Contributions to genomic selection and association mapping in structured and admixed populations : application to maize

Simon Rio

NNT : 2019SACLS097

Thèse de doctorat

Contributions to Genomic Selection and Association Mapping in Structured and Admixed Populations: Application to Maize

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

Ecole doctorale n°581 Agriculture, alimentation, biologie, environnement, santé (ABIES)
Spécialité de doctorat : sciences agronomiques

Thèse présentée et soutenue à Gif-sur-Yvette, le 26 Avril 2019, par

SIMON RIO

Composition du Jury :

Zulma Vitezica Maître de conférence, INP-ENSAT (Sciences Animales)	Rapporteure
Jianming Yu Professeur, Iowa State University (Department of Agronomy)	Rapporteur
Christine Dillmann Professeure, Université Paris-Sud (GQE - Le Moulon)	Présidente
Christina Lehermeier Généticienne, RAGT S.A. (Genetics and Analytics Unit)	Examinatrice
Vincent Segura Chargé de recherche, INRA (BioForA)	Examinateur
Alain Charcosset Directeur de recherche, INRA (GQE - Le Moulon)	Directeur de thèse
Tristan Mary-Huard Chargé de recherche, INRA (GQE - Le Moulon)	Co-encadrant de thèse
Laurence Moreau Directrice de recherche, INRA (GQE - Le Moulon)	Co-encadrante de thèse

Remerciements

Ces quatre années à l'UMR de Génétique Quantitative et Evolution du Moulon ont été d'une très grande richesse, tant du point de vue scientifique qu'humain, et plusieurs personnes ont largement contribué à ce succès.

Je tiens à remercier tout particulièrement mon directeur de thèse Alain Charcosset, en premier lieu pour m'avoir accepté en stage de Master sur un sujet pourtant compliqué, pour m'avoir beaucoup conseillé sur mon orientation lorsque se posait la question de faire une thèse, pour s'être démené à obtenir des financements de thèse auprès de l'INRA et des entreprises partenaires d'Amaizing, pour avoir été mon directeur de thèse tout en m'accordant régulièrement du temps et des conseils précieux, pour m'avoir délégué la prise en charge de TD lors d'une (et bientôt deux) formation professionnelle en sélection génomique, pour m'avoir chargé avec les autres doctorants GQMS de l'organisation d'un symposium qui aura eu un succès certain, pour la pertinence de ses intuitions, pour sa grande créativité scientifique et pour m'avoir accordé sa confiance au cours de ces quatre années pendant lesquelles j'ai bénéficié d'une grande autonomie.

Je tiens également à remercier mes deux co-encadrants de thèse : Tristan Mary-Huard et Laurence Moreau. Tout d'abord Tristan pour m'avoir bâti un socle de connaissance en statistiques (qui était bien fragile à mon arrivée), pour avoir veillé à l'orthodoxie mathématique de mes travaux de thèse, pour avoir grandement contribué à la clarté des messages de mes articles et de mes chapitres de thèse, pour m'avoir appris à coder efficacement sur R, pour sa patience et sa pédagogie envers le biologiste que je suis, et pour m'avoir donné un coup de pouce pour ma première mission doctorale d'enseignement à AgroParisTech. Ensuite Laurence pour m'avoir fait intégrer le réseau R2D2 qui a grandement contribué à élargir ma culture scientifique, pour m'avoir délégué des responsabilités importantes dans le WP8 d'Amaizing, pour avoir fortement contribué à soigner le style de mes articles et de mes chapitres de thèse, pour avoir su faire pencher la balance vers Alain, Tristan (ou aucun) lorsqu'un bras de fer se lançait entre le monde de la biologie et des mathématiques, et pour son franc parler très appréciable pour avancer efficacement. Plus généralement, je tiens à remercier ce triumvirat qui a grandement contribué à la qualité scientifique de mes travaux de thèse, par l'apport de compétences et connaissances très complémentaires, le tout dans une ambiance de travail d'une grande bienveillance, et enfin pour avoir pris de nombreuses heures à relire et corriger mes ébauches d'articles, de chapitres de thèse et de présentations. J'espère sincèrement avoir l'occasion de travailler de nouveau avec eux trois lors de futures collaborations.

Je tiens ensuite à remercier mes rapporteurs de thèse : Zulma Vitezica et Jianming Yu pour avoir accepté d'évaluer ces trois années de travail résumées en un manuscrit. Je remercie également à remercier les examinateurs de ma thèse : Christine Dillmann, Christina Lehermeier et Vincent Segura, ainsi que les membres de mon comité de thèse pour leurs conseils précieux quant à l'orientation de mes recherches : Brigitte Mangin, Anne Ricard, Léopoldo Sanchez et Simon Teyssèdre

Je tiens aussi à remercier les personnes avec qui j'ai noué de forts liens d'amitié pendant ces quatre années. Tout d'abord Clément Mabire et Adama Seye qui sont arrivés le même jour que moi et qui seront restés dans le même bureau jusqu'au bout, partageant les mêmes galères de fin thèse. Certains qui sont partis plus tôt tels que Romain Barbier, Yohan Benoit ou Camille Clipet, ou arrivés plus tard comme Antoine Allier. Avec

eux, il y aura eu de longues discussions devant la machine à café ou à la cafétéria, de grands chantiers pour assurer l'autosuffisance alimentaire au potager (ou presque), l'organisation de deux conférences à succès, un projet de start-up qui ne nous aura pas rendu millionnaire mais qui fut autrement enrichissant et de nombreux fous rires. Ils auront tous largement contribué à rendre agréable ces années de dur labeur.

Je tiens aussi à remercier particulièrement Cyril Bauland pour les séances d'œnologie, les sessions de trail éprouvantes et les expéditions à la montagne, Valérie Combes pour ses grands talents de cuisinière, Philippe Jamin pour de grands débats de société, Julie Fievet pour un beau t-shirt Ganesh que je garde précieusement, mais aussi Stéphane Nicolas, Delphine Madur, Sophie Pin, Fabien Laporte, Héloïse Giraud et tant d'autres qui sont passés par l'équipe GQMS.

Je remercie aussi l'ensemble des personnes travaillant au Moulon qui participent à faire de cette UMR un lieu de travail très agréable, dont la vie de groupe est rythmée par de nombreux évènements : barbecue, chandeleur, repas de Noël,...

Je veux remercier Liliane Bel d'AgroParisTech et Béatrice Albert de l'Université Paris-Sud pour m'avoir confié des missions d'enseignement, ainsi que tous les sympathiques étudiants à qui j'ai eu l'honneur d'enseigner en TD de biostatistique, ce qui restera un de mes souvenirs professionnels les plus marquants.

Je remercie l'ensemble des personnes qui ont contribué à diriger mon orientation professionnelle vers le monde de la recherche, à commencer par l'ensemble de mes professeurs de l'Université de Rennes 1 et d'Agrocampus Ouest.

Et enfin je remercie mes parents et ma famille pour leur soutien indéfectible et leurs encouragements qui étaient bienvenus dans les périodes de travail intense.

Contents

General introduction

Plant breeding is currently one of the main drivers for achieving a productive and sustainable agriculture. Breeding objectives vary from a species to another but generally target productivity traits, disease resistance, stress tolerances, water and nutrient use efficiency, organoleptic properties, or resistances to storage and transport. Among cultivated plant species, maize is the most important cereal and is used for both human and animal consumption. In 2017, the world production exceeded a billion tons on about 200 millions hectares (FAO, 2019). Maize is cultivated in a wide range of environments, from temperate to tropical regions. After its domestication in Mexico 9,000 years ago, its propagation has been facilitated by a large genetic diversity that has allowed its adaptation to local seasonal constraints, notably through a wide spectrum of plant cycle length. Maize breeders further enhanced its propagation in agricultural systems by developing hybrid cultivars that combined homogeneity, productivity and resilience characteristics. Since the adoption of hybrids in the 1930s, maize yield has been constantly increasing in modern agricultural systems, and this gain was for a large part genetic (Hallauer et al., 1988; Duvick, 2005). The evaluation of homogeneous hybrid cultivars greatly improved the accuracy of selection compared to open-pollinated varieties, thanks to better heritability of the experimental design. The breeding scheme was further improved by splitting diversity into "heterotic" groups and by selecting inbred lines based on their combining ability with lines from complementary groups. The example of yield in maize illustrates the necessity to develop complex methods to study and select for traits that have a continuous distribution. This field of study called quantitative genetics, is closely linked to population genetics and faces the same issues, such as dealing with the stratification of the genetic diversity into genetics groups.

Quantitative genetics in the genomic era

Quantitative traits refer to phenotypic traits that are genetically determined by many regions of the genome, also known as quantitative trait loci (QTLs). When a population sample is evaluated for a given quantitative trait, the observed distribution is continuous and results from the combined effect of the QTLs segregating within the sample. As the environment disrupts their observation, the phenotypic values observed are not direct measures of the genetic values of individuals, and they are usually decomposed as: $Y = G + E$ where Y is the phenotype, G is the genetic value including additive, dominance and epistatic effects at QTLs, and E is the environmental effect. Quantitative traits are often opposed to Mendelian traits that are controlled by a single gene and whose observed phenotype can be directly divided into distinct genotypic categories.

Specific experimental designs and statistical methods have been developed to enable partitioning of the phenotypic variance into a genetic componant and a residual component. This partition of phenotypic variance has allowed breeders to select individuals using estimates of breeding values (or genetic values), and has resulted in a higher selection accuracy compared to a selection based on observed phenotypic values. In animal breeding, breeding values are classically predicted by accounting for relatedness between individuals using linear mixed models. This approach, often referred to as the "animal model", relies on the infinitesimal model which supposes a large number of QTLs with small effects (Fisher, 1918). A different methodology was

adopted in plant breeding for which inbred lines, clones or even progenies of candidate individuals can easily be repeated to improve the heritability of the experimental design. With the advent of molecular markers, new possibilities have emerged.

Molecular markers refer to DNA fragments that exhibit polymorphism between individuals and that can be easily typed by geneticists to be used as genetic markers. Since the 1980s, several methods have been developed such as amplified fragment length polymorphism (AFLP), restricted fragment length polymorphism (RFLP) or simple sequence repeats (SSR). These methods were largely used in linkage mapping studies. This approach takes advantage of the non independent segregation of loci physically linked within families. Provided a sufficient density, markers can be clustered into linkage groups corresponding to the number of chromosomes. When a QTL segregates within the family, it can be identified by the markers to which it is linked, if the bi-parental progeny is evaluated for the trait and if the QTL effect accounts for a substantial proportion of variance. This approach relies on the linkage disequilibrium (statistical association) between QTL and markers linked on the same chromosome generated by the co-segregation of QTL and markers alleles originated from the same parent (Fig. 1). Linkage mapping proved its efficiency to identify several QTLs in numerous species (see Mauricio (2001) for a review). However, several limits exist concerning linkage mapping in bi-parental populations. First, the detection of QTLs is limited to those for which polymorphism is observed between parents. The precision concerning the location of the QTLs is often low and does not allow the direct identification of underlying causal polymorphism. The QTL detection power largely depends on the size of the progeny, often generated for the sole purpose of the linkage mapping study, and a limited number of individuals prevents the detection of QTLs with small effects. Similarly, the effects of the detected QTLs are often overestimated, as described by the Beavis effect (Beavis, 1994, 1998). They may also be specific to the genetic background in which they were detected. All these factors limited the use of QTLs in a breeding context with the noticeable exception of disease resistance QTLs, often characterized by medium to large effects (Young, 1996; Pilet-Nayel et al., 2017).

In the early 2000's, new methods have emerged to identify variations at the level of a nucleotide, which were referred to as single nucleotide polymorphism (SNPs). SNPs were first genotyped using microarrays, that involved the hybridization of DNA and fluorescence microscopy (Ganal et al., 2011; Unterseer et al., 2014), or using methods based on sequencing (Elshire et al., 2011). Using these technologies, the density of markers has increased for most species compared to those obtained with prior techniques, ranging from a few thousand to a million of genotyped positions for a given individual. At such densities, some QTLs happen to be genetically linked with given SNPs due to the existence of ancestral LD among the founders of the population. In linkage mapping studies, the co-segregations occurring within the progeny are the only source of LD, whereas in natural populations, LD due to genetic linkage was shaped by the demographic history of the population and ancient recombination events (Fig. 1). Capturing QTL information using ancestral LD between SNPs and QTLs has opened the way to the identification of QTLs in diversity panels with a higher resolution than linkage mapping, by testing for the existence of an association between the studied traits and the SNP alleles. Such studies were referred to as genome-wide association studies (GWAS) and proved their efficiency to identify numerous QTLs in human, animals and plants (see Bush and Moore (2012) and Huang and Han (2014) for reviews).

The identification of many QTLs for a given trait suggested the possibility of using the estimated effects to predict the genetic values of individuals. Such predictions were of clear interest in a breeding perspective as they could either help to shorten the breeding cycle lengths, improve selection accuracy or increase the selection intensity by selecting within a larger set of individuals for which only genotypic information would be known. However, predictions based on the effects of QTLs detected in segregation families were generally disappointing in terms of prediction accuracy (Moreau et al., 2004). This was due to the statistical limitations of linkage mapping studies and suggested a missing heritability embodied by a large number of QTLs with small effects. Whittaker et al. (2000) and Meuwissen et al. (2001) proposed to predict breeding values using all genomic information, rather than testing for the significance of each marker as a prerequisite to be included in the prediction model. To implement this approach, referred to as genomic selection (GS), a sample of

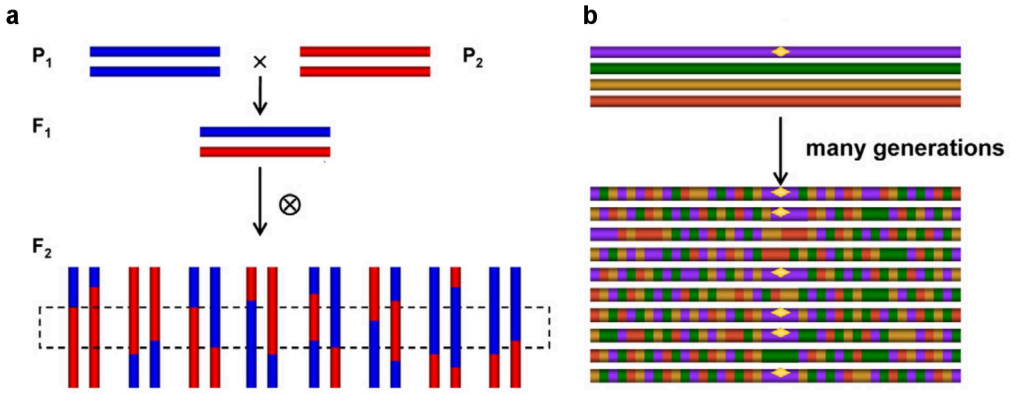


Figure 1: Schematic illustration of the differences between: **a**. linkage mapping and **b**. GWAS, showing the difference of resolution due to the difference in terms of number of recombinations occurring between an F2 and a natural population. The illustration originates from (Zhu et al., 2008)

individuals is typically genotyped and evaluated for a trait, before being used to train a statistical model. Individuals for which the genotypic information is the only source of information can be predicted based on their genomic resemblance with the training population. Numerous methods have emerged including the GBLUP model, classically implemented using linear mixed model and adapted from the "animal model", or a large variety of Bayesian models, making different assumptions on the distribution of allele effects: Bayes-A, Bayes-B or Bayes-Cπ for instance. These techniques often proved a similar efficiency in terms of their predictive ability (Heslot et al., 2012) and GBLUP is still the benchmark thanks to its simplicity. Standard GBLUP model can be written as:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Zg} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is the vector of phenotypes, $\mathbf{1}$ is a vector of "1", μ is the global intercept, $\boldsymbol{Z}$ is an incidence matrix linking phenotypes to breeding values, $\boldsymbol{g}$ is the vector of breeding values with $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{K}\sigma^2_G)$, $\boldsymbol{K}$ is the kinship matrix, σ^2_G is the genetic variance, $\boldsymbol{e}$ is the vector of errors with $\boldsymbol{e} \sim \mathcal{N}(0, \boldsymbol{I}\sigma^2_E)$, $\boldsymbol{I}$ is the identity matrix and σ^2_E the error variance. Compared to a standard "animal model", for which a kinship based on pedigree (i.e. expected) relationships is used, the GBLUP model uses a kinship computed using molecular markers, following for instance formulas proposed by VanRaden (2008) or Astle and Balding (2009). Based on estimates of variance components, the genomic prediction of an individual can be computed using the BLUP of its breeding value, or more efficiently using mixed model equations if the number of phenotypic values is large for each individual (large $\boldsymbol{y}$ compared to $\boldsymbol{g}$) (Henderson, 1975). Note that in the previous model, each individual had a single phenotypic value. GS started to meet an increasing popularity in both animal and plant breeding by the end of the 2000's thanks to the development of dense genotyping arrays in many species. Beyond the prediction of unobserved individuals, other applications have emerged such as a better monitoring of field trials, the prediction of genotype by environment interactions (GxE), benefiting from genetic correlations between traits in multi-trait genomic predictions, or the optimization of crossing designs based on the expected genetic variance (see Heffner et al. (2009) and Heslot et al. (2015) for reviews).

The advent of molecular markers has had a major impact on quantitative genetic studies. Accessing the genomic information underlying the determinism of the evaluated traits opened new fields of research. However, several factors remain complicated to handle, both to identify QTLs and to apply genomic predictions, such as the stratification of breeding populations in genetic groups.

Genetic structure: theory, inference and a maize perspective

Genetic structure refers to the existence of sub-groups of individuals within a population. It is tightly linked to the history of the species in which it is observed. It generally arises when groups of individuals are spatially separated and preferentially mate with individuals from their own group. Following their separation, differential evolutionary forces, such as drift or selection, cause a divergence of the allele frequencies between groups. This divergence is classically measured by the fixation index F_{ST}, first introduced by Wright (1943, 1949). F_{ST} is defined as the correlation of randomly chosen alleles within a given group relative to the entire population, or equivalently as the proportion of genetic diversity due to group differences in allele frequencies (see Holsinger and Weir (2009) for a review). The locus-specific estimates of F_{ST} can be used to identify regions that have been subjected to selection. Such regions may show a particularly high degree of differentiation compared to the genome-wide distribution of F_{ST}, and test procedures are implemented in software like BayesScan (Foll and Gaggiotti, 2008).

From a genome perspective, the genetic structure not only affects the frequency of alleles at given loci, but also the LD between these loci. Group differences in LD extent and linkage phase between physically linked loci can be observed due to specific demographic histories. The extent of LD is tightly linked to the effective population size and to its dynamics, with a strong impact of demographic events such as bottlenecks or expansions. It tends to decline when the effect of recombination is large relative to drift, as observed in populations with a large effective size. Conversely, LD extent tends to increase in small populations, for which the effect of drift is strong compared to that of recombination (Pritchard and Przeworski, 2001; Rogers, 2014). Such group differences in LD have been identified using markers in numerous species including human (Sawyer et al., 2005; Evans and Cardon, 2005), dairy and beef cattle (de Roos et al., 2008; Porto-Neto et al., 2014), pig (Badke et al., 2012), wheat (Hao et al., 2011) or maize (Van Inghelandt et al., 2011; Technow et al., 2012; Bouchet et al., 2013; Rincent et al., 2014b). In addition to group differences in LD, the stratification of a population into genetic groups generates LD between loci that are not genetically linked. These loci are those showing a high differentiation between groups. An extreme example would consist of a set of bi-allelic loci differentially fixed between genetic groups, and for which a global LD of 1 would be estimated between all pairs of loci.

Genetic groups are not always perfectly separated as gene flow may occur between groups. Genetic admixture refer to the existence of DNA fragments of different ancestries within an individual, forming a mosaic of ancestry blocks. Such individuals result from the mating of individuals from different genetic groups and admixture is supposed to be an important factor that allowed the adaptation of species to new environments (Rius and Darling, 2014).

When a population of individuals has been genotyped using molecular markers, one can investigate the existence of population structure within the dataset. This procedure possibly involves different but complementary objectives including the detection of population structure, the determination of the number of genetic groups and the assignation of individuals to these groups. Principal coordinates analysis can simply be applied to a distance matrix computed using marker data, in order to identify clusters on principal components. Model-based methods were also developed, the best known being the STRUCTURE software developed by Pritchard et al. (2000) which models the probability of observed genotypes using ancestry proportions and population allele frequencies. The algorithm was implemented in a Bayesian framework and was further extended to allow for linkage between markers and admixture (Falush et al., 2003). Another sofware, ADMIXTURE, was developed by Alexander et al. (2009) using the same model as STRUCTURE. It was based on maximum likelihood estimation which led to a considerable fastening of computing time. Other methods specialized in the inference of local ancestry information, aiming at locally assigning chromosome blocks to different groups, such as LAMP (Sankararaman et al., 2008), RFmix (Maples et al., 2013) and many others (see Liu et al. (2013) and Padhukasahasram (2014) for reviews).

From a maize perspective, genetic structure is a major component of its existing diversity. Genetic groups have been shaped by various dissemination pathways since maize was domesticated in the Balsas region valley of Mexico around 9,000 years ago from a wild ancestor of the *Zea* genus (Matsuoka et al., 2002). From Mexico, maize would have spread in two main directions: north of Mexico to the United States, and south of Mexico to the Caribbean and South America. These expansions to contrasting environments led to main clusters of diversity including the Mexican highlands, the tropical lowlands, the Andean group or the northern USA group. Each of these clusters can be further divided into distinct genetic groups. For instance, the northern USA group includes the American Northern Flints, which were first introduced to the USA, the Southern Dents later introduced from the Caribbeans and the Corn-Belt Dents originating from their hybridization. The introduction of Maize in Europe probably results from two independent events: tropical material were introduced to Spain and American Northern Flints were introduced to northern Europe, creating the European Northern Flints (Fig. 2). Evidence of admixture events were shown in Europe between genetic materials and led to the creation of new groups such as the European flints or the Italian group (Brandenburg et al., 2017). All these groups were further structured into heterotic groups that are currently used in hybrid breeding. For instance, hybrids between Corn-Belt-Dent and European flints enhanced the productivity of maize in northern Europe and contributed to its propagation in agricultural systems (see Tenaillon and Charcosset (2011) for a European perspective on maize history).

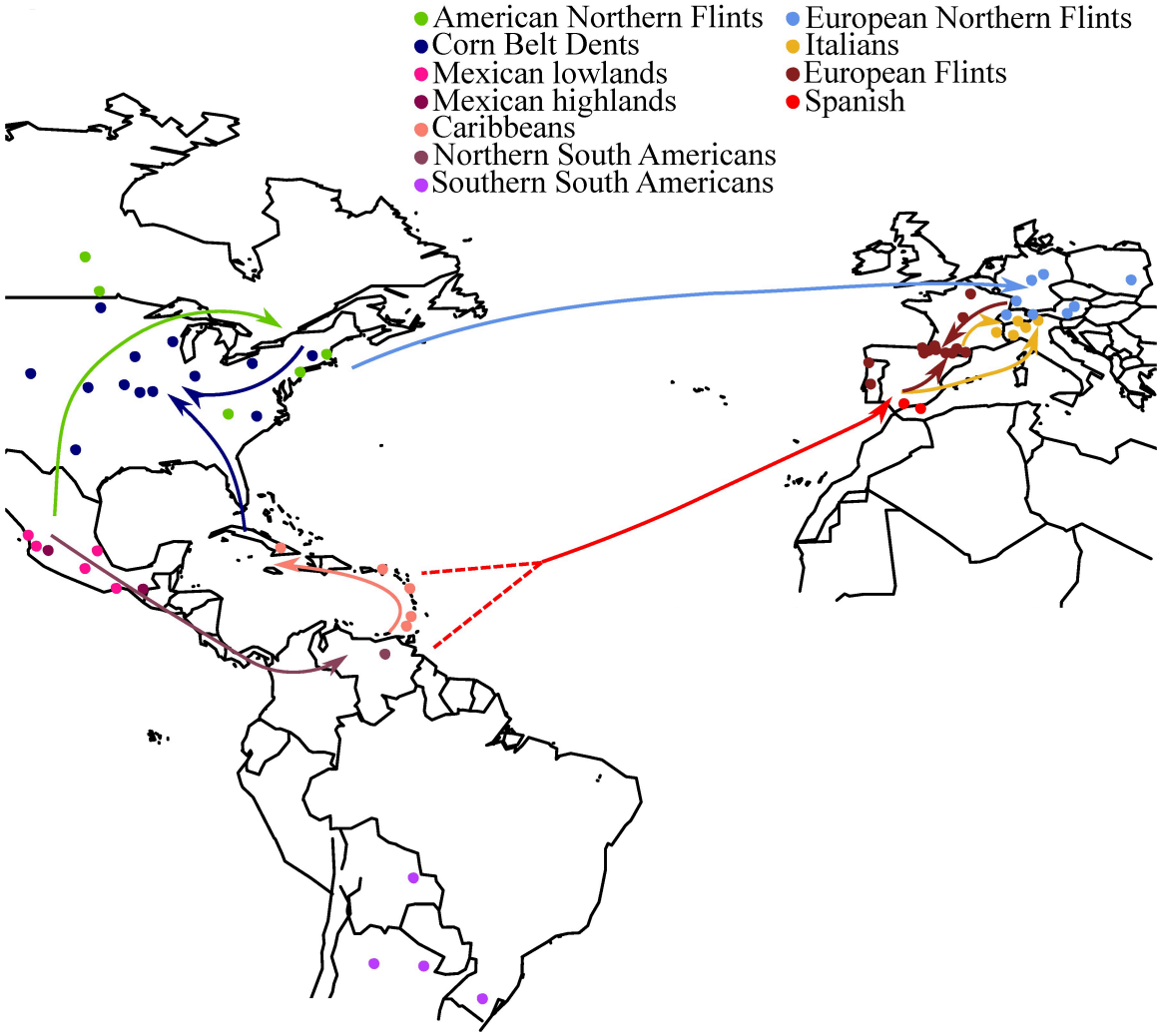


Figure 2: Maize genetic groups and diffusions pathways inferred 66 maize landraces adapted from Brandenburg et al. (2017)

As illustrated by the example of maize, a population stratified into genetic groups seldom involves the

existence of clearly separated groups with similar degrees of differentiation. It is rather the result of a complex phylogeny with hierarchical levels, shaped by demography, migrations or admixture events.

