**HW3, finding eQTLs of immunoglobulin genes**

**Code and results**

[**https://github.com/Yulia-Yakovleva/immunogenomics/tree/master/HW3**](https://github.com/Yulia-Yakovleva/immunogenomics/tree/master/HW3)

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):

|  |  |  |
| --- | --- | --- |
| **Haplotype** | **# individuals** | **Mean usage** |
| Haplotype 2 | 17 | 0.08619130588237642 |
| Haplotype 2-4 | 28 | 0.07745052343832814 |
| Haplotype 2-6 | 6 | 0.08695569960126666 |
| Haplotype 4 | 18 | 0.02789493332605945 |
| Haplotype 4-6 | 14 | 0.07103249923603643 |
| Haplotype 6 | 2 | 0.06957104526499999 |

Table 1.

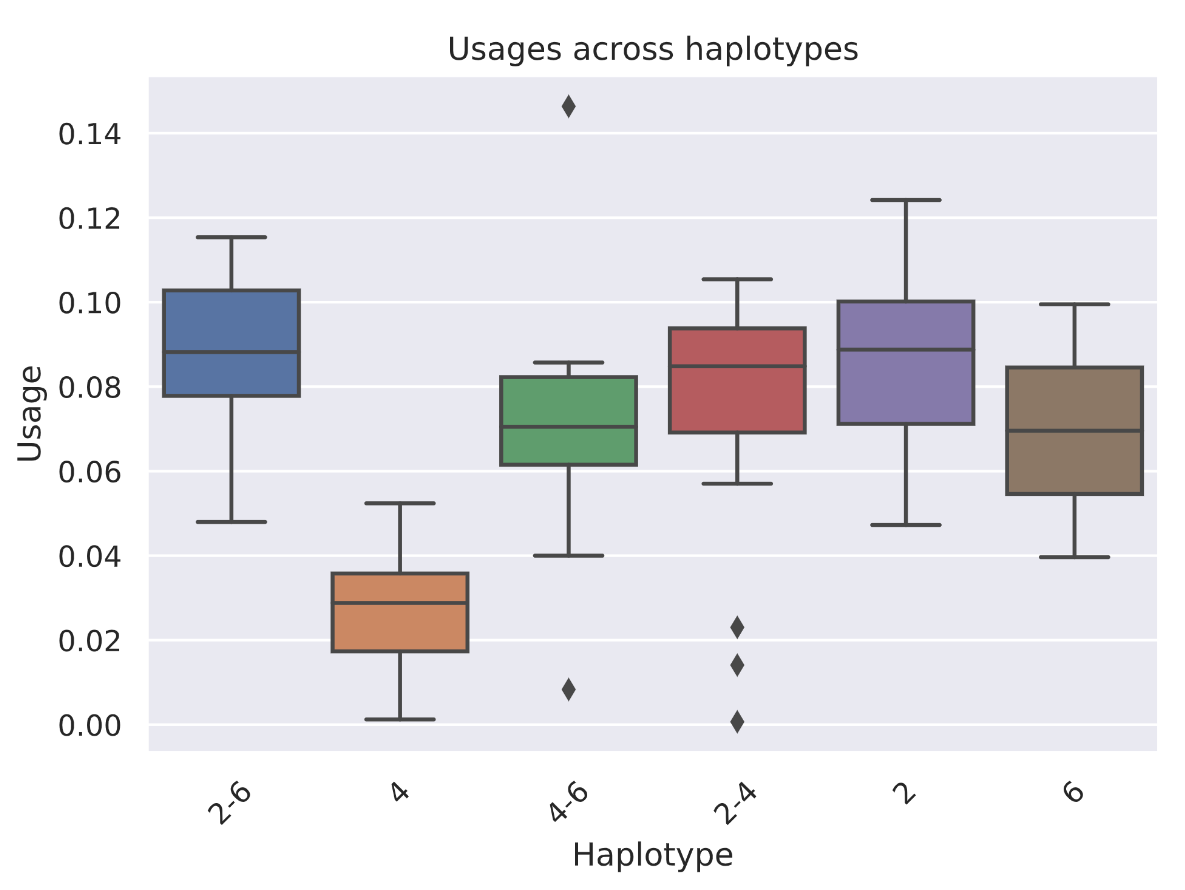
Table 1. <https://github.com/Yulia-Yakovleva/immunogenomics/blob/master/HW3/results/table1.tsv>

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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Haplotype 2-4 | Haplotype 2 | Haplotype 2-6 | Haplotype 4 | Haplotype 4-6 | Haplotype 6 |
| Haplotype 2 | 0.2624716174580812 | - | 0.9461275025566648 | 2.1452088862288644e-10\* | 0.12266063987379809 | 0.38486180480820387 |
| Haplotype 2-4 | - | 0.2624716174580812 | 0.4161316700901836 | 2.829361439989886e-09\* | 0.4755829286991522 | 0.6899600804565442 |
| Haplotype 2-6 | 0.4161316700901836 | 0.9461275025566648 | - | 1.8680666959607151e-07\* | 0.26256894732534597 | 0.47368679718238255 |
| Haplotype 4 | 2.829361439989886e-09\* | 2.1452088862288644e-10\* | 1.8680666959607151e-07\* | - | 6.579080938821483e-06\* | 0.0039040558563635374\* |
| Haplotype 4-6 | 0.4755829286991522 | 0.12266063987379809 | 0.26256894732534597 | 6.579080938821483e-06\* | - | 0.9507894686164409 |
| Haplotype 6 | 0.6899600804565442 | 0.38486180480820387 | 0.47368679718238255 | 0.0039040558563635374\* | 0.9507894686164409 | - |

Table 2.

