Wageningen University

Haplotype-based genomic prediction in polyploids

Master thesis

Student: Pallozzi Simone
Student number: 1038458
MSc Plant Science – Online Master's Plant Breeding
Supervisor:
Dr. Chris Maliepaard
Dr. Roeland Voorrips
Dr. Giorgio Tumino

Haplotype-based Genomic prediction in polyploids

Course
Title: Thesis-research Practice
Code: PBR80464
Credits: 39
Submission date: 4 June 2024

Author
Name: Simone Pallozzi
Student number: 1038458
Programme: MSc Plant Sciences – Master's online Plant Breeding
Email: simone.pallozzi@wur.nl

Supervisor
Name: Dr. Giorgio Tumino
Email: giorgio.tumino@wur.nl
Name: DR. Roeland Voorrips
Email: roeland.voorrips@wur.nl
Name: Dr. Chris Maliepaard
Email: chris.maliepaard@wur.nl

Institute
Name: Wageningen UR Plant Breeding
Address: 6708 PB Wageningen

ABSTRACT

Single nucleotide polymorphisms (SNPs) are widely used in genomic prediction (GP) as bi-allelic markers. Dense genome wide SNP markers are deployed to detect small QTL effects affecting quantitative traits, since it is supposed that all QTLs are in linkage disequilibrium (LD) with at least one marker. However, bi-allelic SNP markers may have some limits in capturing allele effects of multi-allelic QTLs, that are expected to occur more often in polyploids. Therefore, haplotype markers were developed since they are more informative than the bi-allelic SNP and are able to better capture effects of multiple QTL alleles. In this study genomic prediction with bi-allelic SNP and haplotype markers was compared.

We simulated a genotypic and phenotype data of a population of 500 autotetraploid individuals with 4 chromosomes. Phenotypes were derived from six traits architectures with three level of heritability and the causal loci were set as multi-allelic. Shrinkage and variable selection models for genomic prediction were tested with bi-allelic SNPs and multi-allelic markers and predictive abilities were compared.

The performance of haplotype-based GP did not show improvements in comparison to bi-allelic SNP markers, that is not in line with theoretical expectations.

We investigated the reason why haplotypes could not bring improvements in GPs. As first step, we tried to understand the effects of population structure in GP models. Population structure seemed not to interfere significantly with GP. However, SNP markers were denser than loci used for haplotypes and therefore SNPs could presumably capture the variation given by the relation among individuals.

Then, we investigated the effects of single SNP markers and single haplotype marker alleles in GP models. Haplotypes were highly informative and theoretically could capture each single multiallelic QTL effect with each haplotype variant. However, only a few haplotype alleles of the same locus were contributing to estimating the QTL effects, losing the advantage of being multiallelic. Rather, the haplotype marker contributed in GP similarly to biallelic SNP markers.

We recommend either to explore other models which may be more suitable for the haplotype markers or investigate other methods to include haplotypes in the GP models.

Contents

INTRODUCTION

The use of molecular markers has opened new doors to implementing advanced breeding procedures both for plant and animal breeding. At the beginning of this technology, markers were costly and difficult to reproduce (Mammadov et al, 2012). However, the potentiality of the use of markers was high in breeding and, in fact, it was exploited already at the end of the '80s (Paterson et al., 1988; Soller & Plotkin-Hazen, 1977; Soller, 1978). However, there were some pitfalls in marker-assisted selection (MAS), which limited the efficiency of markers such as the low density of marker maps, insufficient prediction accuracy and limited software packages (Collard & Mackill, 2008). Thanks to the advancement in high-throughput DNA sequencing, nowadays the most used markers at DNA level are Single Nucleotide Polymorphisms (SNPs), which allow the researcher to obtain an accurate and dense map of an organism's genome. Nowadays, this advancement of high-throughput DNA sequencing techniques allows a relatively easy and cost-effective way to obtain a large number of molecular markers (Baird et al., 2008; Davey et al., 2011; Elshire et al., 2011; Guepta et al., 2008). SNPs come from the comparison of genomic DNA sequences of different individuals, or within heterozygous individuals, and in some positions two or more bases may differ from each other (Syvänen, 2001). SNPs have the advantage to be abundant and widespread in many species' genomes in both coding and non-coding regions.

Statistical models were developed to efficiently use the genotypic information available such as genome wide association studies (Loos, 2020) and genomic prediction (GP) (Meuwissen et al., 2001). In fact, Meuwissen et al. (2001) suggested to utilize the genome-wide SNP markers to detect small effect QTLs affecting quantitative traits, since it is supposed that all QTLs are in linkage disequilibrium with at least one marker.

However, individual SNPs may show some limits in capturing LD with QTLs. Therefore, the information from biallelic SNPs may be integrated when two or more adjacent SNPs are grouped defining haplotypes. In fact, Rafalski (2002) suggested that the use of haplotypes is more informative than individual SNPs and haplotype-based analysis has more power in analyzing association between markers and phenotypes.

GP is based on using a training population, which consists of observed individuals for which there is genotypic and phenotypic information available. This leads to the possibility of predicting phenotypic values of not phenotyped but genotyped individuals of the population. In fact, GP models figure out estimations of marker effects with which it is possible to predict the genomic estimated breeding values (GEBV) (Crossa et al. 2017;Goddard & Hayes, 2007; Meuwissen et al.,

2001).This approach can have a huge impact on the breeding process because breeders may perform preliminary selections based on genotypic data only and boost the efficiency of their breeding programs. This advantage is given by skipping that cumbersome step of phenotyping which may require lots of time, such as waiting for the right phenological stage, and which often requires many resources.

Despite the very promising use of GP in breeding, there are some complexities. Genomic predictions are made by using the SNP markers available and this causes a high dimensionality of the data. The problem is that for each marker there is a parameter to estimate, but since there is a larger number of markers (p) than the observed individuals (n), this causes an over parameterization of the model (Crossa et al., 2017, Meuwissen et al., 2001). This is known as "large p, small n" problem (p>>n). Consequently, there is the need of using models which can deal with this complication. To this end, a range of parametric, semi-parametric and non-parametric models have been developed and adapted for genomic prediction (de Los Campos et al. 2013).

Parametric models are divided according to the penalty function which is used to perform variable selection, shrinkage of estimates, or a combination of both (De Los Campos et al. 2013).

The effect that the penalty function exhibits on coefficients is different because the components added to the sum of squares would shrink all coefficients toward zero, making them small, or would shrink some coefficients to zero but others not, making therefore selection of variables. The first case refers to Ridge regression (RR) (Endelman, 2011; de Los Campos et al, 2013) and the second to least absolute shrinkage and selection operator (LASSO)(de Los Campos et al., 2013; Tibshirani, 1996). These two methods may be merged giving the so-called Elastic Net (EN), which includes in the model either penalty functions (Hui Zou and Trevor Hastie, 2005). Another class of models is the Bayesian shrinkage methods. In this type of estimations, the model is based on a prior distribution of the unknowns, namely μ, β, $\sigma 2$. The prior densities of marker effects will bring about variable selection or shrinkage. Also, it would determine the amount of shrinkage (de Los Campos et al. 2013). A non-parametric model for genomic prediction, reproducing kernel Hilbert spaces, was proposed by Gianola and van Kaam (2008). This model accounts for non-additive effects instead of only the additive breeding value (Piepho et al., 2008).

However, models depend on several factors which affect prediction accuracy such as trait architecture (de Los Campos et al., 2013; Meher et al., 2022), population structure (Werner et al., 2020), heritability (Crossa et al., 2017; Cuyabano et al., 2015; de Los Campos et al., 2013), linkage disequilibrium (Crossa et al., 2017; Cuyabano et al., 2015; de Los Campos et al., 2013), errors in the

data and the inclusion of prior knowledge in the analysis (Sarinelli et al., 2019: Zhang et al., 2014), just to mention a few.

Because of the huge advantage that genomic prediction may provide for breeding programs, it is an important matter of research for plant breeding studies. But, research is mainly focused on diploid plant species and the application on polyploids has been left behind (Wilson et al. 2021). The genetics of autopolyploids has more complex patterns than diploids, since they have more than two homologous chromosomes. As a consequence, there are more possible allelic dosages and more complex inheritance patterns.

However, autopolyploid plants may benefit from GP to accelerate and render the breeding program more efficient.

Haplotype and haploblock

SNP markers can be grouped into what is called a haplotype. Haplotypes are defined by regions of the genome in which two or more SNP markers are grouped and are in strong linkage disequilibrium, with almost no recombination between them. Haplotypes are detected by the process called haplotyping. These methods are based either on sequence reads or SNP arrays. Some existing methods are described by Browning and Browning (2011), but these are for diploid organisms. However, some methods were proposed for polyploids based on sequence reads such as Compass (Aguiar and Istrail, 2012), HAPLOSWEEP (Clevenger, 2018), Tripoly (Motazedi at Al, 2018) and PopPoly (Motazedi, 2019), and Poly-Harch (He et al, 2018). Other methods based on allele dosages derived from SNP arrays have been implemented for polyploids such as SATlotyper (Neigenfind , 2008), polyHap (Su et al., 2008), SHEsisPlus (Shi and Eh, 2005) and Happy-inf (Willemsen et al., 2018). Since these models do not include pedigree information, Voorrips and Tumino (2022) developed a software package PolyHaplotyper which makes use of full-sib families.

Haploblocks can have different number of SNPs determining the length and the number could be arbitrarily chosen. Many studies agree that the number of SNPs per haplotype should be between 4-10 (Calus et al., 2009, Villumsen and Janss, 2009, Hess et al., 2017). Constructing haplotypes with a fixed number of SNPs is a common method in haplotyping (Calus et al., 2008, Villumsen and Janss, 2009, Hess et al., 2017). Another example of haploblocks construction is given by Thérèse Navarro et al. (2022), who grouped 6 consecutive SNPs with an overlap of 4 SNPs. A further alternative was explored by Cuyabano et al. (2014) who considered linkage disequilibrium to group SNPs in haploblock given that the LD and the pattern of recombination differs across the genome.

One of the advantages of using haplotypes as markers is the feature of these being multiallelic. This provides stronger information for association studies such as QTL mapping or genome-wide association studies (GWAS). The association between marker alleles and QTL alleles is more likely to occur with multi-allelic markers than with bi-allelic SNP markers. This happens because a mutation may alter the haplotype frequency, whereas may not alter the allele frequency of a biallelic SNP, loosing therefore the ability to capture the QTL effect.

Multiallelic markers help in QTL detection in population composed by polyploids or coming from more than two parents. In fact, Thérèse Navarro et al. (2022) developed a model for QTL analysis based on multiallelic haplotypes and suggested that the use of multiallelic models is more powerful in QTL detection in multiparental and/or polyploid populations compared to biallelic models.

Unsurprisingly, haplotypes find applications also in genomic prediction and their contribution seems promising. The incorporation of haplotypes in animal breeding gives higher prediction accuracy as shown by Cuyabano et al. (2014). They incorporated haploblocks into best linear unbiased prediction (BLUP) and Bayesian mixture models, the latter showed the best results. Matias et al. (2017) found that the use of haploblocks increases predictive ability of genomic prediction in breeding population of allogamous plants or plants with high multiallelism.

Population structure

Population structure is an important factor which may affect genomic prediction accuracy. Population structure may occur because of several reasons, for instance geography reasons, natural selection or artificial selection (Yu et al. 2006; Price et al. 2010). In addition, populations in plant breeding may show structure due to diverse genetic background and family structure (Werner et al., 2020). Population structure brings variability of allele frequencies and the degree of relationship between subpopulations is captured by markers, causing inflation of the predictive ability of GP models (Daetwyler et al., 2012; Habier et al., 2007). If the population structure diminishes due to crossing or selection within subpopulations, markers involved in the relationship between individuals will lose their meaning and GP accuracy will increase. Population structure can affect the accuracy of genomic prediction and decline the genetic gain in a breeding program. In fact, in a previous work conducted by de Valk (2023) suggested to investigate population structure for every trait. Therefore, it would be interesting to quantify how much population structure affect the genomic prediction analysis.

Trait architecture

Trait architecture is an important aspect to consider because it affects genomic prediction (de Los Campos, 2013). Many phenotypic traits are quantitative which are controlled by many QTLs. Therefore, several scenarios could occur such as traits affected by one or only a few main effect QTLs, or controlled by many small effect QTLs or even there is a situation that is a mix of the two. Moreover, QTLs may have the same size effect on the trait, or as most probable in real situation, they may have different weight on the phenotype observed. This could happen with one or a few major QTLs plus many more with small effect. In this case, for example, QTL effects may follow a normal distribution. QTLs can also be multiallelic which means they have more than two variants at the same locus which would affect the same trait. In this case to tag multiallelic QTLs, biallelic markers should not be sufficient and therefore it could be advantageous to use multiallelic markers.

Research questions

The main object of this work was to compare the performance of haplotypes-based genomic prediction with SNP-based genomic prediction. Besides this we wanted to explore how trait architecture and heritability could affect genomic prediction based on haplotypes. As consequence of the first results, we moved our attention to evaluate the contribution of population structure in GP. Then we also assessed the effects of single SNPs and haplotype variants in explaining the phenotypic variance and how they were included in GP models.

MATERIALS AND METHODS

Population

For this study, a population of 500 individuals was simulated by using PedigreeSim software (Voorrips and Maliepaard, 2012). We simulated this population starting from a genetic map with 4 chromosomes, 100 cM long and 1001 loci: a locus at every 0.1 cM. The chromosomes were named A, B, C and D. The loci were numbered from 0000 and to 1000. The name of each locus is then given by the combination of the chromosome letter and the position along the chromosome. We used this genetic map for a tetraploid population with 4 chromosomes (*2n = 4x = 16*) and 4004 loci across all chromosomes.

The next step was the random generation of six different, chromosome-long, haplotypes, with 25% haplotype specific SNPs, used as founder haplotypes for the population. These haplotypes were assigned to 10 founder individuals according to the frequencies used in *Vos et al.* (2017) (0.30, 0.25,

0.175, 0.125, 0.10, and 0.05). Founder individuals, identified as generation zero, were used to simulate the first-generation individuals. In this first generation three traits (X, Y, Z) were set up in order to create three subgroups whose names are respectively X, Y and Z. The proportions of these three subpopulations in each generation was 0.5, 0.3 and 0.2 for X, Y and Z, respectively. Then, 15 generations were simulated applying a random mating scheme within the subgroups, with the mating between subgroups limited to 5%. Selection intensity was 0.5 in every generation and for each trait. The target heritability 0.6 for all three traits. Therefore, each subpopulation was subject to selection for one of the trait and this caused a clear population structure.1. Population structure was detected by principal components analysis (PCA) and then the most significant PCs were regressed against the phenotype to estimate the contribution of population structure in explaining the genetic variation. In addition, the ANOVA of phenotype against subpopulation groups was computed. The variance explained by PC or subpopulations, were in relation to the residual variance obtained by the analysis and phenotype variance according to the following formula

1 – Residual variance/phenotype variance.