How does genetic structure affect quantitative traits?

The standard model of quantitative genetics does not explicitly account for the existence of genetic structure, but rather assumes a single population. However, the stratification of a population into genetic groups may impact the genetic components in different manners.

Let us consider a structured population including P genetic groups studied for a given trait. Within each group, the genetic value of each individual is computed as a sum of alleles effects at M bi-allelic QTLs. For the sake of simplicity, each individual is assumed to be inbred (no heterozygosity) and showing no admixture. QTLs are assumed to be in linkage equilibrium within each group and not to interact with other loci (no epistatic interactions). We can model the genetic value of a given individual as:

$$G_i = \sum_{p=1}^{P} Z_{ip} \sum_{m=1}^{M} \left(\beta^0_{mp} + W_{im}\left(\beta^1_{mp} - \beta^0_{mp}\right)\right)$$

where G_i is the genetic value of individual i, Z_{ip} is a variable taking the value "1" if i belongs to group p or "0" otherwise, $(W_{im}|Z_{ip} = 1) \sim \mathcal{B}(f_{mp})$ is the genotype (coded 0/1) of individual i at locus m drawn conditionally to the group ancestry of i in a Bernoulli distribution of parameter f_{mp}, with f_{mp} being the frequency of allele 1 at locus m for group p, β^0_{mp} and β^1_{mp} are the QTL allele effects at locus m for group p for allele 0 and 1, respectively, and all random variables are assumed to be independent from each other.

Using this generative model, it is possible to study how genetic structure affects a quantitative trait in terms of expected value and genetic variance, which are important parameters to characterize a breeding population. The expected value of a given individuals from genetic group p will be:

$$\mathrm{E}\left(G_i|Z_{ip} = 1\right) = \sum_{m=1}^{M} \beta^0_{mp} + f_{mp}\left(\beta^1_{mp} - \beta^0_{mp}\right) = \mu_p$$

where group differences in expected value may result from group-specific QTL allele frequencies but also from differences in terms of QTL allele effects. The same observation can be made concerning the genetic variance of a given individual from genetic group p:

$$\mathrm{V}\left(G_i|Z_{ip} = 1\right) = \sum_{m=1}^{M} f_{mp}(1 - f_{mp})\left(\beta^1_{mp} - \beta^0_{mp}\right)^2 = \sigma^2_{G_p}$$

where group differences in genetic variance may result from group-specific QTL allele frequencies and/or allele effects.

On the one hand, group differences in allele frequencies at QTLs are very likely by definition, as genetic groups are characterized by specific allele frequencies at loci. These differences may result from differential selection pressures in contrasting environments that shift QTL allele frequencies, or may simply be due to an independent drift within each group.

On the other hand, group-specific allele effect at causal QTLs may not be as likely. A possible explanation for their existence lies in epistatic interactions between the QTLs and the genetic background. In this case, the genetic background is represented by one or several loci that are differentially fixed between groups. For a given QTL A interacting with a single locus B, the QTL allele effect at locus A, defined as $\beta^1_{Ap} - \beta^0_{Ap}$, will be conditioned by the allele observed at locus B. If two genetic groups are highly differentiated at locus B, then the mean QTL effect will be different between the two genetic groups, as proposed by Tang (2006) and illustrated in Fig. (3). Another explanation is the appearance of a new genetic mutation very close to

a QTL in a common founder of a given group, resulting in a different effect compared to a group for which the mutation is absent. Evidence of such mutations were found in human, as several Mendelian symptoms of obesity were shown to result from mutations within specific ethnicities (see Stryjecki et al. (2018) for a review)

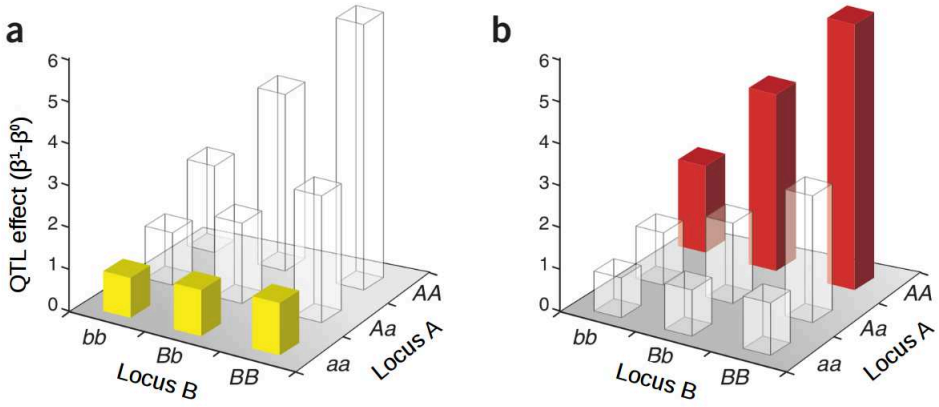


Figure 3: Schematic illustration of two interacting loci adapted from Tang (2006). Filled bars represent common allele combinations while open bars are not observed in the group: **a** in the first group, most individuals have genotype "aa" at locus A, and no QTL effect is observed at locus B, for which all allele combinations are common, **b** in the second group, most individuals have genotype AA at locus A, and the resulting QTL effect at locus B is higher

In analogy to the F_{ST} indicator which quantifies the proportion of genetic diversity due to group differences in allele frequencies, the Q_{ST} indicator was proposed by Spitze (1993) to quantify the proportion of genetic variance that is due to among-group differences when studying quantitative traits. This indicator highlights the existence of a proportion of genetic variance that is not directly accessible to a breeder, unless he generates admixture and segregations by crossing individuals of different groups.

In conclusion, the stratification of a population into genetic groups impacts quantitative traits through differences in QTL allele frequencies and possibly through group-specific QTL allele effects. One should notice that other factors may impact the mean and the genetic variance: group differences in LD between QTLs, in inbreeding, in dominance effects, or even in interactions between QTL allele effects.

Impact of genetic structure on association mapping and genomic selection

The stratification of a population into genetic groups may impact the methods to study quantitative traits, particularly GWAS and GS that involve molecular markers.

Applying GWAS to a structured population raises the issue of spurious associations. They result from the long range LD generated by genetic structure for SNPs and QTLs that are highly differentiated between groups, as previously discussed. If a given trait is characterized by group-specific means, all the SNPs differentiated between groups will correlate to it. An efficient control of these spurious associations can be done by taking structure and kinship into account in the GWAS model (Yu et al., 2006; Price et al., 2006). For each bi-allelic marker m among M loci, a GWAS model can be written in a simplified version of that proposed by Yu et al. (2006) as:

$$Y_{ijk} = \mu + \beta_j^m + \alpha_k + G_{ijk} + E_{ijk}$$

where Y_{ijk} is the phenotype of the individual, μ is the intercept, β_j^m is the effect of the allele j with $j \in \{0, 1\}$

at marker m, α_k is the effect of genetic group k, G_{ijk} is random polygenic effect, $\boldsymbol{g}$ is the vector of random polygenic effects with $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{K}\sigma^2_G)$, $\boldsymbol{K}$ is the kinship matrix, σ^2_G is the genetic variance, E_{ijk} is the error, $\boldsymbol{e}$ is the vector of errors with $\boldsymbol{e} \sim \mathcal{N}(0, \boldsymbol{I}\sigma^2_E)$, $\boldsymbol{I}$ is the identity matrix and σ^2_E is the error variance. This model can account for different levels of structure using α_k for the effect of the main stratification into genetic groups, and by modeling the genetic covariance between individuals using the kinship $\boldsymbol{K}$ for groups of related individuals. If genetic structure is not sufficiently accounted for by the model, false positives may be detected when testing for the existence of a differential effect between alleles ($H_0 : \beta^m_1 - \beta^m_0 = 0$). As an example, the *Dwarf8* locus was found to be associated with maize flowering time in early association studies (Thornsberry et al., 2001), and it was later shown that its effect had been greatly overestimated due to insufficient control of the genetic structure (Larsson et al., 2013). Once structure is accounted for by the GWAS model, a low power of detection is generally observed for the highly differentiated SNPs (Rincent et al., 2014a). QTLs located in differentiated regions happen to be difficult to detect, especially in case of rare alleles. This is why innovative genetic material were developed such as nested association mapping (NAM) (McMullen et al., 2009) or multi-parent advanced generation inter-cross (MAGIC) (Cavanagh et al., 2008). These genetic materials consist in generating progenies from a limited number of founders in order to ensure a high statistical power along with a large diversity studied and a population structure that is either considered as negligeable using MAGIC or that can easily be controled by the familiy structure using NAM.

From a GS perspective, the stratification of a breeding population into genetic groups may impact genomic prediction accuracy in different manners. When a consistency is observed between the training set (TS) and the predicted set (PS) in terms of genetic groups, the group mean differences are well accounted for by the model through the kinship and participate to the accuracy (Guo et al., 2014). Conversely, when targeting a group-specific PS, training a model on a different group can decrease dramatically the accuracy, as shown in several species including dairy and beef cattle (Olson et al., 2012; Chen et al., 2013) and maize (Technow et al., 2013; Lehermeier et al., 2014). The use of multi-group TSs was proposed by de Roos et al. (2009) for several applications including the possibility to apply predictions to a broad range of genetic diversity, the improvement of genomic selection efficiency in genetic groups with limited size or the optimization of resources for traits that are expensive to evaluate. Such multi-group TSs showed a good predictive ability in a wide range of species such as dairy cattle (Brøndum et al., 2011; Pryce et al., 2011; Zhou et al., 2013), maize (Technow et al., 2013) or soybean (Duhnen et al., 2017). However, the gain in precision is often limited compared to what could be obtained by applying predictions separately within groups (Carillier et al., 2014; Hayes et al., 2018).

Structure does not only affect genomic prediction accuracy, but also the ability to forecast this accuracy using *a priori* indicators such as the coefficient of determination (CD) (VanRaden, 2008; Rincent et al., 2012). Forecasting genomic prediction accuracy would allow breeders to evaluate the interest of multi-group TSs and *a priori* indicators could be used as criteria to optimize their constitution. However, when the population features a strong genetic structure, standard *a priori* indicators showed a lack of efficiency to forecast genomic prediction accuracy in multi-breed dairy cattle populations (Hayes et al., 2009) and to optimize TSs in rice populations (Isidro et al., 2015).

A different genetic information captured by SNPs may explain the difficulty to borrow genetic information from one group to another. When a set of molecular markers is available for a trait, the genomic information at QTLs is partially captured by SNPs using LD. Group differences in LD may lead SNPs to capture different genetic information between groups, especially at low to medium genotyping densities. This issue led Wientjes et al. (2015b) to propose a method to estimate the consistency of LD between SNPs and QTLs across genetic groups, which uses the selection index theory and simulated QTL allele effects. The existence of group differences in LD and linkage phases, as well as the possibility of contrasted QTL allele effects between groups, makes the observation of group-specific SNP allele effects in structured populations likely. Such group-specific allele effects may cancel each other out in their overall effect when applying GWAS to a structured panel, making them difficult to detect using standard methods. In GS, accounting for this heterogeneity in QTL allele effects is likely to improve genomic prediction accuracy in a multi-group breeding context. Modeling

group specific SNP allele effects in genomic prediction models was proposed by Karoui et al. (2012) and Lehermeier et al. (2015) by adapting multi-trait models to multi-group predictions. In such models, the SNP allele effects are assumed to be different but correlated between groups. This same formalism was also used to derive new *a priori* indicators of accuracy (Wientjes et al., 2015a) or to propose relevant estimators of relatedness to estimate genetic correlations between groups accurately (Wientjes et al., 2017). Other modelings were proposed such as the decomposition of SNP effects into a main SNP effect and group-specific deviations, as proposed by Schulz-Streeck et al. (2012), de los Campos et al. (2015) or Technow and Totir (2015).

However, all these models were restricted to pure individuals and did not accommodate to the presence of admixed individuals. The interest of such individuals in multi-group TSs was shown by Toosi et al. (2013) using simulations, as they may create connections between groups and allow for more genetic information to be borrowed. The "animal model" was adapted to admixed population before the advent of high density genotyping, by considering pedigree relationships between individuals and global admixture proportions. The genetic variance was split into group-specific and segregation components (Lo et al., 1993; García-Cortés and Toro, 2006). The aim was to account for the additional variance observed in an admixed population compared to parental populations due to the segregation of QTL with differentiated alleles frequencies in admixed individuals. Such methodology was later adapted to genomic prediction by Strandén and Mäntysaari (2013) and Makgahlela et al. (2013) by replacing the pedigree matrix by a standard kinship matrix estimated with SNPs. Alternative methods were also developed to account for various types of heterogeneity between genetic groups, such as computing an alternative covariance matrix based on specific kernel functions (Heslot et al., 2015).

In conclusion, the stratification of a population into genetic groups may affect quantitative genetics studies in several ways. The observation of group-specific allele effects at SNPs, possibly resulting from group-specific allele effects at QTLs, is a major factor affecting both GWAS and GS. While extensive literature exists considering their modeling in a GS context, little attention has been given to their identification using GWAS. In this same perspective, the integration of admixed individuals in GWAS and GS studies has not been much considered so far. They may however be useful to connect genetic group in multi-group TSs or to get some insight concerning the stability of SNP allele effects across genetic backgrounds. As their production requires significant human and material resources, it is important to evaluate their interest according to these objectives.

Objectives of the thesis

From both GS and GWAS perspectives, we studied the impact of genetic structure in quantitative genetics studies using maize structured datasets genotyped at high density. The main objectives were (i) to study the impact of genetic structure on both genomic prediction accuracy and on its *a priori* estimation based on the coefficient of determination (CD), (ii) to identify and unravel group-specific allele effects at SNPs using GWAS and admixed individuals in addition to pure individuals, (iii) to develop genomic prediction models adapted to admixed individuals that account for group-specific SNP allele effects and (iv) to evaluate the interest of using admixed individuals in multi-group TSs.

To achieve these goals, we used two maize inbred diversity panels, involving different levels of genetic structure. The first panel, called "Amaizing Dent", will be presented in Chapter 1. It includes 389 dent lines genotyped for 1M SNP and can subdivided in three genetic groups. This panel was evaluated for hybrid performances, using a common flint tester, for flowering and productivity traits. The second panel will be presented in Chapter 2 and is called "Flint-Dent" panel. It includes 304 flint lines, 300 dent lines included in the "Amaizing Dent" panel and 366 admixed lines. The admixed lines were generated from hybrids, mated

according to a factorial design between the pure dent and flint lines of the panel. All lines were evaluated *per se* for traits related to flowering time and plant heights.

In the Chapter 1, we studied the impact of genetic structure on genomic prediction accuracy using the "Amaizing Dent" panel. For a given size of TS, structure-based scenarios were defined including within-, across- or multi-group predictions. We also evaluated the benefits of adding extra-group individuals to the TS, in order to predict group-specific PSs. All these scenarios were considered to study whether or not genetic information can be borrowed between genetic groups. The genomic prediction accuracy of alternative predicton models, that account for genetic structure explicitly, was also compared to that of standard GBLUP. To study the efficiency of *a priori* indicators of accuracy in structured populations, we compared a standard indicator based on CD to new indicators recently proposed by Wientjes et al. (2015a). The *a priori* estimation of accuracy was compared to the empirical accuracy obtained in the structure-based scenarios. The objective was to evaluate whether *a priori* indicators would be efficient to forecast accuracy within a multi-group breeding population, and could later be used to optimize the composition of multi-group TSs. This study was recently published in Theoretical and Applied Genetics (Rio et al., 2019).

In Chapter 2, we developed a GWAS methodology to test for the existence of a heterogeneity of SNP allele effects between genetic groups, and applied it to the "Flint-Dent" panel evaluated for flowering traits. We showed how including admixed individuals to the analysis can help to disentangle the factors causing the heterogeneity of allele effects across groups: local genomic differences (group differences in LD or group specific mutations) or epistatic interactions between QTLs and the genetic background. A test for directional epistasis was also proposed to support the existence of epistatic interactions in this dataset. The objective was to study if our method can be used to get insight concerning traits in structured populations as well as to analyze the stability of marker effects at main QTLs across genetic groups. This study will soon be submitted to PLOS Genetics.

In Chapter 3, we developed two genomic prediction models that account for the existence of group-specific allele effects in admixed populations. Both models, called Multi-group Admixed (MAGBLUP) 1 and 2, are taking advantage of both genomic data and local admixtures, defined as the group ancestry of SNP alleles. The first model was derived according to the "animal model", for which the genotypes are random, while the second was derived by assuming a random distribution for allele effects. Both models were evaluated for their precision in variance component estimation and genomic prediction accuracy using the "Flint-Dent" panel evaluated for simulated and real traits. In this chapter, we also evaluated the benefits of adding admixed individuals to multi-group TSs. The structure-based scenarios defined in Chapter 1 were adapted by replacing pure lines with admixed lines in TSs. The objective was to evaluate whether admixed individuals would allow for a better genetic connection between genetic groups within structured breeding populations. This study will soon be submitted in a journal to be determined.

Chapter 1

Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel

Simon Rio, Tristan Mary-Huard, Laurence Moreau and Alain Charcosset

Abstract

Key message Population structure affects genomic selection efficiency as well as the ability to forecast accuracy using standard GBLUP.

Genomic prediction models usually assume that the individuals used for calibration belong to the same population as those to be predicted. Most of the *a priori* indicators of precision, such as the Coefficient of Determination (CD), were derived from those same models. But genetic structure is a common feature in plant species and it may impact genomic selection efficiency and the ability to forecast prediction accuracy. We investigated the impact of genetic structure in a dent maize panel ("Amaizing Dent") using different scenarios including within or across group predictions. For a given training set size, the best accuracies were achieved when predicting individuals using a model calibrated on the same genetic group. Nevertheless, a diverse training set representing all the groups had a certain predictive efficiency for all the validation sets, and adding extra-group individuals was almost always beneficial. It underlines the potential of such generic training sets for dent maize genomic selection applications. Alternative prediction models, taking genetic structure explicitly into account, did not improve the prediction accuracy compared to GBLUP. We also investigated the ability of different indicators of precision to forecast accuracy in the within or across group scenarios. There was a global encouraging trend of the CD to differentiate scenarios, although there were specific combinations of target populations and traits where the efficiency of this indicator proved to be null. One hypothesis to explain such erratic performances is the impact of genetic structure through group-specific allele diversity at QTLs rather than group-specific allele effects.

Introduction

Recently, new breeding methods emerged grouped under the term Genomic Selection (GS). Their aim is to predict breeding values using all the genomic markers jointly rather than testing the significance of each of them (Meuwissen et al., 2001). Several models have been proposed in the literature, making different hypotheses on the distribution of QTL effects such as GBLUP, BayesA, BayesB or BayesCπ (Heslot et al., 2012). Most of the time, genomic predictions are calibrated on a Training Set (TS) and then applied to a population to be selected, for which the genotypic data is the only source of information available. This scenario is interesting for traits that are expensive or difficult to evaluate, but genomic prediction accuracy may be limited when the genetic distance is large between the TS and the Validation Set (VS) (Pszczola et al., 2012). Other applications have recently emerged such as a better monitoring of field trials by limiting repetitions (Endelman et al., 2014).

Along with the development of genomic prediction models, there has been a significant effort to develop *a priori* indicators of precision. The forecast of genomic prediction accuracy could allow breeders to evaluate the interest of a generic TS to predict breeding values in a given breeding population. It would also be possible to optimize the TS constitution in order to maximize genomic prediction accuracy and more generally to optimize breeding programs. A first set of approaches using deterministic equations were developed involving different parameters such as the trait heritability, the population size and the effective number of chromosome segments linked to the effective population size (Daetwyler et al., 2008; Goddard et al., 2011; Erbe et al., 2013; Elsen, 2016). Their efficiency depends largely on the ability to estimate this latter parameter accurately, which proved complicated in practice (Brard and Ricard, 2015). An other set of approaches, further referred to as CD, used mixed model equations requiring only the relationships between individuals using genomic or pedigree data and estimates of heritability (VanRaden, 2008; Rincent et al., 2012; Rabier et al., 2016).

The genomic prediction models and their corresponding indicators of precision were first developed while considering one homogeneous population. However, this assumption is often violated as genetic structure is a common feature in human, animal and plants. Genetic structure arises when the allele frequencies of sub-groups of individuals differ when compared to the ancestral population from which they originate. It might be due to a reproductive isolation followed by an independent drift in each group. In maize, genetic structure is found at the level of heterotic groups, that have been selected for their complementarity in order to maximize heterosis of inter-group hybrids, but may also be observed within each of these heterotic groups (Rincent et al., 2014b).

When doing GS, genetic structure can impact the accuracy of predictions (Guo et al., 2014; Albrecht et al., 2014). Most of the models assume a single population, presupposing the conservation of QTLs effects between individuals. However, when a genetic structure is observed, differences between QTLs effects can be observed as well as differences in terms of LD extent (Wientjes et al., 2015c) and linkage phase between SNPs and QTLs across populations (Wientjes et al., 2015b). If the same structure is found within the TS and the VS, it is well taken into account by the kinship, when using a GBLUP model, and contributes to genomic prediction accuracy (Guo et al., 2014). But if the structure is different, the accuracy can be strongly impacted. In dairy cattle, when trying to predict breeding values of a breed with a small population size using information coming from a distantly related breed with a larger size, the gain in prediction accuracy is generally very low and may even be negative (de Roos et al., 2009). Likewise, in maize, a TS combining dent and flint lines allowed marginal gains in terms of accuracy compared to pure dent or pure flint TS (Technow et al., 2013). A substantial gain was nevertheless observed when combining related dairy cattle breeds (Brøndum et al., 2011).

To improve genomic prediction accuracy in structured populations, it is possible to adjust both the experimental design and data modeling. Concerning the experimental design, it is possible to improve accuracy by creating hybrids or admixed individuals allowing to connect the different groups (Toosi et al.,

2013; Esfandyari et al., 2015). Concerning the models, several alternatives have been proposed such as specifying the structure as a fixed effect (Guo et al., 2014) or modeling genetic covariances between individuals from different groups by adapting multi-trait models. The latter led to improvement of genomic prediction accuracy in dairy cattle (Olson et al., 2012; Karoui et al., 2012) and dairy goat (Carillier et al., 2014) when the genetic correlations between groups were sufficiently high. In maize, such types of models were also applied and allowed very limited gains (Lehermeier et al., 2015). In the case of populations resulting from the admixture between groups, there were attempts to take into account the quantitative assignment of individuals to groups by applying random regression allowing limited gains (Strandén and Mäntysaari, 2013; Makgahlela et al., 2013).

Structure does not only affect genomic prediction accuracy, but also the ability to forecast this accuracy using indicators. When the population features a strong genetic structure, *a priori* indicators proved to be inefficient to forecast genomic prediction accuracy in multi-breed dairy cattle populations (Hayes et al., 2009) and to optimize TS in rice populations (Isidro et al., 2015). To tackle this issue, new indicators were recently developed in order to take into account such structures, proving their efficiency on simulated data (Wientjes et al., 2015a, 2016).

The objectives of this article are first to study the impact of genetic structure on genomic prediction accuracy within a diversity panel of maize dent lines. An important question for breeders is whether or not one should combine groups in TS in order to increase accuracy. A second objective is to evaluate the gain in accuracy one may expect when applying models that explicitly account for genetic structure rather than using a standard GBLUP analysis. The last objective is to compare different indicators of precision to study their ability to forecast GBLUP accuracy in presence of genetic structure.

Materials and methods

Genetic material and genotypic data

Genetic material is a panel of 389 dent maize lines assembled within the "Amaizing" project and aiming at representing the diversity of the dent heterotic group that can be used in European breeding. This panel includes most lines from panels assembled for previous projects: "CornFed" (Rincent et al., 2014b) and "Drops" (Millet et al., 2016). This panel was constructed for Genome Wide Association Studies and to apply genomic prediction for traits that are expensive to evaluate. One originality of the panel is to include 49 elite lines coming from seven breeding companies (Fig. 1.1-**a**), all members of the Amaizing project.

