Week 3 Challenge

1) Complete and submit the answers to the questions in the Day 5 Population Structure tutorial (copied below for reference). For reference, these data are those used in Lewanski et al. Molecular Ecology 2022.

    1. Based on the plot, which K appears to minimize the Cross-entropy score?

    2. Do a PCA on these same data using your skills from the first week of class (and include it in the assignment you turn in). Does PCA suggest the same number of clusters as the best sNMF model? How does the PCA assist in your interpretation of the sNMF plots?

    3. What biological processes might explain these patterns? Remember, if you'd like to dig in further you can use the metadata included in evoanalysis/data/Petro_metadata.csv to dig in more.

2) Let's think some more about filtering. The first week of class, you read Hemstrom et al 2024 and we thought in theory about the process of genomic data filtering and its impacts on downstream analyses. Then, you learned some practical tools for how to filter data and VCF files in the second week of class. Now you are working with data and needing to make practical decisions about filtering for your final projects.
    1. Why does one need to filter genomic data? Why not leave it as it comes off the sequencer?
    2. Give four examples of filtering decisions one could make that would bias data in a way that could impact downstream analyses.
    3. Why aren't there best-practice guidelines for filtering that are one-size fits all? Why do different datasets need individual decision making regarding filtering approaches?