GP with no LD with QTL

To further explore the contribution of population structure in GP, a new set of SNP markers were built. For this experiment the monogenic trait was studied. SNP markers in LD with QTL were removed using the distance at which LD drops below 0.1 for 90 % of markers pairs (de Valk, 2023). Genomic prediction were computed with this marker set to detect the effect of population structure and see how contributed the effective LD between marker and QTL-.

Trait architecture

To build the final population, three traits X, Y and Z were generated which then were used to construct the population structure. In fact, the population used for this study had three subpopulations relating to these three traits . Each trait was affected by eight biallelic loci and there were no loci overlapping between these traits. A total of twenty-four loci, involved for the traits X, Y and Z, were removed from the marker set available for simulating the traits for the genomic predictions. For the genomic prediction analysis, we simulated a set of six traits which differed from each other by their genetic architecture. Each trait had different causal loci involved randomly selected from a pool of loci located at integer positions in the genetic map. The simulation consisted of building traits affected by one or more loci, moving from the simplest to a more complex

configuration. The trait architecture reproduced was monogenic, oligogenic, polygenic and a mix of oligogenic and polygenic. The monogenic trait consisted of one multiallelic locus, with six alleles, which affects the phenotype. In our work we denote this architecture as “mono”. For the oligogenic trait there were two configurations. One involves one multiallelic locus, with six alleles, and three biallelic loci, whereas the other consisted of four multiallelic loci each with six alleles. These two trait architectures were named “oligo_1” and “oligo_4”, respectively. The polygenic trait is affected by 100 biallelic loci. These causal loci were distributed equally across the chromosome but randomly within the chromosome,(table 1). With regard to the mixed genetic configuration, there were two types of architectures. One was created by mixing “oligo_1” and “poly” and the other by mixing "oligo_4” and “poly”; their respective names were “mix_1” and “mix_4”. This resulted in 4 major causal loci, 1 multiallelic locus and 3 biallelic loci for “mix_1” and four multiallelic loci for “mix_4”, explaining 50 % of the genetic variance and 100 minor biallelic loci explaining the remaining 50 % of the genetic variance (table 2).

Table 1 Configuration of mono, oligo and poly trait architectures.

	biallelic loci	multiallelic loci
Mono		1
oligo_1	3	1
oligo_4		4
Poly	100	

Table 2 Configuration of mix traits. Major QTL are the ones producing half of the total genetic variance, whereas minor QTLs are producing the other half of the genetic variance.

	major QTLs		minor QTLs	
	biallelic	multiallelic	biallelic	multiallelic
mix_1	1	3	100	
mix_4		4	100	

Phenotypic data

In this work we studied the genomic predictions of individuals whose phenotype is affected by different genetic architectures. To simulate the phenotype, we assigned values to the alleles of the causal loci and these values vary according to the number of alleles. An effect was assigned to each allele, both in the biallelic case and in the multi-allelic case. Then, we assume an additive model, meaning that the effects assigned to a locus over the four homologs were summed up. In case of a biallelic locus in which there are four alleles like 0,0,1,1 and the effects are 0 for the 0s and 1 for the 1s, the phenotype is 2. That principle was also applied to multiallelic QTLs. There was also additivity between the causal loci in case of polygenic traits and therefore the phenotype was formed from the summation of effects of all QTLs involved.

In our population study, there were 6 haplotypes since 6 chromosome-length haplotypes were used for building our population. By using files generated by PedigreeSim, we could track the haplotypes segregation and obtain the genotypes of the individuals in the 15th generation with the right haplotype at each locus (fig. 1).

	G15_X001_1	G15_X001_2	G15_X001_3	G15_X001_4	G15_X002_1	G15_X002_2	G15_X002_3	G15_X002_4
A0000	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0001	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0002	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0003	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0004	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0005	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0006	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0007	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0008	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0009	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0010	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0011	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0012	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0013	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0014	HP_4	HP_1	HP_1	HP_4	HP_1	HP_1	HP_1	HP_1
A0015	HP_4	HP_1	HP_1	HP_4	HP_5	HP_1	HP_1	HP_1
A0016	HP_4	HP_1	HP_1	HP_4	HP_5	HP_1	HP_1	HP_1

Figure 1 Example of genotypic information of individuals at the 15th generation by using haplotypes. At the side of the figure are the loci names. The column names combine the generation number, the name of the individual , and the homolog number.

Knowing the phased haplotypes, it is possible to assign an effect to each haplotype of a multiallelic causal locus and obtain the genotypic value associated to that locus by using a specific PedigreeSim function. In fact, we considered the haplotypes as alleles; in each locus per homolog, there are six possible haplotype alternatives, and we consider them as alleles. The effects assigned to the QTL alleles are 0, 0.33, 0.66 and 1. After having assigned the allele effects and obtained genetic values of the individuals, we added the environmental noise based on three values for the heritability: 0.2, 0.5 and 0.8. The environmental variance σ_ε^2 was obtained using the formula

$\sigma_\varepsilon^2 = (1-H^2)*\sigma_g^2/H^2$

where H^2 is the target heritability and $\sigma_g{}^2$ is the genetic variance of the simulated trait in the fifteenth generation. Once we obtained the environmental variance, the phenotype is given by adding a normal distributed environmental variation with mean 0 and standard deviation $\sqrt{\sigma_\varepsilon^2}$ to the genotypic value.

After computing the phenotype, it is possible to determine the realized heritability $H^2=\sigma_g{}^2/\sigma_p{}^2$. For the mixture trait architecture an oligogenic part, including the main effect QTLs, was combined with a polygenic part consisting of minor effect QTLs. To combine these two components, the QTL effects were adjusted in order to split the genetic variance between the two components equally. To achieve this, the genetic variance of the polygenic side was scaled according to the oligogenic variance. This proportion was used to obtain scaled biallelic QTL effects for the polygenic part of the mixture trait. By doing this, each of the two components of the trait explains 50 % of the genetic variance.

At the end of this process, for each trait architecture three versions were used according to the three different heritability values. In conclusion, we ended up with a set of 18 scenarios which are in fact given by 6 trait architectures times three heritability values.

Marker data

With this study we wanted to investigate if the use of haplotypes in genomic predictions could improve the predictions based on biallelic marker SNPs. To achieve this, we constructed two main marker sets: one made of biallelic SNPs and the other made of multiallelic haplotype markers. Biallelic marker set: PedigreeSim provided phased genotypes of the individuals of our study population. As first data set, PedigreeSim software provides a matrix with markers in the rows and the four homologs of each individual in the columns (fig. 2- A). This data set was adapted to fit the genomic prediction models; therefore, an allele dosage file was built, consisting of the dosages across the homologs in each chromosome (fig. 2 – B). From this last marker set, loci belonging to the marker pool used as causal loci were removed. In addition, markers with a frequency either below 0.05 or above 0.95 were removed. Ultimately, 3339 SNPs were used for genomic prediction and a variant of this marker set was also used for running genomic prediction analysis. In this variant set, haplotype tagging markers were removed; the final number of markers used in this analysis was 2656.

A

marker	G15_X001_1	G15_X001_2	G15_X001_3	G15_X001_4
A0001	1	1	1	1
A0002	1	1	1	1
A0003	1	0	0	1
A0004	0	1	1	0
A0005	0	0	0	0
A0006	1	1	1	1
A0007	1	1	1	1
A0008	1	1	1	1
A0009	1	1	1	1

B

marker	G15_X001	G15_X002	G15_X003	G15_X004
A0001	4	4	4	3
A0002	4	4	2	4
A0003	2	0	2	1
A0004	2	4	3	3
A0005	0	0	1	0
A0006	4	4	3	4
A0008	4	4	1	4
A0009	4	4	2	4

Figure 2 – A) table with phased genotypes with markers in the first column and individual - homologs in the following columns. In the alleles of markers "A0001" of the individual "G15_X001", highlighted in red, has four copies of "1". B) dosage table with individuals as row names and marker names as columns. In this table which derives from the phased genotypes table (A), 4 is the dosage of the individual "G15_X001" for the locus "A0001" which sums up the four copies of "1" in table A for the respective individual and locus.

Haplotype marker set: For each chromosome, six chromosome-long haplotypes (HP_1 to HP_6) were generated and then assigned to 10 founder individuals which formed the base of our final population. Therefore, the phased haplotypes of each founder individual corresponded to a haplotype in all its chromosome length. In our population, corresponding to the 15th generation, there was a mixture of haplotypes per homolog due to recombination. In the end, we obtained a phased file of founder alleles file with 6 possible alleles per locus per homolog (fig. 3 - A). Similarly to biallelic markers, this data format must be adapted to a format fitting the statistical models used for genomic predictions. It must be as similar as the biallelic dosage file. To achieve this, markers were converted into pseudomarkers. Each locus was implemented with 6 alternatives corresponding to each haplotype allele. The names used for the pseudomarkers have been arbitrarily chosen and they are the combination of the locus name and the haplotype allele name. For example, for the locus A0005, the pseudomarker names are A0005_HP_1, A0005_HP_2, A0005_HP_6 to indicate the six haplotypes. Then, the dosage is the sum of the haplotype alleles present in that locus. For example, if we have individual G15_X001 which has haplotypes HP_4, HP_1, HP_1 and HP_4, the dosage would be 2 for the pseudomarkers A0005_HP_1 and A0005_HP_4 and zero for A0005_HP_2, A0005_HP_3, A0005_HP_5 and A0005_HP_6 (fig. 3).

A

	G15_X001_1	G15_X001_2	G15_X001_3	G15_X001_4
A0000	HP_4	HP_1	HP_1	HP_4
A0001	HP_4	HP_1	HP_1	HP_4
A0002	HP_4	HP_1	HP_1	HP_4
A0003	HP_4	HP_1	HP_1	HP_4
A0004	HP_4	HP_1	HP_1	HP_4
A0005	HP_4	HP_1	HP_1	HP_4
A0006	HP_4	HP_1	HP_1	HP_4

B

	A0005_HP_1	A0005_HP_2	A0005_HP_3	A0005_HP_4	A0005_HP_5	A0005_HP_6
G15_X001	2	0	0	2	0	0
G15_X002	4	0	0	0	0	0
G15_X003	1	2	0	0	1	0
G15_X004	2	0	0	1	0	1
G15_X005	2	1	0	0	1	0
G15_X006	1	2	0	0	1	0
G15_X007	1	0	0	0	3	0

Figure 2 – A) table reporting the haplotypes at each locus for each homolog. The row names are the loci names, and the columns are the homolog of each individual. Haplotypes are named as HP in combination with the haplotype number. B) Pseudomarkers table. Row names are the individuals and columns are the pseudomarkers. There are six (amount of haplotypes) pseudomarkers per locus. In this example there are six pseudomarkers for the locus A0005.

Converting the markers into pseudomarkers increases the number of variables to six times as many as the number of single markers. This can bring about to a very large file and increase the time for running the analysis. For this reason, we decided to reduce the loci involved and keep only the ones at point 5 positions in the genetic map. For instance, we used locus A0005, A0015, A0025 and at position 0.5, 1.5, 2.5 cM and so on. In this way the loci kept were 1 cM apart from each other. Then, from our pseudomarker file, rare pseudomarkers with frequencies below 0.5 and above 0.95 were removed. In the end, we obtained a file with 16603 variables in the format as shown in figure 3 - B with individuals in the rows and pseudomarkers in the columns.

Statistical models

rrBLUP

Endelman (2011) developed an R package to run prediction analysis based on Ridge Regression (RR), which is equivalent to best linear unbiased prediction (BLUP) in the context of mixed models (Whittaker et al., 2000; Meuwissen et al.,2001).

The basic rrBLUP model is:

$$\mathbf{y} = \mathbf{WGu} + \varepsilon \qquad [1]$$

where $\mathbf{u} \sim \mathrm{N}(0, \mathbf{I}\sigma^2_{\mathrm{u}})$ is a vector of marker effects, G is the genotype matrix for biallelic SNPs, W is the design matrix relating individuals to observations (y). The BLUP solution for marker effects is $\hat{\mathbf{u}} = \mathbf{Z}'(\mathbf{ZZ}' + \lambda\mathbf{I})^{-1}\mathbf{y}$, where Z = WG and the Ridge parameter is $\lambda = {\sigma^2_{\mathrm{e}}}/{\sigma^2_{\mathrm{u}}}$. The analysis are run using the RR-BLUP package function mixed.solved().

RKHS

From rrBLUP package by Endelman (2011), there is the possibility to use two more kernels in addition to the realized relationship model: one is the Gaussian model and the other is the exponential model. In our experiments, we used the Gaussian model kernel. From the equation 1 used to estimate breeding values based on the marker effects, there is the equivalent kinship-BLUP equation in which the breeding values are predicted based on the individual kinship:

$$\mathbf{y} = \mathbf{Wg} + \varepsilon$$

$$\mathrm{g} \sim \mathrm{N}\left(0, \mathbf{K}\sigma_g^2\right) \quad [2]$$

where g is a vector of genotypic values, and K is the relationship matrix. With equation 2 it is possible to calculate BLUP based on different kernel functions (K). As mentioned, we used the Gaussian model that is given by equation [3]

$$K_{ij} = \exp\left[-\left(\frac{D_{ij}}{\theta}\right)^2\right] \quad [3]$$

where θ is a scale parameter that determine how quickly the genetic covariance decays with distance and D_{ij} is the Euclidean distance between genotypes i and j normalized to the interval [0,1]:

$$D_{ij} = \left[\left(\frac{1}{4}M\right)\sum_{k=1}^{M}\left(G_{ik} - G_{jk}\right)^2\right]^2 \quad [4]$$

where M is the number of markers and G is the genotypic matrix.

In rrBLUP package the function used for RKHS is kinship.BLUP() where it is specified the parameter K.method = “GAUSS”.