The genotyping data, initially assembled for GWAS studies, included SNPs from different technologies: the 50K Illumina MaizeSNP50 BeadChip (Ganal et al., 2011), the 600K Affymetrix Axiom Maize Genotyping Array (Unterseer et al., 2014) and Genotyping-By-Sequencing (Elshire et al., 2011; Glaubitz et al., 2014). The lines from public origin were all genotyped with the three SNP technologies. The lines from private origin were all genotyped with the 50K chip, 28 were also genotyped with the 600K but none with GBS. At each SNP, allele 0 was attributed to the allele carried by B73, the maize inbred line used as reference for sequencing, or to the allele carried by the first line in alphabetic order if B73 genotype was missing or heterozygous. A quality control on SNP data was applied, removing markers featuring heterozygosity above 15% and missing value rate above 20% for 50K and 600K SNPs. For GBS data, heterozygous were transformed into missing values and markers with more than 70% of missing data were discarded. After merging the three datasets, duplicated SNPs were removed from the dataset based on physical position information leaving 986,045 SNPs. The imputation of missing values was done on the whole dataset using Beagle v.3.3.2 and default parameters (Browning and Browning, 2009).

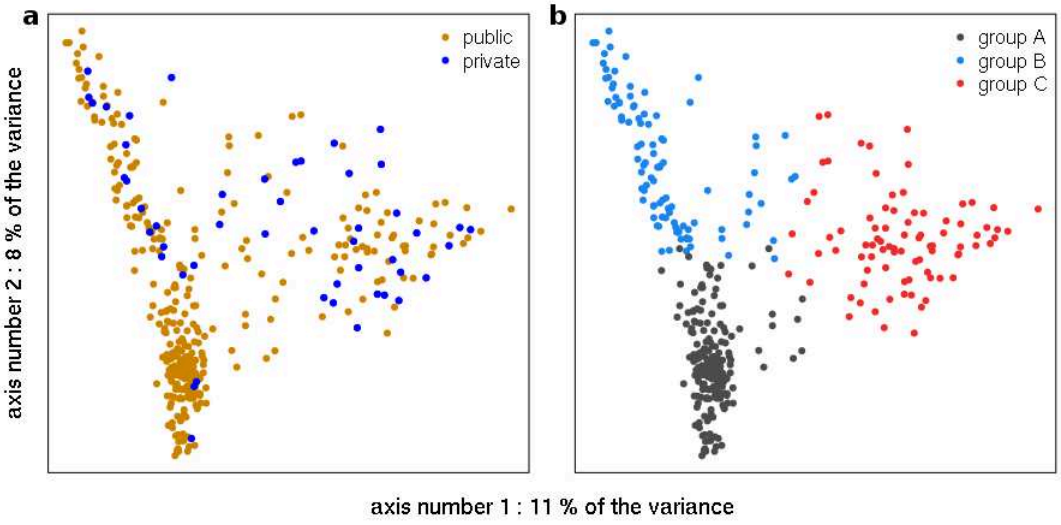


Figure 1.1: PCoA on genetic distances with coloration of individuals depending on **a**. their origin (public/private) or **b**. by assignment to genetic groups.

Structure analysis

We performed a structure analysis using the ADMIXTURE software (Alexander et al., 2009) for different numbers of groups, ranging from 2 to 8 (Supplementary Fig. S1.1, S1.2 and S1.3). The Cross-Validation (CV) error criterion proposed by ADMIXTURE showed an improvement while increasing the number of groups. For the following analyses, we considered three groups which could be linked to well-defined groups in maize breeding which are A: Lancaster and other dent lines (207 lines), B: Stiff-Stalk (98 lines) and C: Iodent (84 lines). Subdividing in more groups would have led to define families or specific pedigree structures rather than well known genetic groups and to insufficient number of lines per group to perform analyses. A PCoA was performed on genetic distances computed as $D_{i,j} = 1 - K^0_{i,j}$ with $K^0_{i,j}$ being the kinship coefficient between lines i and j in Eq. (1.1, see below). This analysis clearly separated individuals based on their maximal admixture coefficient (Fig. 1.1-**b**).

Phenotypic data

All the lines were crossed to the same tester UH007 to produce hybrid progenies for phenotypic evaluation. The 2014 field trials (Supplementary Table S1.1) were conducted in seven locations in standard agronomic conditions including Blois, Mons, Niederhergheim, Souprosse, Villampuy (France), Bernburg (Germany) and Graneros (Chile). Each trial was a latinized alpha design where every genotype was repeated 2 times on average. Grain Moisture (in % of humidity), Grain Yield at 15% of humidity (quintals per hectare) and Male Flowering time were recorded for each plot. Male Flowering time was converted into growing degree-days, considering a base temperature of 6 Celsius degrees, using the mean daily air temperature measured at each location. An economic index, called Yield Index, was also computed as: Yield Index = Grain Yield - 2.5 $\times$ Grain Moisture. This index corresponds to Grain Yield penalized for an excess of humidity at harvest, which would require an expensive drying process, and is used for variety registration in France.

We started the analysis from data collected after correction for within trial spatial effects using different models (Supplementary Table S1.1). Then, we computed Least-Square means (LS-means) over the whole design using model: $Y_{ij} = \mu_i + T_j + E_{ij}$ where Y_{ij} is the performance of individual i in the location j, μ_i is the intercept for individual i, T_j is the j^{th} random trial effect where $T_j \sim \mathcal{N}(0, \sigma^2_T)$ are all independent and identically distributed (i.i.d.), E_{ij} is the error were $E_{ij} \sim \mathcal{N}(0, \sigma^2_E)$ i.i.d. and E_{ij} and T_j are assumed to be

independent.

Genomic prediction models

All the genomic prediction models used in this study can be written as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{g} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is the vector of LS-means which will be further referred to as phenotypes, $\boldsymbol{X}$ is the incidence matrix for fixed effects, $\boldsymbol{\beta}$ is the vector of fixed effects, $\boldsymbol{Z}$ is an incidence matrix linking observations to breeding values, $\boldsymbol{g}$ is the vector of breeding values and $\boldsymbol{e}$ is the vector of errors. All models assume independence between $\boldsymbol{g}$ and $\boldsymbol{e}$.

GBLUP

We used a standard additive GBLUP model as a base model $\mathbf{M_0}$ with the following assumptions: $\boldsymbol{X} = \mathbf{1}_N$ was a vector of 1 of length N (with N the number of individuals), $\boldsymbol{\beta} = \mu$ the general mean of performances, $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{K}\sigma^2_G)$ and $\boldsymbol{e} \sim \mathcal{N}(0, \boldsymbol{I}\sigma^2_E)$ with σ^2_G and σ^2_E being the genetic and residual variances respectively. The kinship between individuals i and j, $K^0_{i,j}$, was estimated using VanRaden (2008):

$$K^0_{i,j} = \frac{\sum_{m=1}^{M}(W_{im} - f_m)(W_{jm} - f_m)}{\sum_{m=1}^{M} f_m(1 - f_m)} \tag{1.1}$$

where W_{im} is the genotype of individual i at locus m (coded 0 ; 0.5 ; 1) and f_m is the allele frequency of allele "1" at locus m, estimated on the whole dataset.

Structured GBLUP

The standard kinship estimation combines relatedness and genetic structure information. In order to test whether modeling the structure as a fixed effect could improve the predictions, we used two adapted GBLUP models.

Model $\mathbf{M_1}$ followed the same assumptions as $\mathbf{M_0}$ except that population structure was added as a fixed effect where $\boldsymbol{X}$ was the $(N \times Q)$ incidence matrix for fixed effects with $X_{iq} = 1$ if i was assigned to the q^{th} genetic group (otherwise $X_{iq} = 0$), $Q = 3$ is the number of genetic groups and $\boldsymbol{\beta} = (\mu_A, \mu_B, \mu_C)^T$ is a vector of fixed group effects. For $\mathbf{M_1}$, the kinship was estimated following Plieschke et al. (2015), by centering the genotypes using group-specific allele frequencies to remove the structure from the kinship and to avoid a redundancy of information in the model:

$$K^1_{i,j} = \frac{\sum_{m=1}^{M}(W_{im} - p_{im})(W_{jm} - p_{jm})}{\sum_{m=1}^{M} f_m(1 - f_m)} \tag{1.2}$$

where $p_{im} = \sum_{q=1}^{Q} X_{iq} f_{mq}$, f_{mq} is the allele frequency of group q as provided by ADMIXTURE.

Model $\mathbf{M_2}$ followed the same assumptions as $\mathbf{M_1}$ except that it considered quantitative assignments of individuals to groups in the prediction model and in the kinship. Thus X_{iq} became the admixture coefficient of individual i for group q. The kinship $K^2_{i,j}$ used in $\mathbf{M_2}$ was estimated using the same expression as for $K^1_{i,j}$ in Eq. (1.2) but with p_{im} being a weighted mean of ancestral group-specific allele frequencies with weights corresponding to admixture coefficients (Thornton et al., 2012).

Multigroup GBLUP

We also used a multivariate model $\mathbf{M_3}$ considering categorical assignments to genetic groups, which is an adaptation of a multi-trait model to the analysis of one trait in different groups proposed by Lehermeier et al. (2015).

In this model, $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are the same as in $\mathbf{M_1}$, $\boldsymbol{Z}$ is an incidence matrix linking phenotypes to the corresponding group-specific breeding value (categorical assignments), $\boldsymbol{g} = \begin{bmatrix} \boldsymbol{g}_A^* \\ \boldsymbol{g}_B^* \\ \boldsymbol{g}_C^* \end{bmatrix}$ is the expanded vector of breeding values of each individual in each group with a size of $3N$ and $\boldsymbol{e} = \begin{bmatrix} \boldsymbol{e}_A \\ \boldsymbol{e}_B \\ \boldsymbol{e}_C \end{bmatrix}$ is the vector of errors of size N where:

- $\begin{bmatrix} \boldsymbol{g}_A^* \\ \boldsymbol{g}_B^* \\ \boldsymbol{g}_C^* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma^2_{G_A} & \sigma_{G_{A,B}} & \sigma_{G_{A,C}} \\ \sigma_{G_{A,B}} & \sigma^2_{G_B} & \sigma_{G_{B,C}} \\ \sigma_{G_{A,C}} & \sigma_{G_{B,C}} & \sigma^2_{G_C} \end{bmatrix} \otimes \boldsymbol{K^0}\right)$
- $\begin{bmatrix} \boldsymbol{e}_A \\ \boldsymbol{e}_B \\ \boldsymbol{e}_C \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \boldsymbol{I}_A\sigma^2_{E_A} & 0 & \\ 0 & \boldsymbol{I}_B\sigma^2_{E_B} & 0 \\ 0 & 0 & \boldsymbol{I}_C\sigma^2_{E_C} \end{bmatrix}\right)$

with $\sigma_{G_{X,Y}}$ being the genetic covariance between groups X and Y (the letters X, Y and Z were further used as group names when not specifically designating group A, B or C). In this model, the kinship between individuals i and j, $K^{0'}_{i,j}$ (Eq. 1.3), was computed following Astle and Balding (2009) as recommended by Lehermeier et al. (2015), although results were very consistent using the kinship defined by VanRaden (2008) (Eq. 1.1).

$$K^{0'}_{i,j} = \frac{1}{M}\sum_{m=1}^{M}\frac{(W_{im} - f_m)(W_{jm} - f_m)}{f_m(1 - f_m)} \tag{1.3}$$

Note that the genetic covariance between groups results from the genetic covariance between allele effects in each group as described in Karoui et al. (2012) and Lehermeier et al. (2015).

We also defined $r_{X,Y} = \dfrac{\sigma_{G_{X,Y}}}{\sigma_{G_X}\sigma_{G_Y}}$ where $r_{X,Y}$ is the genetic correlation between groups X and Y.

For each model, the Genomic Estimated Breeding Values (GEBV) of the VS were computed as: $\widehat{\mathbf{y}}_{VS} = \mathbf{X}_{VS}\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{g}}_{VS}$

Model parameters were estimated using ASReml-R (Butler et al., 2009) for models $\mathbf{M_0}$, $\mathbf{M_1}$ and $\mathbf{M_2}$, using restricted maximum likelihood method. For the last model $\mathbf{M_3}$, a Gibbs sampler implemented in R was used to estimate the parameters [1]. The choice of hyper-parameters was the same as described in Lehermeier et al. (2015). A total of 300,000 MCMC samples were collected with 100,000 discarded as burn-in and thinning was done by keeping one every two samples. The parameter estimates were obtained by computing posterior means.

Evaluation of the precision of genomic predictions

The precision of the models was evaluated using four different CV procedures either neglecting genetic structure or aiming at evaluating its impact on the precision of genomic predictions.

[1] available at https://github.com/QuantGen/MTM

Table 1.1: Scenarios evaluated with the structure-based cross-validations (SHO) where 18 individuals are predicted by 66 other individuals (x100 samples)

Scenario	**TS composition**	**VS composition**
ABC_ABC	$\frac{1}{3}$A + $\frac{1}{3}$B + $\frac{1}{3}$C	$\frac{1}{3}$A + $\frac{1}{3}$B + $\frac{1}{3}$C
ABC_A	$\frac{1}{3}$A + $\frac{1}{3}$B + $\frac{1}{3}$C	A
A_A	A	A
B_A	B	A
C_A	C	A
BC_A	$\frac{1}{2}$B + $\frac{1}{2}$C	A
ABC_B	$\frac{1}{3}$A + $\frac{1}{3}$B + $\frac{1}{3}$C	B
B_B	B	B
A_B	A	B
C_B	C	B
AC_B	$\frac{1}{2}$A + $\frac{1}{2}$C	B
ABC_C	$\frac{1}{3}$A + $\frac{1}{3}$B + $\frac{1}{3}$C	C
C_C	C	C
A_C	A	C
B_C	B	C
AB_C	$\frac{1}{2}$A + $\frac{1}{2}$B	C

The first CV procedure was an averaged Holdout (HO) method and allowed us to study the level of precision that can be obtained when neglecting the role of genetic structure. The initial dataset was split with proportions $\frac{4}{5}$ and $\frac{1}{5}$ for the TS and the VS, respectively. The splitting was repeated 100 times and the precision criteria were averaged over repetitions.

The second CV procedure was a Leave-One-Out method (LOO) where every individual was predicted with a model calibrated using all the remaining individuals. It allowed a simple graphic representation of the quality of prediction of each individual using all the other individuals from the panel, whatever their group of origin. We also used this approach to evaluate the link between the CD (see below) and the prediction of each individual.

The third CV procedure, named Structured Holdout (SHO), allowed us to study the impact of genetic structure in genomic prediction accuracy using different scenarios. It considered samples of restricted sizes where 18 individuals are predicted using the model calibrated with 66 other individuals, repeating sampling 100 times. Those numbers were chosen in order to fit with all the scenarios (Table 1.1), knowing that group C is limited to 84 individuals. The individuals were assigned to the three groups according to their maximal admixture coefficient. All the scenarios are designated as TS_VS, TS and VS referring to the groups represented in the TS and VS respectively. When there were more than one group in the TS or the VS, the composition was always perfectly balanced between groups. As an example, ABC_A referred to a TS equally composed of individuals from the three groups and a VS composed of lines from group A only. Note that the across-group SHO scenarios (i.e. when no individual from the group forming the VS are present in the TS) cannot be evaluated using model $\mathbf{M_1}$ and $\mathbf{M_3}$. They require a TS where all the genetic groups are represented in order to estimate all the group-specific intercept and variance parameters.

The fourth CV procedure, referred to as Structured Holdout + (SHO$^+$), aimed at evaluating the benefits of extra-group individuals (individuals from a group absent of the VS) to improve genomic prediction accuracy. It considered the same samples as in SHO for intra group predictions (scenarios X_X) complemented with

66 individuals of each of the two other groups, reaching a size of 198 individuals. For instance, ABC$^+$_A referred to the SHO$^+$ scenario to predict group A.

Three criteria of precision were used to compare models and scenarios. The first was the predictive ability, defined as the correlation between GEBV and the phenotypes. The second was the accuracy which was computed by dividing the predictive ability by the square root of the heritability. Here, the heritability was computed using the estimated variances obtained by applying $\mathbf{M_0}$ to the whole panel. The third criterion was the Root Mean Square error of Prediction (RMSP), defined as the root mean square of the differences between LS-means and the GEBV.

A priori estimation of accuracy

In mixed models, the accuracy related to the prediction of individual i can be quantified by its associated Coefficient of Determination (CD), using the general formula:

$$CD_i = \text{Cor}(\boldsymbol{g}_i, \widehat{\boldsymbol{g}}_i)^2 = \frac{\boldsymbol{G}_{i,TS}\boldsymbol{\Sigma}^{-1}_{TS,TS}\boldsymbol{G}_{TS,i}}{\boldsymbol{G}_{i,i}} \tag{1.4}$$

where $\boldsymbol{g}_i$ and $\widehat{\boldsymbol{g}}_i$ are the breeding value of individual i and its corresponding BLUP respectively, $\boldsymbol{G}_{i,TS}$ is the covariance matrix between breeding values of i and the TS, $\boldsymbol{G}_{i,i}$ is the genetic variance of i and $\boldsymbol{\Sigma}_{TS,TS}$ is the covariance matrix between phenotypic values within the TS.

The standard CD in a GBLUP model assumes an unstructured population, as described in model $\mathbf{M_0}$ and is computed using Eq. (1.4), where:

- $\boldsymbol{G}_{i,TS} = \boldsymbol{K}^{\mathbf{0}}_{i,TS}\widehat{\sigma}^2_G$
- $\boldsymbol{G}_{i,i} = \boldsymbol{K}^{\mathbf{0}}_{i,i}\widehat{\sigma}^2_G$
- $\boldsymbol{\Sigma}_{TS,TS} = \boldsymbol{K}^{\mathbf{0}}_{TS,TS}\widehat{\sigma}^2_G + \boldsymbol{I}\widehat{\sigma}^2_E$
- $\widehat{\sigma}^2_G$ and $\widehat{\sigma}^2_E$ are the genetic and residual variances respectively, estimated using $\mathbf{M_0}$ calibrated with all the individuals.

We also considered the multigroup CD as proposed by Wientjes et al. (2015a) and derived from $\mathbf{M_3}$. When considering scenario XY_Z, the elements of Eq. (1.4) become:

- $\boldsymbol{G}_{i,TS} = \left[\boldsymbol{K}_{Z,X}\widehat{\sigma}_{G_{Z,X}} \quad \boldsymbol{K}_{Z,Y}\widehat{\sigma}_{G_{Z,Y}}\right]_{i,TS}$
- $\boldsymbol{G}_{i,i} = \left[\boldsymbol{K}_{Z,Z}\widehat{\sigma}^2_{G_Z}\right]_{i,i}$
- $\boldsymbol{\Sigma}_{TS,TS} = \begin{bmatrix} \boldsymbol{K}_{X,X}\widehat{\sigma}^2_{G_X} + \boldsymbol{I}\widehat{\sigma}^2_{E_X} & \boldsymbol{K}_{X,Y}\widehat{\sigma}_{G_{X,Y}} \\ \boldsymbol{K}_{Y,X}\widehat{\sigma}_{G_{X,Y}} & \boldsymbol{K}_{Y,Y}\widehat{\sigma}^2_{G_Y} + \boldsymbol{I}\widehat{\sigma}^2_{E_Y} \end{bmatrix}_{TS,TS}$
- $\widehat{\sigma}_{G_{Z,X}}$, $\widehat{\sigma}^2_{G_X}$ and $\widehat{\sigma}^2_{E_X}$ are the genetic covariance between group Z and X, the genetic variance in group X and the residual variances in group X respectively estimated on the whole dataset.

Two versions of this multigroup CD were computed using different kinships $\boldsymbol{K}$. The first version called $CDgp^1$ used $\boldsymbol{K}^0$ defined in Eq. (1.1) while the second called $CDgp^2$ used a new estimator recommended by

Wientjes et al. (2017):

$$K^3_{i,j} = \frac{\sum_{m=1}^{M}(W_{im} - p_{im})(W_{jm} - p_{jm})}{\sqrt{\sum_{m=1}^{M}(p_{im}(1-p_{im}))}\sqrt{\sum_{m=1}^{M}(p_{jm}(1-p_{jm}))}} \quad (1.5)$$

with $p_{im} = \sum_{q=1}^{Q} X_{iq} f_{mq}$ where X_{iq} considered categorical assignment of individuals to groups as defined in $\mathbf{M_1}$.

$CDgp^1$ is computed using variances estimated with $\mathbf{M_3}$ and $\boldsymbol{K^{0'}}$ (computed using Eq. (1.3) as recommended by Lehermeier et al. (2015)) on the whole dataset, although parameters estimates were very consistent using $\boldsymbol{K^0}$ (Eq. 1.1). $CDgp^2$ is computed using variances estimated with $\mathbf{M_3}$ and $\boldsymbol{K^3}$ on the whole dataset as summarized in Table 1.2.

After computing the CD values of every individual of the VS, we averaged them and computed the square root to obtain an *a priori* indicator of accuracy.

The *a priori* estimates of accuracy were compared to empirical accuracies, obtained from model $\mathbf{M_0}$ with the SHO method, for the different scenarios described above using two criteria of precision. The first criterion was the correlation between the *a priori* estimates of accuracy and the empirical accuracies. The second criterion was the Root Mean Square error of Estimation (RMSE) which is defined as the root mean square of the differences between *a priori* and empirical accuracies.

We also computed standard CD for each predicted individual in the context of LOO CV.

Table 1.2: Summary table of empirical and *a priori* accuracies using different CDs, describing the statistical model, the kinship matrix and the variance estimates used to compute them. The statistical model and the kinship used to estimate variances are shown between braces

Accuracy	**Model**	**Kinship**	**Variances**[a]
Empirical	$\mathbf{M_0}$	$\boldsymbol{K^0}$	CV est. $\{\mathbf{M_0}$,$\boldsymbol{K^0}\}$
CD	$\mathbf{M_0}$	$\boldsymbol{K^0}$	Whole data est. $\{\mathbf{M_0}$,$\boldsymbol{K^0}\}$
$CDgp^1$	$\mathbf{M_3}$	$\boldsymbol{K^0}$	Whole data est. $\{\mathbf{M_3}$,$\boldsymbol{K^{0'}}\}$
$CDgp^2$	$\mathbf{M_3}$	$\boldsymbol{K^3}$	Whole data est. $\{\mathbf{M_3}$,$\boldsymbol{K^3}\}$

[a] Variances were estimated for each cross-validation training set (CV est.) or through a single estimation using the whole dataset (Whole data est.).

Results

Global, within and across group precision of genomic predictions

We first estimated variances for the four traits by applying model $\mathbf{M_0}$. The estimated heritabilities were very high (Table 1.3), between 0.86 and 0.95 and consistent with the high heritabilities computed in each trial without considering kinship (Supplementary Table S1).

The estimates of accuracy obtained with the HO method and $\mathbf{M_0}$ model ranged from 0.77 for Yield Index to 0.84 for Grain Yield (Table 1.4). The accuracy estimates were close to the predictive abilities as a consequence of high heritabilities. We also studied the ability of all the lines from public origin to predict all the private lines. The accuracies obtained were 0.64, 0.49, 0.33 and 0.76 for Grain Moisture, Grain Yield, Yield Index and Male Flowering respectively.

Table 1.3: Mean, genetic variance, environmental variance and heritability estimated on all the data using $\mathbf{M_0}$. Standard errors for variances estimates are shown between brackets

	μ	σ^2_G	σ^2_E	h^2
Grain Moisture	27.50	2.52 (0.26)	0.14 (0.09)	0.95
Grain Yield	84.51	50.27 (5.75)	6.87 (2.34)	0.88
Yield Index	15.80	57.77 (6.77)	9.46 (2.94)	0.86
Male Flowering	891.13	518.09 (50.11)	26.64 (15.43)	0.95

Table 1.4: Average of precision criteria evaluated with the standard cross-validations (HO) using $\mathbf{M_0}$. Standard deviations are shown between brackets

	Predictive ability	**Accuracy**	**RMSP**
Grain Moisture	0.76 (0.02)	0.78 (0.03)	1.33 (0.06)
Grain Yield	0.78 (0.02)	0.84 (0.03)	6.20 (0.38)
Yield Index	0.73 (0.03)	0.79 (0.03)	6.71 (0.43)
Male Flowering	0.75 (0.02)	0.77 (0.02)	19.45 (0.82)

In Fig. 1.2, where predictions were obtained with LOO, we observed a differentiation between groups for the four traits. For instance, groups B and C were almost perfectly separated along the two axes for Male Flowering. These mean differences may be partly responsible for the high level of correlation obtained in Table 1.4.

In order to study more carefully the impact of genetic structure on genomic predictions, we performed a third type of CV named SHO using $\mathbf{M_0}$ with different scenarios defined in Table 1.1. Accuracies (Table 1.5) and predictive abilities (Supplementary Table S1.2) showed similar trends. Scenario ABC_ABC displayed in general a higher accuracy than scenarios ABC_X but its RMSP was generally not the lowest (Table 1.6).

Considering group-specific VS, the best predictions, both in terms of accuracy and RMSP, were achieved when predicting one group with individuals from the same group. The only exception was Grain Moisture and Male Flowering in group C for which the lowest RMSP and the highest accuracies were obtained by using a TS consisting of individuals from the three groups (scenario ABC_C). The worst predictions were always achieved when trying to predict one group using only one other group (scenarios X_Y) while using the two other groups allowed intermediate accuracies and RMSP (scenarios XY_Z). Except for Yield Index, trying to predict group C using scenario AB_C was among the best options, conversely to what is observed for the symmetric scenarios in the other group for which BC_A and AC_B were outperformed by A_A and B_B respectively.

In general, group-specific accuracies tended to be higher in group A than in group B and C, regardless of the trait considered. The opposite was generally observed when considering RMSP, for which group A presented a higher prediction error, meaning a lower precision.

