# Study of *Oryza sativa* phenotypes using genome-wide association to improve rice yield

Adib HABBOU, Alae KHIDOUR – Group 1 – Bio Project

## Abstract:

In a context of demographic explosion where the world population has just exceeded 8 billion people, the demand for food continues to increase considerably. One of the solutions to meet this demand is to improve the yield of food production through the study of the genome. Here we will look at *Oryza sativa* rice to understand its genetic basis for improving its yield, quality, and sustainability.

SNPs are single base-pair changes in the DNA sequence with the highest frequency of occurrence in the genome. *[1]* Here we present the genome-wide association study based on genotyping 44,100 SNP variants in 413 diverse *Oryza sativa* accessions collected from 82 countries that were systematically phenotyped over two consecutive seasons for 36 morphological, developmental, and agronomical criteria.

The 44,100 SNPs come from two data sources: the Oryza SNP project and OMAP. The panel includes 87 indica, 57 aus, 96 temperate japonica, 97 tropical japonica, 14 groupV/aromatic and 62 highly mixed accessions. *[2]*

## Dataset presentation:

We have three datasets *pheno.df*, *genom.df* and *Xmat*:

**-** *pheno.df* contains 413 rows and 38 columns such that the first two columns are used to identify every individual and each of the next 36 columns represent a phenotype which is an observable trait in a rice plant *Oryza sativa*. For example, the height of the plant, the length and width of its flag leaf...

**-** *genom.df* contains 36 901 rows and 416 columns, it represents the genome of *Oryza sativa* for 413 individuals, the genome being the whole genetic material of a species coded in its DNA, we are interested here in the different alleles of the various genes.

- *marker:* short DNA sequence surrounding a SNP
- *chrom:* location of the gene among the 12 chromosomes of the *Oryza sativa*
- *pos:* position of the gene in the DNA sequence
- *1 - 413:* alleles of each gene: 0 - homozygous major allele, 1 - heterozygous, 2 - homozygous minor allele

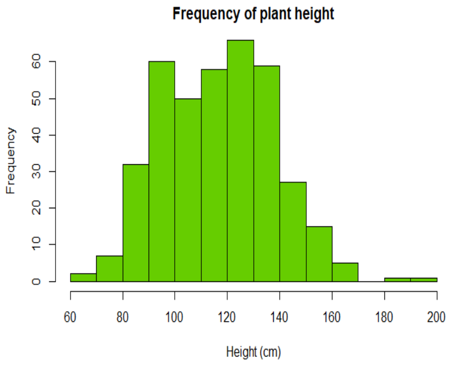
**-** *Xmat* contains 413 rows and 36 901 columns, which is the transpose of *geno.df* dataset without the *marker*, *chrom* and *pos* columns. Each cell gives us the information about the allele type of a specific gene for a specific individual.

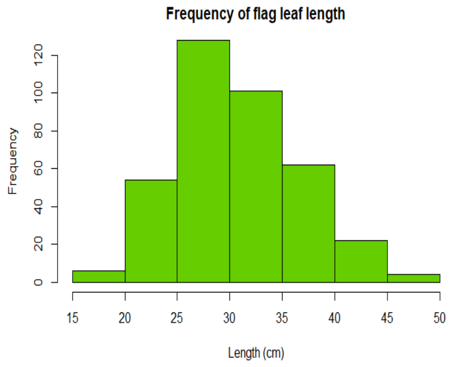
## Choice of target variables:

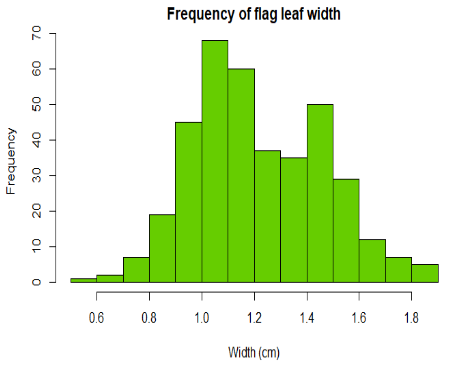
The height of a plant generally means its good health and thus its increased productivity. Plant height is an important developmental and yield characteristic. Dozens of genes regulating plant height in rice have already been identified. *[2]*

During the transition between the growth of the rice plant and the time when it begins to produce grain, the last leaf of the plant emerges which is called the flag leaf. Most of the carbohydrates required for grain filling are supplied by photosynthesis in this leaf, making it the most important leaf for the yield of the rice plant. *[3]*

Therefore, the *Plant.height*, *Flag.leaf.length* and *Flag.leaf.width* variables were chosen since modifying the genotypes responsible for these phenotypes can greatly improve the yield and quality of the rice plant.







## Missing values:

We can also see that the variables *Plant.height*, *Flag.leaf.length* and *Flag.leaf.width* are among those with the least number of missing values. The solution chosen to deal with the missing values in our target variables is to replace them by the mean value and creating a vector of the same size that will tell us whether the value was missing or not.