Elastic Net

The R package glmnet developed by Friedman et al. (2010) calculate penalized regression model that combines LASSO and RR by using weighted average of the ℓ1 and ℓ2 norms as penalty function. The Elastic Net regression model estimates the parameters mean and marker effects by solving the optimization problem given in the equation 5:

$$\left(\hat{\mu}, \hat{\beta}\right)_{argmin} = \left\{\sum\left(y_i - \mu - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda j(\beta)\right\} \quad [5]$$

$$j(\beta) = (1-\alpha)\frac{1}{2}\left|\left|\beta\right|\right|_{\ell 2}^{2} + \alpha||\beta_j||_{\ell 1} = \sum_{j=1}^{p}[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|] \quad [6]$$

Where $\sum(y_i - \mu - \sum_{j=1}^{p} x_{ij}\beta_j)^2$ of the equation [5] is the sum of squared residuals and $j(\beta) = (1-\alpha)\frac{1}{2}\|\beta\|_{\ell 2}^2 + \alpha\|\beta\|_{\ell 1}$ [Eq. 6] is the penalty function which is a weighted average of ℓ1 and ℓ2 norms. The two parameters controlling the penalty function are lambda (λ) and alpha (α).

Lambda controls the severity of the penalty function and is internally optimized by using the glmnet function cv.glmnet(). This function computes the 10-fold-cross validation to find the best lambda over a grid of 100 values of lambda. The function returns two values of lambda for which the user can choose from: lambda.min and lambda.1se. From the manual of glment written by Hastie at. All (2023), the two lambda provided by cv.glmnet is explained as "lambda.min is the value of lambda that gives minimum mean cross-validated error, while lambd.1se is the value of lambda that gives the most regularized model such that the cross-validated error is within one standard error of the minimum". In previous analysis (de Valk, 2023), Elastic Net was too stringent for certain combinations of training and tests sets when predicting for polygenic traits with low heritability. This brought about the incapability of calculating the Pearson correlation for the predictive ability because NA were generated (we refer to this problem as NA problem). To minimize this NA problem it was decided to use lambda.min as it provides a less stringent model. As consequence, for the present work we used the same parameter.

The α parameter controls the mix between LASSO (α = 1) and Ridge regression (α = 0) in the penalty function.

For this parameter there is no an internal cross validation cycle provided by the glmnet functions and therefore we set up a cross-validation cycle to select the best α from a range of 11 values (0 to 1 by 0.1) for our genomic predictions. Just after having selected the best α, there is a second five-fold cross-validation cycle run m times to obtain m means of the predictive ability. The fold which returned NA were kept out when calculating the predictive ability mean of the inner-cross validation.

Bayes B

In our study we include a model which performs variable selection from the R package BGLR (Pérez & de Los Campos, 2014).

BGLR implement the linear model

$$y = \mu + X\beta + \varepsilon \qquad [7]$$

$$\sigma_\varepsilon^2 = \chi^{-2}(df_\varepsilon, S_\varepsilon)$$

$$P(\theta|y,\omega) \propto P(y|\theta)P(\theta)$$

Where y is a vector of responses, μ is the overall mean, β is a vector of random marker effects, X is an allelic dosage matrix and ε is a vector of residuals. The residual variance σ_ε^2 has a scaled-t distribution with degree of freedom df_ε (>0) and scale parameter S_ε (>0). The third expression of equation [7], is the conditional distribution of the data and θ represents the unknowns such as the intercept, regression coefficient, random effects and the residual variance.

The prior density for the marker effects in Bayes B is a combination of a point of mass at zero and the scaled-t slab summarized by the expression:

$$P\left(\beta_j\ ,\sigma_\beta^2,\pi\ \right) = \prod_k \beta_{jk} \sim N\left(0,\sigma_\beta^2\right) + (1+\pi)1(\beta_{jk}=0) \qquad [8]$$

Where $P\left(\beta_j\ ,\sigma_\beta^2,\pi\ \right)$ is a function of the scaled-t slab and the point of mass at zero $(1+\pi)1\left(\beta_{jk}=0\right)$. The effects of important markers are sampled from the scaled-t slab which has a large variance that comes from a scaled-inverse χ^2 distribution $\sigma_\beta^2 = \chi^{-2}(df_\beta, S_\beta)$ while the effects of unimportant markers are equated to zero by sampling from the point of mass at zero.

In Bayes B three hyperparameters are used to control the severity of the penalty: df_β is the degrees of freedom for marker effects, S_β is a scaling parameter that is sampled from the gamma distribution $S_\beta \sim G(r,s)$ and π is the proportion of non-zero markers allowed in the model which is sampled from the beta distribution $\pi \sim B(p_0,\pi_0)$ and optimized using an internal validation method. The prior densitie for the remaining unknown model parameters are a flat prior for the overall mean and a scaled-inverse $\chi 2$ density for the residual variance $\sigma_\varepsilon^2 = \chi^{-2}(df_\epsilon, S_\varepsilon)$.
The R package BGLR utilizes a Gibbs sampler to optimize the posterior density. In total we used to 12500 iterations of this Gibbs sampler, of which 2500 were burn in iterations.

Experimental setup

The experiment aimed to compare SNP-based GP with haplotype-based GP under different scenarios consisting of the combination of trait architecture and heritability. Each trait was replicated ten times and the loci for the same trait were different in each replication. For the SNP-based GP was used two set of markers: one complete with 3339 and the other without the haplotype-tagging SNP with a total of 2656 markers. For the haplotype-based GP a set of 16603 pseudomarkers was used. SNP markers and pseudomarkers, for the haplotypes, were normalized to [-1, - 0.5, 0, 0.5, 1], then centered and scaled. In addition, phenotypic data was centered and scaled.

RESULTS

Haplotype vs SNP markers

To investigate the contribution that multiallelic haplotypes may have on genomic predictions, we compared haplotype-based genomic prediction with biallelic SNP-based genomic prediction for a set of traits characterized by different level of heritability and genetic structure. There are several methods to construct haplotypes and together with other factors, such as models, trait architecture etc., can have an impact on prediction accuracies. Some studies are present for diploid species (Cuyabano et al., 2014; Yang Da 2015; Hess et al. 2017; Matias et al.,2017; Ballesta et al., 2019; Won et al. 2020; Lin et al.,2024) and the research on this topic for polyploid species has lagged behind. Multiallelic markers have a high potential to detect the effect of a linked QTL, especially in case of multiallelism (two or more effects). We hypothesized that the use of multiallelic haplotype markers could have a positive contribution on genomic prediction compared to biallelic SNP markers for capturing the effects of multiallelic causal loci.

To this aim, we generated multiallelic haplotype markers composed of one locus with six alleles (corresponding to the founder alleles), then we ran genomic prediction analysis with four different statistical models for each trait at three different levels of heritability. Haplotype- genomic prediction was compared with SNP-based genomic prediction. 14400 observations were collected, and the predictive ability ranged from -0.02 to 0.89. On average Bayes B and Elastic Net yielded higher prediction ability than RKHS and rrBLUP regardless the type of marker used. Looking at the difference between haplotype-based and SNP-based genomic predictions, RKHS and rrBLUP had

nearly the same average predictive ability, while Bayes B and Elastic Net were slightly higher in favor of the biallelic SNP markers.

Predictive ability of haplotype-based GP in relation to the heritability, trait architecture and models had the same pattern of the SNP-based GP analysis and the predictive ability values were quite similar. The main difference was between shrinkage models and variable selection models (Figure 4 and appendix 1), that was in line with previous work (de Valk, 2023). For the simplest genetic configuration, monogenic trait, variable selection models yielded better predictions than shrinkage models at all heritability levels and for both marker sets. Similarly, predictions for "oligo_1" had a general clear difference between the two types of models, however this difference was less remarkable at low heritability. In “oligo_4, where the four causal loci were multiallelic, the gap between variable selection and shrinkage models varied among the heritability levels. In fact, at low heritability all four models gave on average similar predictive abilities, whereas the difference became larger at middle and high heritability. Moving into a more complex configurations, the gap between models is almost nullified and the models performed similarly. The variable selection models were influenced by trait architecture and, in fact, the prediction ability dropped for the polygenic traits at the level of shrinkage models. On the other hand, shrinkage models had a consistent performance among traits despite their genetic complexity. However, for the polygenic traits, BB generally performed slightly better than the other models at all heritability levels for both the marker sets.

Looking at the differences between haplotypes and SNPs (fig. 5), generally SNP-based genomic predictions performed moderately better than haplotype-based genomic predictions. Still there was a distinction between variable selection models and shrinkage models because the differences were clearer for the variable selection model and mainly at middle and high heritability (fig. 4). On the other hand, shrinkage models (rrBLUP and RKHS) on average performed equally between the two marker sets (fig. 5-B). In fact, the gap was in the range of 0.001, an imperceptible difference.

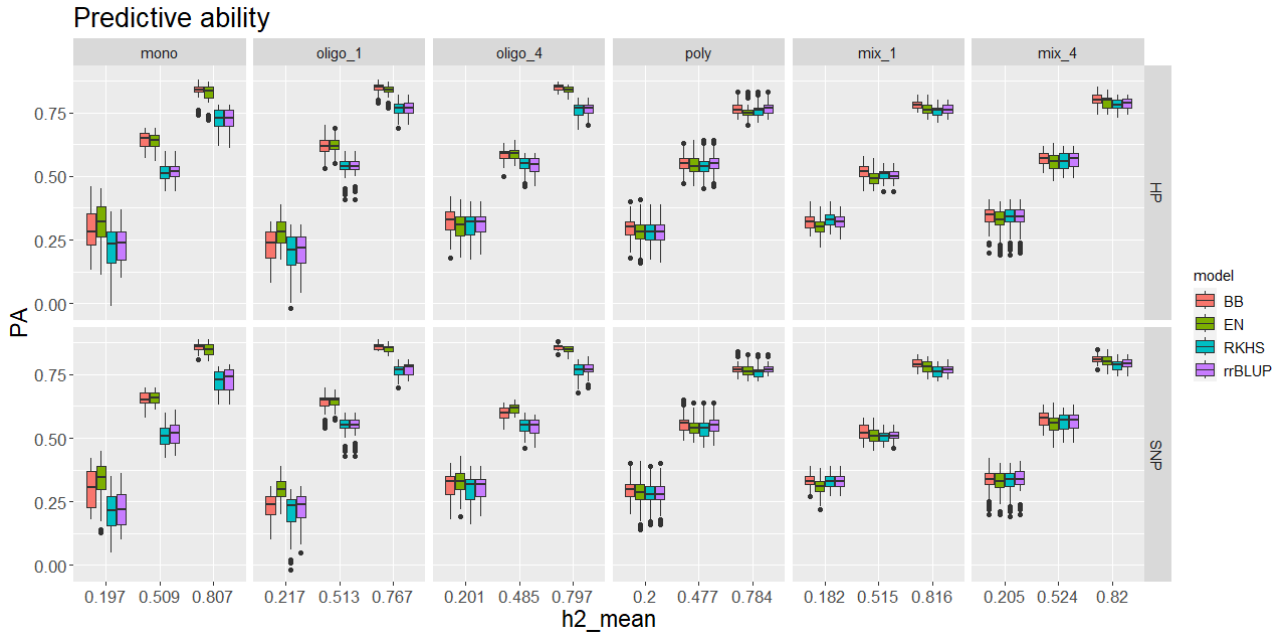

Figure 4 – Boxplots showing the mean genomic prediction results per genomic prediction model. The models represented are BB for Bayes B, EN for elastic net, RKHS for reproducing kernel Hilbert space and rrBLUP for Ridge regression BLUP. Each panel represents the trait architecture as "mono", "oligo_1", "oligo_4", "poly", "mix_1" and "mix_4" in combination with the marker set used for genomic prediction (HP for the multiallelic marker set and SNP for the biallelic SNP marker set). In the x-axis there are the mean heritability values per trait architecture obtained from ten replications; the y-axis shows the predictive ability. On average BB and EN performed better than RKHS and rrBLUP for the monogenic and oligogenic traits. Whereas genomic predictions from BB and EN dropped at shrinkage models outputs for the polygenic traits. However, variable selection models yielded higher predictive ability in almost all scenarios.

Figure 5 – Box plot showing the same genomic prediction results from fig. 1, but grouped by marker sets: HP stands for multiallelic haplotype marker set and SNP for biallelic SNP marker set. Each panel is the combination of trait architecture and statistical model used for the genomic prediction. In the x-axis there are the mean heritability values per trait architecture obtained from ten replications; the y-axis shows the predictive ability. In A) is shown the results for variable selection models (Bayes B and Elastic Net) in which it is visible that SNP-based genomic predictions had a slight higher predictive abilities than haplotype-based genomic predictions in almost al scenarios. Whereas in B) results from shrinkage models (Ridge regression BLUP and Reproducing Kernel Hilbert Space) are shown. In this case the predictive abilities were equivalent and indeed the marker sets used did not entail any difference.

Removing haplotype-tagging SNPs

A Further analysis consisted of removing haplotype-tagging SNPs from the SNP marker set. Since the results of haplotype-based GP were similar to SNP-based GP, we wanted to explore whether haplotype-tagging SNPs had a significant impact on genomic predictions. To this aim, only the monogenic trait was observed, and the average level of heritability values were 0.197, 0.509 and 0.807.

Looking at the overall averages, the predictive ability based on SNP marker set was 0.547 (standard deviation 0.23), for the haplotype-marker set was 0.54 (standard deviation 0.224) and for the non-haplotype-tagging SNP marker set the predictive ability was 0.539 (standard deviation 0.229). The pattern was consistent with the previous analysis: predictive ability was positively related to the heritability values and there was a clear distinction between variable selection and shrinkage models. Within models, the differences are subtle between markers groups. By removing haplotype-tagging SNPs from the SNP marker set, it did not bring any important changes and the average prediction ability was approximately the same as the ones obtained by using the complete SNP marker set.