In order to study the impact of adding extra-group individuals on the accuracy of genomic predictions, we performed SHO$^+$ CV (Table 1.7). Adding individuals always increased accuracy and decreased RMSP except in group C for Yield Index. Generally, the gain in precision was greater in group C than in group B, itself greater than in group A.

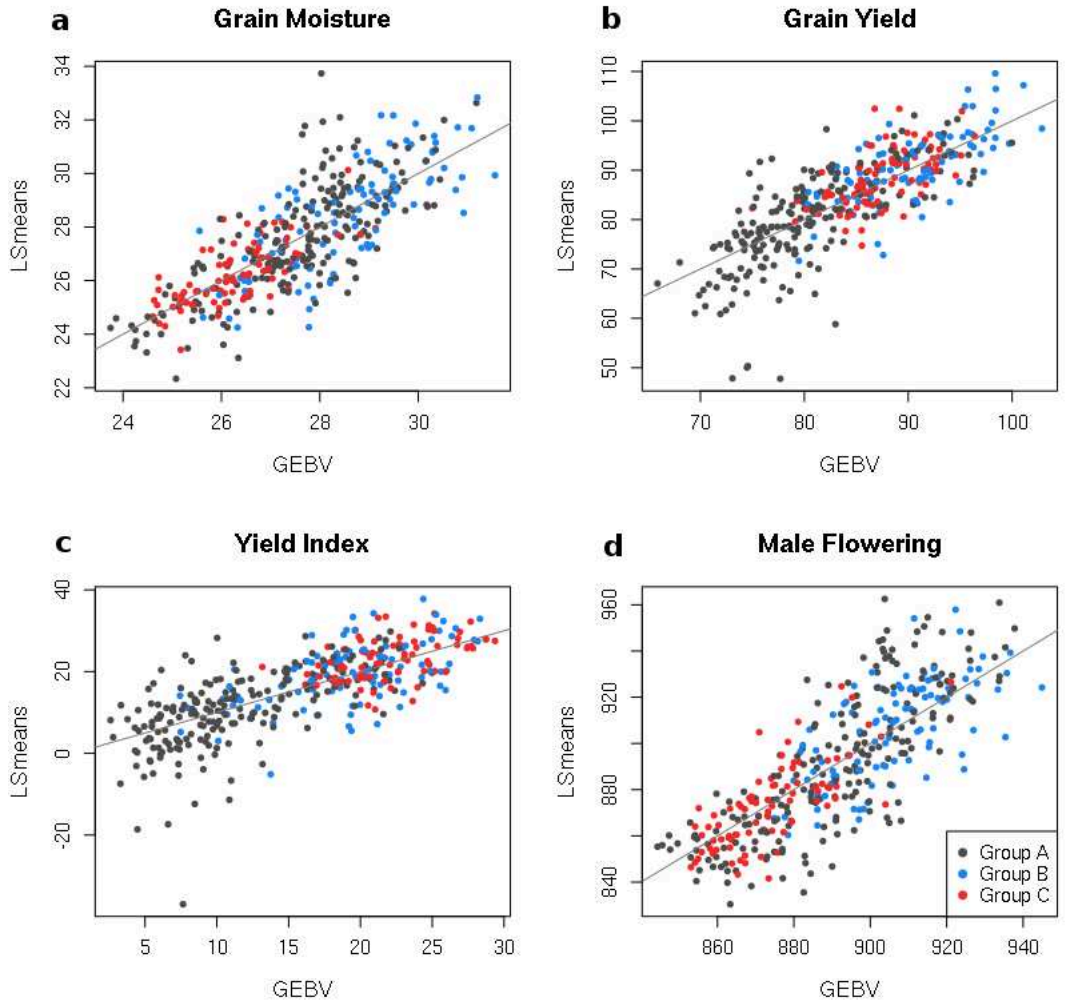


Figure 1.2: LS-means values plotted against GEBV obtained by LOO cross-validations using $\mathbf{M_0}$ and coloration of individuals dots using assignment to groups for **a** grain Moisture, **b** Grain Yield, **c** Yield Index and **d** Male Flowering

Accounting for structure in genomic prediction models

We tested three other models, taking into account genetic structure, to compare them to model $\mathbf{M_0}$ on scenario ABC_ABC and on scenarios ABC_X (SHO CV). In general, the four models tended to reach similar performances when considering accuracy as a criterion (Fig. 1.3). Model $\mathbf{M_3}$ reached performances below the other models in all the scenarios for Male Flowering and for some scenarios in the other traits such as scenario ABC_C for Grain Moisture. However it allowed better accuracies in scenario ABC_ABC and ABC_C for Yield Index. The same conclusions could be made considering predictive ability or RMSP as criteria (Supplementary Fig. S1.4 and S1.5). The across-group scenarios were also tested to compare $\mathbf{M_0}$ and $\mathbf{M_2}$ (model considering quantitative assignment to groups) showing no improvement when using the latter (Supplementary Fig. S1.6 , S1.7 and S1.8).

A priori estimation of precision

To compute the CD, variances were estimated within the whole population using $\mathbf{M_0}$ (Table 1.3). To compute $CDgp^1$ and $CDgp^2$, variances and genetic correlations were estimated between groups using $\mathbf{M_3}$ with $\boldsymbol{K^0}$ (Table 1.8) and $\boldsymbol{K^3}$ respectively (Supplementary Table S1.3). For all traits, the genetic variance estimates

Table 1.5: Average of accuracies obtained with the structure-based cross-validations (SHO) using $\mathbf{M_0}$. Standard deviations are shown between brackets

Scenario	Grain Moisture	Grain Yield	Yield Index	Male Flowering
ABC_ABC	0.74 (0.13)	0.74 (0.13)	0.70 (0.16)	0.70 (0.12)
ABC_A	0.63 (0.17)	0.64 (0.18)	0.55 (0.20)	0.62 (0.15)
A_A	0.70 (0.12)	0.70 (0.14)	0.62 (0.16)	0.70 (0.12)
BC_A	0.56 (0.18)	0.50 (0.19)	0.25 (0.25)	0.55 (0.16)
B_A	0.50 (0.21)	0.54 (0.18)	0.30 (0.21)	0.54 (0.16)
C_A	0.59 (0.16)	0.48 (0.18)	0.15 (0.24)	0.52 (0.16)
ABC_B	0.68 (0.13)	0.65 (0.15)	0.45 (0.19)	0.51 (0.15)
B_B	0.71 (0.11)	0.69 (0.12)	0.51 (0.16)	0.57 (0.14)
AC_B	0.59 (0.17)	0.53 (0.19)	0.43 (0.19)	0.51 (0.15)
A_B	0.50 (0.18)	0.51 (0.18)	0.29 (0.22)	0.46 (0.18)
C_B	0.62 (0.13)	0.57 (0.16)	0.33 (0.21)	0.46 (0.16)
ABC_C	0.63 (0.14)	0.64 (0.15)	0.41 (0.20)	0.61 (0.12)
C_C	0.53 (0.13)	0.63 (0.14)	0.55 (0.14)	0.55 (0.15)
AB_C	0.63 (0.15)	0.61 (0.13)	0.24 (0.24)	0.62 (0.15)
A_C	0.57 (0.16)	0.53 (0.17)	0.06 (0.28)	0.58 (0.14)
B_C	0.57 (0.17)	0.46 (0.18)	0.23 (0.20)	0.41 (0.20)

were lower in group C than in the other groups. The genetic correlations between groups were very high, around 0.90, except for Grain Moisture were they ranged from 0.72 to 0.76 using $\boldsymbol{K^0}$ and from 0.62 to 0.72 using $\boldsymbol{K^3}$. The group-specific heritabilities obtained from these estimates were also high (results not shown).

Before studying the ability of CD values to reflect the accuracy in the VS, we observed how CD values were connected to the prediction of individual performances obtained with LOO CV (Fig. 1.4). For Grain Moisture and Male Flowering (Fig. 1.4a and d), the individuals featuring low CD values were predicted close to the mean and were more likely to have important observed errors of prediction. Conversely, individuals featuring high CD values had a broader range of predicted values and the predictions were more accurate. A different situation was observed for Grain Yield and Yield Index which are submitted to directional selection (Fig. 1.4b and c). For these traits, individuals with low CD values were predicted to have low performances. Conversely, those with high CD values were predicted to have high performances.

The *a priori* accuracy of the different SHO scenarios were estimated by computing the square root of the average of the CDs over the individuals of the VS. The *a priori* estimates of accuracy were compared to the empirical accuracies using the correlation between *a priori* estimates and empirical accuracies and the RMSE between these two accuracies.

When first looking at the different plots between *a priori* and empirical accuracies (Supplementary Fig. S1.9, S1.10, S1.11 and S1.12), one could notice that there was a high variability of the empirical accuracies for a defined scenario (see also Table 1.5). All the indicators (CD, $CDgp^1$ and $CDgp^2$) led to either positive or null values of correlation between the *a priori* estimates of accuracy and the empirical accuracies (Table 1.9). There were different abilities to forecast accuracy depending on the trait and the group considered. It was harder to predict the level of accuracy in group C than in groups A and B for all the traits except Yield Index. For instance, the correlation was almost null regardless of the indicator used for Grain Moisture in

Table 1.6: Average of RMSP obtained with the the structure-based cross-validations (SHO) using $\mathbf{M_0}$. Standard deviations are shown between brackets

Scenario	Grain Moisture	Grain Yield	Yield Index	Male Flowering
ABC_ABC	1.41 (0.27)	6.58 (1.43)	7.06 (1.60)	20.21 (3.54)
ABC_A	1.63 (0.29)	8.70 (1.99)	8.79 (2.35)	26.00 (3.34)
A_A	1.56 (0.35)	7.58 (1.78)	7.71 (2.22)	24.58 (3.61)
BC_A	1.77 (0.30)	11.72 (2.32)	12.27 (2.44)	27.59 (3.90)
B_A	1.84 (0.27)	11.44 (2.23)	10.99 (2.27)	29.33 (4.10)
C_A	1.93 (0.34)	14.08 (2.23)	15.83 (2.21)	28.08 (4.13)
ABC_B	1.46 (0.24)	5.83 (0.91)	6.88 (1.06)	18.99 (2.69)
B_B	1.46 (0.20)	5.75 (0.92)	6.91 (1.17)	18.73 (2.46)
AC_B	2.05 (0.35)	7.50 (1.47)	7.31 (1.07)	22.61 (3.71)
A_B	2.03 (0.30)	9.79 (1.67)	9.10 (1.40)	21.16 (3.02)
C_B	2.33 (0.33)	7.42 (1.30)	9.73 (1.74)	25.04 (4.04)
ABC_C	0.95 (0.16)	4.85 (0.99)	5.16 (0.90)	15.16 (2.38)
C_C	0.97 (0.15)	4.81 (0.84)	4.63 (0.64)	15.89 (3.23)
AB_C	1.26 (0.28)	5.49 (1.13)	7.27 (1.43)	23.89 (3.99)
A_C	1.29 (0.26)	8.31 (1.74)	10.45 (1.64)	24.45 (4.27)
B_C	1.39 (0.29)	5.59 (0.85)	5.44 (0.84)	28.98 (3.55)

this group. In contrast, the ability to forecast the level of precision was up to 0.56 in group C for Yield Index. When comparing the three indicators using the correlation between empirical and *a priori* accuracies, it was difficult to assess which one performed best. The differences were very low between CD, $CDgp^1$ and $CDgp^2$ which might be explained by the high genetic correlation between groups (except for Grain Moisture), as well as a limited impact of the kinship used to compute the CDs. Along with the correlation, using the RMSE between *a priori* and empirical accuracies did not allow us to identify a better indicator of accuracy (Supplementary Table S.4).

Discussion

The impact of genetic structure on genomic prediction accuracy

We investigated genomic prediction accuracy using LS-means corrected for trial effects as observed phenotypes to minimize environmental effects. As a consequence, the estimates of heritabilities obtained when fitting additive model $\mathbf{M_0}$ were very high and consistent with the high heritabilities obtained for each trial without considering kinship. Along with the high heritabilites, the predictive abilities and the accuracies were high when neglecting population structure, revealing both the relevance of model $\mathbf{M_0}$ to make predictions and the quality of the data.

In this dataset, structure participated to genomic prediction accuracy, as the accuracy was generally higher in scenario ABC_ABC than in scenarios ABC_X for a given TS size. The standard kinship matrix contains information about the structure of the population in genetic groups. When there is a difference of mean between groups and the same structure is found in the TS and the VS, this difference is well taken into

Table 1.7: Average of accuracies and RMSP obtained with the within-group cross-validations (SHO) compared to those obtained with the cross-validations adding extra-group individuals to the TS (SHO$^+$). Standard deviations are shown between brackets

	Grain Moisture		**Grain Yield**		**Yield Index**		**Male Flowering**	
	Accuracy	**RMSP**	**Accuracy**	**RMSP**	**Accuracy**	**RMSP**	**Accuracy**	**RMSP**
A_A	0.70 (0.12)	1.56 (0.35)	0.70 (0.14)	7.58 (1.78)	0.62 (0.16)	7.71 (2.22)	0.70 (0.12)	24.58 (3.61)
ABC$^+$_A	0.72 (0.13)	1.48 (0.33)	0.72 (0.14)	7.37 (2.00)	0.64 (0.17)	7.65 (2.25)	0.73 (0.11)	23.19 (3.65)
B_B	0.71 (0.11)	1.46 (0.20)	0.69 (0.12)	5.75 (0.92)	0.51 (0.16)	6.91 (1.17)	0.57 (0.14)	18.73 (2.46)
ABC$^+$_B	0.76 (0.10)	1.33 (0.19)	0.72 (0.11)	5.45 (0.79)	0.56 (0.15)	6.59 (0.95)	0.67 (0.11)	16.71 (2.43)
C_C	0.53 (0.13)	0.97 (0.15)	0.63 (0.14)	4.81 (0.84)	0.55 (0.14)	4.63 (0.64)	0.55 (0.15)	5.89 (3.23)
ABC$^+$_C	0.71 (0.11)	0.83 (0.13)	0.69 (0.12)	4.66 (0.87)	0.52 (0.15)	4.82 (0.71)	0.70 (0.13)	13.49 (2.62)

Table 1.8: Posterior mean of group-specific genetic variances, genetic correlations and environmental variances estimated using $\mathbf{M_3}$ and $\boldsymbol{K^0}$ on all the data. Posterior standard deviations, obtained on Gibbs samples, are shown between brackets

	Grain Moisture	**Grain Yield**	**Yield Index**	**Male Flowering**
$\sigma^2_{G_A}$	2.11 (0.24)	52.78 (6.03)	56.55 (6.32)	508.88 (52.79)
$\sigma^2_{G_B}$	3.02 (0.57)	44.81 (9.73)	54.44 (17.24)	490.26 (87.32)
$\sigma^2_{G_C}$	1.36 (0.28)	42.81 (9.37)	30.84 (10.25)	453.57 (69.18)
r_{AB}	0.76 (0.10)	0.96 (0.05)	0.97 (0.05)	0.99 (0.01)
r_{AC}	0.73 (0.10)	0.95 (0.07)	0.93 (0.09)	0.99 (0.01)
r_{BC}	0.73 (0.11)	0.96 (0.04)	0.95 (0.08)	0.99 (0.00)
$\sigma^2_{E_A}$	0.34 (0.10)	5.02 (2.19)	4.34 (1.93)	45.42 (17.43)
$\sigma^2_{E_B}$	0.52 (0.20)	6.82 (3.49)	14.57 (7.23)	36.61 (24.52)
$\sigma^2_{E_C}$	0.31 (0.09)	3.94 (2.47)	7.91 (3.76)	9.37 (6.38)

account by the GBLUP model and participates to the accuracy (Guo et al., 2014). At the extreme, one could imagine a trait for which there would be a global positive correlation between predicted and true breeding values but with a null accuracy within each group composing the VS. The RMSP criterion is not impacted by genetic structure like the accuracy, as RMSP is not lower in scenario ABC_ABC than in scenarios ABC_X. RSMP is thus complementary to the accuracy to evaluate the precision of the predictions.

The main interest of a plant breeder is to know the level of precision that can be reached within each genetic group, as selection will be often applied on individuals derived from crosses between related elite lines from a same group. In this dataset, to predict a group-specific VS for a given size of TS, it was generally better to use a TS from the same group (scenarios X_X), as previously shown in soybean (Duhnen et al., 2017). Depending on the trait, accuracy could be severely impacted by not representing relatives of the VS in the TS (scenarios X_Y and XY_Z). This suggests an inconsistency of allele effects between groups, or different LD extent between SNPs and QTLs. However these hypotheses are not supported by the high genetic correlations estimates between groups for all the traits.

Group C showed interesting results as it was best predicted with a diverse TS, except for Yield Index. Simultaneously, the accuracy of scenario AB_C was as high or higher than the accuracy achieved in scenario C_C. We can hypothesize that the allele effects are conserved between groups and that there are none or few QTLs specifically polymorphic in group C for these three traits. This hypothesis is supported by SNP data as group C features less specifically polymorphic SNP and a lower genome-wide genetic diversity than

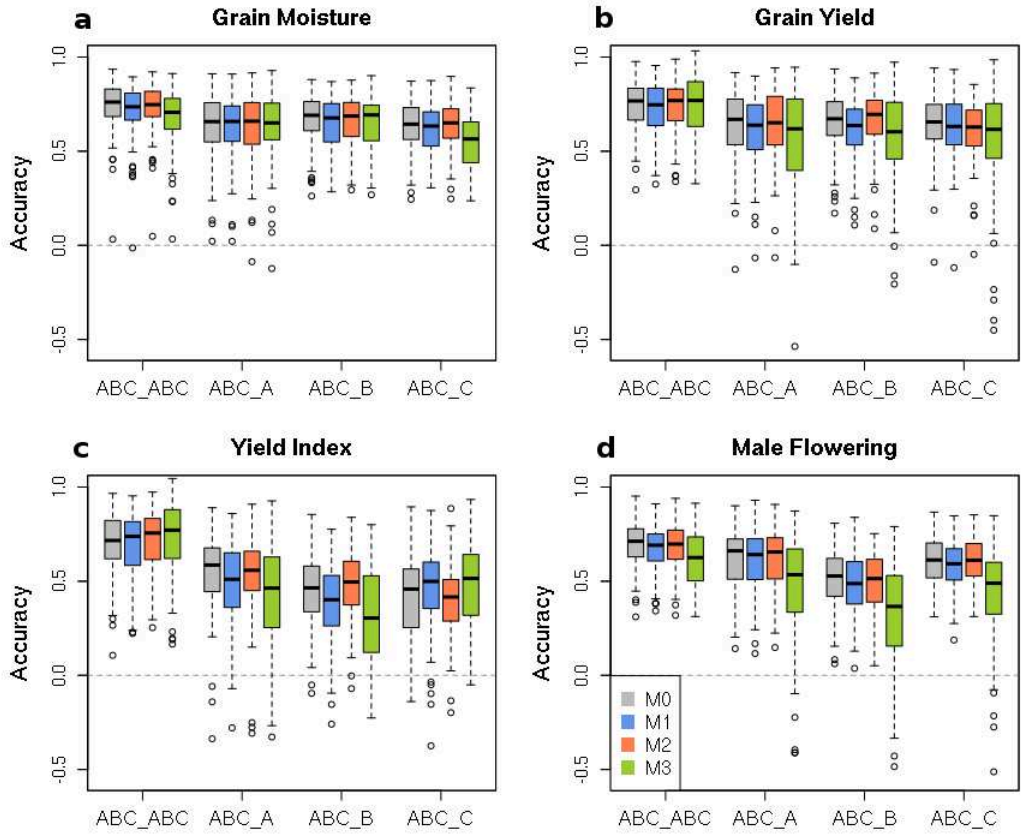


Figure 1.3: Box-plots of accuracies obtained with the structure-based cross-validations (SHO) for scenarios ABC_ABC and ABC_X using different models of prediction for **a**. grain Moisture, **b**. Grain Yield, **c**. Yield Index and **d**. Male Flowering

the other two groups (results not shown). As this group is the less diverse, with a high degree of relatedness between individuals, it might be beneficial to calibrate a model on a diverse set of individuals. Indeed, from a statistical point of view, the precision of the estimation of the effects of each locus or each breeding value gets worse as the diversity decreases within the TS. The situation is different for groups A and B which possibly presented specifically polymorphic QTLs. Thus, when trying to predict group A or B by the two other groups (using scenarios B_A, C_A , BC_A, A_B, C_B and AC_B), the effects associated to these group-specific polymorphisms cannot be taken into account by the model. The particular situation of group C (Iodents) is supported by its recent history as it was initially derived from individuals from group A for its complementarity with group B (Stiff Stalks) 50 to 70 years ago. Yield Index presented a different behavior and possibly involves yield QTLs independent from precocity that are specifically polymorphic in each group. We also checked that this different behavior was not due to predicting Yield Index directly rather than computing it using the individual predictions of Grain Yield and Grain Moisture (results not shown).

Table 1.9: Correlations between *a priori* estimates of accuracy and empirical accuracies using $\mathbf{M_0}$ and structure-based cross-validations (SHO) for group-specific VS (e.g. A includes ABC_A, A_A, BC_A, B_A and C_A)

	Grain Moisture			**Grain Yield**			**Yield Index**			**Male Flowering**		
VS	A	B	C	A	B	C	A	B	C	**A**	B	C
CD	0.47	0.46	-0.03	0.36	0.38	0.30	0.46	0.31	0.55	0.44	0.25	0.13
$CDgp^1$	0.42	0.41	-0.05	0.37	0.38	0.30	0.52	0.31	0.55	0.43	0.25	0.13
$CDgp^2$	0.40	0.38	-0.07	0.39	0.41	0.28	0.56	0.31	0.56	0.44	0.24	-0.07

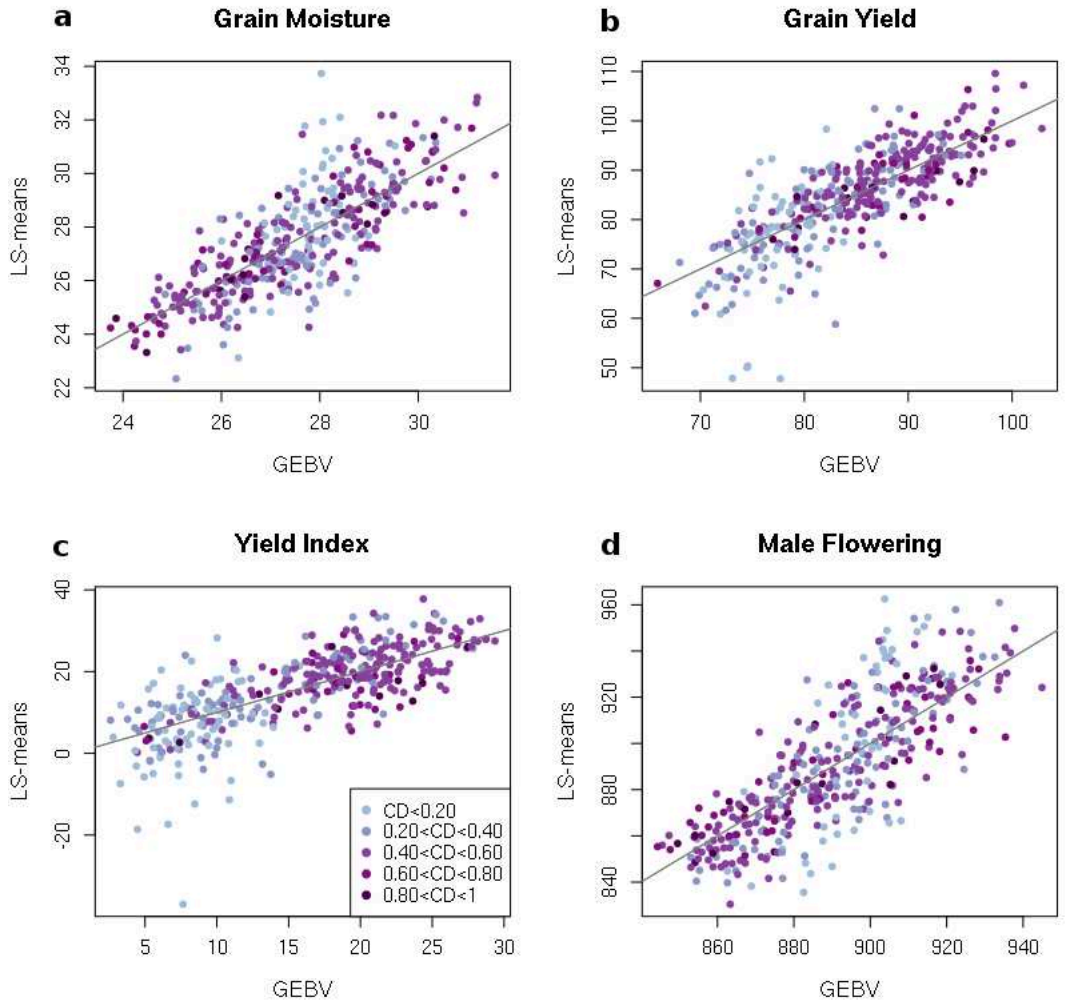


Figure 1.4: LS-means values plotted against GEBV obtained by LOO cross-validations using $\mathbf{M_0}$ and coloration of individuals dots using standard CD values for **a**. grain Moisture, **b**. Grain Yield, **c**. Yield Index and **d**. Male Flowering

One could notice that QTLs specifically polymorphic in one group is the extreme case of QTLs having group-specific allele diversities.

