Study of *Oryza sativa* genotypes involved in plant height using genome-wide association

Adib HABBOU, Alae KHIDOUR

Table of contents

Introduction

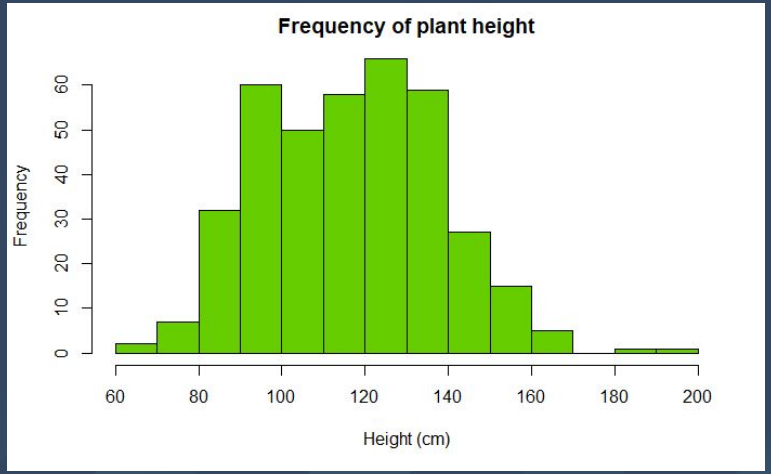
- Data Science can now model real life problems
- The growing demand for food is one of the greatest global challenges
- Studying *Oryza sativa* genotypes can help to improve rice yield
- Our database is the mix of “Oryza SNP” and “OMAP” projects where the information were collected from 82 countries [1]
- The final goal is to identify and locate the most relevant genes involved in a specific phenotype

[1] Keyan Zhao “Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa”

Data Exploration

- Our database contains:
  - 44 100 SNP variants in 413 *Oryza sativa* accessions
  - 36 morphological, developmental, and agronomical phenotypes

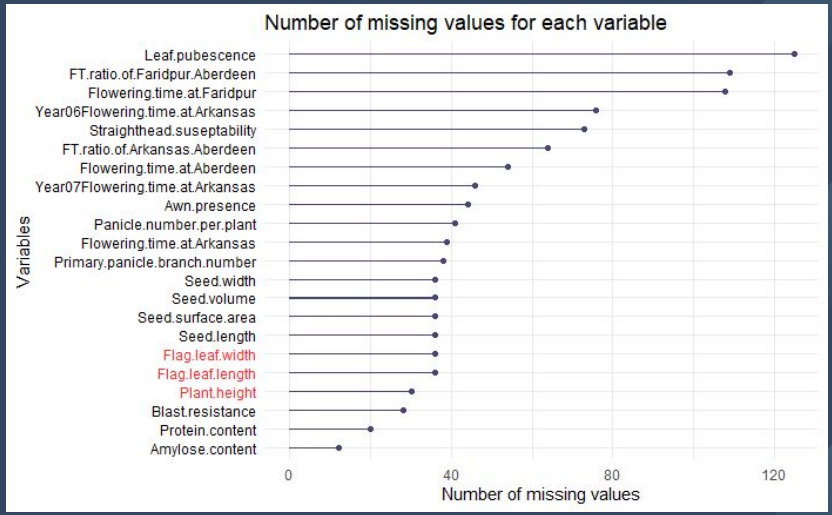
- We have chosen to study the plant height because it's an important developmental and yield characteristic [1]



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

Data Preparation

- We delete all individuals with less than 0.1% of homozygotes with minor allele [2]
- We delete all individuals with more than 10% of missing values [2]
- We replace remaining missing values in each gene by the most frequent value
- For plant height we replace missing values with the mean



