**FINAL ASSIGNMENT   
FOR THE EVOLUTIONARY AND POPULATION GENETICS COURSE – WINTER EXAMINATION PERIOD 2024/2025**

**Introduction**

This assignment consists of two parts, one concerning **the evolutionary topics** and based on mtDNA markers, and one based on autosomal SNP addressing **population genetic topics**. In this assignment you’ll need to combine the data analysis skills gained during the exercises with the theoretical knowledge acquired during the lectures (some of the questions are purely theoretical and not connected to the dataset). Please prepare your file as a report, with a story-like structure, enough details to be understood by someone who does not know the dataset and a clear distinction between the Evolutionary and Population parts. Please be aware that for the SNP dataset, you got a random subset, so numbers in the publication are of no use.

Data for the evolutionary part of the exam are available on the website <https://www.ncbi.nlm.nih.gov/>. You have received the SNP analysis data by email, along with instructions and exam questions. Each student receives their own data sets for both the evolutionary and the population part, so all answers must contain data about the performed analyses related to your data set. We would like to inform you that the submitted exams will be thoroughly screened for the use of artificial intelligence. Any violation will be subject to appropriate disciplinary sanctions, according to the faculty's regulations.

Submit the exam report in Word format, via the e-classroom, under the section "Exam report submission - 1st term". You have 7 days to prepare the exam report (no later than **February 2, 2024, at 23:59**). After this deadline, submission will no longer be possible, and it is considered that you have not taken the exam. When submitting, you must confirm the submission twice, with the first confirmation you submit the assignment, with the second you confirm that the product is your independent author's work.

Keep in mind there are multiple ways to approach an analysis, using different software and/or R packages, so make sure to comment what tools you used, if you changed any default settings and why. Make good use of the combination of knowledge and experience you’ve gained in other courses or found in various articles. If necessary, add images and tables to your final document, but make sure to add a proper description/interpretation to each of them, as an imago or table alone is insufficient. Don’t forget to properly prepare all input files for your analyses as well.

Good luck.

**Evolutionary Genetics**

Conduct a phylogenetic analysis and answer the questions.

1) Construct a phylogenetic tree using the Maximum likelihood method and explain your results (relations between observed taxa). How does the method you used differ from the Bayesian inference method? 2T

2) While working in the field you found an unknown eDNA sample. After a thorough genetic analysis, you’ve discovered the following sequence:

1 tactctttac cttttattcg gtgcctgagc tggcatggtg gggactgctc ttagtcttct

61 aatccgggcc gaactgggcc aacctggtac actactagga gatgatcaga tttacaatgt

121 aatcgtcact gcccatgctt ttgtaatgat cttttttatg gtgatgccta ttataattgg

181 agggttcgga aactgattgg tcccattaat aattggagct cctgacatag catttccccg

241 aataaacaac atgagcttct gactcctccc tccatccttt ctactcttac tcgcctcatc

301 tatggtagaa gccggagcgg gaactggctg aacagtatac ccacccctag ccggcaacct

361 ggctcatgca ggagcatccg tagacctaac tattttttca ctacacctgg caggtgtctc

421 ctcaatcttg ggtgctatta atttcattac tactattatt aatataaaac ctcctgccat

481 gtcccaatat caaacacctc tatttgtctg atcagtctta atcactgctg tcttactact

541 tctatcactt ccagtcttag cagcgggaat cactatatta ttaacagatc gaaacctaaa

601 caccacattc tttgaccccg ctgggggagg agatcctatc ttataccaac acttattc

Which species does the sample belong to? Does it have a reference genome published? Is the species endangered? What is its whole taxonomic classification? 2T

3) Take a look at the provided haplotype network. What conclusions can you draw about the observed species based on the information provided by the network? 2T