In a configuration, where allele effects seem conserved between groups (based on genetic correlations estimates in Table 1.8) which is consistent with a moderate and recent genetic structure, one should recommend the constitution of a TS as diverse as possible where all the genetic groups are represented. This is supported by the high level of accuracy reached for scenarios ABC_X and by the accuracy gains obtained by adding extra-group individuals to the TS in SHO$^+$ CV. Such diverse TS should be efficient to calibrate prediction models for a wide range of genetic material. This underlines its value as a generic TS for expensive traits evaluated on high-throughput phenotyping platforms or through extensive field trials (Millet et al., 2016). We showed that the accuracies, obtained when predicting the elite private lines by lines from public origin, were moderate or high depending on the trait. The accuracy was lower for traits submitted to directional selection such as Grain Yield or Yield Index for which the variability is limited among the elite lines. As elites lines were distributed in the three groups, we checked that accuracies were not entirely explained by the genetic structure (results not shown). As more data were imputed for elite lines, we also checked the impact of imputation by performing a subset of analyses (within-group or across-group SHO scenarios) on 50K data that were available for all the lines, and found very limited changes in terms of accuracy.

Modeling genetic structure to improve predictions

When performing genomic predictions within a structured population, one may wish to improve accuracy by using specific models taking into account this structure. There are different possibilities such as specifying structure as a fixed effect considering categorical or quantitative assignments of individuals to groups. It is also possible to model group-specific random effects, with group-specific variances and covariances between groups.

For this dataset, applying different models did not allow to improve the accuracy on the scenarios tested (ABC_ABC and ABC_X). For model $\mathbf{M_1}$ and $\mathbf{M_2}$, structure was removed from the kinship in Eq. (1.2) after being included as a fixed effect thanks to a genotypic centering using group-specific allele frequencies. One could expect that modeling groups as fixed effect would be advantageous if the differences between groups are larger than what can be attributed to differences in allele frequencies (Plieschke et al., 2015). Such benefits were not observed on our data. In $\mathbf{M_3}$ the assumption is not only that groups differ in terms of mean but that allele effects may be potentially correlated between groups, leading to group specific genetic variances and specific covariances between groups. This model did not improve genomic prediction accuracy and sometimes reached substantially lower accuracies than $\mathbf{M_0}$. When applying $\mathbf{M_3}$, there was a variability of genetic correlations estimates probably due to small TS sizes (results not shown) which might explain its poorer performances. One should also notice that the inference procedure differed between model $\mathbf{M_3}$ (bayesian inference) and the three other models (REML), which could possibly explain part of the differences in terms of performance.

It is important to note that for $\mathbf{M_1}$ and $\mathbf{M_3}$ which consider categorical assignments, all the parameters are not estimable for across group predictions. For instance, scenario AB_C requires to estimate μ_C for both models, and $\sigma^2_{G_C}$, $\sigma_{G_{AC}}$ and $\sigma_{G_{BC}}$ for model $\mathbf{M_3}$. These quantities cannot be estimated if group C is absent of the TS. In such a context, one should either neglect genetic structure or take advantage of the admixed individuals to connect groups. We tested model $\mathbf{M_2}$ which take into account such admixture as fixed effects showing no clear improvement in terms of accuracy compared to $\mathbf{M_0}$.

Is it possible to forecast accuracy using CD?

Being able to forecast the accuracy of predictions would allow many applications such as the optimization of TS or the anticipation of genetic gain in breeding programs. Many indicators were developed in order to get an *a priori* estimation of accuracy. Among them, the CD is well known (VanRaden, 2008) and can be easily derived as the square correlation between the breeding value of an individual and its corresponding BLUP in a standard GBLUP model. The estimate obtained is supposed to quantify the amount of information available from the TS to correctly predict the breeding values, with a value scaled between 0 and 1. In theory, when the CD value is low, the prediction is more likely to be inaccurate (high expected errors of prediction) and will be strongly shrunk toward the mean and conversely for a high CD value. This situation was observed for Grain Moisture and Male Flowering on the LOO plot. A different situation was observed for Grain Yield and Yield Index and could be interpreted as an effect of modern breeding. In this panel, the lines featuring a high CD were recent and related to several older ones featuring lower CDs. As those two traits are those of major interest in breeding and this panel included lines that have been under directional selection, it yielded this gradient of CD values along the prediction-axis. Individuals featuring high CD values were predicted to have high performances and conversely for low CD values. Male Flowering and Grain Moisture on opposite, were not submitted to directional selection. Thus individual CD value is an interesting criterion in breeding but one could question whether it could be informative about the accuracy of a set of individuals.

The CD value is linked to the accuracy as it represents its square value. For a defined TS and VS, one can easily compute each individual CD of the VS and compute the square root of the average CD to get an

a priori estimation of accuracy. Such CD-based accuracy usually succeeded in differentiating the structure-based scenarios in terms of accuracy. However, there were differences depending on the trait and the genetic group considered. For instance the null correlation between empirical and *a priori* accuracies encountered in group C for Grain Moisture was probably due to the overestimation of within group *a priori* accuracy (Supplementary Fig. S1.9 and scenario C_C).

The use of standard CD in the presence of a genetic structure was already criticized in the literature for TS optimization (Isidro et al., 2015) or to forecast accuracy (Hayes et al., 2009). One hypothesis to explain the poor performance of the CD in such context is the existence of different but correlated allele effects between groups. To tackle this problem, a new CD indicator was derived from a multigroup model (Wientjes et al., 2015a). The authors also recommended the use of an alternative Kinship estimator, using group-specific allele frequencies to standardize the genotypic data. In this study we tested two versions of this indicator, one using a standard Kinship: $CDgp^1$, and the other using the Kinship $\boldsymbol{K^3}$ in Eq. (1.5) recommended by Wientjes et al. (2017): $CDgp^2$.

Both indicators required estimates of genetic correlations between groups. These correlations were originally defined as resulting from the correlation between allele effects in the different groups (Karoui et al., 2012; Lehermeier et al., 2015). Each correlation may also be considered at the level of one individual where its breeding value in one group could be correlated to its potential breeding value in another group, assuming that these two groups showed correlated allele effects. The estimated correlations obtained in this study by applying $\mathbf{M_3}$ using $\boldsymbol{K^{0'}}$ for $CDgp^1$ or $\boldsymbol{K^3}$ for $CDgp^2$, were very high except for Grain Moisture. One could wonder how accurate they were as these values did not allow us to explain the differences in prediction accuracies observed between traits. In a recent study, Wientjes et al. (2017) investigated the impact of different genomic relationship matrices on the estimation on genetic correlations between groups. While they showed that using the genomic relationship matrix $\boldsymbol{K^0}$ of VanRaden (2008) or $\boldsymbol{K^3}$ of Wientjes et al. (2017) estimated genetic correlations between groups unbiasedly, they warned about the use of $\boldsymbol{K^{0'}}$ defined in Astle and Balding (2009) that we used in $\mathbf{M_3}$. However, we could check that the kinship of VanRaden (2008) and the one of Astle and Balding (2009) gave very close estimates (results not shown).

As a consequence of high genetic correlations, CD and $CDgp^1$ gave very similar results. One could notice that using genetic correlations of 1 and equal genetic and residual variances for each group would result in $CDgp^1$ and CD being perfectly equal. Using $\boldsymbol{K^3}$ and a set of parameters estimated using $\mathbf{M_3}$ with $\boldsymbol{K^3}$ to compute $CDgp^2$ also led to very similar results which indicated us that the kinship estimator did not have much of an impact to forecast accuracy in this dataset. Grain Moisture is the only trait featuring lower genetic correlations between groups but both multi-group CD did not outperform standard CD for this trait.

As $CDgp^1$ and $CDgp^2$ assume an $\mathbf{M_3}$ genetic model, one could wonder whether they would not be better correlated to empirical accuracies obtained with the $\mathbf{M_3}$ model instead of $\mathbf{M_0}$. As mentioned in the previous part of the discussion, $\mathbf{M_3}$ cannot be used as a predictive model for across group scenarios. However, we predicted breeding values and computed empirical accuracies obtained with $\mathbf{M_3}$, using parameters estimated on the whole dataset. Multi-group CDs did not better forecast these empirical accuracies (Supplementary Fig. S1.13 and S1.14).

Once again, these results supported the hypothesis that the allele effects were indeed highly correlated between groups and the impact of genetic structure would mostly be due to different group-specific allele diversity at QTLs. The CD indicators are based on macro-parameters such as the global genetic variance, but they do not take into account more detailed information like the number of QTLs and their localization along the genome. Simple simulations, using real genotypic data, of traits similar in terms of genetic variance and heritability with the ones measured in our study, showed an important variability of the impact of genetic structure on accuracy considering SHO results (Supplementary Fig. S1.15). The differences between traits were only due to allele effects sampling and they could not be captured by CD indicators, thus supporting this hypothesis. The impact of QTLs specifically polymorphic in DH bi-parental families on genomic prediction

accuracy was recently shown by Schopp et al. (2017) when performing across family predictions. The authors also discussed the impact of such QTLs on CD-based estimations of accuracy and recommended to use $\boldsymbol{K^3}$, the kinship estimator we used in $CDgp^2$. However, we showed that $CDgp^2$ did not improve estimations of accuracy in our study. We also tested other CDs such as the CD of contrast (Rincent et al., 2012, 2017) between breeding values and the mean breeding value of the TS, or the new proxy developed by Rabier et al. (2016). Both approaches did not give better results than the individual CD (results not shown).

Conclusion

In conclusion, genetic structure impacted genomic prediction accuracy in this dent maize panel. For a given size of TS, the highest accuracies were often achieved when the TS and the VS were consistent in terms of group composition. However, a diverse TS remained efficient for every VS and adding extra-group individuals almost always improved accuracy. These results are encouraging concerning the use of this panel as a generic TS to be characterized on high-throughput phenotyping platforms or through extensive field trials. Using alternative prediction models, taking genetic structure into account, did not allow any precision gain compared to GBLUP. Finally, the use of CD, an priori indicator derived from mixed model equations, proved to be sometimes but not always effective to forecast the level of precision in a set of predicted individuals. New indicators taking structure into account did not achieve better performances. This study has highlighted that, in groups that diverged recently, the impact of group structure is likely due to differences in group specific allele diversity instead of differences in allele effects that cannot be captured by global parameters such as genetic covariances between groups used in indicators proposed so far. As the distribution of allele effects along the genome is probably of great importance, new *a priori* indicators of precision taking such information into account need to be developed.

Acknowledgments

This research was supported by the "Investissement d'Avenir" project "Amaizing". S. Rio is jointly funded by the program AdmixSel of the INRA metaprogram SelGen and by the breeding companies partners of the Amaizing project: Caussade-Semences, Euralis, KWS, Limagrain, Maisadour, RAGT and Syngenta. We thank Valerie Combes, Delphine Madur and Stephane Nicolas for DNA extraction, analysis and assembly of genotypic data. We thank Cyril Bauland and Carine Palaffre (INRA Saint-Martin de Hinx) for the panel assembly and the coordination of seed production, all breeding companies partners of the Amaizing project and Biogemma for field trials and Pierre Dubreuil (Biogemma) for the assembly and the analysis of phenotypic data.

Chapter 2

Disentangling group specific QTL allele effects from genetic background epistasis using admixed individuals in GWAS: an application to maize flowering

Simon Rio, Tristan Mary-Huard, Laurence Moreau, Cyril Bauland, Carinne Palaffre, Delphine Madur, Valérie Combes, Alain Charcosset

Abstract

When handling a structured population in association mapping, group-specific allele effects may be observed at quantitative trait loci (QTLs) for several reasons: (i) a different linkage disequilibrium (LD) between SNPs and QTLs across groups, (ii) the apparition of group-specific genetic mutations in QTL regions, and (iii) epistatic interactions between QTLs and other loci that have differentiated allele frequencies between groups. In this study, we proposed to apply genome wide association studies (GWAS) jointly in different genetic groups while taking into account and testing the heterogeneity between their marker allele effects. Including admixed individuals in the analysis, with known genome wide ancestries (local admixtures), can help to disentangle the factors causing the heterogeneity of allele effects across groups: local genomic differences (group differences in LD or group specific mutations) or epistatic interactions between QTLs and the genetic background. This methodology was applied to a "Flint-Dent" maize inbred panel including admixed individuals that was evaluated for flowering traits. Several associations were detected revealing a wide range of configurations of allele effects, especially at known flowering QTLs (*Vgt1*, *Vgt2* and *Vgt3*). We found several QTLs whose effect depended on the group ancestry of alleles while others interacted with the genetic background. The existence of directional epistasis was highlighted using admixed individuals and was consistent with the existence epistatic interactions at QTLs. Our GWAS approach provides useful information on the stability of QTL effects across genetic groups and could be applied to a wide range of species.

Introduction

Quantitative traits are genetically determined by numerous regions of the genome, also known as quantitative trait loci (QTLs). The advent of high density genotyping of single nucleotide polymorphism (SNPs) has opened the way to the identification of QTLs in diversity panels. These studies, referred to as genome wide association studies (GWAS), use the linkage disequilibrium (LD) between the SNPs and the QTLs underlying the traits of interest. The panels evaluated in GWAS often include sets of individuals with complex pedigrees or genetic structure (Yu et al., 2006). The latter is a common feature in plant and animal species and arises when groups of individuals cease to mate with each other and start to be subjected to different evolutionary forces.

Applying GWAS in a diversity panel including individuals from different groups raises the issue of spurious associations. The stratification of a population into genetic groups generates LD between loci that are differentiated between groups but not necessarily genetically linked. When a given trait is characterized by contrasted group-specific means, all these SNPs will correlate to it and may be detected as false positives. An efficient control of these spurious associations can be done by taking structure and kinship into account in the statistical model (Yu et al., 2006; Price et al., 2006). This procedure will however limit the statistical power at differentiated SNPs, making them difficult to detect in multi-group GWAS, especially in case of rare alleles (Rincent et al., 2014a).

In a structured population, group-specific allele effects can be observed at SNPs, and testing an overall effect using a standard GWAS model may not be effective if the QTL effect is of opposite sign in the different groups. Such effects can result from group differences in LD between SNPs and QTLs across genetic groups. A different LD extent or linkage phase between linked loci can be explained by specific dynamics of population size such as bottlenecks or expansions (Pritchard and Przeworski, 2001; Rogers, 2014). Such patterns of LD were identified in numerous species including human (Sawyer et al., 2005; Evans and Cardon, 2005), dairy and beef cattle (de Roos et al., 2008; Porto-Neto et al., 2014), pig (Badke et al., 2012), wheat (Hao et al., 2011) and maize (Van Inghelandt et al., 2011; Technow et al., 2012; Bouchet et al., 2013; Rincent et al., 2014b). A genetic mutation appearing in a QTL region may also lead to group-specific allele effects if it occurred in a founder specific of the genetic group. Several Mendelian syndromes of obesity where shown to result from mutation within specific ethnicities in human, as reviewed by Stryjecki et al. (2018). Studying obesity in a multi-ethnic GWAS would probably lead to identifying QTLs with group-specific allele effects nearby these mutations. Another possibility consists in QTLs interacting with other loci that have differentiated allele frequencies between groups. In human, Tang (2006) discussed this possibility for a candidate gene associated with a higher risk of myocardial infarction in African American than in European population, as shown by Helgadottir et al. (2006). Another example is a SNP in the promoter region of *HNF4A* gene which was associated with a higher risk of developing type 2 diabetes in Askenazi compared to United Kingdom populations (Barroso et al., 2008). This locus was later proven to be interacting with another gene in the Askenazi population (Neuman et al., 2010). In maize, evidence of QTLs with group-specific allele effects can also be found, even though the cause of these differences remains unclear. The presence of allelic series has been demonstrated for QTLs associated with flowering time, including *Vgt1* (Buckler et al., 2009). A QTL with group-specific allele effects was also identified in a maize diversity panel for a phenology trait (Durand et al., 2012). More generally, studying the stability of QTL allele effects across genetic backgrounds is an important issue. In human, it determines the ability of a genetic marker to predict the predisposition of an individual to develop a genetic disease across ethnic groups. In plant or animal breeding, it conditions the success of introgressing a favorable allele coming from a source of diversity into an elite genetic material.

Different GWAS strategies were adopted to address this issue depending on the species. In human, GWAS mostly focused on a specific genetic group, and these group-specific studies were compared later through meta-analyses (Evangelou and Ioannidis, 2013; Li and Keating, 2014). Some of these meta-analyses revealed highly conserved effects between populations (Ioannidis et al., 2004; Marigorta and Navarro, 2013)

while other put in evidence more differences (Ntzani et al., 2012). In dairy cattle, the first GWAS studies focused on a specific breed (Cole et al., 2009; Hayes et al., 2010; Cole et al., 2011). More recently, multi-breed GWAS were conducted to refine QTLs locations by taking advantage of the low LD extent observed in such composite populations (Raven et al., 2014; van den Berg et al., 2016; Sanchez et al., 2017). In maize, the possibility to use seeds from different origins and generations led geneticists to assemble GWAS panel with a broad range of genetic materials (Flint-Garcia et al., 2005; Camus-Kulandaivelu et al., 2006; Romay et al., 2013). These panels often include a limited proportion of admixed individuals that were derived from crosses between individuals from different genetic groups. The genome of these admixed individuals consist in a mosaic of fragments with different ancestries. Admixture events are a common feature in living species and can contribute to the successful colonization of new environments (Rius and Darling, 2014; Brandenburg et al., 2017). In plants, innovative admixed genetic materials were created to enable high statistical power of QTL detection along with a wide spectrum of genetic diversity studied, such as nested association mapping (NAM) (McMullen et al., 2009) or multi-parent advanced generation inter-cross (MAGIC) (Cavanagh et al., 2008). Both NAM and MAGIC populations are of great interest to study the stability of QTL effects in a wide range of genetic backgrounds. However, they generally include a limited number of founders and do not address the stability of QTL allele effects across genetic groups.

This study aimed at evaluating the interest of producing admixed individuals, derived from a large set of parents, in order to decipher the genetic determinism of a trait using innovative GWAS models. The objectives were (i) to demonstrate the interest of multi-group analyses to identify new QTLs, (ii) to highlight the interest of applying multi-group GWAS models to identify group-specific allele effects at QTLs and (iii) to show how admixed individuals can help to disentangle the factors causing the heterogeneity of allele effects across groups: local genomic differences or epistatic interactions between QTLs and the genetic background. This methodology was applied to a maize inbred population evaluated for flowering traits, including dent, flint and admixed lines. Maize flowering time is an interesting trait to analyze in quantitative genetics studies. It is considered as a major adaptive trait by tailoring vegetative and reproductive growth phases to local environmental conditions. Admixed individuals were also used to investigate the existence of directional epistasis using a test based on the mean of admixed individuals relative to that of their progenitors.

Materials and methods

Genetic material and genotypic data

Genetic material consisted in a panel of 970 maize inbred lines assembled within the "Amaizing" project. It gathered 300 dent lines, 304 flint lines and 366 admixed doubled haploid, further referred to as admixed lines. The dent lines were those included in the "Amaizing Dent" panel (Rio et al., 2019) and the flint lines were those included in the "CF-Flint" panel (Rincent et al., 2014b). The dent and flint lines aimed at representing the diversity of their respective heterotic group used in European breeding and included several breeding generations. The admixed lines were derived from 206 hybrids between flint and dent lines, mated according to a sparse factorial design (Fig. 2.1), followed by in situ gynogenesis (Bordes et al., 1997) to produce fixed admixed inbred lines. Each dent or flint line was involved in 0 to 11 hybrids (1.21 in average), each leading to 1 to 4 admixed lines (1.77 in average). In total, 172 dent lines and 171 flint lines were involved as parents of admixed lines.

All the flint and dent lines were genotyped using the 600K Affymetrix Maize Genotyping Array (Unterseer et al., 2014). Heterozygous data were treated as missing and all missing values were imputed independently within each group using Beagle v.3.3.2 and default parameters (Browning and Browning 2009). The admixed lines were genotyped with a 15K chip provided by the private company Limagrain which included a reduced set of SNPs from the 50K Illumina MaizeSNP50 BeadChip (Ganal et al., 2011). Eight check lines were included

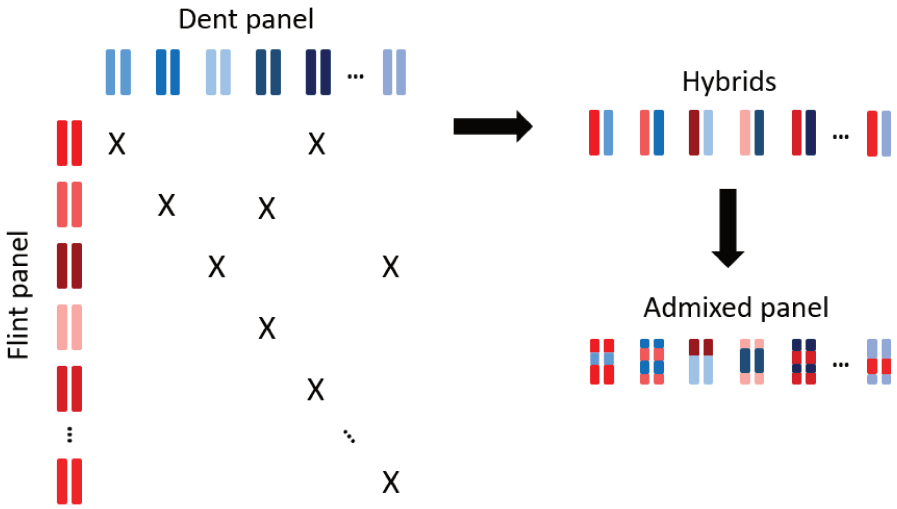


Figure 2.1: Diagram of admixed lines production from hybrids obtained by mating dent and flint lines according to a sparse factorial design

in both datasets to standardize the allele coding (0/1) on the common SNPs (around 9,000). The positions of recombination breakpoints and the parental origin of the alleles for admixed lines were determined with these common SNPs. A smoothing of parental allele origins was performed when a SNP indicated discordant information with respect to the chromosome block in which it was located. In this case, we considered the underlying genotypic datapoint as missing. In admixed lines, parental origins of alleles were imputed up to 600K using adjacent SNP information. If a set of SNPs to be imputed was located within a recombination interval, the new position of the breakpoint was positioned at half of that ordered set, according to the physical position of the SNPs along the chromosome. Alleles at SNPs were then imputed based on their origin using parental genotypic data. The whole procedure is illustrated in Supplementary Fig. S2.1. The MITE associated with the flowering QTL *Vgt1* (Salvi et al., 2007; Ducrocq et al., 2008) was also genotyped for all the individuals (0: absence, 1: presence). There was a total of 482,013 polymorphic SNPs in this dataset, for which we had information for each individual concerning the SNP allele (0/1), its ancestry (dent/flint) and the genetic background (dent/flint/admixed) in which it was observed.

The dent genome proportion of the admixed lines ranged from 0.16 to 0.86 with a mean equal to 0.51 (Supplementary Fig. S2.2). Possible selection biases were studied along the genome by comparing the observed allele frequencies with the expected allele frequencies. No major patterns could be observed suggesting no or minor selection biases among the admixed lines (Supplementary Fig. S2.3). A PCoA was performed on genetic distances computed as $D_{i,j} = 1 - K_{i,j}$, with $K_{i,j}$ being the kinship coefficient between lines i and j (computed following Eq. 2.2, see below). The flint and dent lines were clearly distinguished on the first two components, with a small overlapping region in the center of the graph, while the admixed lines were filling the genetic space between the two groups (Fig. 2.2).

LD between pairs of loci was estimated separately in the dent and the flint datasets using two estimators. The first was the standard measure of LD, computed as the square correlation between pairs of loci r^2. The second was the estimator r^2_K proposed by Mangin et al. (2011) accounting for relatedness using Eq. (2.2). We only considered SNPs for which at least ten individuals carried the minor allele in both dent and flint datasets. LD extent was first compared between groups using both estimators. A sliding window of 1Kbp up to 2 Mbp was used to group pairs of loci that were similar in terms of physical distance. The average LD was computed within each group-specific windows and revealed a higher LD extent in the dent than in the flint genetic group (Supplementary Fig. S2.4), which was consistent with previous studies (Van Inghelandt et al., 2011; Technow et al., 2012; Bouchet et al., 2013; Rincent et al., 2014b). As suggested by de Roos et al. (2008), the persistence of LD linkage phases across flint and dent genetic groups was evaluated by

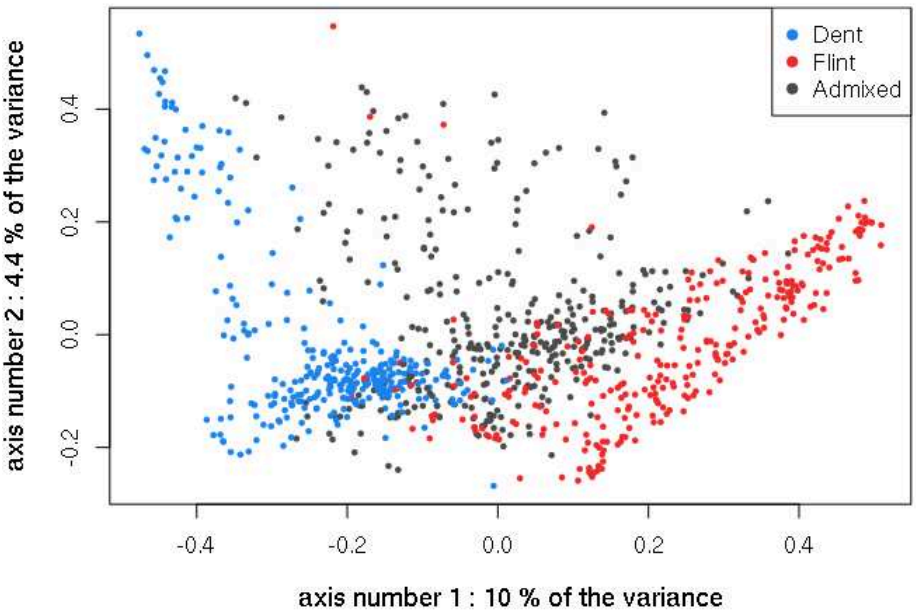


Figure 2.2: PCoA on genetic distances with coloration of individuals depending on their type: dent, flint or admixed lines

computing the correlation between the r (and r_K) estimated in each group using a sliding window of 1Kbp up to 2 Mbp. We also studied the consistency of marker phases between group by computing, for each LD estimator, the correlation between their signs in the two groups. LD phases were very consistent over short physical distances but began to diverge dramatically when the loci were distant by more than 100-200 Kbp (Supplementary Fig. S2.5).

