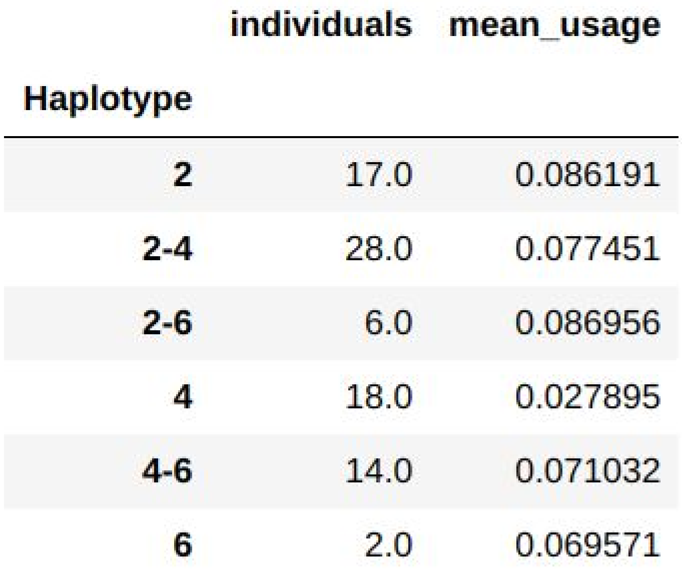
HW3, finding eQTLs of immunoglobulin genes
Changalidi Anton, IB 2020

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

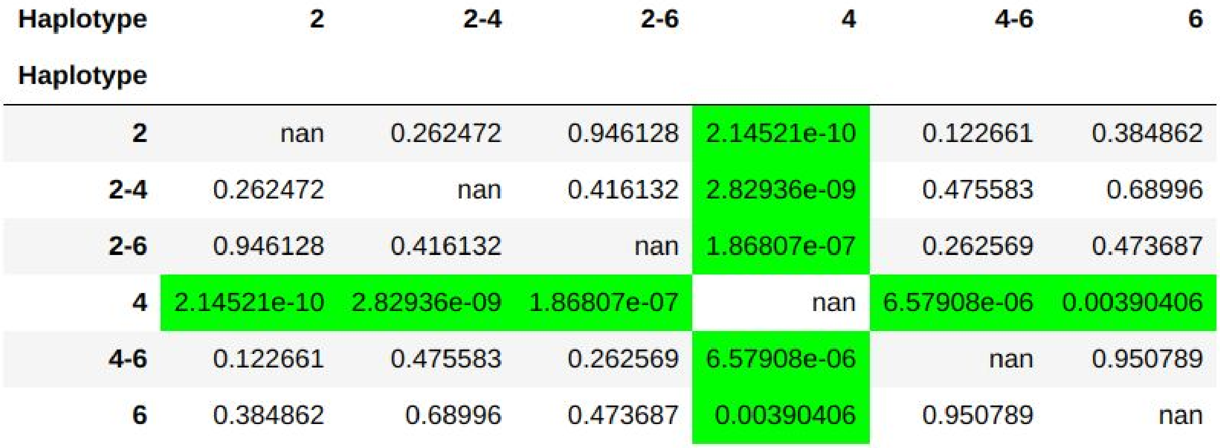
1. Download a dataframe containing usage values of gene IGHV1-2 collected across 85 healthy individuals. Usage values are provided in the "Usage" column. For each individual, haplotypes of IGHV1-2 were also computed and written to the "Haplotype" column. Haplotypes are described by IDs of alleles of 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):



	individuals	mean_usage
Haplotype		
2	17.0	0.086191
2-4	28.0	0.077451
2-6	6.0	0.086956
4	18.0	0.027895
4-6	14.0	0.071032
6	2.0	0.069571