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

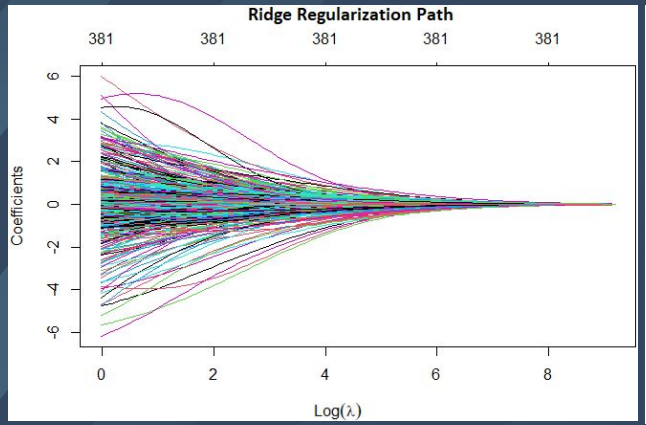
Variable Selection

- After dealing with missing values we went from 413 to 382 observations, our dataset have now 36 902 variables for 382 observations so we need to do variable selection
- Simple regression model cannot compute coefficients for variables which are highly correlated therefore we select only the non correlated variables (1144)
- Forward and Stepwise regression gives the same model, so we are going to keep only the variables they selected (381)

