**HW3, finding eQTLs of immunoglobulin genes**

The goal of this homework assignment is to learn techniques for finding eQTLs of antibody repertoires. To complete this assignment, perform the following steps:

1. Download a [dataframe](https://docs.google.com/spreadsheets/d/1gdmu6LhfaJ0Rzjj6ZrLsppdep0oqcAL8EHWDAr-J4RY/) containing usage values of gene IGHV1-2 collected across 85 healthy individuals. Usage values are provided in the “Usage” column. For each individual, haplotypes of IGHV1-2 were also computed and written to the “Haplotype” column. Haplotypes are described by IDs of alleles of IGHV1-2. For example, while a homozygous haplotype of individual 2 is described by allele IGHV1-2\*04, a heterozygous haplotype of individual 1 is described by two alleles: IGHV1-2\*02 and IGHV1-2\*06.
2. For each unique haplotype, compute the number of individuals representing it and the mean usage of IGHV1-2. Fill Table 1 (add rows if needed):

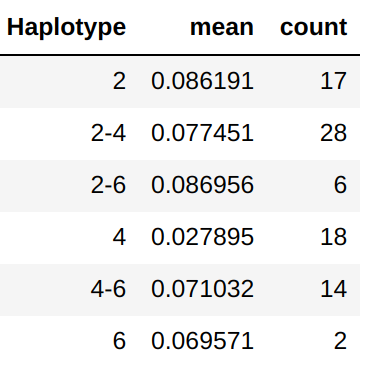


Table 1.

1. For each pair of haplotypes (H1, H2), compare their usages (U1 and U2) and compute a p-value showing the probability that U1 and U2 have the same means. For computing p-value, use the one-way ANOVA test. Fill Table 2 (add rows and columns if needed) and mark statistically significant pairs with \* (e.g., H2-H3). Visualize usages across all haplotypes as a boxplot and add it below.

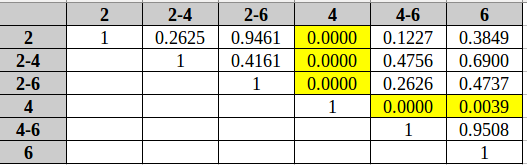
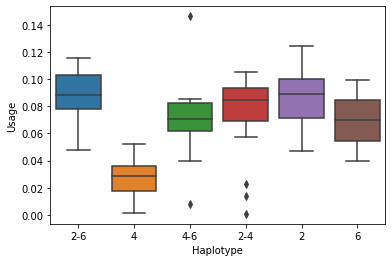


Table 2.



5. Extract sequences of alleles forming haplotypes in Table 1 from [IGHV.fa](https://drive.google.com/file/d/1UUzlIkK4AUcytNtMc_Po-t6acncexCQi/) and compute their multiple alignment. Identify SNPs (=differences) between alleles and, for each allele, describe them as pairs (N, P), where N is the nucleotide at position P in the multiple alignment. Fill Table 3 (add rows if needed). Clustal Omega:

|  |  |
| --- | --- |
| Allele 1 | A list of pairs (N, P) for all positions of SNPs |
| IGHV1-2\*02 | - |
| IGHV1-2\*04 | (T,199) |
| IGHV1-2\*06 | (C,148) |

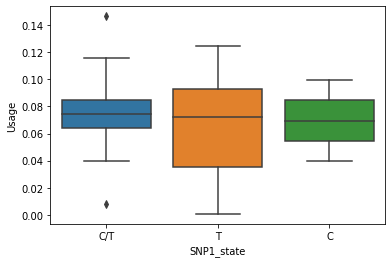
Table 3.

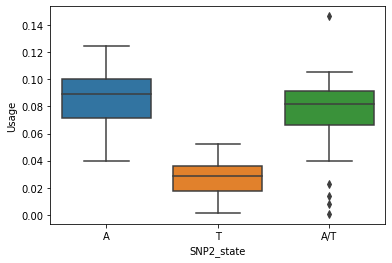
1. For each haplotype, compute a state for each SNP as a list of allele nucleotides. If a haplotype is homozygous, then its state N. If a haplotype is heterozygous, then its state is either N (if two alleles have the same nucleotide N), or N1/N2 (if two alleles have different nucleotides N1 and N2). Note that N1/N2 = N2/N1. Fill Table 4 (add rows if needed).

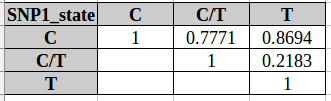
|  |  |
| --- | --- |
| Haplotype 1 | A list of states for all SNPs |
| 2 | T A |
| 2-4 | T A/T |
| 2-6 | C/T A |
| 4 | T T |
| 4-6 | C/T A/T |
| 6 | C A |

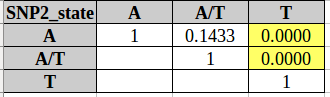
Table 4.

1. As a result, each SNP is described by a set of states (e.g., A, A/C, C) across all haplotypes. For each SNP, add a boxplot showing the distribution of usages across its states. Compute a p-value showing association between SNP states and usages using the one-way ANOVA test. Comment on statistical significance of such association.









**It seems that antibodies with SNP in IGHV1-2\*04 allele (nucleotide T on 199 position, starting from 1) are presented less often than antibodies without this SNP. This polimorfism might make antibody less effective, but it still works.**

**If we have heterozygous haplotype for this SNP, this polimorfism doesn’t make any difference.**

**Deadline:** Dec 6 (Sunday), 11:59 pm PST. Please send you reports directly to Nastya Vinogradova (@vinogradovana).

**Useful links:**

One-way ANOVA in Python:

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html>

Visualizing boxplots via seaborn:

<https://seaborn.pydata.org/generated/seaborn.boxplot.html>