1 Introduction

Answer all the questions below and provide the R-code necessary to obtain your answer, as well as the graphs you have produced. The results should be saved as pdf and submitted via the moodle platform

2 re-analysis of the experimental data

You will analyse a subset (roughly 5%) of the data presented in the paper by Gompert etal (2014). In this article, the authors carry out an "evolve and resequence" experiment on stick insects over a short period of 8 days. They intend to identify loci under selection, ie, those that show unduly large changes in allele frequencies.

First, lets load the data, located at "http://www2.unil.ch/popgen/teaching/R14". The files are `subdat.txt` and `indpopsurv.txt`. You should have already downloaded the file `subdat.txt`. It consists of 500 columns and 10000 rows, each column representing the molecular "phenotype" of one individual at 10000 Rad-Tag loci.

1. Load this file in R, call it `dat`
2. What are the dimensions of the data set?
3. What is the range of values in this data set?
4. Draw a histogram of these values and comment it (you might need to convert `dat` into a matrix).

   The file `indpopsurv.txt` is made of three columns.

5. Load it into R and call it `idp`.

   The first column is an individual identifier, the second the population on which it was assigned (this population is defined by a location (1 to 5) and a host plant (*A* or *C*)). Finally the third column tells whether the individual was recaptured. As a brief reminder, stick insects were collected from one host plant (*Adenostoma fasciculatum*) and transplanted to either the same host plant (treatment *A*) or the plant *Ceanothus* (treatment *C*). Individuals recaptured and thus deemed surviving are given a 1, those not recaptured and deemed dead a 0. Using the informations in `idp`, answer the following questions:

6. create a variable `treat` assigning individuals to one of two categories, treatment *A* or treatment *C*
7. create a variable `loc` assigning individuals to their respective locality, not accounting for treatment
8. what is the number of surviving and missing individuals in each population?
9. what is the number of surviving and missing individuals in each treatment?

10. what is the number of surviving and missing individuals in each locality?

11. assign names to the columns of `dat`, corresponding to the individual identifier in `idp`

    Next, we will have a look at the distribution of the molecular *phenotypes* (numbers between 0 and 2). There is some uncertainty about the exact genotypes obtained from high throughput sequencing (HTS), thus the real numbers.

12. create a new matrix called `datn` with the same dimension as `dat`. Store in `datn` the data recoded so that anything below 0.2 is encoded 0, anything above 1.8 is encoded 2, anything between 0.5 and 1.5 is encoded 1, and what remains is considered as missing data. You might need to use logical OR (`|`) and AND (`&`) for this

13. produce a histogram of what is contained in `datn` and describe it.

14. create a vector containing the sum of each row, and make an histogram of it (you may find function `rowSums` or `colSums` useful)

15. store in a vector `missing` the number of `NA` on each row. For this you might either use a loop, or use the function `apply`

16. estimate the mean of each row and divide it by two. This is the allele frequency in the total data set. Store this in vector `tot.freq`. Here and in the following questions, if you fail to create the matrix `datn`, answer the questions using `dat`.

17. create a data frame consisting only of individuals in the treatment C, store it in `datn.C`. Estimate the mean of each row and divide it by two. Store this is vector `C.freq`

    Next, we focus on the surviving individuals, ie, after selection has been operating.

18. Create a new data set, `datn.C.s`, containing only the genotypes of surviving individuals from individuals put in treatment C.

19. estimate the mean of each row and divide it by two. This is the allele frequency in the surviving individuals. Call it `C.s.freq`

20. in a 2 panel windows, produce a boxplot, and then a histogram, of the allele frequency changes and comment the figure.

21. discuss what would be needed next to identify loci under the influence of selection