Phenotypic data

All the lines were evaluated *per se* at Saint-Martin-de-Hinx (France) in 2015 and 2016 for male flowering (MF) and female flowering (FF), in calendar days after sowing. Each plot consisted in a row of 25 plants. MF and FF were measured as a median value within the whole plot. Each trial was a latinized alpha design where every line was evaluated two times on average. Field trials were divided in blocks of 36 plots each. To avoid competition between genetic backgrounds, dent, flint and admixed lines were sown in different blocks. Three check individuals were repeated in all blocks (B73, F353 and UH007).

Variance components were estimated using model:

$$
\begin{aligned}
&Y_{jklrc} = \mu + \beta_j + \alpha_k + G_{l(k)} + (G \times \beta)_{l(k)j} + X_{r(j)} + Z_{c(j)} + E_{jklrc} \\
&G_{l(k)} \sim \mathcal{N}\left(0, \sigma^2_{G_k}\right) \text{ independent} \\
&(G \times \beta)_{l(k)j} \sim \mathcal{N}\left(0, \sigma^2_{(G\times\beta)_{jk}}\right) \text{ independent} \\
&E_{jklrc} \sim \mathcal{N}(0, \sigma^2_{E_j}) \text{ independent}
\end{aligned}
$$

Y_{jklrc} is the phenotype, μ is the intercept, β_j is the fixed effect of trial j, α_k is the fixed effect of genetic background k (dent, flint, admixed, or the different checks: B73, F353 and UH007), $G_{l(k)}$ is the random genotype effect of line l nested within the genetic background k (not for checks) with $\sigma^2_{G_k}$ being the genotypic variance in genetic background k, $(G \times \beta)_{l(k)j}$ is the random Genotype x Environment (GxE) interaction of line l nested within the genetic background k and the trial j, with $\sigma^2_{(G\times\beta)_{jk}}$ being the GxE variance in the genetic background k for trial j, E_{ijklc} is the error with $\sigma^2_{E_j}$ being the error variance for trial j, $X_{r(j)}$ and $Z_{c(j)}$ are the row and column random effects respectively, as defined by the field design, both nested within the trial j. The row and column effects were modeled as independent or using an autoregressive model (AR1)

as determined based on the AIC criterion (Supplementary Table S2.1). Least squares means, further referred to as phenotypes, were computed over the whole design using the same model, with genotypes as fixed effects. Models parameters were estimated using ASReml-R (Butler et al., 2009) using restricted maximum likelihood (ReML).

Global assessment of directional epistasis

This panel allowed us to test for the existence of directional epistasis, which refers to epistatic interactions that are biased toward high or low genetic values, as reviewed by Le Rouzic (2014). In the presence of directional epistatic interactions and provided no selection, we can expect the genetic mean of the admixed lines to be different from its expected value, obtained by considering only additive effects (see proof in Appendix A). The existence of directional epistasis was investigated using a test based on the comparison between the mean of a progeny and the means of parental populations. The following model was applied on the joint dent, flint and admixed dataset:

$$Y_{kl} = \mu + \alpha_k + G_{kl} + E_{kl} \tag{2.1}$$

where Y_{kl} is the phenotype of the line among the N individuals of the sample, μ is the intercept, α_k is the genetic background effect with $k \in \{D, F, A\}$ for dent, flint and admixed genetic background respectively. G_{kl} is the random genetic value of the line where $\boldsymbol{g}$ is the vector of genetic values with $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{K}\sigma^2_G)$, $\boldsymbol{K}$ is the kinship matrix computed following Eq. (2.2) using allele frequencies estimated on the joint dent, flint and admixed dataset, σ^2_G is the genetic variance, E_{kl} is the residual error of the line where $\boldsymbol{e}$ is the vector of residuals with $\boldsymbol{e} \sim \mathcal{N}(0, \boldsymbol{I}\sigma^2_E)$, $\boldsymbol{I}$ is the identity matrix and σ^2_E is the residual variance. For each trait, the linear combination $H_0 : \frac{1}{2}(\alpha_D + \alpha_F) - \alpha_A = 0$ was tested to identify directional epistasis.

The kinship between individuals i and j, K_{ij}, was computed following VanRaden (2008):

$$K_{ij} = \frac{\sum_{m=1}^{M}(W_{im} - f_m)(W_{jm} - f_m)}{\sum_{m=1}^{M} f_m(1 - f_m)} \tag{2.2}$$

where W_{im} is the genotype of individual i at locus m coded 0/1 and f_m is the frequency of allele 1 at locus m.

GWAS models

In this study, three GWAS models were applied to every locus on different population samples (Table 2.1). The GWAS strategies were (i) to analyze dent and flint lines separately using a standard GWAS model $\mathbf{M_1}$, (ii) to analyze dent and flint lines jointly using a GWAS model $\mathbf{M_2}$ accounting for allele ancestry (confounded with the genetic background) and (iii) to analyze dent, flint and admixed lines in a GWAS model $\mathbf{M_3}$ accounting for both allele ancestry and the genetic background of the individuals. Models all aimed at detecting a SNP effect, defined as a contrast effect between alleles 0 and 1 at a given SNP.

Standard GWAS model $\mathbf{M_1}$

The first GWAS model $\mathbf{M_1}$ (Yu et al., 2006) was applied separately to the dent and flint datasets. For each SNP among the M loci, one has:

$$Y_{il} = \mu + \beta^m_i + G_{il} + E_{il}$$

where β^m_i is the effect of the SNP allele i at locus m (Table 2.2). All other terms are identical to those described Eq. (2.1), and the kinship was computed following Eq. (2.2) using allele frequencies estimated for each dataset. The existence of a SNP effect was tested using hypothesis $H_0 : \Delta^m = \beta^m_1 - \beta^m_0 = 0$.

Table 2.1: Population sample to which each model was applied with the corresponding number of SNPs conserved for analysis shown between brackets. Note that the number of SNP in multi-group GWAS ($\mathbf{M_2}$, $\mathbf{M_3}$) is higher than the minimum of the number of SNPs in single group GWAS ($\mathbf{M_1}$ (Dent)). SNPs carrying redundant information within a single group were indeed reduced to a single SNP for M1 and may no longer carry redundant information when datasets are pooled ($\mathbf{M_2}$, $\mathbf{M_3}$)

	Dent	Flint	Dent + Flint	Dent + Flint + Admixed
$\mathbf{M_1}$	✓ (248,747)	✓ (282,951)	✗	✗
$\mathbf{M_2}$	∅	∅	✓ (288,101)	✗
$\mathbf{M_3}$	∅	∅	∅	✓(256,951)

✓ : model was applied to the sample
✗ : model was not applied to the sample but could theoretically have been, provided the addition of a genetic background effect
∅ : model could not have been applied to the sample or would have simplified to another model

Table 2.2: Allelic states observed in each GWAS model, resulting from a combination of SNP alleles, their ancestry and the genetic background in which they are observed

	SNP	Ancestry	Genetic background	Allelic states
$\mathbf{M_1}$	$\{0,1\}$	-	-	$\{0,1\}$
$\mathbf{M_2}$	$\{0,1\}$	$\{D,F\}$	-	$\{0D,1D,0F,1F\}$
$\mathbf{M_3}$	$\{0,1\}$	$\{D,F\}$	$\{D,A,F\}$	$\{0DD,1DD,0DA,1DA,0FA,1FA,0FF,1FF\}$

0 : SNP reference allele
1 : SNP alternative allele
D : Dent ancestry or genetic background
F : Flint ancestry or genetic background
A : Admixed genetic background

Multi-group GWAS model $\mathbf{M_2}$

We applied a multi-group GWAS model $\mathbf{M_2}$ jointly to the flint and dent datasets, specifying the allele ancestry (confounded with the genetic background). For a given SNP m, one has:

$$Y_{ijl} = \mu + \beta_{ij}^m + G_{ijl} + E_{ijl}$$

where β_{ij}^m is the effect of the SNP allele i at locus m with ancestry j, as defined in Table 2.2. All other terms are identical to those described Eq. (2.1), and the kinship was computed following Eq. (2.2) using allele frequencies estimated on the joint dent and flint dataset. At a given SNP, the following hypotheses were tested:

- $H_0 : \Delta_D^m = \beta_{1D}^m - \beta_{0D}^m = 0$
- $H_0 : \Delta_F^m = \beta_{1F}^m - \beta_{0F}^m = 0$
- $H_0 : \Delta_{D+F}^m = \Delta_D^m + \Delta_F^m = 0$
- $H_0 : \Delta_{D-F}^m = \Delta_D^m - \Delta_F^m = 0$

Hypotheses Δ_D^m and Δ_F^m tested the existence of a dent and a flint SNP effect respectively. Hypothesis Δ_{D+F}^m tested for a general SNP effect while Δ_{D-F}^m tested for a divergent SNP effect between the dent and flint ancestries

Multi-group GWAS model $\mathbf{M_3}$

We applied a multi-group GWAS model $\mathbf{M_3}$ jointly to the flint, dent and admixed datasets, specifying the allele ancestry and the genetic background of the individual. For a given SNP m, one has:

$$Y_{ijkl} = \mu + \beta^m_{ijk} + G_{ijkl} + E_{ijkl}$$

where β^m_{ijk} is the effect of the SNP allele i at locus m with ancestry j in genetic background k, as defined in Table 2.2. All other terms are identical to those described Eq. (2.1), and the kinship was computed following Eq. (2.2) using allele frequencies estimated on the joint dent, flint and admixed dataset. At a given SNP, 16 hypotheses were tested (Table 2.3). Hypotheses referred to as "simple" (Δ^m_{DD}, Δ^m_{DA}, Δ^m_{FA} and Δ^m_{FF}) were tested to identify QTLs with a significant SNP effect for each combination of ancestries and genetic backgrounds. For instance, Δ^m_{DD} tested whether a dent SNP effect (differential effect between alleles 0 and 1 of dent ancestry) existed in the dent genetic background. Hypotheses referred to as "general" (Δ^m_{FA+FF}, Δ^m_{DD+DA}, Δ^m_{DA+FA}, Δ^m_{DD+FF} and, $\Delta^m_{DD+DA+FA+FF}$) were used to identify QTLs with a mean SNP effect over ancestries and genetic backgrounds. For instance, Δ^m_{FA+FF} tested for a general flint SNP effect between the flint and the admixed genetic backgrounds. Hypotheses referred to as "divergent" (Δ^m_{DA-FA}, Δ^m_{DD-DA}, Δ^m_{FA-FF}, Δ^m_{DD-FF}, Δ^m_{DA-FF}, Δ^m_{DD-FA}, $\Delta^m_{(DD+DA)-(FA+FF)}$, $\Delta^m_{(DD+FF)-(DA+FA)}$, $\Delta^m_{(DD-DA)-(FF-FA)}$) were tested to identify QTLs with a contrasted SNP effect between ancestries and/or genetic backgrounds. For instance, Δ^m_{DA-FA} tested for a divergent SNP effect between the dent and the flint ancestries in the admixed genetic background.

Table 2.3: Linear combination tested with $\mathbf{M_3}$ compared to hypotheses tested using other GWAS models

	Type	Δ^m_{DD} [a]	Δ^m_{DA} [b]	Δ^m_{FA} [c]	Δ^m_{FF} [d]	$\mathbf{M_1}$	$\mathbf{M_2}$	
Δ^m_{DD}	simple	+1	0	0	0	✓	✓	
Δ^m_{DA}	simple	0	+1	0	0	-	-	
Δ^m_{FA}	simple	0	0	+1	0	-	-	
Δ^m_{FF}	simple	0	0	0	+1	✓	✓	
Δ^m_{FA+FF}	general	0	0	+1	+1	-	-	
Δ^m_{DD+DA}	general	+1	+1	0	0	-	-	
Δ^m_{DA+FA}	general	0	+1	+1	0	-		-
Δ^m_{DD+FF}	general	+1	0	0	+1	-	✓	
$\Delta^m_{DD+DA+FA+FF}$	general	+1	+1	+1	+1	-	-	
Δ^m_{DA-FA}	divergent	0	+1	-1	0	-	-	
Δ^m_{DD-DA}	divergent	+1	-1	0	0	-	-	
Δ^m_{FA-FF}	divergent	0	0	+1	-1	-	-	
Δ^m_{DD-FF}	divergent	+1	0	0	-1	-	✓	
$\Delta^m_{(DD+DA)-(FA+FF)}$	divergent	+1	+1	-1	-1	-	-	
$\Delta^m_{(DD+FF)-(DA+FA)}$	divergent	+1	-1	-1	+1	-	-	
$\Delta^m_{(DD-DA)-(FF-FA)}$	divergent	+1	-1	+1	-1	-	-	

[a] $\Delta^m_{DD} = \beta^m_{1DD} - \beta^m_{0DD}$
[b] $\Delta^m_{DA} = \beta^m_{1DA} - \beta^m_{0DA}$
[c] $\Delta^m_{FA} = \beta^m_{1FA} - \beta^m_{0FA}$
[d] $\Delta^m_{FF} = \beta^m_{1FF} - \beta^m_{0FF}$
✓ : hypothesis also tested using the corresponding GWAS model
- : hypothesis not tested using the corresponding GWAS model

On a biological standpoint, we could expect a QTL with contrasted SNP effects between groups to be caused by (i) a local genomic difference due to a group-specific genetic mutation or to group differences in LD or (ii) an interaction with the genetic background. According to the first hypothesis, we expect a SNP effect,

for a given allele ancestry, to be conserved between different genetic backgrounds. According to the second hypothesis, we expect a SNP effect, for a given ancestry, to vary depending on the genetic background. One example would be a QTL with a strong SNP effect in a dent genetic background, but none in the flint genetic background, while the SNP effects would be of intermediate size for alleles of both ancestries in the admixed genetic background. The two biological hypotheses were illustrated in Fig. 2.3. Note that other complex configurations are possible, justifying the inclusion of all tests in the analysis.

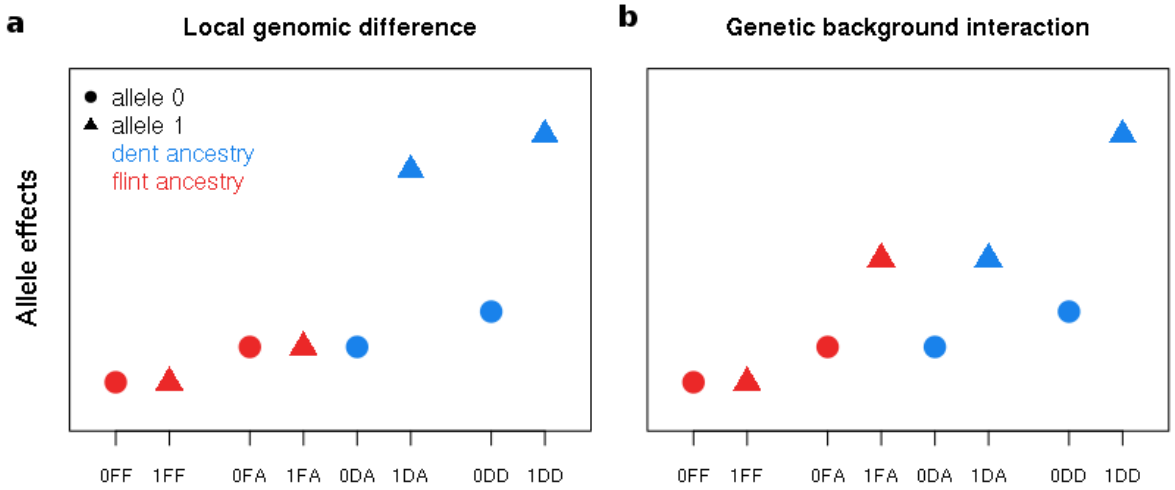


Figure 2.3: Schematic of allele effects when divergent SNP effects are observed between groups, depending on the biological hypothesis: **a.** local genomic difference between groups and **b.** allele effects interacting with the genetic background. The denomination of the allelic states on the x-axis include the SNP allele (0/1), its ancestry (D/F) and the genetic background in which it is observed (D/A/F), as presented in Table 2.2

For the three GWAS models, a SNP was discarded if its minor allelic state (Table 2.2) was carried by less than 10 individuals, or if it carried a redundant information with another SNP. Model parameters were estimated using ReML and the linear combinations of fixed effects were tested using Wald tests, both implemented in the R-package MM4LMM (Laporte et al., 2019). The false discovery rate (FDR) was controlled by applying the procedure of Benjamini and Hochberg (1995) jointly to the whole set of tests defined by each GWAS strategy, and repeatedly for each trait. For a given hypothesis tested, significant SNPs were clustered into QTLs if they were located within a physical windows of 3 Mbp, leading to a LD below 0.05 between markers of different QTLs.

Results

Phenotypic analysis and directional epistasis

We observed a substantial phenotypic variability within the dent, flint and admixed genetic backgrounds. The variance components and mean values estimated in the phenotypic analysis were summarized in Supplementary Table S2.1. Similar trends were observed for both MF and FF. The dent phenotypic mean value was higher than the flint mean value, while admixed lines were intermediate. The admixed genotypic variance was lower than the dent and flint genotypic variances, which were themselves comparable. GxE were limited and the broad sense heritabilities were high for each genetic background, ranging from 0.88 in the admixed lines to 0.96 in the dent and flint lines for both MF and FF.

The presence of admixed lines allowed us to test the existence of directional epistasis. It was evaluated by comparing the mean of admixed lines to their expected mean, in a model accounting for relatedness. The test

was significant for both MF and FF (Table. 2.4), for which the mean of admixed lines differed significantly from the one expected without directional epistatic interactions. In average, admixed lines flowered as late as dent lines while the flint lines flowered earlier.

Table 2.4: Information regarding the directional epistatic test with group-specific means estimated by the model (Eq.2.1) and the p-value (pval) of the directional epistatic deviation

	Dent	Flint	Admixed	pval
MF	68.26	66.26	68.44	$3.14\ 10^{-10}$ ***
FF	69.84	67.87	70.16	$1.99\ 10^{-11}$ ***

*** : pval $< 10^{-3}$; ** : $< 10^{-3}$ pval $< 10^{-2}$; * : $< 10^{-2}$ pval $< 5 \times 10^{-2}$

Associations detected and GWAS strategies

For each GWAS model a conservative FDR of 5% was applied, as well as a less conservative FDR of 20%. The number of significant SNPs detected and the corresponding number of QTLs were summarized in Table 2.5 for both traits. The location of QTLs detected using a FDR of 20% was represented along the genome in Fig. 2.4 for MF and in Supplementary Fig. S2.6 for FF. All associations were listed in Supplementary Tables S2.2 and S2.3. Note that major QTLs detected by a model (e.g. $\mathbf{M_1}$) may be discarded with another model (e.g. $\mathbf{M_3}$) because of the filtering on allele frequencies.

First, a standard GWAS model $\mathbf{M_1}$ was applied separately to the dent and the flint datasets. Based on a 20% FDR, 35 SNPs were associated with MF in the dent dataset while 21 SNPs were associated in the flint dataset. These SNPs could be clustered into 12 QTLs in the dent dataset and into 13 QTLs in the flint dataset. Interestingly, none of these SNPs were detected in both datasets and they only pointed to one common QTL between datasets, which was located in the vicinity of Vgt2 on chromosome 8 (Bouchet et al., 2013).

Secondly, dent and flint datasets were analyzed jointly using model $\mathbf{M_2}$, which takes into account the dent or flint ancestry of the allele. Note that the allele ancestry was confounded with the genetic background in this model. The existence of dent (Δ^m_D) and flint (Δ^m_F) SNP effects could be tested like with $\mathbf{M_1}$, and the model allowed in addition to test a general (Δ^m_{D+F}) and a divergent (Δ^m_{D-F}) SNP effect between the two ancestries. Based on a 20% FDR, 57 SNPs were associated with MF and were significant for Δ^m_D (37 SNPs), Δ^m_F (3 SNPs), Δ^m_{D+F} (11 SNPs) and Δ^m_{D-F} (18 SNPs). Note that a given SNP could display more than one significant test, which explains why the total number of SNPs did not sum to 57 over the four tests. These SNPs could be clustered in 22 QTLs that were significant for Δ^m_D (9 QTLs), Δ^m_F (2 QTLs), Δ^m_{D+F} (7 QTLs) Δ^m_{D-F} (9 QTLs). Note that some QTLs were already detected using $\mathbf{M_1}$ such as the QTL located in the vicinity of *Vgt3* on chromosome 3 Salvi et al. (2011, 2017) detected in the dent dataset. Other QTLs were specific to $\mathbf{M_1}$ such as the QTL located on chromosome 2 detected in the flint dataset, or specific to $\mathbf{M_2}$ such as the QTL located chromosome 5 detected using Δ^m_{D-F}. Based on a 20% FDR, a similar number of QTLs was detected between $\mathbf{M_1}$ and $\mathbf{M_2}$ for MF, while more QTLs was detected using $\mathbf{M_1}$ than $\mathbf{M_2}$ for FF.

Finally, the dent, flint and admixed lines were analyzed jointly using model $\mathbf{M_3}$ which distinguished the allele ancestry and the genetic background. The existence of a dent SNP effect could indeed be tested in the dent (Δ^m_{DD}) and in the admixed genetic backgrounds (Δ^m_{DA}), and similarly for the flint SNP effect (Δ^m_{FF} and Δ^m_{FA}). Several hypotheses on general and divergent SNP effects were also tested between ancestries and genetic backgrounds (Table 2.3). Based on a 20% FDR, 116 SNPs were associated with MF and were significant for Δ^m_{DD} (32 SNPs), Δ^m_{DD+FF} (34 SNPs), Δ^m_{DD-DA} (4 SNPs), $\Delta^m_{(DD+DA)-(FA+FF)}$ (5 SNPs) and others. These SNPs could be clustered in 41 QTLs that were significant for Δ^m_{DD} (8 QTLs), Δ^m_{DD+FF} (12 QTLs), Δ^m_{DD-DA} (4 QTLs), $\Delta^m_{(DD+DA)-(FA+FF)}$ (3 QTLs) and others. Note that some of the QTLs were

Table 2.5: Number of SNPs associated with each trait, depending on the GWAS strategy, using a FDR of 5% and 20%. The number of corresponding QTLs is also indicated

	MF				**FF**			
	5%		20%		5%		20%	
	SNP	QTL	SNP	QTL	SNP	QTL	SNP	QTL
$\mathbf{M_1}$ [a]	**7**	**2**	**56**	**24**	**8**	**3**	**38**	**14**
- Δ^m (Dent)	4	1	35	12	4	1	22	6
- Δ^m (Flint)	3	1	21	13	4	2	16	8
$\mathbf{M_2}$ [a]	**4**	**1**	**57**	**22**	**4**	**1**	**7**	**3**
- Δ^m_D	4	1	37	9	4	1	4	1
- Δ^m_F	-	-	3	2	-	-	2	1
- Δ^m_{D+F}	1	1	11	7	2	1	2	1
- Δ^m_{D-F}	-	-	18	9	-	-	1	1
$\mathbf{M_3}$ [a]	**7**	**3**	**116**	**41**	**6**	**2**	**11**	**6**
- Δ^m_{DD}	4	1	32	8	4	1	4	1
- Δ^m_{DA}	-	-	1	1	-	-	-	-
- Δ^m_{FA}	2	1	10	2	-	-	1	1
- Δ^m_{FF}	-	-	1	1	-	-	-	-
- Δ^m_{FA+FF}	-	-	4	4	-	-	-	-
- Δ^m_{DA+DD}	-	-	10	4	-	-	-	-
- Δ^m_{DA+FA}	-	-	11	5	-	-	-	-
- Δ^m_{DD+FF}	2	2	34	12	2	2	6	2
- $\Delta^m_{DD+DA+FA+FF}$	-	-	19	6	1	1	2	1
- Δ^m_{DA-FA}	-	-	2	2	-	-	-	-
- Δ^m_{DD-DA} [b]	-	-	4	4	-	-	-	-
- Δ^m_{FA-FF} [b]	-	-	1	1	-	-	-	-
- Δ^m_{DD-FF}	-	-	15	5	-	-	-	-
- $\Delta^m_{(DD+DA)-(FA+FF)}$	-	-	5	3	-	-	-	-
- $\Delta^m_{(DD+FF)-(DA+FA)}$ [b]	-	-	5	4	-	-	2	2
- $\Delta^m_{(DD-DA)-(FF-FA)}$ [b]	-	-	2	2	-	-	-	-

[a] number of SNPs detected over the set of tests (a given SNP could be detected using different tests)

[b] hypothesis testing an interaction between the QTL and the genetic background

already detected using $\mathbf{M_1}$ and $\mathbf{M_2}$ such as the QTL located in the vicinity of Vgt3 on chromosome 3, while several QTLs were specific to $\mathbf{M_3}$ such as the two QTLs detected in chromosome 2 using Δ^m_{FA}. Several QTLs were detected as showing a divergent SNP effect, including hypotheses testing an interaction with the genetic background. Based on 5% and 20% FDRs, the number of QTLs detected with $\mathbf{M_3}$ was the highest for MF and intermediate between $\mathbf{M_1}$ and $\mathbf{M_2}$ for FF.

