Project by NHLBI. Write a short custom script to evaluate f(N)/f(S) value for each gene (remove all genes without synonymous variants).

1) Draw the distribution of the f(N)/f(S) value across all genes. What is the mean value of this metric?

2) Provide a list of top-100 genes with lowest and highest values of the calculated metric. Which list contains (a) most and (b) least conserved genes?

3) Run MSigDB (http://software.broadinstitute.org/gsea/msigdb/annotate.jsp) analysis using each list of genes (separately) against canonical pathway (CP) list. What are the functional terms associated with highest and lowest conservation? Can you give any biological reasoning for this observation?