Statistical analysis: rice phenotype

Houda Aiboud Benchekroun and Thomas Roiseux (group 8)

2022-11-20

General information

Genetic data hold important information about living beings. These information can be used to understand and even predict the ability of a species, or an individual, to protect itself from a specific disease. Using this and statistical tools, we may be able to find some predispositions for individuals of a population to have or a not specific disease, and by making predictions, to find the most suitable gene that will grant the best protection against several problems.

To explore this, we will use a data set made with genetic information of rice. These have been extracted from the UK Bio-Bank. As rice is one of the basics in food in several countries, it production must be protected to avoid any starvation in these countries. That's why studying its genetic data to protect it is quite important nowadays.

To understand that, we are going to study phenotypic characters and explain them using genetic data coming from the genotype of this rice species, named *Oryza sativa*.

Variables available

Genotype

The data set we have is made of two different parts: we have a genotype set and a phenotype set. The phenotype data set holds data about expressed characters of the rice, whereas the genotype dataset holds all available alleles for rice genes. 413 accessions have been sampled in this data set, referenced in their genotype. The genotype and its variables will be used as explanatory variables, as genes and alleles grant different visible characteristics to the rice. The 413 samples included in this data set reflect the 413 accessions from the total of genotyping rice species, that have grown across the planet (Figure 1).

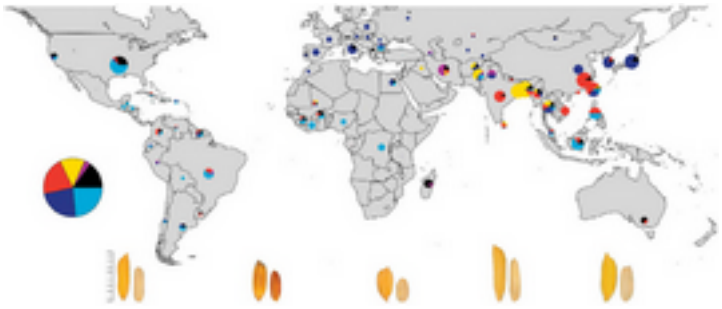

Figure 1: World map with all locations where the rice has grown

This map shows the repartition of the 413 species used to make the data set.

These variables, crossed with the genetics markers will provide explanations to the rice phenotype. With these explanatory variables, we have three other ones:

`marker`	Name of the marker, used to differentiate the genes.
`chrom`	Reference to the chromosome that holds the gene.
`pos`	Position of the gene among the chromosome.

Table 1: Details of explanatory variables hold in `geno.df`

As shown in Table 1, the genotypic data set includes three other variables, that are used to understand which genetic marker is present or not on each rice species, as these markers will be very useful to understand the phenotypes.

Then, all explanatory variables are grouped into a single matrix.

Phenotype

We also have 36 available phenotypic variables, coming from several visible characters of this rice.