Highlighted QTLs

Among the 41 QTLs detected for MF with $\mathbf{M_3}$, six QTLs were selected and studied in further details. The five first QTLs had (i) at least one significant test among $\mathbf{M_3}$ hypotheses based on a FDR of 20%, and (ii) a large frequency for each allele with a minimum of 30 lines carrying the minor allelic state (*QTL7.2*). Among them, one SNP was located in the vicinity of *Vgt2* (Bouchet et al., 2013) and another in the vicinity of *Vgt3* (Salvi et al., 2011, 2017). In addition to these five QTLs, we also considered a MITE polymorphism known to be associated with *Vgt1*, a flowering QTL detected in several studies (Salvi et al., 2007; Buckler et al.,

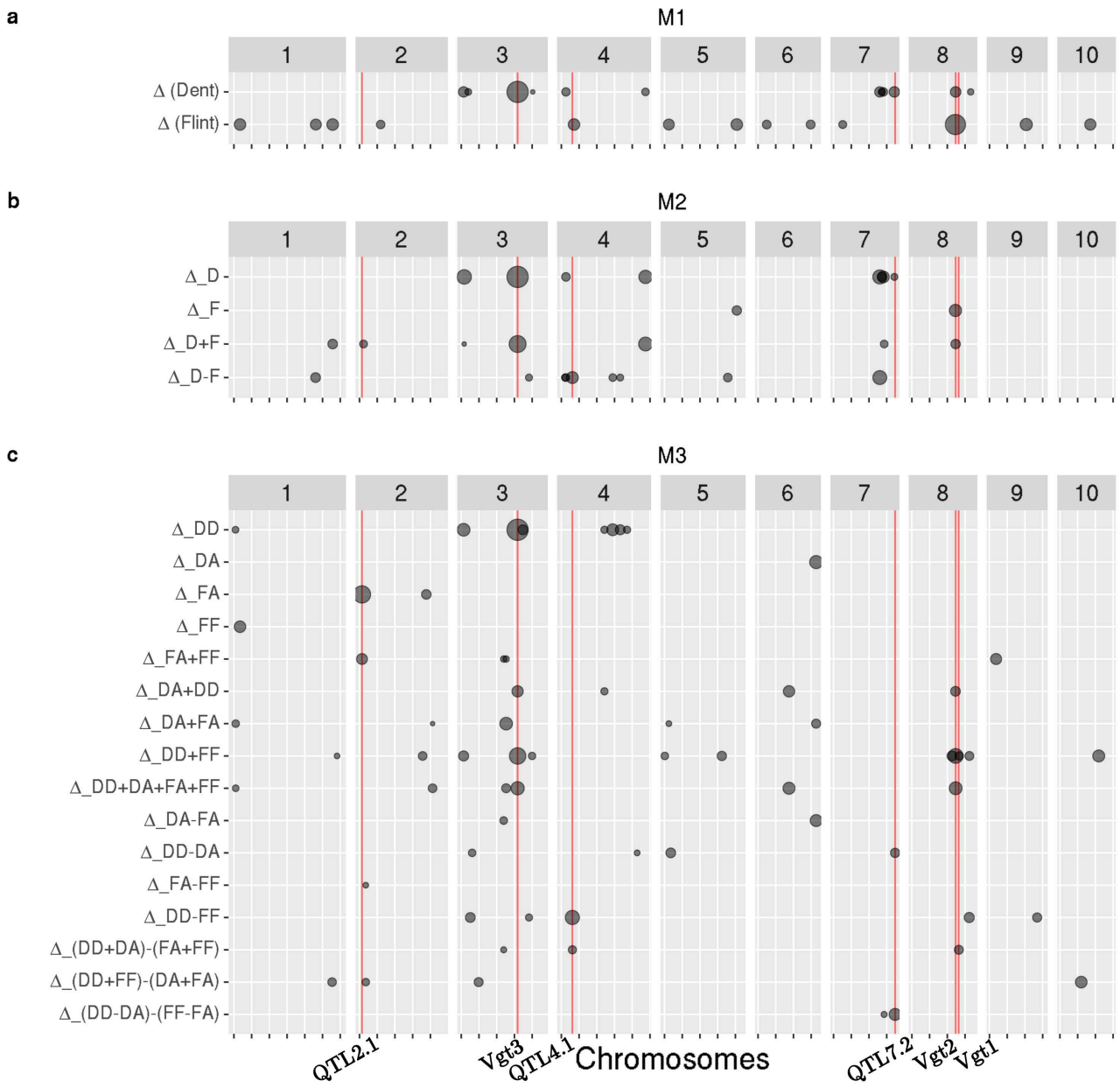


Figure 2.4: Position of QTLs detected by each GWAS strategy for MF using a FDR of 20%. The size of the grey dots is proportional to the -$\log_{10}$(pval) of the test at the most significant SNP of the region. Red vertical lines and names below correspond to QTL discussed in section "Highlighted QTLs". Note that major QTLs detected by a model may be discarded with another model because to filtering on allele frequencies

2009; Ducrocq et al., 2008). For all QTLs, information concerning their physical position along the genome, the frequency of each allelic state and their -$\log_{10}$(pval) at each test was summarized in Table 2.6. The distribution of the phenotypes was illustrated for each allele after correcting for relatedness in Fig. 2.5, and their location along the genome was indicated by red vertical lines in Fig. 2.4. Other QTLs had interesting profiles, showing either group-specific allele effects conserved between ancestries or interactions with the genetic background, and were presented in Supplementary Fig. S2.8 and Table S2.3.

The SNP matching *Vgt2* region on chromosome 8 was detected as associated with MF (5% FDR) using Δ^m_{DD+FF} ($\mathbf{M_3}$). This QTL showed a conserved effect across ancestries and genetic backgrounds (Fig. 2.5-**a**). This observation was supported by a high -$\log_{10}$(pval) for tests relating to a general SNP effect: Δ^m_{D+F} (5.25), Δ^m_{DD+DA} (5.35), Δ^m_{DA+FA} (3.20), Δ^m_{DD+FF} (7.15) and $\Delta^m_{DD+DA+FA+FF}$ (6.46), and a low -$\log_{10}$(pval) for tests relating to divergent SNP effects.

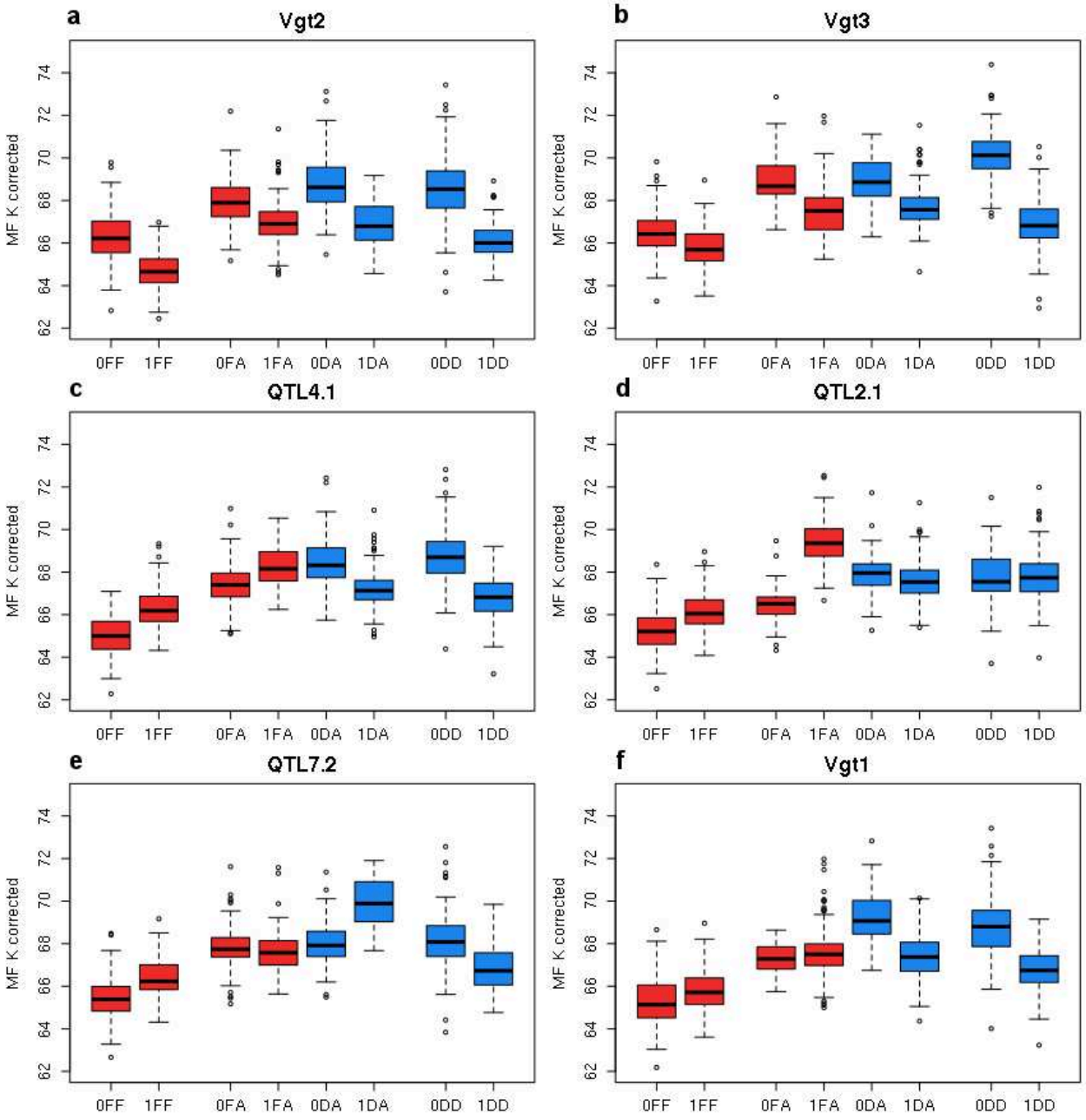


Figure 2.5: Boxplots of phenotypes for the different alleles of the six highlighted QTLs: **a**. *Vgt2*, **b**. *Vgt3*, **c**. *QTL4.1*, **d**. *QTL2.1*, **e**. *QTL7.2*, **f**. *Vgt1*, after correcting for relatedness using $\mathbf{M_3}$. The denomination of the allelic states on the x-axis include the SNP allele (0/1), its ancestry (D/F) and the genetic background in which it is observed (D/A/F), as presented in Table 2.2

The SNP matching *Vgt3* region on chromosome 3 was detected as associated with MF (5% FDR) using Δ^m_{DD} ($\mathbf{M_3}$). This QTL showed a large effect in the dent genetic background, a medium effect in the admixed genetic background regardless of the allele ancestry and a small effect in the flint genetic background (Fig. 2.5-**b**). This observation was supported by a high -$\log_{10}$(pval) for the tests relating to the dent SNP effect in the dent genetic background: Δ^m (Dent, 10.99), Δ^m_D (9.65) and Δ^m_{DD} (10.53), and a low -$\log_{10}$(pval) for the tests relating to the flint SNP effect in a flint genetic background. Like for *Vgt2*, a high -$\log_{10}$(pval) was detected for tests relating to a general SNP effect: Δ^m_{D+F} (7.47), Δ^m_{DD+DA} (6.01), Δ^m_{DD+FF} (7.86) and $\Delta^m_{DD+DA+FA+FF}$ (6.59), but a high -$\log_{10}$(pval) was detected for the test relating to a divergent SNP effect between the dent and the flint genetic backgrounds: Δ^m_{DD-FF} (3.86). There was also a high -$\log_{10}$(pval) for a divergent dent SNP effect between different genetic backgrounds: Δ^m_{DD-DA} (3.03). All these results support the existence of a QTL effect that tends to be higher when the dent genome proportion increases

Table 2.6: Information regarding the six highlighted QTLs

	Vgt2	***Vgt3***	***QTL4.1***	***QTL2.1***	***QTL7.2***	***Vgt1***
Trait	MF	MF	MF	MF	MF	MF
SNP	AX-91100620	AX-91583310	AX-91218190	AX-90601996	AX-91744673	MITE
Chromosome	8	3	4	2	7	8
Position (Mbp)	123.50	158.97	31.10	7.04	173.73	131.99
Allele frequency						
- 0DD	230	97	115	75	243	151
- 1DD	70	203	185	225	57	149
- 0DA	119	48	53	50	161	70
- 1DA	58	141	127	134	30	108
- 0FA	81	92	107	74	113	17
- 1FA	108	85	79	108	62	171
- 0FF	162	158	161	102	210	49
- 1FF	142	146	143	202	94	255
-$\log_{10}$(pval)						
$\mathbf{M_1}$						
- Δ^m (Dent)	4.26 *	10.99 ***	4.96 *	0.05	1.00	3.34 *
- Δ^m (Flint)	2.74 .	0.88	0.31	1.24	1.20	0.86
$\mathbf{M_2}$						
- Δ^m_D	4.16 *	9.65 ***	4.01 *	0.03	0.96	3.37 *
- Δ^m_F	1.96	1.10	2.16 .	1.29	0.77	0.80
- Δ^m_{D+F}	5.25 **	7.47 ***	0.56	0.69	0.10	0.42
- Δ^m_{D-F}	0.57	3.17 *	5.76 **	0.82	1.47	3.04 *
$\mathbf{M_3}$						
- Δ^m_{DD}	5.21 **	10.53 ***	4.42 *	0.00	1.79	4.62 *
- Δ^m_{DA}	2.95 .	1.38	1.47	0.31	2.64 .	2.96 .
- Δ^m_{FA}	1.09	2.12 .	0.97	8.24 ***	0.15	0.15
- Δ^m_{FF}	2.85 .	0.92	2.34 .	1.23	1.51	0.41
- Δ^m_{FA+FF}	2.38 .	2.00 .	2.00 .	5.91 **	0.49	0.33
- Δ^m_{DD+DA}	5.35 **	6.01 **	3.44 *	0.19	0.32	4.96 *
- Δ^m_{DA+FA}	3.20 *	2.93 .	0.23	3.09 *	1.47	0.85
- Δ^m_{DD+FF}	7.15 ***	7.86 ***	0.42	0.70	0.14	1.07
- $\Delta^m_{DD+DA+FA+FF}$	6.46 **	6.59 **	0.39	2.45 .	0.63	1.25
- Δ^m_{DA-FA}	0.69	0.11	2.11 .	4.84 *	2.00 .	1.52
- Δ^m_{DD-DA} [b]	0.35	3.03 *	0.59	0.29	5.58 **	0.07
- Δ^m_{FA-FF} [b]	0.60	0.69	0.48	3.70 *	1.51	0.10
- Δ^m_{DD-FF}	0.58	3.86 *	6.93 **	0.73	2.93 .	2.93 .
- $\Delta^m_{(DD+DA)-(FA+FF)}$	0.82	1.25	5.39 **	3.51 *	0.04	2.91 .
- $\Delta^m_{(DD+FF)-(DA+FA)}$ [b]	0.73	0.94	0.06	1.35	1.66	0.04
- $\Delta^m_{(DD-DA)-(FF-FA)}$ [b]	0.08	2.96 .	0.86	2.60 .	6.21 **	0.13

*** : -$\log_{10}$(pval) > 7 ; ** : 7 > -$\log_{10}$(pval) > 5 ; * : 3 > -$\log_{10}$(pval) > 5 ; . : 2 > -$\log_{10}$(pval) > 3

[b] hypothesis testing an interaction between the QTL and the genetic background

within individuals. It suggests that *Vgt3* interacts with the genetic background for MF.

The SNP matching a region further referred to as *QTL4.1* on chromosome 4 was detected as associated with MF (20% FDR) using Δ^m_{DD-FF} ($\mathbf{M_3}$). This QTL showed a contrasted effect between alleles of different ancestries with an apparent inversion of effects (Fig. 2.5-**c**). This observation was supported by a high -$\log_{10}$(pval) for the tests relating to a divergent SNP effect between ancestries: Δ^m_{D-F} (5.76), Δ^m_{DD-FF} (6.93) and $\Delta^m_{(DD+DA)-(FA+FF)}$ (5.39). Conversely a low -$\log_{10}$(pval) was detected for tests Δ^m_{DD-DA} and Δ^m_{FA-FF}, which would have otherwise suggested an interaction with the genetic background. These results support the

existence of a local genomic difference at *QTL4.1* between the dent and the flint genetic groups for MF, but no interaction with the genetic background.

The SNP matching a region further referred to as *QTL2.1* on chromosome 2 was detected as associated with MF (5% FDR) using Δ^m_{FA} ($\mathbf{M_3}$). This QTL showed a flint effect in the admixed genetic background (Fig. 2.5-**d**), which was supported by a high -$\log_{10}$(pval) for the test Δ^m_{FA} (8.24). Although there was a high -$\log_{10}$(pval) for a general flint SNP effect across genetic backgrounds: Δ^m_{FA+FF} (5.91), a high -$\log_{10}$(pval) was observed for a divergent SNP effect between those same alleles: Δ^m_{FA-FF} (3.70). A high -$\log_{10}$(pval) was also observed for a divergent SNP effect between different ancestries in the admixed genetic background: Δ^m_{DA-FA} (4.84). All these results support the existence of a QTL effect existing only for alleles of flint ancestry in the admixed genetic background. It suggests that *QTL2.1* interacts with the genetic background for MF.

The SNP matching a region further referred to as *QTL7.2* on chromsome 7 was detected as associated with MF (20% FDR) using $\Delta^m_{(DD-DA)-(FF-FA)}$ ($\mathbf{M_3}$). This QTL showed contrasted dent effects between the dent and the admixed genetic background (Fig. 2.5-**e**). This observation was supported by a high -$\log_{10}$(pval) for the test relating a divergent dent SNP effect between genetic backgrounds: Δ^m_{DD-DA} (5.58). A high -$\log_{10}$(pval) was also observed for the hypothesis testing the equality between the divergent dent SNP effect and the divergent flint SNP effect: $\Delta^m_{(DD-DA)-(FF-FA)}$ (6.21). All these results support the existence of a QTL with opposite effects between the dent and the admixed genetic backgrounds. It suggests that *QTL7.2* interacts with the genetic background for MF.

The MITE known to be associated with *Vgt1* was never detected for MF using a FDR of 5% or 20%. However, it showed a dent effect that was conserved between the dent and the admixed genetic background, and no flint effect (Fig. 2.5-**f**). This observation was supported by a high -$\log_{10}$(pval) for tests relating to the dent SNP effect: Δ^m (Dent) (3.34), Δ^m_D (3.37), Δ^m_{DD} (4.62) and Δ^m_{DD+DA} (4.96), and a low -$\log_{10}$(pval)l for tests relating to flint SNP effects. These results supported the existence of a local genomic difference at *Vgt1* between flint and dent genetic groups. Apparently, this QTL did not interact with the genetic background for MF.

Discussion

In GWAS, the stratification of the population sample into distinct genetic groups is a common feature. Such structure challenges the methods to detect QTLs because (i) spurious associations may be detected if the genetic structure is not accounted for by the statistical model, (ii) QTLs whose polymorphism is correlated with the genetic structure generally have a low probability of being detected, and (iii) group-specific allele effects can be observed due to group differences in LD, group-specific genetic mutations, or epistatic interactions with the genetic background.

Accounting for genetic groups in GWAS

A simple way to deal with genetic groups is to analyze them separately. In our study, a standard GWAS model $\mathbf{M_1}$ was applied separately within the dent and the flint datasets. These datasets included inbred lines genotyped at high density and evaluated simultaneously for flowering traits. High heritabilities were estimated for each genetic group in the phenotypic analysis, forming ideal datasets to detect QTLs. While few QTLs were detected using a 5% FDR, several were detected using a 20% FDR, especially for MF. Interestingly, only one QTL was detected in both dent and flint datasets for MF, and not at the same SNPs, while none were detected in common for FF. One could question whether observing such differences between datasets indicated group specific allele effects, or simply group differences in terms of statistical power. This

question often arises when GWAS is applied separately to genetic groups, as in maize (Rincent et al., 2014b; Revilla et al., 2016) or dairy cattle (Buitenhuis et al., 2014, 2015), and is very difficult to answer except for obvious configurations such as associations at SNPs segregating only in one group.

Another way to handle genetic groups is to analyze them jointly. One possibility is to apply model $\mathbf{M_1}$ while specifying genetic structure as a global fixed effect, in order to prevent the detection of spurious associations. In dairy cattle, this strategy generally improved the precision concerning QTL locations by taking advantage of the low LD extent observed in multi-group datasets. However, while Sanchez et al. (2017) and van den Berg et al. (2016) observed a gain in statistical power due to a larger population size, Raven et al. (2014) detected less QTLs by combining breeds compared to separate analyses. They attributed this finding to the limited amount of QTLs segregating within both Holstein and Jersey breeds, but also reported that QTLs detected in both breeds showed only small to medium correlations between within-breed estimates of SNP effects (e.g. 0.082 for milk yield). Obviously, applying $\mathbf{M_1}$ jointly to genetic groups does not help to answer the previous question, which is whether or not QTL effects are conserved between genetic groups.

A model specifying group specific allele effects was referred to as $\mathbf{M_2}$ in this study. As with $\mathbf{M_1}$, the existence of a dent (Δ^m_D) or a flint (Δ^m_F) SNP effects can be tested, but $\mathbf{M_2}$ also allows us to test the existence of a general (Δ^m_{D+F}) and a divergent (Δ^m_{D-F}) SNP effects between flint and dent ancestries. Note that testing Δ^m_{D+F} is similar, although not strictly equivalent, to testing a SNP effect by applying $\mathbf{M_1}$ to a multi-group dataset. Using the hypotheses specifically tested in $\mathbf{M_2}$ (Δ^m_{D+F} and Δ^m_{D-F}), it was possible to detect new QTLs that were not detected with $\mathbf{M_1}$, which indicated a gain of power for some genomic regions. In particular, QTLs were detected as having a divergent SNP effect between the dent and flint genetic groups, proving the existence of group-specific QTLs effects in this dataset. Several QTLs were detected in common with $\mathbf{M_1}$ but each strategy allowed the detection of specific QTLs, demonstrating the complementarity between the models. For a comparable effect in $\mathbf{M_1}$ and $\mathbf{M_2}$ (e.g. Δ^m (Dent) and Δ^m_D), the lower number of associations detected with $\mathbf{M_2}$ could often be attributable to a more conservative filtering on allele frequencies. It could also be due to the use of different kinship matrices, as the allele frequencies used for their computation were re-estimated for each dataset. In conclusion, $\mathbf{M_2}$ was efficient to identify QTLs with either conserved or specific allele effects between ancestries, but observing group-specific allele effects is not giving any insight concerning the cause of this specificity. Admixed individuals can help to tackle that issue.

Benefits from admixed individuals

For this study, admixed individuals were generated by mating pure individuals of each group according to a sparse factorial design. Integrating these admixed individuals in GWAS can be done by simply analyzing the joint multi-group dataset using $\mathbf{M_1}$ or $\mathbf{M_2}$, as it would probably lead to a gain in statistical power, due to an increase in population size. Additionally, admixed individuals can be used to disentangle the factors causing the heterogeneity of allele effects across groups: (i) local genomic differences (e.g. group-specific LD between SNPs and QTLs) or (ii) epistatic interactions between QTLs and the genetic background.

A model distinguishing the allele ancestry (dent/flint) and the genetic background (dent/flint/admixed) was referred to as $\mathbf{M_3}$. Using this model, the existence of a SNP effect can be tested for a given ancestry and genetic background (e.g. Δ^m_{DD} to test the dent SNP effect in the dent genetic background). An overall SNP effect can also be tested across ancestries or genetic backgrounds (e.g. Δ^m_{DD+DA} for an overall dent SNP effect), as well as many hypotheses concerning a divergent SNP effect (e.g. Δ^m_{DD-DA} for an interaction of the dent SNP with the genetic background). Several QTLs were detected for both traits, especially when considering a 20% FDR for MF with a total of 41 QTLs. While many of these QTLs were previously detected using $\mathbf{M_1}$ and $\mathbf{M_2}$, the new hypotheses tested allowed us to discover new interesting regions. These new QTLs resulted from a gain in statistical power by (i) testing an overall SNP effect across ancestries and/or

genetic backgrounds, for which individual effects were not large enough to be detected, or by (ii) testing hypotheses for complex configurations between allele effects that would have prevented the detection of their overall effect. A QTL detected using $\mathbf{M_1}$ and $\mathbf{M_2}$ but no longer with $\mathbf{M_3}$ may again be attributable to a more conservative filtering on allele frequencies, which suggests a complementarity between GWAS models.