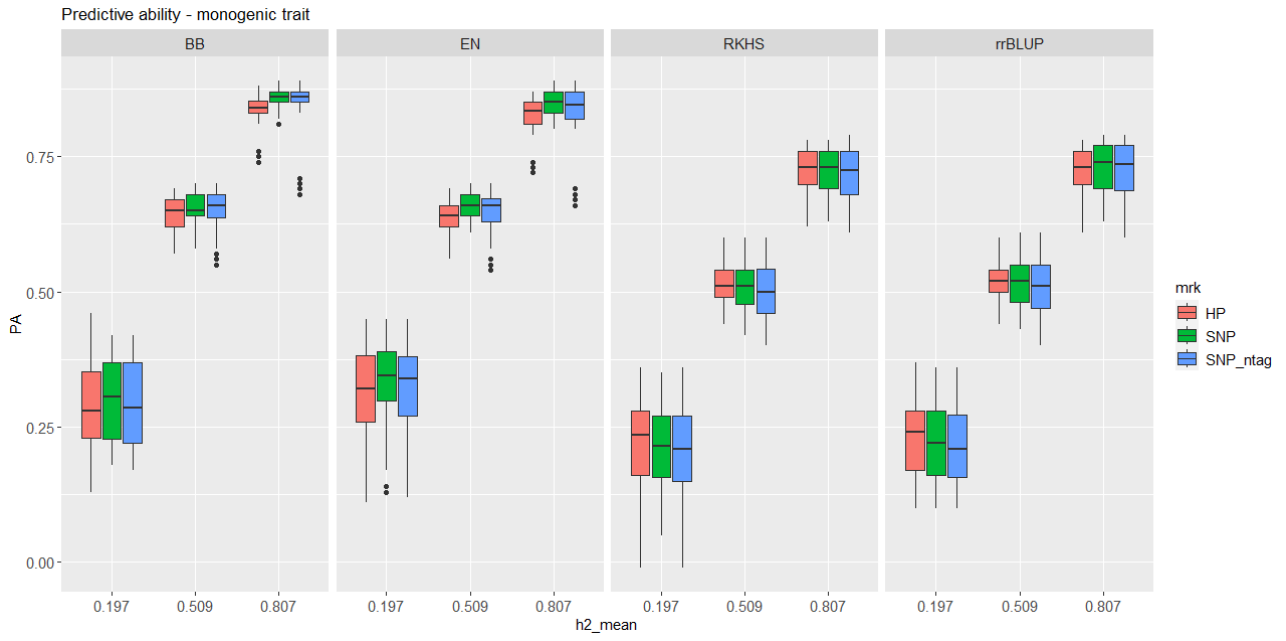

Figure 6 – Boxplot of genomic predictions of the "mono" trait grouped by typology of marker set: HP for multiallelic haplotype marker set, SNP for biallelic marker set and SNP_ntag for biallelic marker set with haplotype-tagging SNPs removed. Each panel represents the statistical model used which are Bayes B (BB), elastic net (EN), reproducing kernel Hilbert space (RKHS) and Ridge regression BLUP (rrBLUP). In the y-axis there is the predictive ability and x-axis the heritability means for the monogenic trait.

Population structure

The comparison between haplotype-based genomic prediction with SNP-based genomic prediction did not meet our hypothesis. In a previous work carried out on the same population by de Valk (2023), it was showed that the population structure had an impact on genomic predictions. Therefore, we first analyzed the genetic structure of the population through PCA (principal components analysis). Subpopulations were clearly separated by means of the first principal components that explained 20.1 % of the genotypic variance. PC1 could cluster the population in two main groups. Subpopulation Y was well separated from the other two subgroups X and Z, which showed to be closely related. The other principal components could not further explain the population structure since they were not able to further separate the population in subgroups (fig. 7).

The pattern generated by PC1 was generally in line with the means of subpopulation genetic values of each trait in most replications (Appendix 2). That was confirmed by the ANOVA in which almost all replications had at least one subpopulation mean differing statistically from the others (tab. 3). After that, we wanted to analyze the variance explained by principal components for the monogenic trait. To achieve this, we computed the analysis of variance on the phenotype at the three level of heritability in relation to an increasing number of principal components and then plot the R-squared against the number of PCs (fig. 8). This result showed that for this population and for this trait the variance explained by PCs had a logarithm fashion path and the initial slop is increasing in relation to the heritability level. However, the curve had a lower increase from 100 PCs onward and the proportion of variance approaches to the heritability levels (0.2, 0.5 and 0.8).

Principal components were fit in genomic predictions as fixed effects to account for population structure as proposed by Daetwyler et al. (2012) in sheep population. With this method, these authors obtained a decrease in predictive abilities as the number of PCs increased in the model till reaching a plateau. They speculate that that plateau is given from LD of markers and QTL as the majority of the population structure effect is accounted for. We carried out a similar analysis including in rrBLUP model up to 350 principal components (fig. 9) for predicting the monogenic trait. For the predictive abilities at low and middle level of heritability, there was a clear decrease from 100 PCs whereas at high heritability the decrease started from around 150 PCs. By contrast, we did not obtain a plateau but a continuous decrease in predictions which became greater as the number of PCs increased.

A further analysis to check if the population structure could affect predictive ability, rrBLUP with a SNP marker set without loci in LD with QTL was launched for the monogenic trait. In fig. 10 the predictive ability is shown and it is notable how much the predictive ability dropped if markers in LD were removed. In table 3 instead are reported the proportion of the across subpopulation groups over the heritability. These values could give a weight of the variance explained by population structure over the genetic variance. It is interesting to see the replicates which had high proportion of heritability across subpopulation had a smaller drop in predictive ability when SNPs in LD with the QTL were removed. The other way round, when this proportion was smaller the drop in predictive ability was greater.

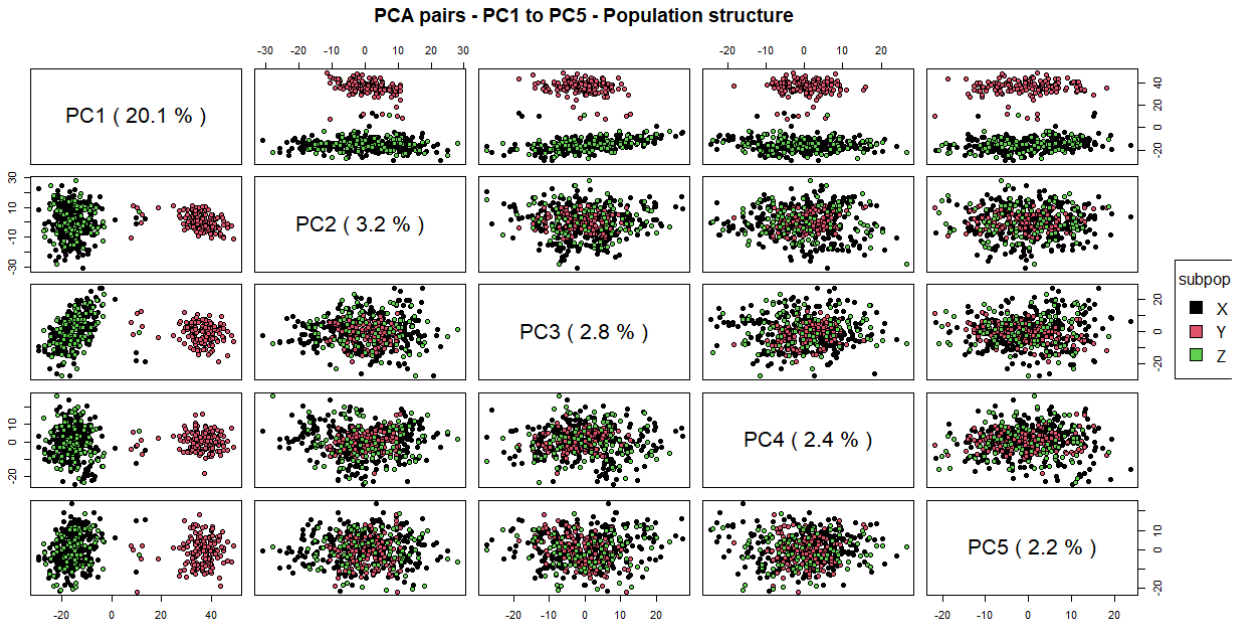

Figure 7 – Pairs of the first 5 Principal components of a principal components analysis (PCA) for the genetic structure of the population used for genomic prediction. The subgroups of the population are X, Y and Z which are clustered by the first principal component accounting for 20.1% of the genotypic variance.

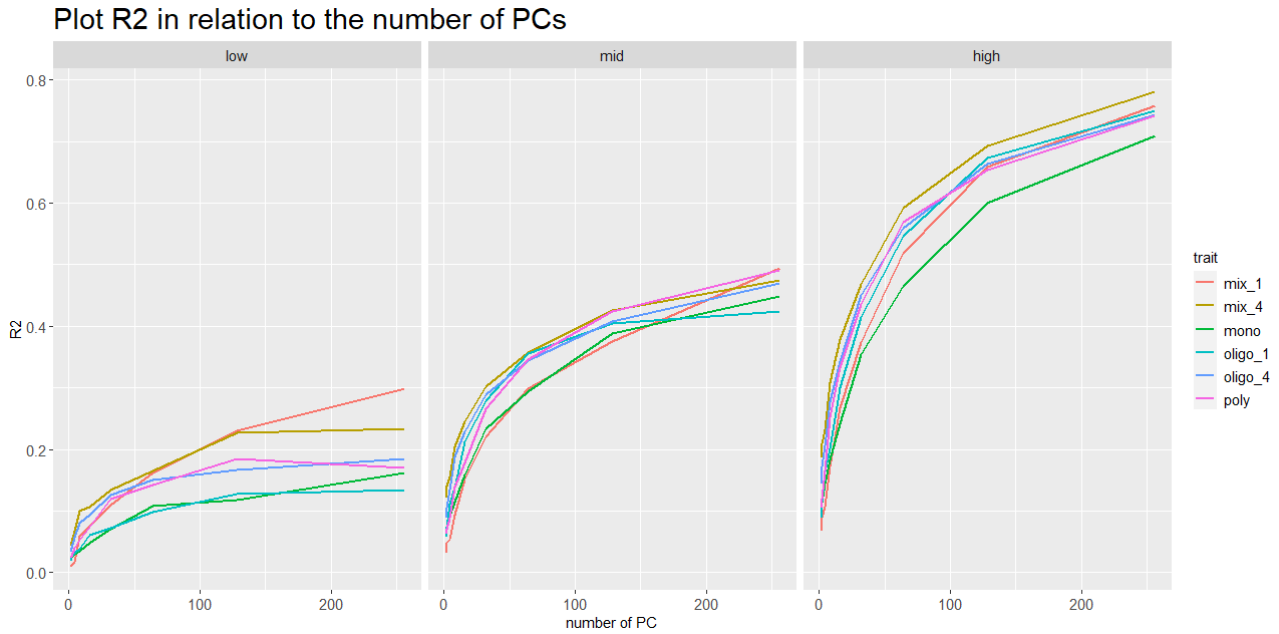

Figure 8 – Plot of the proportion of phenotypic variation explained by an increasing number of PCs. In the X axis there are the number of PCs and in the Y-axis the R2 (proportion of phenotypic variance explained by PCs); each panel corresponds to the low, mid and high heritability which values corresponded to 0.2, 0.5 and 0.8 respectively.

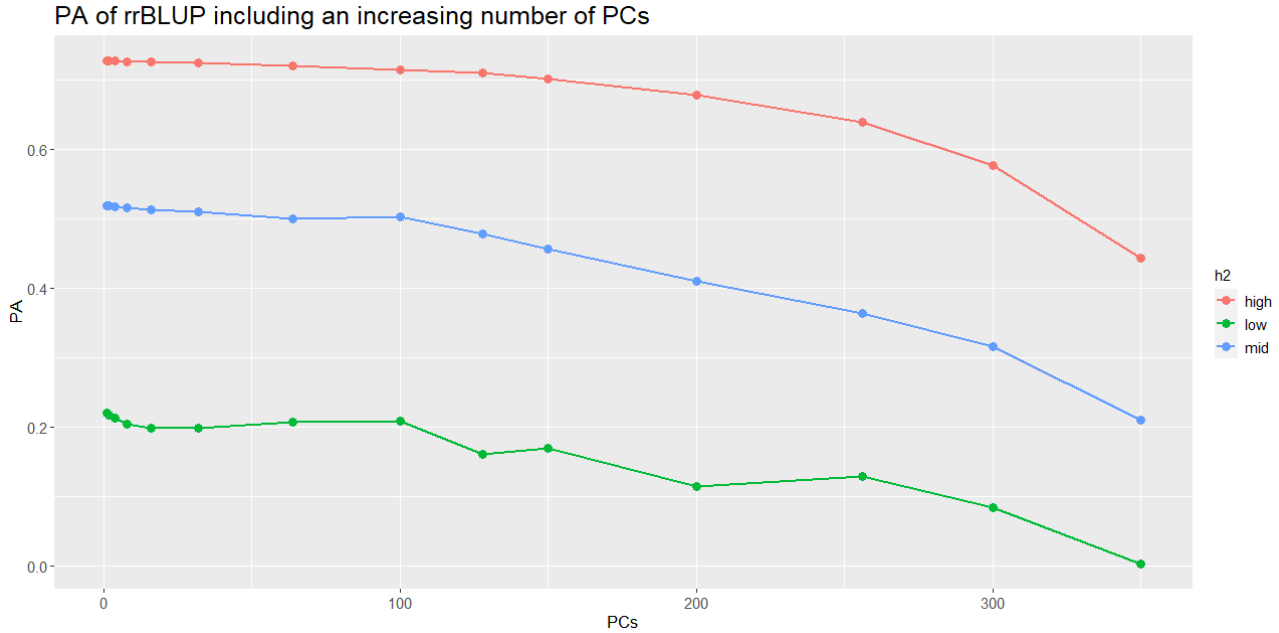

Figure 9 – Plot of predictive ability resulted from including an increasing number of PCs in the base rrBLUP model. Each line corresponds to a level of heritability which are approximately 0.2, 0.5 and 0.8 for low, mid and high respectively.