Table 2. <https://github.com/Yulia-Yakovleva/immunogenomics/blob/master/HW3/results/anova_rslt.tsv>



Boxplot https://github.com/Yulia-Yakovleva/immunogenomics/blob/master/HW3/results/boxplot.pdf

1. 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).

|  |  |
| --- | --- |
| Allele | 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 | A list of states for all SNPs |
| Haplotype 2 | (T, 148); (A, 199) |
| Haplotype 2-4 | (T, 148); (A/T, 199) |
| Haplotype 2-6 | (T/C, 148); (A, 199) |
| Haplotype 4 | (T, 148); (T, 199) |
| Haplotype 4-6 | (T/C, 148); (Т/A, 199) |
| Haplotype 6 | (C, 148); (A, 199) |

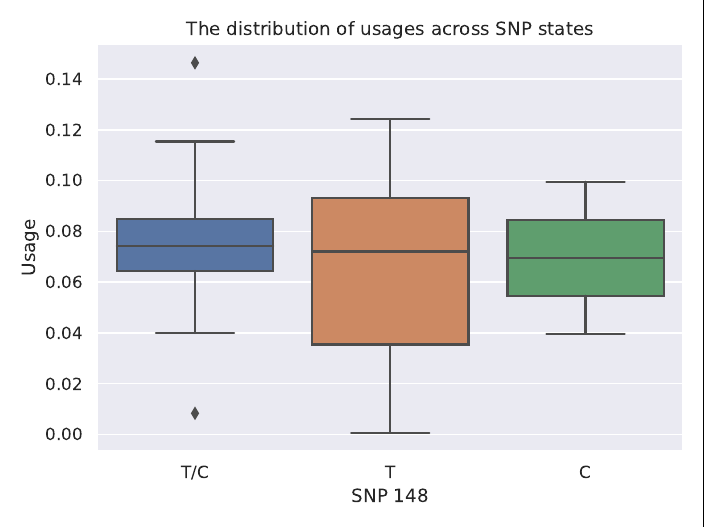
Table 4.

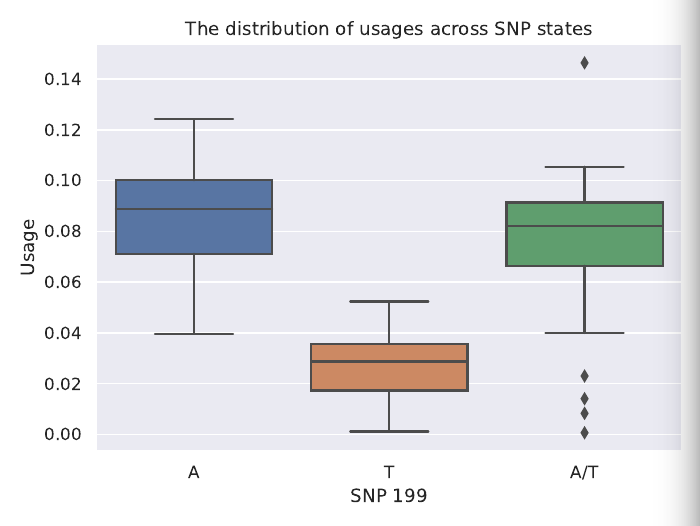
1. As a result, each SNP is described by a set of states (e.g., A, A/C, C) across all haplotypes. For each SNP, add a boxplot showing the distribution of usages across its states. Compute a p-value showing association between SNP states and usages using the one-way ANOVA test. Comment on statistical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

Plots

https://github.com/Yulia-Yakovleva/immunogenomics/blob/master/HW3/results/boxplot3.pdf

<https://github.com/Yulia-Yakovleva/immunogenomics/blob/master/HW3/results/boxplot2.pdf>





Anova results for SNP 148 (no statistically significant association between SNP states and usages):

https://github.com/Yulia-Yakovleva/immunogenomics/blob/master/HW3/results/anova\_rslt2.tsv

Anova results for SNP 199 (statistically significant association between A and T; and between T and A/T states and usages). То есть по какой-то причине варианты A и A/T действительно сильно отличаются по usage от T.

https://github.com/Yulia-Yakovleva/immunogenomics/blob/master/HW3/results/anova\_rslt3.tsv

**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>