Here is the Exercise for Problem C:

The idea is that you want to see if you can correctly classify the individuals in the test set based on the information in the training set.

1. At each locus, use the training set to estimate the allele frequencies (ie proportions) in each of the four subpopulations.

Assume for the remainder of this exercise that these allele frequencies from the training set are the "true" frequencies in each population.

2. For each individual in the test data set, compute the posterior probability that it arose from each of the four populations, assuming that all four populations are equally likely a priori. You can assume that the 12 loci contribute independently to the likelihood. That is, the likelihood is defined by multiplying the likelihood across loci.

3. If you ``assign'' each individual in the test set to the population that maximizes its posterior probability, what is the error rate? (ie how many individuals are misassigned vs correctly assigned?)

4. Comment on any problems you came across as you did this exercise, and how you solved them. Your answer should include all your R code in a format that can be run to reproduce your results (I recommend using RStudio and the knitr package to produce your report).