

# Exercises on data preparation

*M. Humberto Reyes-Valdés*

*June 23, 2019*

In some cases, data cleaning in marker tables involves selecting for MAF, which is “minimum allele frequency”, where all loci containing an allele with a frequency lower than the established MAF are eliminated.

Use the table `CleanWheat.csv`, and eliminate all loci containing an allele with a frequency lower than 0.1, i.e., the selected table will be chosen under a criterion of  $MAF = 0.1$ . Remember that each loci comprises two rows, and by no means you should select only one row in a given locus.

To select the table you can use either base code or *dplyr*.

As an aid, you can get the vector of allele frequencies across rows as follows

```
freq<-apply(dat[-c(1,2)],1,function(x) mean(x,na.rm=T))
```

How many loci are in your selected table? What fraction of the whole table was selected?

Store the table as `CleanWheatMAF10.csv`, without row names.