Statistical Analysis and Description MRR Project

Teddy ALEXANDRE, Arthur BABIN (Group 9, Bio Project)

18 Novembre 2022

Introduction

Over half of the world's population is fed by the cultivars of *Oryza sativa*, also referred to as Asian rice, which is a form of cultivated rice. Hence, rice production and productivity is essential for guaranteeing food security, alleviating hunger, and setting the stage for long-term economic growth.

Furthermore O.Sativa is a model organism for the botany of cereals (small size, ease of growth, high fecundity, short generation time, small genome and amenability to genetic manipulation, including crossing, mutagenesis and gene modification) and has a genome that is *easy to genetically manipulate*, with 430 Mbp spread across 12 chromosomes that can be sorted into five sub-populations: Indica, Aus, Temperate Japonica, Tropical Japonica and Aromatic.

The goal is to understand the genetic basis of various physiological, developmental, and morphological traits in order to lay the groundwork for improving rice yield, quality, and sustainability.

Genotype dataset :

- Rows (number of 36901) : each row corresponds to a SNP (Single-Nucleotide Polymorphism)
- Columns (number of 413+marker+chrom+pos so 416): marker allele data : value of each accession for the corresponding SNP : 0=homozygous (major+major), 1=heterozygous (major+minor), 2=homozygous (minor+minor) and NA (represents missing data)
  - marker: SNP identifier
  - chrom: chromosome where the SNP is located on
  - pos: position of the SNP in the ADN sequence of the associated chromosome

Phenotype dataset :

- Rows (number of 413) : an observation corresponds to an accession
- Columns (number of 38) : corresponds to the traits used to describe the accession (means are calculated to represent the accession in the data set, from the differents plants of the accession)
- "HybID": accession number (unique identifier of the accession)
- "NSFTVID": local identifier of the accession (from 1 to 413)
- "Flowering.time.at.Arkansas", "Flowering.time.at.Faridpur", "Flowering.time.at.Aberdeen", "Year*Flowering.time.at.*": flowering time of the plant in days
- "FT.ratio.of.Arkansas.Aberdeen", "FT.ratio.of.Faridpur.Aberdeen": flowering time ratio

Plant Morphology attributes

- "Culm.habit": estimated average angle of inclination of the base of the main culm (from 0 to 9)
- "Leaf.pubescence": large pubescence (=1) and small pubescence (=0)
- "Flag.leaf.length": length in cm of the flag leaf (the top leaf on a rice stem, closest to the grain-producing panicle)
- "Flag.leaf.width": width in cm of the flag leaf
- "Awn.presence": if the plant has awn (=1) or not (=0)

Yield-related attributes

- "Panicle.number.per.plant": number of panicle per plant
- "Plant.height": height of the plant in cm
- "Panicle.length": length of the panicle in cm
- "Primary.panicle.branch.number": number of panicle branches attached to the main central axis
- "Seed.number.per.panicle": number of seed per panicle
- "Florets.per.panicle": number of florets per panicle
- "Panicle.fertility": fertility ratio of the panicles

Seed morphology attributes

- “Seed.length”: seed length in mm
- “Seed.width”: seed width in mm
- “Seed.volume”: seed volume in dozens of mm^3
- “Seed.surface.area”: seed surface area in dozens of mm^2
- “Brown.rice.seed.length”: brown rice (the seed without the husk) length in mm
- “Brown.rice.seed.width”: brown rice width in mm
- “Brown.rice.surface.area”: brown rice surface area in dozens of mm^2
- “Brown.rice.volume”: brown rice volume in dozens of mm^2
- “Seed.length.width.ratio”: ratio between seed length and seed width
- “Brown.rice.length.width.ratio”: ratio between brown rice length and brown rice width
- “Seed.color”: descriptor of the color of the seed (1 or 0)
- “Pericarp.color”: descriptor of the color of the pericarp (1 or 0)