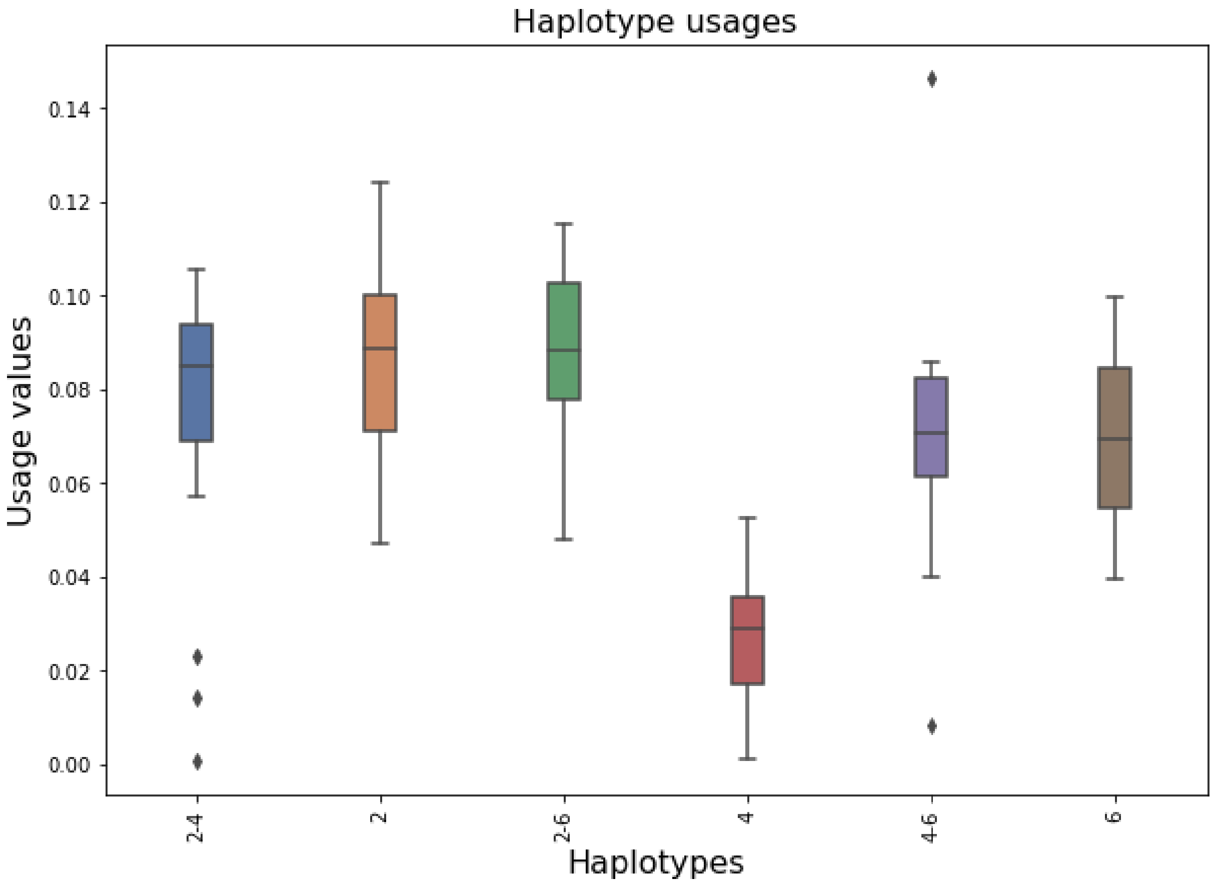
Table 1.

3. 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.