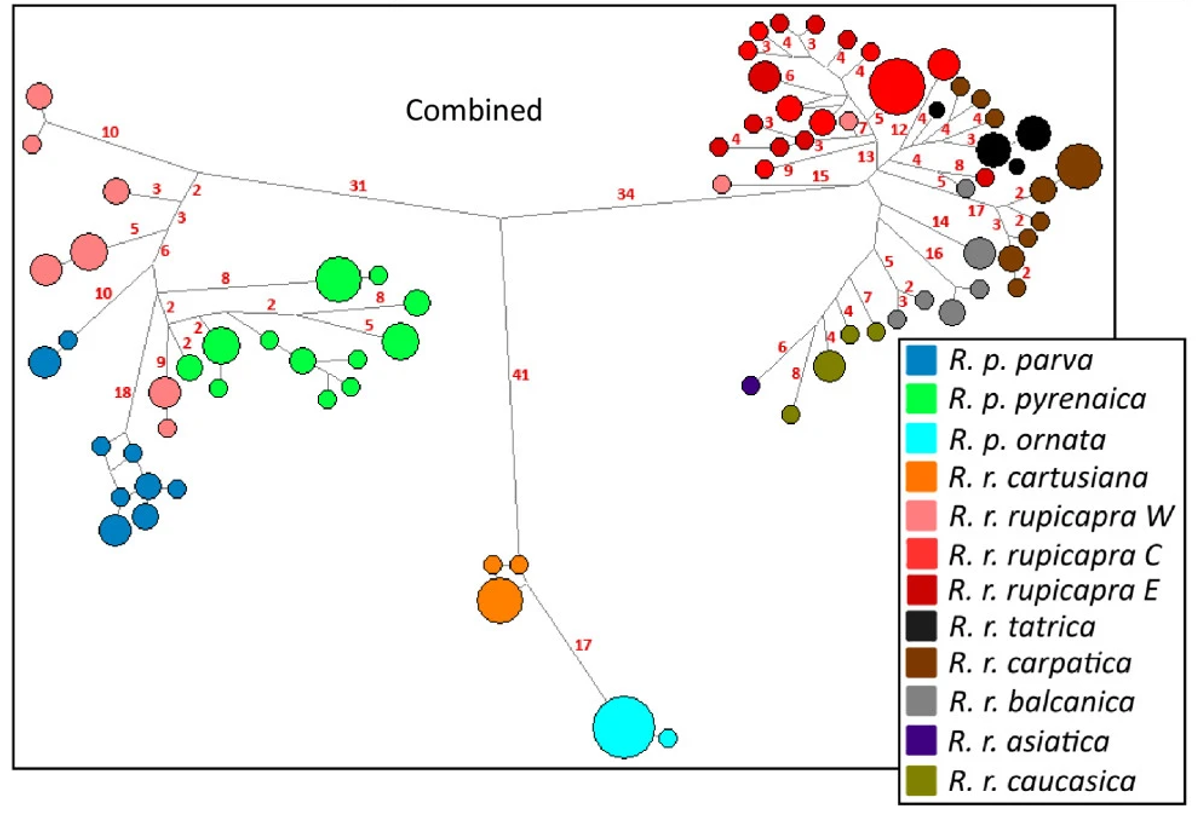
4) Construct a haplotype network and explain your results. Include the following key terms: **mutation steps**, **recent expansion**, **highest** and **lowest genetic variability. 2T**

5) Your dataset is defined by a single locus or molecular marker. Which molecular markers would you use if you wanted to examine closer or more distant phylogenetic relationships within a species or genus? Would you change the model in this case, and do you think it is possible to use multiple markers? If so, which markers are available and would be useful for this purpose? 2T

6) How many S (singleton), C (conserved), V (variable) and Pi (Parsimony – informative) sites did your alignment have? What do these values tell us? For the purpose of our analyses, are singleton or parsimony-informative sites more informative? Why? 2T

7) How does the ratio of transitions (Ts) to transversions (Tv) in your dataset affect the evolutionary history or the topology of the phylogenetic tree? How would the tree change in the following cases: 2T

1. More transitions (Ts)
2. Fewer transversions (Tv)



**Population genetics**

Analyse the data set based on SNPs and answer the questions below.

1. After sequencing and bioinformatics analysis, you have a set of SNPs that you want to filter for further analysis. The species you are studying has 39 chromosome pairs. You want to exclude all loci that have more than 5% missing data and all individuals that have more than 4% missing genotype. You also want to exclude all loci that have a minor allele frequency below 1%. Construct a PLINK command that will filter your hypothetical binary file "data\_SNP\_2.bed" by the desired parameters. 2P
2. How many individuals, populations, loci do you have in your dataset? Which populations do you have, please give the names. Which command did you use to do this? Before proceeding you need to run another command to check the validity of the data. What is this command? 2P
3. How many individuals, populations, loci do you have in your dataset? Which populations do you have, please give the names. Which command did you use to do this? Before proceeding you need to run another command to check the validity of the data. What is this command? 2P
4. Calculate the FST between populations and explain genetic differentiation. If present, give possible causes. 2P
5. Calculate and then produce a PCA plot. Explain what the plot represents (interpret the results of the analysis). What can you tell about the structure of your dataset? 2P
6. How many populations are your study specimens included in the analysis divided into according to the molecular data? On the basis of which parameter can you make this conclusion? Use the arguments: k.range=1:10, num.k.rep=2, numreps=1000, burnin=1000, remember to specify other arguments. Graphically show the analysis of the structure of the populations. Interpret the results of the analysis of the structuring of the populations. 4P
7. Submit your own RMarkdown scripts. 2P