Table 3 – P-values for the significance of at least one subpopulation mean be different from the other means. When the P-value is below 0.001, for simplicity is indicated as "<0.001". Column names indicate the trait architecture and row names are the corresponding replications.

	mono	oligo_1	oligo_4	poly	mix_1	mix_4
rep_1	< 0.001	< 0.001	< 0.001	0.681	< 0.001	< 0.001
rep_2	< 0.001	< 0.001	< 0.001	0.041	< 0.001	< 0.001
rep_3	< 0.001	0.293	< 0.001	< 0.001	< 0.001	< 0.001
rep_4	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
rep_5	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
rep_6	< 0.001	0.054	< 0.001	0.016	0.933	< 0.001
rep_7	0.154	0.083	0.304	< 0.001	< 0.001	0.116
rep_8	< 0.001	0.106	< 0.001	< 0.001	< 0.001	< 0.001
rep_9	< 0.001	< 0.001	< 0.001	< 0.001	0.313	< 0.001
rep_10	< 0.001	0.083	< 0.001	< 0.001	< 0.001	< 0.001

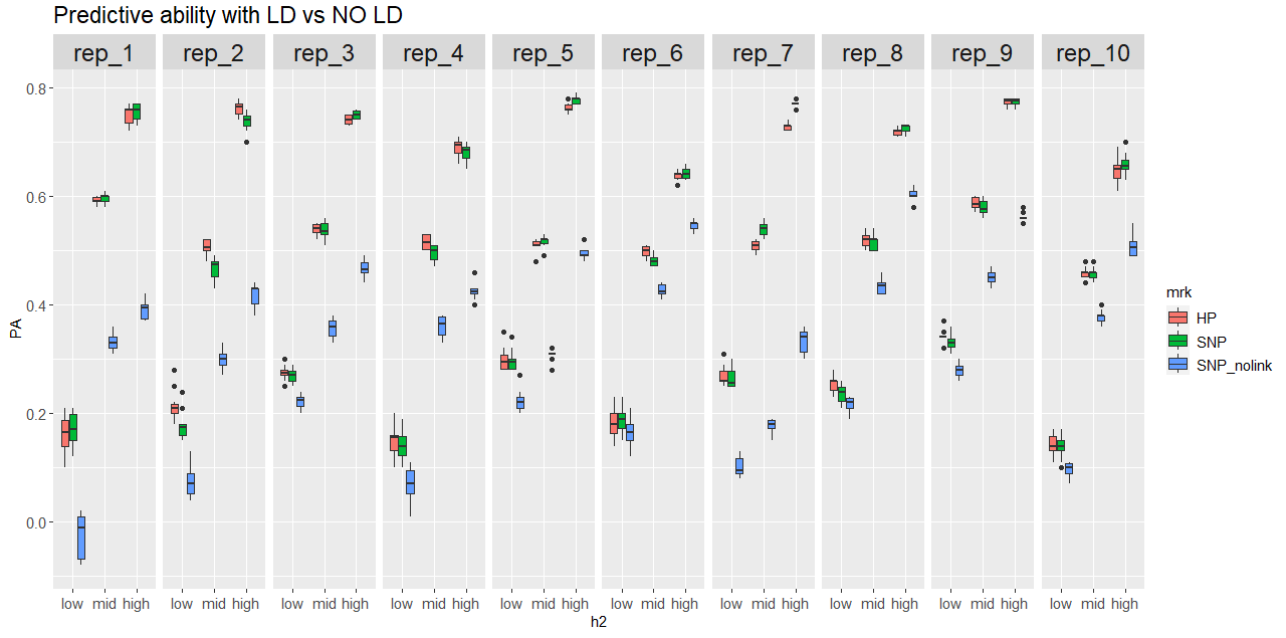

Figure 10 – Comparison of predictive ability of rrBLUP run with haplotype, SNP and markers with SNP in LD with the QTL removed. The x-axis are reported the level of heritability and low corresponds to heritability around 0.2, mid for 0.5 and high for 0.8. in the y-axis there is the predictive ability.

Table 3 – proportion of heritability across subpopulation over genetic heritability per replication for the monogenic trait

rep_1	rep_2	rep_3	rep_4	rep_5	rep_6	rep_7	rep_8	rep_9	rep_10
0.041	0.066	0.104	0.054	0.108	0.172	0.003	0.301	0.226	0.167

Investigation at locus level

To dig into the reason why multiallelic haplotype did not bring improvements to genomic predictions, we wanted to analyze into more detail each locus, for both marker sets and see the contribution that each locus could have in predicting the phenotype. Since we were dealing with simulated data, we knew all about every locus of the genome including the causal ones. To see what happens at locus level, we opted to analyze in depth the simplest case like the "mono" trait whose phenotype was affected only by one multiallelic locus. Specifically, we considered the monogenic trait of replication one with heritability 0.775. The causal locus was at position D0700 and the assigned allele effects were 0.33, 0, 0.66, 1, 0.66, and 0 for the alleles HP_1, up to HP_6, respectively. The allele frequencies were 0.33, 0.29, 0.27, 0.015, 0.08 and 0.012 for HP_1 to HP_6, respectively.

As first analysis, we estimated the allele effects of the causal locus and see if the alleles corresponded to the values assigned. To achieve this, we computed a linear regression with phenotype as dependent variable and pseudomarkers of the causal locus as independent variables. The estimated values were 0.359, 0.013, 0.688, 0.958, 0.679, NA for HP_1 to HP_6 respectively and the R-squared of the regression was 0.799.

We next moved to investigating the contribution of the surrounding loci in explaining the phenotypic variation. To obtain this information, we regressed the phenotype against variables like haplotype-markers and SNP-markers. Haplotype-markers were regressed against the phenotype in two ways: first all pseudomarkers of a locus together and, secondly a single pseudomarker as single independent variable (fig. 11).

Regression for haplotypes with the six alleles belonging to the same locus, the R-squared resulted high for the two loci closest to the causal locus, which were D0695 and D0705 whose values were 0.667 and 0.736 respectively. The region where loci gave the R-squared above 0.4 ranged from D0675 to D0725. Out of these boundaries R-squared decreased quite rapidly toward zero. From

D0625 and D0815 the R-squared started being below 0.1. On the other hand, when single pseudomarkers were regressed against the phenotype the pattern of the R-squared was like the previous result, it was approximately a bell shape with the pick around the causal locus. The highest R-squared values were attributable to the two loci surrounding the causal locus and specifically for allele HP_2 and HP_3. Pseudomarkers D0695_HP_2 and D0695_HP_3 showed a R-squared equal to 0.451 and 0.342 respectively. The other important pseudomarkers were D0705_HP_2 and D0705_HP_3 with R-squared of 0.484 and 0.339. Conversely, the other pseudomarkers of these two loci like D0705_HP_1, D0705_HP_4, D0705_HP_5, D0705_HP_6, D0695_HP_1, D0695_HP_4, D0695_HP_5 and D0695_HP_6 had R-squared ranging from 0.008 to 0.062.

R-squared obtained by regressing the phenotype over each single biallelic SNP marker, had a bell distribution similar to the previous analysis made on haplotypes. The presence of loci was approximately regular, one every 0.1 cM, but the effects which they had on the phenotype were not always consistent with the distance from the casual locus. The highest R-squared values belonged to the loci closer to the casual locus such as D0704, D0699 and D0693 and their respective values were 0.483, 0.478 and 0.458. However, some loci close to the casual locus had low R-squared values as for example D0701, D0702, D0697 with values equal to 0.001, 0.067 and 0.0002.

After that, we wanted to see which markers were selected by a model. This allowed us to explored if there was a correspondence between selected markers of the two marker sets and how haplotype markers were used by the model. To this end, we looked at the elastic net outputs generated by glmnet R package because it delivered higher predictive abilities together with Bayes B model, and delivered a selection of markers with their respective estimated values (appendix 3 and 4). For the bi-allelic SNP markers, the model selected 20 SNPs on the chromosome D, which is the chromosome where the casual locus sat. Seven markers had high R-squared ranging from 0.232 to 0.483 and their estimated effects ranged from -0.249 to 0.239. Interestingly, among the loci which had the highest R-squared, there were two haplotype-tagging SNPs (D0703 and D0704) which tagged the founder haplotype three and haplotype two. These two markers had R-squared values of 0.336 and 0.483. Focusing on marker D0703 for which allele “1” was the tagging allele for haplotype two, and haplotype two at causal locus corresponded to HP_2 allele whose effect was zero. The “0” allele instead represented other haplotypes whose alleles at casual locus D0700 were HP_1, HP_3, HP_4, HP_5 and HP_6 and their respective effects were 0.33, 0.66, 1, 0.66 and 0. Therefore, the allele “1” seemed to link to the null effect, while allele “0” linked mostly to higher effects.

Another relevant marker was D0699 whose R-squared was 0.478. The allele “1” belonged to haplotype one and two and the “0” for haplotype three, four, five and six. Therefore, allele “1” was linked to both HP_1 and HP_2, whereas the “0” allele was linked to HP_3, HP_4, HP_5 and HP_6 of the causal locus. Also in this case seemed to be present a connection between marker alleles and casual locus alleles: “1” for lower effects and “0” for the higher effects.

By contrast, looking at a marker with a poor fit in linear regression, despite its closeness to the casual locus, it does not provide any useful information. For instance, marker D0708 had the R-squared equal to 0.00006 and the allele “0” is connected to HP_1, HP_5 and HP_6, whereas allele “1” is linked to HP_2, HP_3 and HP_4. That means that allele “0” is linked to value 0.33, 0.66, 0 and allele “1” to value 0, 0.66 and 1.

Regarding the haplotype data set, the Elastic Net model selected and estimated the effects of twenty-five pseudomarkers for chromosome D. The heaviest effects were attributed to HP_2 and HP_3 for the loci D0695 and D0705. The values assigned to D0695_HP_2 and D0695_HP_3 were -0.12 and 0.11, instead for the pseudomarkers D0705_HP_2 and D0705_HP_3 were -0.29 and 0.30. For locus D0705 there was also a value for the haplotype HP_5 corresponding to 0.144. The remaining estimated effects were around zero and their R-squared computed previously was approximately zero.

With this analysis we could see how SNP markers and haplotype markers were used by the model. Alleles of SNP markers selected by the model, for example, were generally associated with a high or low effect to the multiallelic causal locus. On the other hand, we expected that haplotype markers could be associated with the casual locus alleles more precisely delivering a more accurate estimation of the causal locus allele effects and, consequently, delivering a higher predictive ability. However, the allele effects captured by the model were not complete, because only some alleles of the most significative loci were estimated. In fact, broadly speaking, the predictive abilities were on par or slightly lower for the haplotype marker set in comparison to the biallelic SNP markers.

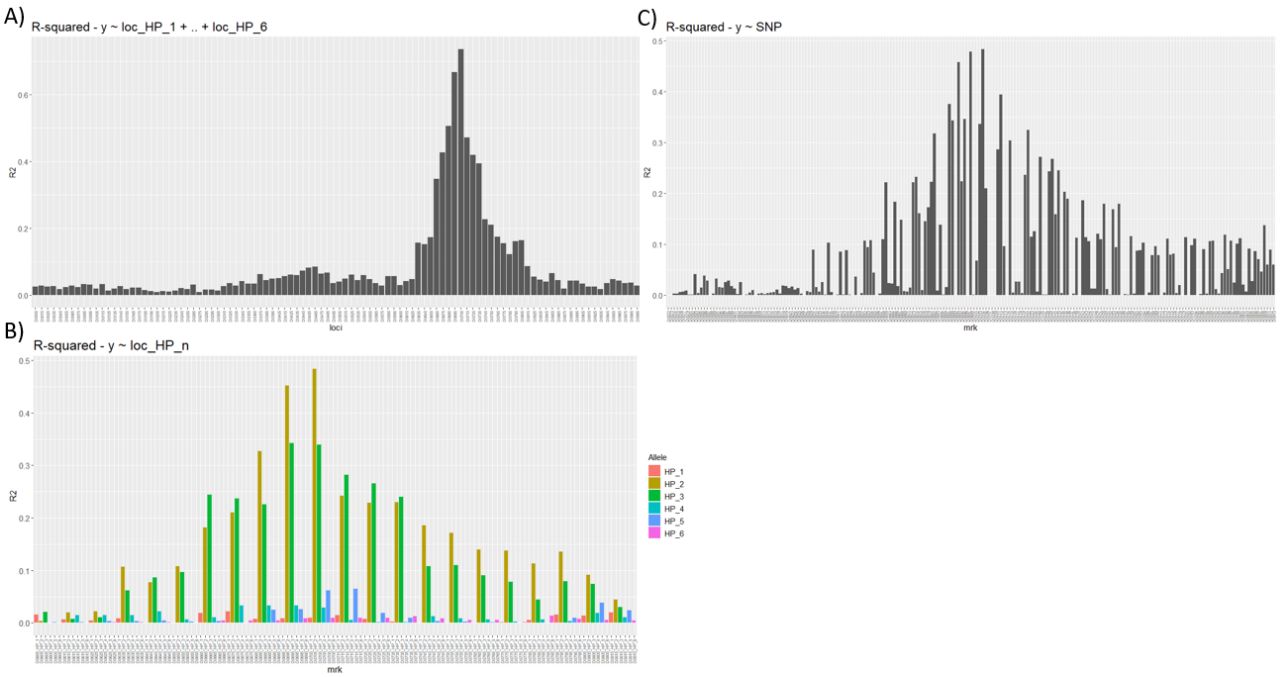

Figure 11 – A) Bar chart for the R-squared of multiallelic loci regressed against the phenotype. Phenotype was regressed against the six haplotypes of a locus according to the formula y ~ loc_HP_1 + loc_HP_2 + loc_HP_3 + loc_HP_4 + loc_HP_5 + loc_HP_6. On the x-axis there is the loci names and on the y-axis the R-squared. B) Bar chart of R-squared obtained by regressing each single pseudomarker of the haplotype marker set according to the formula y ~ loc_HP_n (for example y ~ D0769_HP_4). On the x-axis the pseudomarker names from D0605_HP_1 to D0815_HP_6, on the y-axis the R-squared. Each color represents an alleles (HP_1 to HP_6) of the loci. C) Bar chart of R-squared obtained by regressing each single SNP marker of the biallelic marker set according to the formula y ~ SNP. On the x-axis the SNP marker names from D0572 to D0813, on the y-axis the R-squared.

DISCUSSION

Comparison between bi-allelic SNP marker and haplotype markers

Haplotype may have a positive contribution to genomic prediction compared to single SNP due to its multiallelic nature and may contribute to better capture the LD with multi-allelic casual loci (Cuybano et al., 2014; Meuwissen et al., 2014). In addition, haplotypes are able to track ancestral information of genotypes (Bath at al. 2021). With this study, we wanted to acquire an insight into the application of haplotypes in genomic prediction for polyploids individuals and compare the results with SNP-based genomic prediction. Haplotype-based GP gave very similar results to SNP-based GP in terms of response to trait architecture, heritability and statistical model. The performance of haplotype-based GP may depend on several factors as for example traits, genetic structure of the population and the method used for haplotype construction (Lin et al., 2024).