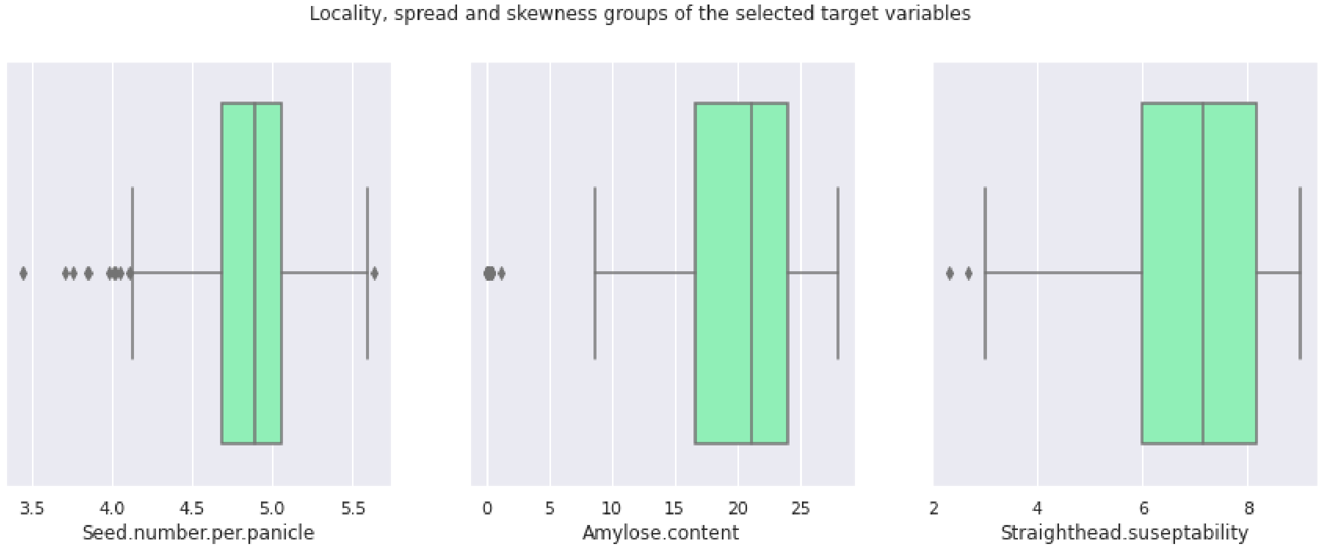
Stress tolerance attributes

- “Straighthead.susceptability”: rice susceptibility to straighthead (disease) from 1 (no symptoms) to 9 (severe damages with plants not heading)
- “Blast.resistance”: rice resistance to foliar blast (disease) from 0 (Highly resistant) to 9 (Very susceptible)

Quality-related attributes

- “Amylose.content”: Amylose content in percentage
- “Alkali.spreading.value”: Alkali Spreading Value (ASV) related to the gelatinization temperature of the seed
- “Protein.content”: Protein content in percentage

Concerning our next steps, we chose three target variables to explain and reflect the measure of the quality, durability and yield of rice production. We will focus on *Seed.number.per.penicle*, *Amylose.content* and *Straighthead.susceptability*. The following plots describe the partition between the target variables.



Locality, spread and skewness groups of the selected target variables

Indeed, we see that for the variable *Seed.number.per.panicle*, the median is located around 4.8, with a bit of outliers on the left, which we do not want. Concerning the *Straighthead.susceptability* distribution, half of the values seem to be contained between 6 and 8, which has to be moved much more to the left, in order to have a much more resistant rice. Indeed, the perfect situation would be to have an indicator of 1. Finally, when looking at the *Amylose.content* variable, the distribution seems to be satisfaying, even with a couple of outliers. The goal will be to maximize the amount of amylose, which will make more rice in volume, so much more yield.

As said, the goal is the durability, quality and yield of O. Sativa production. To achieve this goal, we will have to use statistical methods to identify genomic regions (from the explanatory data) associated with the three presented complex traits (target variable). It is also called *Gene mapping* and it can be used to develop rice varietes.