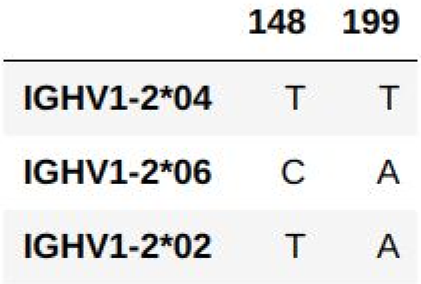
Haplotype	2	2-4	2-6	4	4-6	6
Haplotype						
2	nan	0.262472	0.946128	2.14521e-10	0.122661	0.384862
2-4	0.262472	nan	0.416132	2.82936e-09	0.475583	0.68996
2-6	0.946128	0.416132	nan	1.86807e-07	0.262569	0.473687
4	2.14521e-10	2.82936e-09	1.86807e-07	nan	6.57908e-06	0.00390406
4-6	0.122661	0.475583	0.262569	6.57908e-06	nan	0.950789
6	0.384862	0.68996	0.473687	0.00390406	0.950789	nan

Table 2.



Ну да, тут видно, что 4-й очень сильно выбивается и на боксплоте, и pval у него ххороший по сравнению со ВСЕМИ другими.

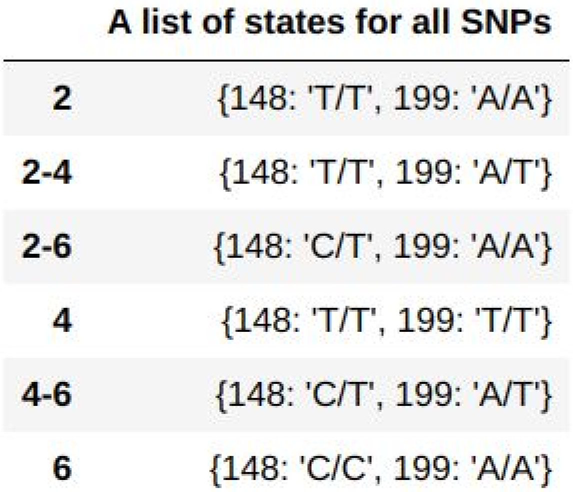
4. Extract sequences of alleles forming haplotypes in Table 1 from IGHV.fa and compute their multiple alignment. Identify SNPs (=differences) between alleles and, for each allele, describe them as pairs (N, P), where N is the nucleotide at position P in the multiple alignment. Fill Table 3 (add rows if needed).



	148	199
IGHV1-2*04	T	T
IGHV1-2*06	C	A
IGHV1-2*02	T	A

Table 3.