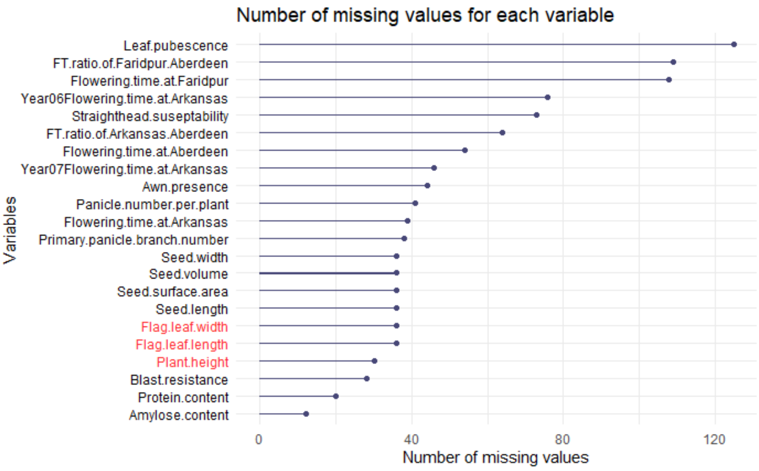


Table 1: Sample of our final target variable

| | Plant.height | Plant.height.is.missing |
|---|---|---|
| 10 | 116.5826 | 1 |
| 11 | 135.1667 | 0 |
| 12 | 117.8889 | 0 |
| 13 | 116.5826 | 1 |
| 14 | 161.8333 | 0 |
| 15 | 88.0000 | 0 |
| 16 | 132.0000 | 0 |
| 17 | 130.5000 | 0 |
| 18 | 120.6667 | 0 |

Concerning the *Xmat* dataset we find 660 751 missing values randomly distributed on all the columns and all the rows. To manage this, we will apply the following protocol: we delete all individuals that contain more than 10% of missing values or less than 0.1% of homozygotes with minor allele since they will not be relevant for the model, then we replace the remaining missing values in each gene by the most frequent value (either 0 or 2) in the considered gene. *[4]*

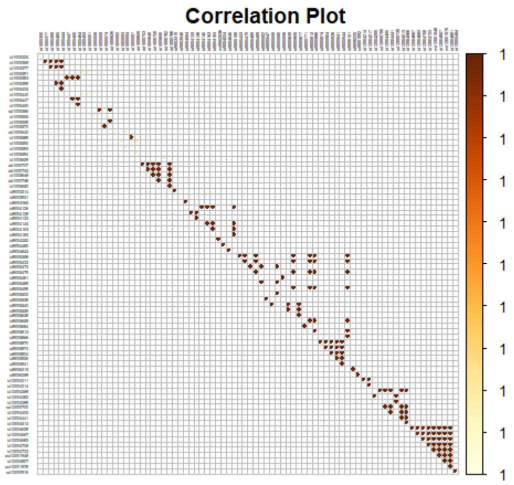
Table 2: Sample of our final dataset

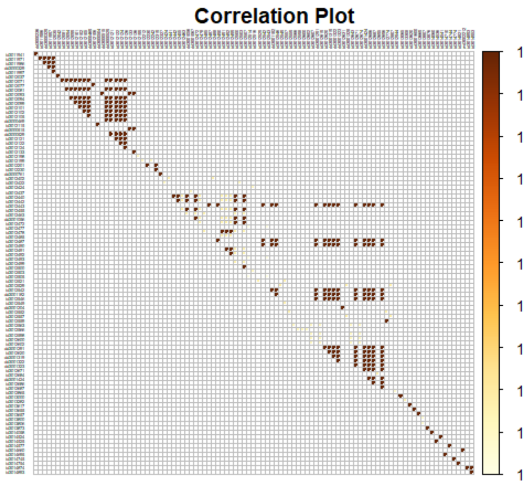
| X | id1000001 | id1000003 | id1000005 | id1000007 |
|---|---|---|---|---|
| 3 | 2 | 2 | 0 | 2 |
| 4 | 2 | 2 | 0 | 2 |
| 5 | 2 | 2 | 2 | 0 |
| 6 | 2 | 2 | 0 | 2 |
| 7 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 |
| 13 | 2 | 2 | 0 | 2 |

## Correlation:

The physical characteristics of the rice grain are also strongly correlated with region and environment. This means that genotypic and phenotypic variations in *Oryza sativa* are all correlated to some extent. *[2]*

Given the very large number of variables available we will keep in the correlation plot only the variables correlated to more than 99% which will be symbolized by the red points. The correlation plot will be computed on subsets of our dataset, the aim is to detect if there is strong correlation between some of the explanatory variables.





## Issues considered:

It is clear from the correlation plots that there is a high correlation between some of the variables, so a simple multiple regression will not be sufficient to obtain a good model. We will need regularized regression methods to select and penalize the explanatory variables.

## References:

*[1]* William S. Bush, Jason H. Moore "*Chapter 11: Genome-Wide Association Studies*"

*[2]* Keyan Zhao "*Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa*"

*[3]* R.A. Sperotto, F.K. Ricachenevsky "*Rice grain Fe, Mn and Zn accumulation: How important are flag leaves and seed number?*"

*[4]* Qian, Hastie "*A fast and scalable framework for large-scale and ultrahigh-dimensional spare regression with application to the UK Biobank*"