In our simulation, haplotypes were made of one single locus for which there were six variants corresponding to the ancestral haplotypes used as base of our population study. While the traits were replicated ten times with different random position of the causal loci. The haplotypes we employed were supposed to be highly informative because they brought accurate information of both the genome structure and LD through the genome.

In contrast to the potential improvement that haplotypes may provide (Cuybano et al., 2014; Lin et al., 2024), our results showed no improvement in predictive ability with haplotype markers in comparison to a complete set of bi-allelic SNP markers, despite the model used and heritability level. Rather, shrinkage models run with SNP markers, Bayes B and Elastic Net, conferred slightly higher predictive abilities than haplotype-based genomic prediction for the monogenic and oligogenic traits. Elastic net outputs were still a bit higher using SNP markers for the polygenic traits, whereas Bayes B gave similar results for both marker sets. Shrinkage models were not affected by trait architecture and predictive abilities were steady for both marker sets. The trend of the predictive ability in relation to models, traits and heritability levels were the same for the two marker sets and these results were in accordance with what was observe by de Valk (2023) in GP computed with bi-allelic SNP markers on traits with bi-allelic QTLs. In fact, the effects of traits architecture, model and heritability on predictive ability are not touched by the type of marker used.

Therefore, we can confirm what was observed by de Valk (2023) that variable selection models are more appropriate for traits with few QTLs, whereas they lose predictive power as the number of QTL increases since the model assigns an error margin to the selected markers (Daetwyler et al., 2010). In the polygenic traits the differences between models were minimal. Variable selection models lose predictive power due to the high number of QTLs, whereas shrinkage models had a steady performance among traits. Shrinkage models showed this consistency because they include all markers in the model and no error is associated to the identification of markers. Noise stems from heritability of the trait and the extent of LD between markers and QTL (Daetwyler et al., 2010; de Valk, 2023).

Since the SNP markers included haplotype-tagging markers, we speculated that this could have conferred a stronger power to detect the effects of multiallelic QTLs. Therefore, we ran the experiment with the monogenic trait and made a new set of SNP markers without the haplotype-tagging SNPs. In the overall results this did not affect the predictive ability which remained equivalent to the complete SNP marker set. The complete SNP marker set may have been

sufficiently dense providing the sufficient information that the model needed to predict the phenotype. That means that there were highly informative non-haplotype-tagging SNPs in LD with the QTL. However, zooming into the replicates, in a few cases the haplotype-tagging SNPs had a relevant impact in estimating the QTL effects and removing them from the marker set the predictive ability dropped. This effect happened because for the trait of that replicate the haplotype-tagging SNPs were also in strong LD with the QTL and therefore they were highly informative. In fact, Alemu et al. (2023) applied a preselection of haplotype-tagging SNPs to include in the prediction model obtaining a higher predictive ability for FHB (fusarium head blight) resistance and yield-related traits. They argued that the advantage of preselection of haplotype-tagging markers can reduce the dimensionality of the models and potentially reduce the cost of genotyping, targeting the ones which tag important haplotypes. However, we did not investigate further this topic in the current work, because we were interested in understanding the reason why the haplotypes did not outperform the bi-allelic SNP markers. For the current study the haplotype-tagging SNP did not level up the prediction of the SNP marker set, but surly there are potentialities to exploit and that can be topic for future investigations.

Effect of population structure

Since the loci used to build the haplotypes were about six times smaller than the number of bi-allelic SNPs, we speculate that the variation of population structure could have been captured better with the dense SNP marker set and inflate the predictive ability. Population structure was assessed through the principal components analysis using the SNP-markers. The subgroups (X, Y and Z) were clustered in two main groups only in relation to the first PC. The rest of PCs did not further cluster the population. Then, we wanted to correct the genomic prediction for population structure by including in the model an increasing number of PCs as fixed effects. We focused on monogenic trait and the genomic prediction were computed with rrBLUP model. The method of correcting the prediction, including in the model PCs as fixed effects, was proposed by Deatwyler et al. (2012). They obtained a plateau that would represent the prediction accuracy due to LD of markers and QTL as the main population structure effect was accounted for. In contrast, we obtained a continuous decrease in predictive ability with steeper decline as the number of PC became greater than 100. In addition, no plateau was reached. The striking point was that the predictive ability obtained with the correction of the first PCs remained almost unchanged. Whereas it would be expected that

already at the first PCs the predictive ability would have dropped because PC1 explained the variation of the population structure. To further explore the effect of population structure, a simulation was carried out for the monogenic trait removing the markers in LD with the QTL, according to the LD values assessed by de Valk (2013). The overall means showed that the predictive ability significantly dropped for the markers with no LD as there was not a strong population structure. De Valk (2013) compared the effect of marker set with and without LD with QTLs for both a trait strongly correlated to the population structure and one not correlated to population structure. The predictive ability obtained with markers without LD to QTLs had a significant drop for the trait not correlated with population structure, whereas the trait correlated with the population structure had a light drop. That significant drop in predictive ability was observed on this work and therefore it seemed that, in the overall mean, the population structure was not significantly inflating the predictions. Zooming into the replicates, it is possible to see different cases in which traits had different correlation to the population structure and population structure variation was estimated as suggested by Guo et al. (2014). A few replicates had high population structure effect and consequently the drop of predictive ability using SNP without LD with QTL was less than the others which had very low population structure effect. However, for the trait with no population structure and predicted with markers not in LD with the QTL, there was still a residual variation that must be explained.

The remaining variation of the population used in the present work could come from a strong relation among individuals. That could be speculated from the fact that only one PC could cluster 2 main groups and the group composed of subpopulation X and Z was the most numerous. Then, when correcting the genomic prediction for PCs, there was a continuous decrease in predictive ability meaning that PCs were accumulatively accounting for genetic variation. Based on these results, we suggest to implement the population study in order to augment the genetic distance among subgroups or individuals to investigate if that remaining variation could partially derive from the relation among individuals.

Effects of single loci

Haplotype markers did not improve genomic prediction and SNP markers seemed to be in some cases more informative. We knew the true founder haplotypes of each locus and therefore we constructed the haplotypes in a very theoretical way: a haplotype block was formed of one locus

and each haplotype had six variants corresponding to the founder haplotypes, which were the origin of the population studied. Then, to assign a phenotype to each individual of the population, four effects of the causal loci were assigned to their respective haplotypes. Therefore, haplotype markers used as variants in genomic prediction were supposed to be highly informative and, the ones in LD with the QTL, strictly connected to the allele effects of the causal locus. Therefore, we investigated the contribution of each locus in explaining the genetic variation for the monogenic trait.

By regressing each single variable singularly against the phenotype, we could observe which variables were informative for genomic predictions. Starting with the bi-allelic SNP markers, the R-squared was higher as its position approached the causal locus and, vice versa, R-squared decreased as the distance increased. However, not all the closest SNP markers were significant in estimating the locus effects. That most probability was due to the lack of association of “0” and “1” allele of the bi-allelic SNP marker to a high or low value of the multiallelic QTL effects. Whereas the markers with high R-squared could capture either high or low QTL allele values with the 1 or 0 allele of the marker. Therefore, a biallelic marker allele could be associated with more than one QTL allele effect. In this way the marker allele 0, for example, could be associated with three QTL haplotypes with a relative high value, vice versa could be for the allele 1 that could be associated with the other three QTL effects with low values.

The most significant SNP markers, with a sufficiently high R-squared were located within approximatly 4 cM from each side of the casual locus. In fact, 4 cM corresponds to the short-range LD estimated by de Valk (2023) for this population. Beyond 4 cM distance from the causal locus, the R-squared values were closer to zero.

The elastic net model, in fact, selected and assigned heavier effects to markers with high R-squared values and they sat around the causal locus region, which is 8 cM long. Whereas, the other markers selected by the model falling beyond the region of the causal locus had low R-sqaured and the values assigned by the Elastic Net model were approximately zero.

When regressing the haplotype variables singularly, the trend was similar to that of the SNP markers. The more the distance from the causal locus the smaller R-squared. Zooming into a specific case as example, see replication one, the variants HP_2 and HP_3 of some loci in the QTL region explained a good extent of genetic variation. Whereas, The other loci variants near to the casual locus (HP_1, HP_4, HP_5, HP_6) were not able to explain genetic variation. The tendency that some loci variants explained much more genetic variation than others was quite common through the

replications. Then, the variants whose R-squared was high, corresponded to either a high effect QTL allele or low effect.

As expected, Elastic Net model assigned the heaviest values for those variants which had the highest R-squared. However, also less significant loci and variants were selected by the model despite their low R-squared values, but their estimated effects were approximately zero.

Therefore, the phenotype was estimated based only on a few variants per selected locus. For example, the model estimated effects for the allele HP_2 and HP_3 of the locus D0695, whereas the other alleles of the same locus were not considered.

Haplotype variables included in the genomic prediction models seemed to behave similarly to the bi-allelic SNP markers, since a single significant haplotype variable captured more than one casual QTL effect. In this way the haplotype marker set loses that advantage given by the multiallelism. Haplotypes in LD with the QTL would have provided specific alleles for each QTL alleles providing a more accurate estimation of the phenotype. Multiallelic haplotypes have the advantage of having more variants that could be associated with the multiallelic variants of a QTL, but the rearrangement of haplotypes alleles as single variable in the design matrix may have increased the multicollinearity causing the loss this advantage (Aschard et al., 2015; Matias et al., 2017).

The potentiality that haplotypes can furnish to the genomic prediction was confirmed by the results of regressing the phenotype against a single locus with all its variants. Similarly to single SNP markers and single haplotype variants, the R-squared was higher as the loci approaches the causal locus and diminished as the distance increased. The only difference was that the values were higher than the single haplotype variants and single SNPs. The most significant loci had R-squared close to the heritability of the trait. That proofs what we hypothesized that the haplotype markers are highly informative. Probably, the method used to fit the haplotype markers in the predictive models was not suitable for exploiting all the advantages that haplotypes can bring. That is ground for future research to test other models or implement models and design matrix to better exploit the advantages brought by haplotypes.

In this work we have to consider that the comparison between the two marker sets was not fair. We reduced the number of loci to convert in haplotype markers in order to reduce the dimension of the data and consequently to reduce the time needed for computing the analysis, especially for the variable selection models. In fact, the total number of bi-allelic SNP markers were 2556 while

the loci used for haplotypes were 395. That could be an aspect to consider for future works since in this way loci strongly linked to the QTL would be present also for the haplotypes. However, equating the number of loci for both marker sets would not change the fact that haplotype alleles of a locus would be exploited properly by the models and rather the collinearity problem would increase as the number of variables would be grater. But this must be tested.

CONCLUSION

In this work haplotypes did not bring improvements in genomic predictions for all traits architectures, heritability levels and GP models. These results are not in line with some works in literature which confirmed GP accuracy being improved. On the other hand, some other studies did not show improvements similarly to what showed in the present work. However, the haplotypes used for genomic prediction in other studies were built from a highly dense SNP marker set, in contrast to our haplotypes which included only a marker from a very limited number of loci. In addition, the comparison between the marker set in this work was unfair because the number of loci implied were different between the sets being greater for bi-allelic SNP markers. Despite the different number of loci implied, the haplotypes conceived in this work were highly informative, but they were not well used by models since they could not precisely capture the alleles effect of the causal QTL. But, closest haplotypes to the QTL when regressed against the phenotype as whole locus, including all its variants, provided an R-squared close to the genetic heritability of the trait. This indicates that haplotypes are very informative and could explain more genetic variance than the single SNP or single haplotype variable (as pseudomarker). Future studies should aim to explore models which can better use haplotypes or explore methods for including haplotypes in the models as locus, with its respective alleles, and not as single variable that would behave similarly to bi-allelic SNP markers losing the advantage of being multiallelic.

Population structure was explored and quantified for the monogenic trait to see if could affect the predictive ability. In the population object of this study there was a genetic structure that was assessed with PCA analysis, but the monogenic trait did not show to strongly relate to the population structure. Therefore, the predictive ability was not inflated by population structure effect. However, when assessing the correlation of the trait to the population structure removing markers in LD, still there was a part of genetic variation explained by unknown factors. From the results of this study, we could speculate that part of that remaining genetic variation, not connected to the LD between markers and QTL, could be due to a strong relation between individuals. We recommend for future investigations, to implement the population with more distantly related individuals in order to explore that remaining genetic variation not explained by the LD with the QTL.

LITERATURE

Alemu, A., Batista, L., Singh, P. K., Ceplitis, A., & Chawade, A. (2023). Haplotype-tagged SNPs improve genomic prediction accuracy for Fusarium head blight resistance and yield-related traits in wheat. *Theoretical and Applied Genetics*, *136*(4), 92.

Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L., & Kraft, P. (2015). Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *The American Journal of Human Genetics*, *96*(2), 329-339.

Guo, Z., Tucker, D. M., Basten, C. J., Gandhi, H., Ersoz, E., Guo, B., ... & Gay, G. (2014). The impact of population structure on genomic prediction in stratified populations. *Theoretical and applied genetics*, *127*, 749-762.

He, D., Saha, S., Finkers, R., & Parida, L. (2018). Efficient algorithms for polyploid haplotype phasing. *BMC genomics*, *19*, 171-180.

Shen, J., Li, Z., Chen, J., Song, Z., Zhou, Z., & Shi, Y. (2016). SHEsisPlus, a toolset for genetic studies on polyploid species. *Scientific reports*, *6*(1), 24095.

Calus, M. P., Meuwissen, T. H., Windig, J. J., Knol, E. F., Schrooten, C., Vereijken, A. L., & Veerkamp, R. F. (2009). Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. *Genetics Selection Evolution*, *41*, 1-10.

Calus, M. P. L., Meuwissen, T. H. E., De Roos, A. P. W., & Veerkamp, R. F. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics*, *178*(1), 553-561.

Da, Y. (2015). Multi-allelic haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers. *BMC genetics*, *16*, 1-12.

Hess, M., Druet, T., Hess, A., & Garrick, D. (2017). Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. *Genetics Selection Evolution*, *49*, 1-14.

Matias, F. I., Galli, G., Correia Granato, I. S., & Fritsche-Neto, R. (2017). Genomic prediction of autogamous and allogamous plants by SNPs and haplotypes. *Crop Science*, *57*(6), 2951-2958.

Ballesta, P., Maldonado, C., Pérez-Rodríguez, P., & Mora, F. (2019). SNP and haplotype-based genomic selection of quantitative traits in Eucalyptus globulus. *Plants*, *8*(9), 331.