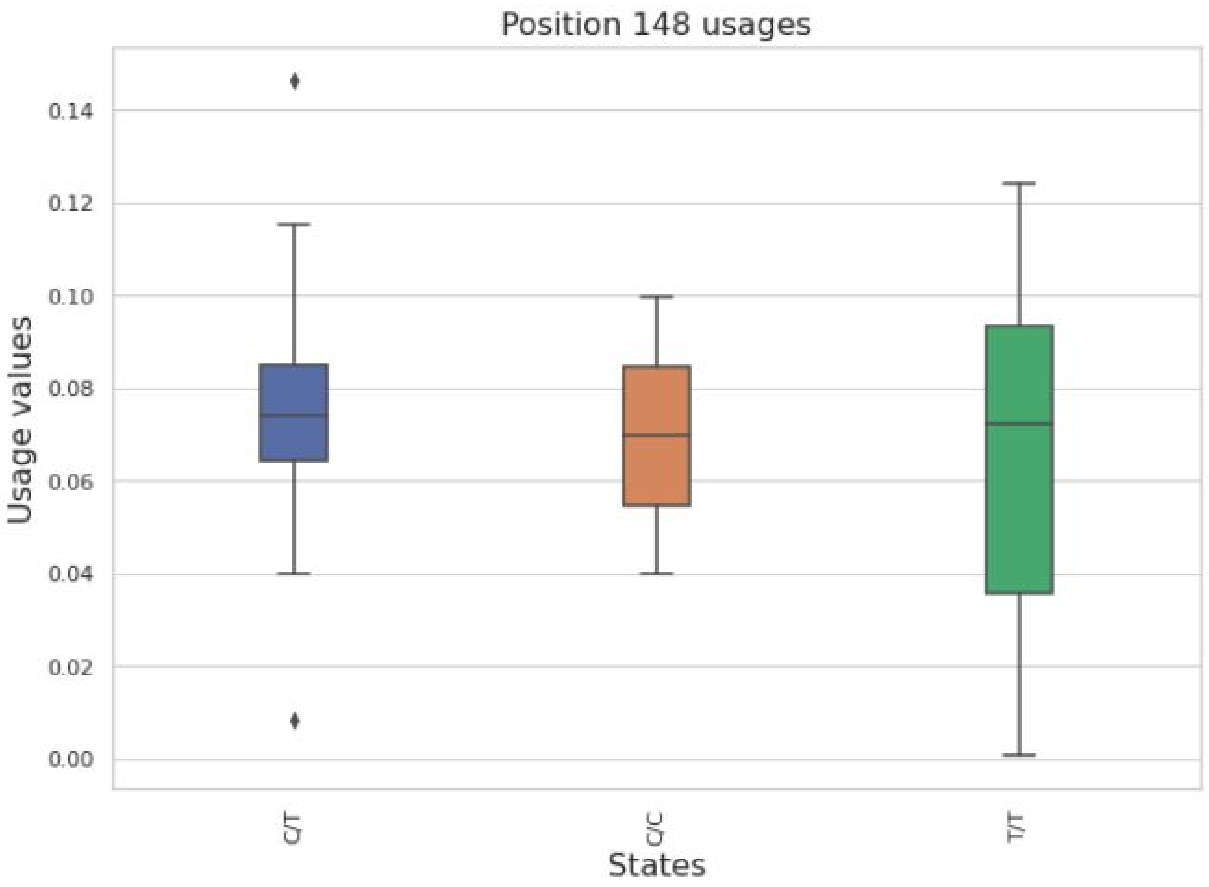
5. For each haplotype, compute a state for each SNP as a list of allele nucleotides. If a haplotype is homozygous, then its state N. If a haplotype is heterozygous, then its state is either N (if two alleles have the same nucleotide N), or N1/N2 (if two alleles have different nucleotides N1 and N2).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

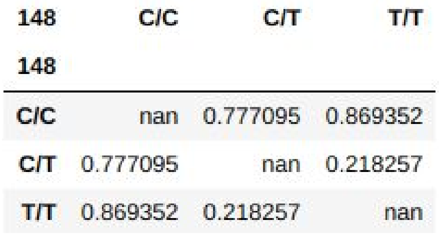


	A list of states for all SNPs
2	{148: 'T/T', 199: 'A/A'}
2-4	{148: 'T/T', 199: 'A/T'}
2-6	{148: 'C/T', 199: 'A/A'}
4	{148: 'T/T', 199: 'T/T'}
4-6	{148: 'C/T', 199: 'A/T'}
6	{148: 'C/C', 199: 'A/A'}