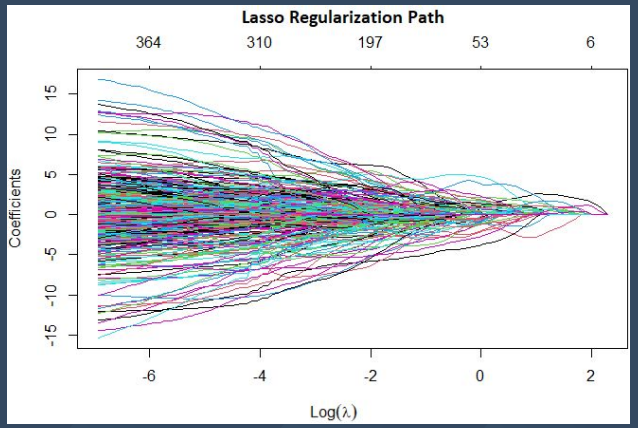
Penalized Methods

- In order to avoid overfitting we need to penalize our objective function

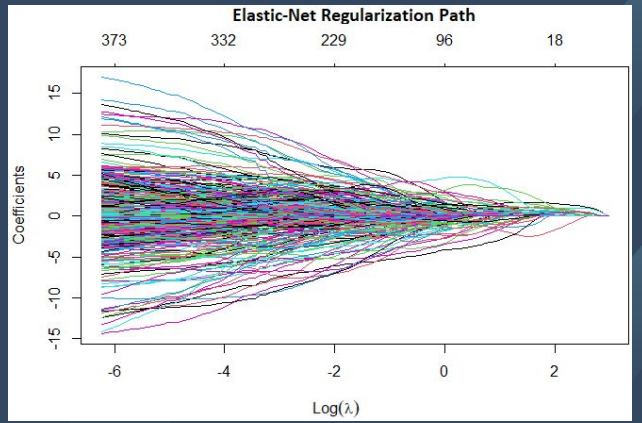
Ridge: we penalize with a L2-term



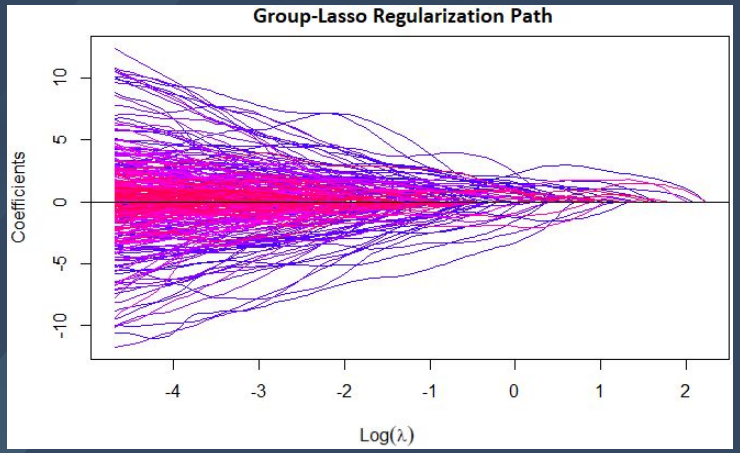
Lasso: we penalize with a L1-term



Elastic-Net: we penalize with a L1-term and L2-term

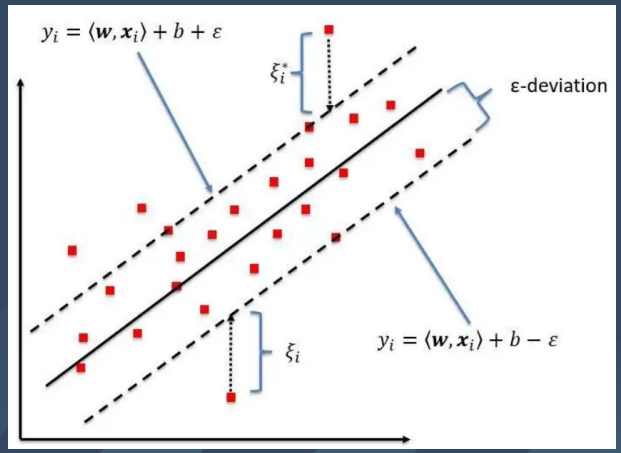


Sophisticated Methods



- Group-Lasso applies a penalization with L1-term and takes in consideration the group structure of the variables

- Support Vector Regression find the hyperplane that separates our observations and maximizes the margin which is the distance that separates a hyperplane from the closest observation



Results

Coefficient of determination (R^2)

RIDGE	LASSO	ELASTIC-NET	GROUP-LASSO	SVR
0.51	0.92	0.85	0.91	0.81

- Ridge is clearly not relevant because it does not do variable selection
- Elastic-Net was computed with $\alpha = 0.5$ which implies that we keep more variable at the end than with a Lasso
- Support Vector Regression could be optimised by looking for the best value for the deviation
- Lasso and Group-Lasso seem to be the most suitable models due to the implicit variable selection: they only keep the most significant variables

Results

- Knowing that we have chosen a Lasso Model, we had two different choices:
  - λ_{min} : which is the model with the best predictive power
  - λ_{1se} : which is the model with a standard deviation equal to 1
- The first model keeps 207 genes, meanwhile the second model keeps only 164 genes which are mainly involved in plant height

List of genes mainly involved in plant height obtained using Lasso Regression with λ_{1se}

"id1000852" "ud7001348" "id7004052" "ud1000154" "id1004393" "id1000051" "id1000994" "id3015622" "wd7002954" "wd3001877" "id7004166" "id3012431"
"id3008445" "id7002916" "id7003169" "id3008476" "ud7001215" "id3014217" "id3011218" "id7003112" "ud7001352" "id3013258" "id1000026" "id1003248"
"id7003361" "id7003332" "id7003377" "ud7001201" "id3009600" "id7002708" "id1000399" "id7002717" "id1000660" "id1002920" "id7002946" "id1000685"
"id1000858" "id1001692" "id3011075" "id4000010" "id3015885" "id7003201" "id1002509" "id7003704" "ud7001582" "id3009868" "id7004086" "id1000528"
"id1000529" "id3008355" "ud7001515" "id1003839" "id3012912" "id3008721" "id7002795" "id7003442" "id1003840" "id7002799" "id3009602" "id1001049"
"ud7001237" "id7002775" "id7003947" "id3011059" "id4000756" "wd1000189" "id7004032" "id1002087" "id7003555" "wd7003065" "id3016604" "id3009692"
"id7003576" "id1002919" "id7002918" "id3017376" "id7003475" "id3010236" "id7003195" "id3011813" "id3009243" "id3009276" "id1000007" "id3011640"
"id3013989" "id7002923" "id3012384" "id1001915" "id1001509" "id1000857" "id4000164" "ud7001180" "id1000980" "id3008682" "ud3000950" "id3013397"
"id3011115" "id1001516" "id3012248" "id7002838" "id3017406" "dd3000953" "id3017627" "id1000661" "id3010370" "id3017179" "id3011960" "id1000027"
"id7003931" "id1003919" "id3011044" "id1005297" "id1000030" "id3008411" "ud1000187" "id7005398" "id1001266" "id1001638" "wd7002914" "id3017002"
"id7002749" "id7003060" "id3009947" "id1000731" "id7003641" "id7004040" "id7003039" "id7003877" "id3008674" "id1003877" "id7003036" "id3017136"
"id3016668" "id7004170" "id7003467" "id3014850" "id1001128" "id7002788" "wd7002589" "fd10" "id1003368" "id3015084" "id3016043" "id1004633"
"id1003465" "id1001003" "id7003078" "id3009379" "id3009824" "id7002999" "id3016447" "id1001332" "id7003043" "id1001936" "id1004872" "id1002730"
"id1000423" "id3009934" "id1000841" "ud7001431" "id3016565" "id7003040" "id1001318" "id3008844"