Won, S., Park, J. E., Son, J. H., Lee, S. H., Park, B. H., Park, M., ... & Lim, D. (2020). Genomic prediction accuracy using haplotypes defined by size and hierarchical clustering based on linkage disequilibrium. *Frontiers in genetics*, *11*, 134.

Lin, Y. C., Mayer, M., Valle Torres, D., Pook, T., Hölker, A. C., Presterl, T., ... & Schön, C. C. (2024). Genomic prediction within and across maize landrace derived populations using haplotypes. *Frontiers in Plant Science*, *15*, 1351466.

Bhat, J. A., Yu, D., Bohra, A., Ganie, S. A., & Varshney, R. K. (2021). Features and applications of haplotypes in crop breeding. *Communications biology*, *4*(1), 1266.

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, *185*(3), 1021-1031.

De Valk, J. (2023). Genomic prediction in polyploids crops. MSc thesis, Wageningen University.

Los Campos, G., ... & Varshney, R. K. (2017). Genomic selection in plant breeding: methods, models, and perspectives. Trends in plant science, 22(11), 961-975.

Cuyabano, B. C. D., Su, G., & Lund, M. S. (2014). Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics, 15(1), 1–11

Cuyabano, B. C., Su, G., & Lund, M. S. (2015). Selection of haplotype variables from a high-density marker map for genomic prediction. Genetics Selection Evolution, 47(1), 1-11.

Daetwyler, H. D., Kemper, K. E., van der Werf, J. H. J., & Hayes, B. J. (2012). Components of the accuracy of genomic prediction in a multi-breed sheep population. Journal of Animal Science, 90(10), 3375–3384.

Daetwyler, H. D., Villanueva, B., & Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. PloS one, 3(10), e3395.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 2011, 12(7), 499–510.

De los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. L. (2013). Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics, 193(2), 327–345.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLOS ONE, 6(5), e19379.

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome, 4(3), 250–255.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. JSS Journal of Statistical Software, 33(1).

Gianola, D., & Van Kaam, J. B. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics, 178(4), 2289-2303.

Habier, D., Fernando, R. L., & Dekkers, J. C. M. (2007). The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics, 177(4), 2389–2397.

Hess, M., Druet, T., Hess, A. & Garrick, D. 2017. Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. Genetics Selection Evolution, 49**,** 54.

Loos, R. J. F. (2020). 15 years of genome-wide association studies and no signs of slowing down. Nature Communications 2020 11:1, 11(1), 1–3.

Matias, F.I., Galli, G., Correia G., I. S. & Fritsche-Neto, R. 2017. Genomic prediction of autogamous and allogamous plants by SNPs and haplotypes. Crop Science, 57**,** 2951-2958.

Meher, P. K., Rustgi, S., & Kumar, A. (2022). Performance of Bayesian and BLUP alphabets for genomic prediction: analysis, comparison and results. Heredity, 128(6), 519–530.

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4), 1819–1829.

Motazedi, E., de Ridder, D., Finkers, R., Baldwin, S., Thomson, S., Monaghan, K., & Maliepaard, C. (2018). TriPoly: haplotype estimation for polyploids using sequencing data of related individuals. Bioinformatics, 34(22), 3864-3872

Motazedi, E., Maliepaard, C., Finkers, R., Visser, R., & De Ridder, D. (2019). Family-based haplotype estimation and allele dosage correction for polyploids using short sequence reads. Frontiers in Genetics, 10, 335.

Neigenfind, J., Gyetvai, G., Basekow, R., Diehl, S., Achenbach, U., Gebhardt, C., ... & Kersten, B. (2008). Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. BMC genomics, 9, 1-26.

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., & Tanksley, S. D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature, 335(6192), 721–726.

Piepho, H. P. (2009). Ridge regression and extensions for genomewide selection in maize. Crop Science, 49(4), 1165-1176.

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nature reviews genetics, 11(7), 459-463.

Sarinelli, J. M., Murphy, J. P., Tyagi, P., Holland, J. B., Johnson, J. W., Mergoum, M., ... & Brown-Guedira, G. (2019). Training population selection and use of fixed effects to optimize genomic predictions in a historical USA winter wheat panel. Theoretical and Applied Genetics, 132(4), 1247-1261.

Soller, M. (1978). The use of loci associated with quantitative effects in dairy cattle improvement. Animal Science, 27(2), 133–139.

Soller, M., & Plotkin-Hazan, J. (1977). The use marker alleles for the introgression of linked quantitative alleles. TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik, 51(3), 133–137.

Syvänen, A. C. (2001). Accessing genetic variation: genotyping single nucleotide polymorphisms. Nature Reviews Genetics, 2(12), 930-942.

Thérèse Navarro, A., Tumino, G., Voorrips, R. E., Arens, P., Smulders, M. J., van de Weg, E., & Maliepaard, C. (2022). Multiallelic models for QTL mapping in diverse polyploid populations. BMC bioinformatics, 23(1), 1-16.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Source: Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288.

Villumsen, T. M. & Janss, L. Bayesian genomic selection: the effect of haplotype length and priors. BMC proceedings, 2009. Springer, S11.

Voorrips, R. E., & Maliepaard, C. A. (2012). The simulation of meiosis in diploid and tetraploid organisms using various genetic models. BMC Bioinformatics, 13(1), 1–12

Voorrips, R. E., & Tumino, G. (2022). PolyHaplotyper: haplotyping in polyploids based on bi-allelic marker dosage data. BMC bioinformatics, 23(1), 442.

Werner, C. R., Gaynor, R. C., Gorjanc, G., Hickey, J. M., Kox, T., Abbadi, A., ... & Stahl, A. (2020). How population structure impacts genomic selection accuracy in cross-validation: implications for practical breeding. Frontiers in plant science, 11, 592977.

Willemsen JH. Happy-haplotype-inference V1 download link. https://git.wageningenur.nl/wille094/Happy-haplotype-inference/-/tree/master/V1. Accessed 14 September 2020.

Wilson, S., Zheng, C., Maliepaard, C., Mulder, H. A., Visser, R. G., van der Burgt, A., & van Eeuwijk, F. (2021). Understanding the effectiveness of genomic prediction in tetraploid potato. Frontiers in plant science, 12, 672417.

Yu, J., Pressoir, G., Briggs, W. H., BI, I. V., Yamasaki, M., Doebley, J. F., Mcmullen, M. D., Gaut, B. S., Nielsen, D. M. & Holland, J. B. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics, 38, 203-208.

Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., ... & Simianer, H. (2014). Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PloS one, 9(3), e93017.

Zou, H. & Hastie, T. 2005. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67**,** 301-320.

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. *Genetics*, *185*, 1021–1031. https://doi.org/10.1534/genetics.110.116855

Daetwyler, H. D., Villanueva, B., & Woolliams, J. A. (2008). Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach.

Pérez, P., & de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, *198*(2), 483–495. https://doi.org/10.1534/GENETICS.114.164442

Appendix 1. Average predictive ability.

The averages are given by grouping marker type, trait architecture, model and heritability level. In addition, these averages are obtained from the prediction of ten replication per trait. Column mrk indicate the typology of marker used: HP stands for haplotype marker set and SNP for biallelic marker set. Tr_arc column indicate the trait architecture analyzed which are mono, oligo_1, oligo_4, poly, mix_1 and mix_4. Under the column model there are the abbreviations of the prediction model used for the analysis which are Bayes B (BB), elastic net (EN), reproducing kernel Hilbert space (RKHS) and Ridge regression BLUP (rrBLUP). In h2-mean column there is the average heritability for each trait; PA-mean is the prediction ability means and PA-sd is the standard deviation of the PA means.

mrk	tr_arc	heritability	model	h2 - mean	PA - mean	PA - sd
HP	mono	low	BB	0.197	0.2933	0.0832
SNP	mono	low	BB	0.197	0.2979	0.0795
HP	mono	low	EN	0.197	0.3176	0.0829
SNP	mono	low	EN	0.197	0.3341	0.073
HP	mono	low	RKHS	0.197	0.2194	0.0755
SNP	mono	low	RKHS	0.197	0.2129	0.0709
HP	mono	low	rrBLUP	0.197	0.2289	0.0695
SNP	mono	low	rrBLUP	0.197	0.2213	0.0678
HP	mono	mid	BB	0.509	0.6423	0.0294
SNP	mono	mid	BB	0.509	0.6537	0.0265
HP	mono	mid	EN	0.509	0.6396	0.0293
SNP	mono	mid	EN	0.509	0.6601	0.0263
HP	mono	mid	RKHS	0.509	0.5147	0.0399
SNP	mono	mid	RKHS	0.509	0.511	0.0452
HP	mono	mid	rrBLUP	0.509	0.5231	0.0398
SNP	mono	mid	rrBLUP	0.509	0.5182	0.0452
HP	mono	high	BB	0.807	0.8355	0.0331
SNP	mono	high	BB	0.807	0.8578	0.0189
HP	mono	high	EN	0.807	0.8248	0.0379
SNP	mono	high	EN	0.807	0.8504	0.0235
HP	mono	high	RKHS	0.807	0.7205	0.0457
SNP	mono	high	RKHS	0.807	0.7224	0.0479
HP	mono	high	rrBLUP	0.807	0.7212	0.0473
SNP	mono	high	rrBLUP	0.807	0.7267	0.0487
HP	oligo_1	low	BB	0.217	0.2265	0.0647
SNP	oligo_1	low	BB	0.217	0.2315	0.0485
HP	oligo_1	low	EN	0.217	0.2817	0.0521
SNP	oligo_1	low	EN	0.217	0.3024	0.0429
HP	oligo_1	low	RKHS	0.217	0.1959	0.075
SNP	oligo_1	low	RKHS	0.217	0.2091	0.0682
HP	oligo_1	low	rrBLUP	0.217	0.2075	0.0649
SNP	oligo_1	low	rrBLUP	0.217	0.22	0.0595
HP	oligo_1	mid	BB	0.513	0.6219	0.0391
SNP	oligo_1	mid	BB	0.513	0.6413	0.0363
HP	oligo_1	mid	EN	0.513	0.6209	0.0333
SNP	oligo_1	mid	EN	0.513	0.6427	0.0273
HP	oligo_1	mid	RKHS	0.513	0.5361	0.0401
SNP	oligo_1	mid	RKHS	0.513	0.5444	0.0349
HP	oligo_1	mid	rrBLUP	0.513	0.5384	0.039
SNP	oligo_1	mid	rrBLUP	0.513	0.5444	0.0336
HP	oligo_1	high	BB	0.767	0.8478	0.0206
SNP	oligo_1	high	BB	0.767	0.8621	0.013
HP	oligo_1	high	EN	0.767	0.8362	0.0221
SNP	oligo_1	high	EN	0.767	0.8529	0.0139
HP	oligo_1	high	RKHS	0.767	0.7639	0.0285
SNP	oligo_1	high	RKHS	0.767	0.7654	0.0257
HP	oligo_1	high	rrBLUP	0.767	0.7687	0.0276
SNP	oligo_1	high	rrBLUP	0.767	0.7721	0.0248
HP	oligo_4	low	BB	0.201	0.3205	0.0484
SNP	oligo_4	low	BB	0.201	0.314	0.0514
HP	oligo_4	low	EN	0.201	0.305	0.0474
SNP	oligo_4	low	EN	0.201	0.3248	0.0528
HP	oligo_4	low	RKHS	0.201	0.307	0.052
SNP	oligo_4	low	RKHS	0.201	0.302	0.0544

mrk	tr_arc	heritability	model	h2 - mean	PA - mean	PA - sd
HP	oligo_4	low	rrBLUP	0.201	0.3118	0.0508
SNP	oligo_4	low	rrBLUP	0.201	0.3071	0.0521
HP	oligo_4	mid	BB	0.485	0.586	0.0232
SNP	oligo_4	mid	BB	0.485	0.5978	0.0246
HP	oligo_4	mid	EN	0.485	0.5886	0.0234
SNP	oligo_4	mid	EN	0.485	0.6157	0.0181
HP	oligo_4	mid	RKHS	0.485	0.5459	0.031
SNP	oligo_4	mid	RKHS	0.485	0.5458	0.0324
HP	oligo_4	mid	rrBLUP	0.485	0.5411	0.0296
SNP	oligo_4	mid	rrBLUP	0.485	0.5423	0.0308
HP	oligo_4	high	BB	0.797	0.8489	0.0111
SNP	oligo_4	high	BB	0.797	0.8557	0.012
HP	oligo_4	high	EN	0.797	0.8369	0.0132
SNP	oligo_4	high	EN	0.797	0.8469	0.0126
HP	oligo_4	high	RKHS	0.797	0.7621	0.0293
SNP	oligo_4	high	RKHS	0.797	0.7653	0.0291
HP	oligo_4	high	rrBLUP	0.797	0.7672	0.0256
SNP	oligo_4	high	rrBLUP	0.797	0.7703	0.0259
HP	poly	low	BB	0.2	0.2964	0.0433
SNP	poly	low	BB	0.2	0.2985	0.0461
HP	poly	low	EN	0.2	0.284545	0.05
SNP	poly	low	EN	0.2	0.2864	0.0561
HP	poly	low	RKHS	0.2	0.2785	0.0476
SNP	poly	low	RKHS	0.2	0.2828	0.0492
HP	poly	low	rrBLUP	0.2	0.279	0.0474
SNP	poly	low	rrBLUP	0.2	0.2851	0.0488
HP	poly	mid	BB	0.477	0.5491	0.0344
SNP	poly	mid	BB	0.477	0.5565	0.0339
HP	poly	mid	EN	0.477	0.5419	0.0378
SNP	poly	mid	EN	0.477	0.5444	0.0353
HP	poly	mid	RKHS	0.477	0.5414	0.0399
SNP	poly	mid	RKHS	0.477	0.541	0.0389
HP	poly	mid	rrBLUP	0.477	0.5506	0.0372
SNP	poly	mid	rrBLUP	0.477	0.5524	0.0368
HP	poly	high	BB	0.784	0.7678	0.0248
SNP	poly	high	BB	0.784	0.774	0.0239
HP	poly	high	EN	0.784	0.7531	0.029
SNP	poly	high	EN	0.784	0.7664	0.0242
HP	poly	high	RKHS	0.784	0.7578	0.0268
SNP	poly	high	RKHS	0.784	0.762	0.0256
HP	poly	high	rrBLUP	0.784	0.7661	0.0245
SNP	poly	high	rrBLUP	0.784	0.7737	0.0229
HP	mix_1	low	BB	0.182	0.319	0.0309
SNP	mix_1	low	BB	0.182	0.3325	0.0271
HP	mix_1	low	EN	0.182	0.2998	0.0355
SNP	mix_1	low	EN	0.182	0.309	0.0343
HP	mix_1	low	RKHS	0.182	0.3273	0.0278
SNP	mix_1	low	RKHS	0.182	0.3287	0.0262
HP	mix_1	low	rrBLUP	0.182	0.3192	0.0296
SNP	mix_1	low	rrBLUP	0.182	0.3306	0.0268
HP	mix_1	mid	BB	0.515	0.5172	0.0267
SNP	mix_1	mid	BB	0.515	0.5262	0.0286
HP	mix_1	mid	EN	0.515	0.4923	0.026
SNP	mix_1	mid	EN	0.515	0.5108	0.0284
HP	mix_1	mid	RKHS	0.515	0.5029	0.0231
SNP	mix_1	mid	RKHS	0.515	0.5078	0.0211
HP	mix_1	mid	rrBLUP	0.515	0.5053	0.0226
SNP	mix_1	mid	rrBLUP	0.515	0.5109	0.0196
HP	mix_1	high	BB	0.816	0.7815	0.0188
SNP	mix_1	high	BB	0.816	0.7946	0.0188
HP	mix_1	high	EN	0.816	0.7653	0.0228
SNP	mix_1	high	EN	0.816	0.7791	0.0218
HP	mix_1	high	RKHS	0.816	0.7578	0.02
SNP	mix_1	high	RKHS	0.816	0.7614	0.0201
HP	mix_1	high	rrBLUP	0.816	0.7648	0.0193
SNP	mix_1	high	rrBLUP	0.816	0.7703	0.0193
HP	mix_4	low	BB	0.205	0.3424	0.0418
SNP	mix_4	low	BB	0.205	0.3376	0.0419
HP	mix_4	low	EN	0.205	0.3258	0.0471