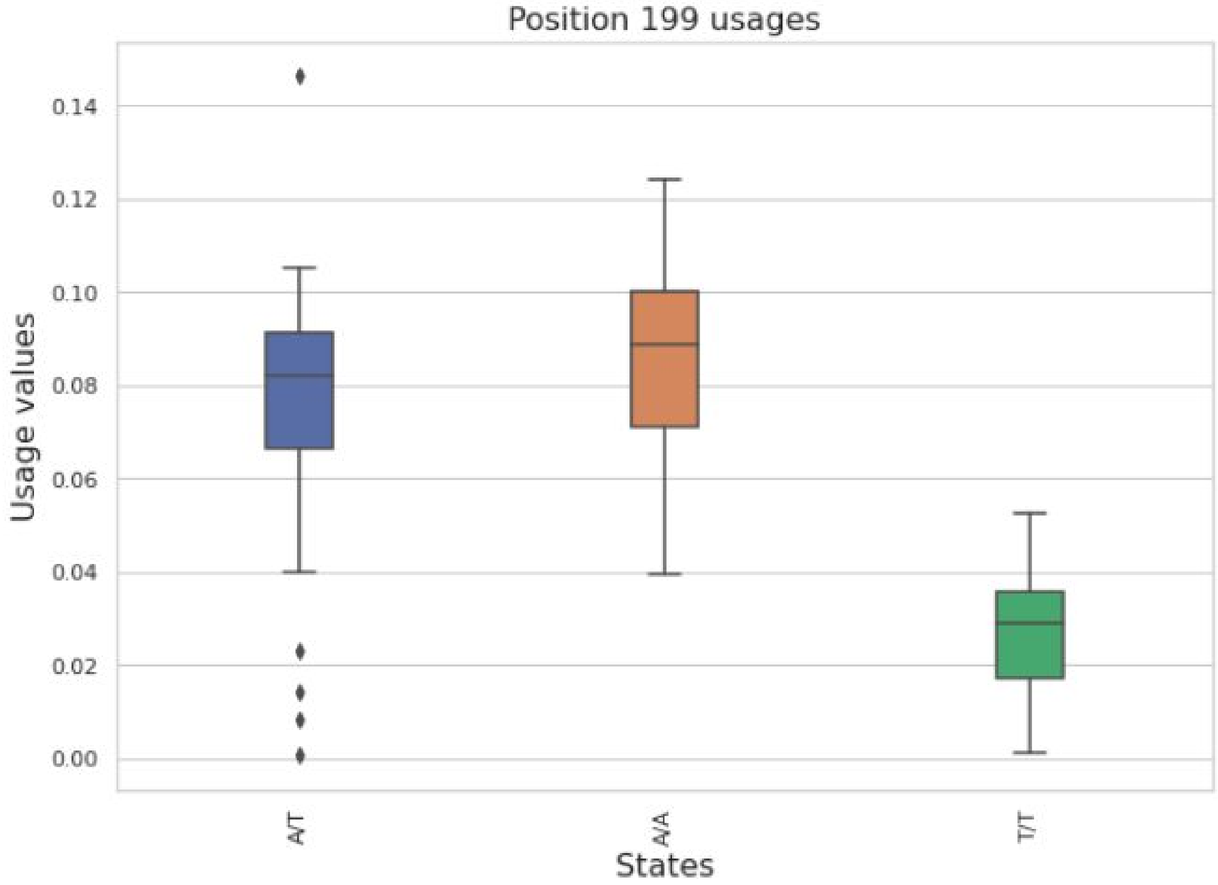
Table 4.

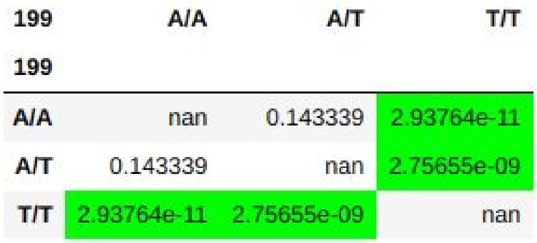
6. As a result, each SNP is described by a set of states (e.g., A, A/C, C) across all haplotypes. For each SNP, add a boxplot showing the distribution of usages across its states. Compute a p-value showing association between SNP states and usages using the one-way ANOVA test. Comment on statistical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





148	C/C	C/T	T/T
148			
C/C	nan	0.777095	0.869352
C/T	0.777095	nan	0.218257
T/T	0.869352	0.218257	nan





199	A/A	A/T	T/T
199			
A/A	nan	0.143339	2.93764e-11
A/T	0.143339	nan	2.75655e-09
T/T	2.93764e-11	2.75655e-09	nan

Можно сделать вывод, что в 148 позиции чего-то статистически значимого нет (не поддерживается отбором), однако в 199 позиции гомозигота по T имеет значительно более маленький usage, чем гомозигота по A и гетерозигота. Этот SNP, по все видимости, и влияет на весь eQTL, меняя его usage.