Conclusion

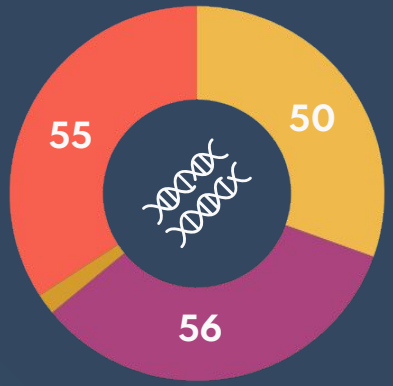
- In the field of genome-wide association, dozens of genes regulating plant height in rice have been identified related to harvest index and yield [1]
- The 164 genes we were able to identify are spread out such that:

31% Chrom 1

34% Chrom 7

33% Chrom 3

2% Chrom 4

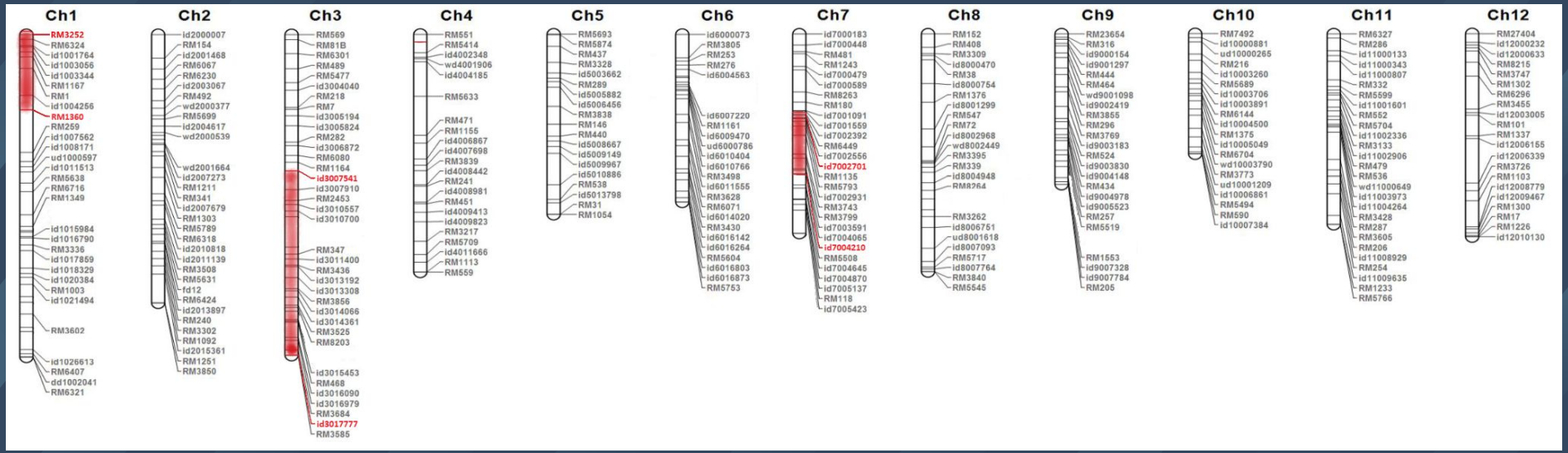


	Position	
Chrom	Min	Max
Chrom 1	75 852	7 016 392
Chrom 3	16 865 092	35 956 835
Chrom 7	16 956 824	23 473 128

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

Conclusion

- Our project leads us to conclude that the most influential genes for plant height are mainly located in the chromosomes 1, 3 and 7



Jong-Min Jeong, Youngjun Mo, Ung-Jo Hyun, Ji-Ung Jeung "Identification of Quantitative Trait Loci for Spikelet Fertility at the Booting Stage in Rice (Oryza sativa L.) under Different Low-Temperature Conditions"