mrk	tr_arc	heritability	model	h2 - mean	PA - mean	PA - sd
SNP	mix_4	low	EN	0.205	0.3271	0.0438
HP	mix_4	low	RKHS	0.205	0.3333	0.0464
SNP	mix_4	low	RKHS	0.205	0.331	0.0458
HP	mix_4	low	rrBLUP	0.205	0.3365	0.0452
SNP	mix_4	low	rrBLUP	0.205	0.3345	0.0445
HP	mix_4	mid	BB	0.524	0.5704	0.0284
SNP	mix_4	mid	BB	0.524	0.5732	0.0314
HP	mix_4	mid	EN	0.524	0.5566	0.0364
SNP	mix_4	mid	EN	0.524	0.5588	0.0379
HP	mix_4	mid	RKHS	0.524	0.5612	0.0328
SNP	mix_4	mid	RKHS	0.524	0.5621	0.0327
HP	mix_4	mid	rrBLUP	0.524	0.5655	0.0318
SNP	mix_4	mid	rrBLUP	0.524	0.566	0.0321
HP	mix_4	high	BB	0.82	0.8023	0.0221
SNP	mix_4	high	BB	0.82	0.8105	0.0211
HP	mix_4	high	EN	0.82	0.7908	0.0273
SNP	mix_4	high	EN	0.82	0.8007	0.0235
HP	mix_4	high	RKHS	0.82	0.78	0.0224
SNP	mix_4	high	RKHS	0.82	0.7863	0.0222
HP	mix_4	high	rrBLUP	0.82	0.7871	0.0208
SNP	mix_4	high	rrBLUP	0.82	0.793	0.0206

Appendix 2.- Means of subpopulation genetic values

Table with means of genetic values for each subpopulation and for each trait architecture in each replication. Under trait architecture (Trait arch.) column there are the names of the trait analyzed such as Mono, Oligo_1, Oligo_4, Poly, Mix_1 and Mix_4. Rep column indicate the replication number. Column X, Y and Z are the subpopulations names.

Trait arch.	Rep	X	Y	Z
Mono	rep_1	1.34	1.56	1.39
	rep_2	1.58	1.85	1.58
	rep_3	1.94	1.45	1.97
	rep_4	2.44	2.82	2.60
	rep_5	1.58	1.13	1.56
	rep_6	2.09	2.72	2.00
	rep_7	2.29	2.18	2.31
	rep_8	1.55	2.46	1.63
	rep_9	2.66	1.61	2.65
	rep_10	3.38	3.77	3.39
Oligo_1	rep_1	6.86	4.12	6.72
	rep_2	10.41	7.86	10.64
	rep_3	7.67	7.95	7.85
	rep_4	6.41	8.11	6.47
	rep_5	8.75	7.25	8.76
	rep_6	5.35	5.57	5.08
	rep_7	7.15	6.83	7.19
	rep_8	11.02	11.22	11.23
	rep_9	10.64	9.99	10.70
	rep_10	11.97	11.64	11.80
Oligo_4	rep_1	6.31	9.08	6.46
	rep_2	8.25	7.49	8.35
	rep_3	7.66	8.56	7.44
	rep_4	7.29	5.83	7.35

Trait arch.	Rep	X	Y	Z
	rep_5	6.51	7.81	6.51
	rep_6	9.34	8.51	9.29
	rep_7	7.18	7.35	7.12
	rep_8	8.88	9.46	9.06
	rep_9	7.67	10.07	7.73
	rep_10	8.87	7.43	8.74
Poly	rep_1	202.02	202.63	202.67
	rep_2	202.51	201.18	203.50
	rep_3	192.09	196.07	193.25
	rep_4	198.78	179.86	198.76
	rep_5	202.89	215.17	202.83
	rep_6	217.00	215.72	218.68
	rep_7	227.36	223.81	228.32
	rep_8	204.76	207.85	204.57
	rep_9	205.25	209.89	205.59
	rep_10	219.00	211.59	218.87
Mix_1	rep_1	59.16	56.58	59.18
	rep_2	67.46	64.53	67.97
	rep_3	49.67	50.82	50.10
	rep_4	41.68	40.03	41.74
	rep_5	46.45	47.23	46.44
	rep_6	48.62	48.59	48.68
	rep_7	48.09	47.13	48.31
	rep_8	41.31	41.97	41.48
	rep_9	56.74	57.12	56.87
	rep_10	47.93	46.38	47.73
Mix_4	rep_1	54.94	57.85	55.24
	rep_2	47.86	46.84	48.16
	rep_3	41.82	43.42	41.80
	rep_4	32.09	28.27	32.14
	rep_5	39.31	42.59	39.30
	rep_6	46.68	45.62	46.91
	rep_7	41.81	41.44	41.90
	rep_8	44.93	46.06	45.08
	rep_9	49.99	53.34	50.11
	rep_10	49.90	47.07	49.75

Appendix 3.-Table with SNP information selected by Elastic net for monogenic trait and replication one

Table reporting the biallelic SNP markers selected by elastic net model and their respective information such as: frequence of allele “1” (freq_1), founder haplotype tagged (hap_tag), the tagging alleles (tagall), the founder haplotype frequences for the locus (HP_1_freq to HP_6_freq), marker estimated effects (EN_SNP_effect), R-squared (R2) and the founder haplotype alleles (H1 to H6). Causal locus is D0700 and under the columns of founder haplotypes is reported the respective allele name and its effect assigned.

marker	freq_1	hap_tag	tagall	HP_1_freq	HP_2_freq	HP_3_freq	HP_4_freq	HP_5_freq	HP_6_freq	EN_SNP_effect	R2	H1	H2	H3	H4	H5	H6
D0295	0.568	2	1	0.104	0.568	0.212	0.077	0.013	0.026	-0.010	1.364E-06	0	1	0	0	0	0
D0345	0.428			0.046	0.541	0.3	0.018	0.064	0.031	0.012	1.518E-02	1	0	1	1	1	0
D0579	0.605			0.29	0.285	0.298	0.022	0.105	0	0.001	9.188E-03	0	1	1	1	0	0
D0586	0.389			0.296	0.292	0.274	0.022	0.116	0	-0.012	3.803E-02	0	0	1	0	1	1
D0602	0.688			0.234	0.312	0.282	0	0.172	0	-0.006	2.048E-03	1	0	1	0	1	1
D0664	0.723			0.263	0.24	0.338	0.02	0.124	0.014	0.049	2.323E-02	0	1	1	1	1	0
D0674	0.626			0.284	0.232	0.348	0.012	0.11	0.015	-0.011	2.325E-01	1	1	0	0	1	0
D0682	0.716			0.298	0.268	0.302	0.015	0.102	0.015	0.005	3.171E-01	1	0	1	1	1	0
D0693	0.291			0.302	0.278	0.286	0.015	0.108	0.012	-0.249	4.579E-01	0	1	0	0	0	1
D0699	0.578			0.304	0.274	0.297	0.015	0.098	0.012	-0.107	4.779E-01	1	1	0	0	0	0
D0700	0.284			0.33	0.294	0.27	0.015	0.08	0.012			HP_1 = 0.33	HP_2 = 0	HP_3 = 0.66	HP_4 = 1	HP_5 = 0.66	HP_6 = 0
D0702	0.11			0.326	0.286	0.278	0.015	0.084	0.012	0.036	6.764E-02	0	0	0	1	1	1
D0703	0.723	3	0	0.332	0.281	0.278	0.015	0.084	0.012	-0.089	3.363E-01	1	1	0	1	1	1
D0704	0.282	2	1	0.332	0.281	0.277	0.015	0.083	0.012	-0.205	4.831E-01	0	1	0	0	0	0
D0708	0.561			0.326	0.248	0.282	0.029	0.102	0.012	0.003	5.634E-05	0	1	1	1	0	0
D0712	0.339			0.313	0.278	0.258	0.058	0.082	0.012	0.239	3.942E-01	0	0	1	0	1	0
D0751	0.18			0.264	0.264	0.283	0.043	0.138	0.009	0.044	1.170E-02	0	0	0	1	1	0
D0755	0.257			0.264	0.248	0.281	0.058	0.141	0.009	-0.002	1.791E-01	0	1	0	0	0	1
D0992	0.329			0.294	0.377	0.07	0.12	0.086	0.052	0.008	2.109E-02	0	0	1	1	1	1

Appendix 4. Table with haplotype markers information selected by Elastic net and for monogenic trait replication one

Table reporting the multiallelic haplotype markers selected by elastic net model and their respective information such as: locus names (locus), the founder haplotype frequences for the locus (HP_1_freq to HP_6_freq), the estimated effects of each allele (HP_1_EN_eff to HP_6_EN_eff) and the R-squared of each allele per locus (HP_1_R2 to HP_6_R2). In correspondence the causal locus (D0700), it is reported the respective allele name and its effect assigned under the columns HP_1_EN to HP_6_EN.

locus	HP_1_freq	HP_2_freq	HP_3_freq	HP_4_freq	HP_5_freq	HP_6_freq	HP_1_EN_eff	HP_2_EN_eff	HP_3_EN_eff	HP_4_EN_eff	HP_5_EN_eff	HP_6_EN_eff	HP_1_R2	HP_2_R2	HP_3_R2	HP_4_R2	HP_5_R2	HP_6_R2
D0065	0.112	0.49	0.259	0.066	0.034	0.038				0.0025			0.00258	0.01237	0.00993	0.01280	0.00038	0.00043
D0205	0.101	0.479	0.248	0.106	0.052	0.015	-0.0037						0.00210	0.00141	0.00282	0.00280	0.00194	0.00046
D0265	0.068	0.494	0.229	0.164	0.025	0.018		-0.0294					0.00862	0.02106	0.00357	0.01086	0.00426	0.00020
D0315	0.051	0.576	0.288	0.013	0.052	0.021		-0.0162					0.01269	0.00597	0.00295	0.00274	0.00768	0.00002
D0585	0.298	0.289	0.274	0.022	0.116	0		0.0026					0.00574	0.00735	0.02962	0.02221	0.00396	0.00000
D0605	0.232	0.315	0.293	0	0.16	0			-0.0097				0.01590	0.00300	0.02038	0.00000	0.00134	0.00000
D0635	0.231	0.296	0.296	0.005	0.16	0.011			0.0205				0.00802	0.10692	0.06177	0.01505	0.00329	0.00132
D0655	0.321	0.232	0.266	0.018	0.147	0.016	-0.0020						0.00025	0.10770	0.09690	0.00664	0.00267	0.00010
D0665	0.254	0.24	0.339	0.02	0.123	0.023			0.0194				0.01875	0.18202	0.24369	0.01015	0.00297	0.00466
D0675	0.281	0.23	0.35	0.014	0.11	0.015	-0.0040						0.02169	0.21063	0.23689	0.03283	0.00003	0.00404
D0695	0.296	0.281	0.298	0.015	0.098	0.012		-0.1054	0.1102				0.00845	0.45148	0.34198	0.03283	0.02592	0.00889
D0700	0.33	0.294	0.27	0.015	0.08	0.012	HP_1 = 0.33	HP_2 = 0	HP_3 = 0.66	HP_4 = 1	HP_5 = 0.66	HP_6 = 0						
D0705	0.332	0.282	0.277	0.015	0.083	0.012		-0.2900	0.3013		0.1444		0.00980	0.48396	0.33902	0.02930	0.06178	0.01002
D0715	0.287	0.276	0.284	0.058	0.084	0.012					0.1167		0.01456	0.24159	0.28228	0.00565	0.06505	0.01002
D0745	0.274	0.216	0.312	0.052	0.136	0.01				0.0489			0.00041	0.18573	0.10806	0.01219	0.00355	0.00895
D0755	0.264	0.248	0.281	0.058	0.141	0.009		-0.0070		0.0171	0.0243		0.00020	0.17107	0.11036	0.00862	0.00252	0.00561
D0775	0.246	0.239	0.283	0.043	0.139	0.05						-0.0395	0.00182	0.13747	0.07798	0.00211	0.00028	0.00103
D0805	0.132	0.212	0.357	0.054	0.224	0.021		-0.0151	0.0172				0.01323	0.09111	0.07417	0.01901	0.03822	0.00574
D0895	0.278	0.336	0.07	0.08	0.096	0.14	0.0172						0.01497	0.01828	0.00060	0.02037	0.00612	0.00268
D0995	0.301	0.372	0.076	0.121	0.078	0.052		-0.0086					0.00739	0.00134	0.00483	0.00831	0.00128	0.00744