In conclusion, producing admixed individuals from a large number of parents is an innovative and useful genetic material to analyze the genetic determinism of traits in structured populations. Using our GWAS models, admixed individuals can give interesting insight concerning the stability of QTL allele effects across genetic groups. The new hypotheses tested with $\mathbf{M_2}$ and $\mathbf{M_3}$ did not lead to an increase in false positive rate, based on the observation of the QQ-plots of the test p-values (results not shown). Only homozygous inbred lines were considered in this study, but the methodology may be easily generalized to heterozygous individuals, in order to be applied in a wide range of species. The main drawback of this genetic material is that it does not solve the issue of rare alleles, for which statistical power is limited, compared to other genetic material such as NAM (McMullen et al., 2009) or MAGIC (Cavanagh et al., 2008). Increasing the number of admixed individuals derived from each cross (below 2 on average in our study) while maintaining a high number of founders could be a way to improve the properties of the design.

Heterogeneity of maize flowering QTL allele effects

From a global perspective, the high number of QTLs detected in our study (41 QTLs for MF) is consistent with previous studies on maize flowering time (Chardon et al., 2004; Buckler et al., 2009; Romay et al., 2013; Rincent et al., 2014b; Li et al., 2016b). When evaluating the American NAM, Buckler et al. (2009) suggested that flowering time was trait controlled by a large number of QTLs with additive effects. Our results would rather suggest a large number of QTLs with either additive or interacting profiles. This heterogeneity of QTL allele effects may however be consistent with their observation of allelic series at several QTLs, which might resulted from interactions with the genetic background of the NAM founders. In addition to this genome-wide scan, it is also useful to look more closely at specific QTLs.

When doing GWAS in a multi-group population, geneticists usually prospect for QTLs featuring a conserved effect between groups. Such QTLs were detected in our study with the example of the SNP associated with MF in the vicinity of *Vgt2* on chromosome 8, previously identified by Bouchet et al. (2013). At this SNP, all hypotheses that tested a general SNP effect had a high -$\log_{10}$(pval), and conversely for hypotheses testing a divergent SNP effect. When simultaneously interpreting all tests, *Vgt2* appeared to have an effect that is conserved between genetic groups. Such QTL should easily be detected in a multi-group population sample using a standard GWAS model (Yu et al., 2006).

When group-specific allele effects are due to group differences in LD or group-specific mutations at the QTL, their difference should be conserved between the pure and the admixed genetic backgrounds. A first QTL matching this situation (*QTL4.1*) was detected by a SNP located on chromosome 4 and showed an opposite effect between ancestries. High -$\log_{10}$(pval) were observed for tests relating to a divergent SNP effect between the dent and flint genetic backgrounds (Δ^m_{DD-FF}) or to a divergent SNP effect between ancestries ($\Delta^m_{(DD+DA)-(FA+FF)}$). These tests supported the existence of a contrasted QTL effect between the dent and flint groups due to a local genomic difference. To validate this hypothesis, nearly isogenic lines could be produced at this SNP for the two alleles and both ancestries. Individuals inheriting a dent allele in a flint genetic background, and conversely, would give a definite proof of a local genomic difference. It also remains complicated to disentangle the effect of LD from that of a genetic mutation without complementary analysis. LD was shown to be different between groups, with a higher LD extent in the dent group (Supplementary Fig. S2.4), while LD phases appeared well-conserved at short distances (Supplementary Fig. S2.5). However, a strong overall conservation of LD phases at short distances does not exclude a specific configuration for a given SNP-QTL pair.

Another example is the MITE that we selected based on the *a priori* knowledge that it is associated with *Vgt1* (Salvi et al., 2007; Buckler et al., 2009; Ducrocq et al., 2008). A high -$\log_{10}$(pval) was observed for tests relating to a dent SNP effect (Δ^m_{DD} and Δ^m_{DD+DA}) but not for tests relating to a flint SNP effect. Note that another SNP (AX-91103145) was detected close to the MITE (548 Kbp further), based on 20% a FDR, for $\Delta^m_{(DD+DA)-(FA+FF)}$ (see *QTL8.4* in Supplementary Fig S2.7-**a** and Supplementary Table S2.4). This SNP also showed evidence for a contrasted QTL effect between the dent and flint groups due to a local genomic difference. However these two loci were in very low LD with each other (below 0.05). We could reasonably suggest that the MITE and the SNP were both capturing a partial but different genetic information of the causal genetic variant at *Vgt1*. Ducrocq et al. (2008) already showed the existence of other genetic variants being more associated with maize flowering than the MITE in the vicinity of *Vgt1*, such as CGindel587.

Group-specific allele effect may also be due to an interaction with the genetic background. A first QTL matching this profile was detected by a SNP in the vicinity of *Vgt3* on chromosome 3 (Salvi et al., 2011, 2017). The QTL effect increased with the dent genome proportion, suggesting an interaction with the genetic background. A high -$\log_{10}$(pval) was observed for tests that supported this hypothesis: a dent SNP effect in the dent genetic background (Δ^m_{DD}), a divergent SNP effect between the dent and the flint genetic backgrounds (Δ^m_{DD-FF}) and a divergent dent SNP effect between genetic backgrounds (Δ^m_{DD-DA}). Provided interactions with numerous loci, this QTL could lead to a disappointment if a breeder tried to introgress its alleles from a dent to a flint genetic background, as its effect would probably vanish with repeated back-cross generations. Provided interactions with a single locus, its effect would be conditioned by the allele at the other locus. Using nearly isogenic lines that cumulated an early mutation at *Vgt1* (Chardon et al., 2005) and the early allele at *Vgt3*, the effect of *Vgt3* was shown to vanish in presence of the early allele of *Vgt1* (A. Charcosset pers. comm.), which supported the hypothesis of *Vgt3* interacting with the genetic background. The existence of such interactions is consistent with flowering time being controlled by a network of interacting loci, as now well established in model species arabidopis (Bouché et al., 2015). Recently, Liang et al. (2019) demonstrated the action of *ZmMADS69*, located in the region of *Vgt3*, as being a flowering activator of *ZmRap2.7-ZCN8*, a major regulator of maize flowering time located in the region of *Vgt2*.

Other examples of QTLs interacting with the genetic background were identified. Two of them featured a similar profile in the sense that they mainly exhibited a QTL effect in the admixed genetic background. One was located on chromosome 2 (*QTL2.1*) and showed a flint effect in the admixed genetic background, while the other QTL was located on chromosome 7 (*QTL7.2*) and showed an opposite dent effect between the dent and the admixed genetic backgrounds. Such QTLs are interesting as they are mainly revealed when creating admixed genetic material. They are also suggesting complex epistatic interactions between QTLs for these traits.

The existence of epistatic interactions was also evaluated globally by a test that aimed at detecting directional epistasis (Le Rouzic, 2014). This test was specifically developed to benefit from our admixed genetic material and revealed important directional epistasis for both flowering traits with admixed lines flowering closer to the dent than the flint group. Such epistasis may imply that (i) the effects of early alleles from flint origin tend to decrease in presence of common dent alleles and/or (ii) the effect of late alleles from dent origin tends to be promoted by common flint alleles. Alternatively, this epistasis can be interpreted as late QTL alleles (common in dent lines but rare in flint lines) interacting in a duplicate way (Mather, 1967), i.e. the presence of a late allele at one QTL is sufficient to confer a late phenotype. This hypothesis is equivalent to early QTL alleles (common in flint lines but rare in dent lines) interacting in a complementary way (Mather, 1967), i.e. early alleles are needed at both loci to confer an early phenotype. We also tested global epistasis that is not directional by decomposing the genetic variance into an additive and an epistatic component, as suggested by Vitezica et al. (2017). This confirmed the existence of epistatic interactions for FF and MF (Supplementary Table S2.5). In conclusion, the assessment of global epistasis supported the possibility of QTLs interacting with the genetic background, resulting from epistatic interactions with loci that have differentiated allele frequencies between groups. It would be interesting to test the existence of epistatic interactions between pairs of loci. However, a filtering on crossed allele frequencies between pairs

of loci would lead to discard most SNPs from the analysis. Other possibilities would be to apply GWAS procedures that are based on testing the epistatic variance each SNP (Jannink, 2007; Crawford et al., 2017).

Conclusion

In this study, we proposed an innovative multi-group GWAS methodology which accounts and tests for the heterogeneity of QTL allele effects between groups. The addition of admixed individuals to the dataset was useful to disentangle the factors causing the heterogeneity of allele effects, being either a local genomic differences or epistatic interactions with the genetic background. Both methodology and genetic material open new perspectives to evaluate the stability of QTL effects across genetic backgrounds in a wide range of species.

Acknowledgments

This research was supported by the "Investissement d'Avenir" project "Amaizing". S. Rio is jointly funded by the program AdmixSel of the INRA metaprogram SelGen and by the breeding companies partners of the Amaizing project: Caussade-Semences, Euralis, KWS, Limagrain, Maisadour, RAGT and Syngenta. We thank Valerie Combes, Delphine Madur and Stéphane Nicolas (GQE - Le Moulon) for DNA extraction, analysis and assembly of genotypic data. We thank Cyril Bauland (GQE - Le Moulon), Carine Palaffre, Bernard Lagardère, Jean-René Loustalot (INRA Saint-Martin de Hinx) for the panel assembly and the coordination of seed production, all the breeding companies partners of the Amaizing project for the production of admixed lines and the company Limagrain for the genotyping of admixed lines.

Appendix A

Effect of directional epistasis on the mean of a progeny

Genetic values are modeled using allele effects at bi-allelic QTLs and deviation effects for each pair of alleles coming from different loci, leading to epistatic interactions between QTLs:

$$G_i = \sum_{m=1}^{M}[(1-W_{im})\beta_m^0 + W_{im}\beta_m^1] + \sum_{m=1}^{M}\sum_{m'>m}^{M}[(1-W_{im})(1-W_{im'})\delta_{mm'}^{00} + W_{im}(1-W_{im'})\delta_{mm'}^{10} + (1-W_{im})W_{im'}\delta_{mm'}^{01} + W_{im}W_{im'}\delta_{mm'}^{11}]$$

where G_i is the genotype of individual i, M is the number of QTLs, W_{im} is the QTL genotypes of individual i at locus m (coded 0/1) with $W_{im} \sim \mathcal{B}(f_m)$ independent, f_m is allele frequency of allele 1, β_m^0 is effect of allele 0 at locus m, β_m^1 is effect of allele 1 at locus m, $\delta_{mm'}^{00}$ is the deviation effect specific to the pair of alleles 0 at locus m and m', $\delta_{mm'}^{10}$ is the deviation effect specific to the pair of allele 0 at locus m and allele 1 at locus m', $\delta_{mm'}^{01}$ is the deviation effect specific to the pair of allele 1 at locus m and allele 1 at locus m' and $\delta_{mm'}^{11}$ is the deviation effect specific to the pair of alleles 1 at locus m and m'.

We can compute μ, the expected value of G_i:

$$\mu = \sum_{m=1}^{M}[\beta_m^0 + f_m(\beta_m^1 - \beta_m^0)] + \sum_{m=1}^{M}\sum_{m'>m}^{M}[\delta_{mm'}^{00} + f_m(\delta_{mm'}^{10} - \delta_{mm'}^{00}) + f_{m'}(\delta_{mm'}^{01} - \delta_{mm'}^{00}) + f_m f_{m'}(\delta_{mm'}^{11} + \delta_{mm'}^{00} - \delta_{mm'}^{10} - \delta_{mm'}^{01})]$$

Let us consider two genetic groups D and F of pure lines and a group of admixed lines A. Both group D and F have specific allele frequencies f_{mD} and f_{mF} at each locus m. Let us suppose that admixed lines were obtained by mating pure lines from each group randomly into across-group hybrids, before generating one inbred admixed line from each hybrid. In absence of epistatic interactions between loci, we could expect that $\mu_A = \frac{\mu_D + \mu_F}{2}$. The absence of selection involves $f_{mA} = \frac{f_{mD} + f_{mF}}{2}$ at a given locus m.

In presence of epistatic interactions between loci and without selection, we expect a deviation between the observed mean μ_A and the expected mean $\frac{\mu_D + \mu_F}{2}$:

$$\mu_A - \frac{\mu_D + \mu_F}{2} = -\frac{1}{4}\sum_{m=1}^{M}\sum_{m'>m}^{M}(f_{mD} - f_{mF})(f_{m'D} - f_{m'F})\left(\delta_{mm'}^{11} + \delta_{mm'}^{00} - \delta_{mm'}^{10} - \delta_{mm'}^{01}\right)$$

Such deviation becomes large if the allele frequencies are highly differentiated between group D and F at both loci. It also requires that epistasis is directional, meaning that QTL deviation effects do not cancel each other out among loci.

Chapter 3

Accounting for group-specific allele effects and admixture in genomic predictions: theory and experimental evaluation in maize

Simon Rio, Laurence Moreau, Alain Charcosset and Tristan Mary-huard

Abstract

When a population is stratified into genetic groups, a heterogeneity of single nucleotide polymorphism (SNP) allele effects may be observed across groups. To account for this heterogeneity, we developed two genomic prediction models: Multi-group Admixed GBLUP (MAGBLUP) 1, which was derived according to the "animal model", and MAGBLUP 2 modeling group-specific distributions of SNP allele effects. Both models are adapted to the prediction of admixed individuals and account for local admixture through the group ancestry of alleles. In terms of complementarity, MAGBLUP 1 can be used to identify the segregation variance generated by admixture while MAGBLUP 2 can be used to disentangle the variance that is due to main SNP allele effects from that due to group-specific deviations. MAGBLUP 1 and 2 were evaluated for their precision in estimating variance components and for their genomic prediction accuracy, using a Flint-Dent maize inbred panel including admixed individuals. Based on simulated traits, both models were accurate to estimate their respective variance components and proved their efficiency to improve genomic prediction accuracy compared to a standard GBLUP model. However the gain in genomic prediction accuracy was very low for real traits, due to a limited contribution of group-specific SNP deviations effects. The benefits of using admixed individuals in multi-group training sets (TSs) was also investigated using this panel. Their interest was confirmed using simulated traits, but was variable using real traits. The discrepancy between the results obtained using the simulated and the real traits may be due to the existence of epistatic interactions between QTLs and the genetic background, as already shown using this dataset. However, both MAGBLUP models and admixed individuals should find interesting applications, especially if the existence of group-specific SNP allele effects is suspected for a given species.

Introduction

Genomic prediction was proposed by Meuwissen et al. (2001) and have since become a central tool in many plant and animal breeding programs. In its simplest application, a set of individuals is evaluated for a given trait and genotyped at a high density using single nucleotide polymorphisms (SNPs). A statistical model is trained on this dataset, referred to as the training set (TS), and is used to predict the breeding value of individuals for whom only genomic information is known, referred to as the predicted set (PS). The breeding values of PS individuals are predicted based on their genomic resemblance with TS individuals by taking advantage of the linkage disequilibrium (LD) between the SNPs and the quantitative trait loci (QTLs) underlying the trait. Several models have been developed making different assumptions on the distributions of QTL allele effects (Heslot et al., 2012), and among them, the GBLUP model is probably the simplest and the most used by geneticists. However, most genomic prediction models, including GBLUP, do not consider explicitly the possible existence of a genetic structure within the population.

When a population is stratified into genetic groups, genomic prediction accuracy can be impacted in different manners. First, when the same structure is observed within the TS and the PS, the group mean differences are implicitly taken into account by the model through the kinship and participate to the accuracy as shown by Guo et al. (2014) and in Chapter 1 (Rio et al., 2019). Conversely, when targeting a group-specific PS, training a model on a different group can decrease accuracy dramatically as shown in several species including dairy and beef cattle (Olson et al., 2012; Chen et al., 2013) and maize (Technow et al., 2013; Lehermeier et al., 2014). The ability to "borrow" genetic information across genetic groups is generally inferior to what can be observed within a given group, but may be substantial when the genetic groups diverged recently (Rio et al., 2019). The combination of genetic groups in a multi-group TS has been proposed to apply predictions to a wide range of genetic diversity (de Roos et al., 2009). This solution is particularly interesting for genetic groups with a limited population size or to optimize resources for traits that are expensive to evaluate, so that a same TS can be used for different group-specific PSs. Such multi-group TSs showed a good predictive ability in a wide range of species such as dairy cattle (Brøndum et al., 2011; Pryce et al., 2011; Zhou et al., 2013), maize (Technow et al., 2013; Rio et al., 2019) or soybean (Duhnen et al., 2017). However, the gain in precision is often limited compared to what could be obtained by applying predictions separately within groups (Carillier et al., 2014; Hayes et al., 2018). Based on simulations, Toosi et al. (2013) showed that including admixed individuals into multi-group TSs should be beneficial. The genome of admixed individuals is a mosaic of chromosome fragments from different group ancestries. These individuals create connections between genetic groups and including them in the TS should improve the properties of multi-group TSs.

Across group prediction often leads to limited performance, which may result from differences in genetic information captured by SNPs. An obvious configuration consists in QTLs segregating only in a given group, which cannot be accounted for when training the model on other groups. Group differences in genetic information captured by SNPs may also be due to group specific SNP effects. Such heterogeneity may result from differences in LD between SNPs and QTLs, as observed in several species including dairy cattle (de Roos et al., 2008) and maize (Technow et al., 2012). They may also be due to the apparition of group-specific genetic mutations nearby QTLs, or to epistatic interactions between QTLs and the genetic background. Such heterogeneity in SNP allele effects was shown in Chapter 2 by studying a "Flint-Dent" maize panel including admixed individuals and evaluated for flowering traits. Specifying these group-specific SNP allele effects in genomic prediction models thus appears as an appealing solution to improve genomic prediction accuracy in admixed populations.

Modeling group specific SNP allele effects in genomic prediction models was proposed by Karoui et al. (2012) and Lehermeier et al. (2015) by adapting multi-trait models to multi-group predictions. In such models, the SNP allele effects are assumed to be different but correlated between groups. Another possibility to account for the heterogeneity of allele effects is to decompose them as a sum between a main SNP effect

and group-specific deviations, as proposed by Schulz-Streeck et al. (2012), de los Campos et al. (2015) or Technow and Totir (2015). While proving their efficiency in structured datasets, these models have been applied to pure individuals only. Before the advent of genomic data, the "animal model", which considers pedigree relationships between individuals, had been adapted to multi-group populations including admixed individuals. The aim was to account for the additional variance observed in an admixed population compared to parental populations, by splitting the genetic variance into group-specific and segregation components using global admixture proportions (Lo et al., 1993; García-Cortés and Toro, 2006). Such methodology was later adapted to genomic prediction by Strandén and Mäntysaari (2013) and Makgahlela et al. (2013) by replacing the pedigree-based kinship matrix by a kinship matrix estimated with SNPs. However, to our knowledge, no model was proposed which accounted both for group specific allele effects and local admixture, in terms of group ancestry of alleles. This information concerning local admixtures is relatively easy to obtain when admixed individuals were generated from a controlled mating design, and can even be inferred in natural admixed populations using softwares such as STRUCTURE (Pritchard et al., 2000), LAMP (Sankararaman et al., 2008) or RFmix (Maples et al., 2013).

In this study, we present two genomic prediction models that meet these characteristics and are easy to implement using the linear mixed model. The two models, called Multi-group Admixed GBLUP (MAGBLUP) 1 and 2, were evaluated for their precision in estimating variance components as well as for their genomic prediction accuracy. Both models were applied to a "Flint-Dent" maize dataset including admixed individuals using simulated traits and real traits. In this study, we also evaluated the benefits of using admixed individuals in multi-group TSs, along with these new models. Different scenarios were investigated by leveraging the proportion of pure and admixed individuals within the TS.

Statistical context

To develop a relevant genomic prediction model, our general strategy was to express an infinitesimal generative model for genetic values into a variance component model. Studying the expected genetic value, the variance and the covariance between genetic values can help to identify which parameters need to be estimated, and which incidence and covariance matrices are required for their estimation. We considered two statistical formalisms that are classically found in the genomic prediction literature and first presented them for GBLUP as an illustration. MAGBLUP 1 was derived according to the first formalism by modeling of the distribution of the genotypes at QTLs, as commonly done in the "animal model" (Henderson, 1984; Kruuk, 2004). MAGBLUP 2 was derived according to the second formalism by modeling the distribution of QTL allele effects, as proposed by (Meuwissen et al., 2001), along with a re-parametrization of QTL allele effects into main effects and group-specific deviations.

GBLUP

Let us consider a population of homozygous inbred lines without stratification into genetic groups. If we suppose a polygenic trait with bi-allelic QTLs and no epistatic interactions among loci, we can model the genetic value of an individual as:

$$G_i = \sum_{m=1}^{M} \left(\beta_m^0 + W_{im}\left(\beta_m^1 - \beta_m^0\right)\right)$$

where G_i is the genetic value of individual i, M is the number of QTLs controlling the trait, W_{im} is the QTL genotype at locus m, taking the value "1" if individual i has the allele 1 and "0" otherwise, β_m^0 and β_m^1 refer to the effects of alleles 0 and 1 at locus m, respectively.

First formalism

According to the first statistical formalism, QTL genotypes are modeled as being drawn in a Bernoulli distribution: $W_{im} \sim \mathcal{B}(f_m)$ where f_m is the frequency of allele 1 at locus m. An absence of LD is assumed between QTLs, which amounts to assuming independence between QTLs.

For a given trait, let $\mathrm{E}(G_i)$ and $\mathrm{cov}(G_i, G_j|\alpha_{ij})$ be the expected genetic value and the genetic covariance assuming a kinship α_{ij} (being formally defined as $\alpha_{ij} = \mathrm{cor}(W_{im}, W_{jm})$ for all m) between individual i and j, One has (see Appendices A and B):

$$\mathrm{E}(G_i) = \mu$$

where the intercept $\mu = \sum_{m=1}^{M} \left(\beta_m^0 + f_m \left(\beta_m^1 - \beta_m^0\right)\right)$ corresponds to the sum of QTL means over all loci, and:

$$\mathrm{cov}(G_i, G_j|\alpha_{ij}) = \alpha_{ij}\sigma_G^2$$

where $\sigma_G^2 = \sum_{m=1}^{M} f_m(1-f_m)\left(\beta_m^1 - \beta_m^0\right)^2$ is the genetic variance, corresponding to the sum of QTL variances over all loci. Note that when $i = j$, the genetic covariance simplifies to the genetic variance $\mathrm{V}(G_i) = \sigma_G^2$.

From this formalism, we can model the phenotypic value of a set of individuals as the sum between a fixed intercept and two random components: a genetic component and an error component including environmental effects as well as other genetic effects (e.g. genotype by environment interactions), the two components being independent of each other:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{g} + \boldsymbol{e} \tag{3.1}$$

where $\boldsymbol{y}$ is the vector of phenotypes, N is the number of individuals, $\mathbf{1}$ is a vector of 1, $\boldsymbol{g}$ is the vector of genetic values and $\boldsymbol{e}$ is the vector of errors. The normality of the errors is classically assumed: $\boldsymbol{e} \sim \mathcal{N}(0, \boldsymbol{I}\sigma_E^2)$ where $\boldsymbol{I}$ is the identity matrix. Assuming an infinitesimal model, the normality of the genetic values is also assumed: $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{K}\sigma_G^2)$ where $\boldsymbol{K}$ is the kinship matrix with $(\boldsymbol{K})_{ij} = \alpha_{ij}$.

To apply this model, the kinship matrix can be computed following VanRaden (2008):

$$(\boldsymbol{K})_{ij} = \frac{\sum_{m=1}^{M}(W_{im} - f_m)(W_{jm} - f_m)}{\sum_{m=1}^{M} f_m(1-f_m)} \tag{3.2}$$

Second formalism

According to the second statistical formalism, QTL allele effects are modeled as being drawn in a normal distribution: $\beta_m^0 \sim \mathcal{N}(0, \sigma_\beta^2)$ independent and identically distributed (IID) and $\beta_m^1 \sim \mathcal{N}(0, \sigma_\beta^2)$ IID, with β_m^0 and $\beta_{m'}^1$ independent for all m and m'

Let $\mathrm{E}(G_i|\boldsymbol{w}_i)$ and $\mathrm{cov}(G_i, G_j|\boldsymbol{w}_i, \boldsymbol{w}_j)$ be the expected genetic value and the covariance between genetic values of individuals i and j. Here the genotypes $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ are assumed to be fixed, and the expectation is computed over an infinite sampling of allele effects. One has (see Appendices C and D):

$$\mathrm{E}(G_i|\boldsymbol{w}_i) = 0$$

and:

$$\mathrm{cov}(G_i, G_j|\boldsymbol{w}_i, \boldsymbol{w}_j) = \phi_{ij}\sigma_U^2$$

where $\sigma_U^2 = M\sigma_\beta^2$ is the variance due to QTL effects and ϕ_{ij} is the identity by state (IBS) between i and j. Note that when $i = j$, the covariance simplifies to the variance $\mathrm{V}(G_i|\boldsymbol{w}_i) = \sigma_U^2$.

The IBS expression is explicitly indicated in the expression of the covariance:

$$\phi_{ij} = \frac{1}{M} \sum_{m=1}^{M} \left((1 - W_{im}) (1 - W_{jm}) + W_{im} W_{jm} \right) \tag{3.3}$$

From this formalism, we can also model the phenotypic value of a set of individuals as being a sum between a genetic component and an error component. While not specified by the generative model, a fixed intercept is generally assumed:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{u} + \boldsymbol{e}$$

where $\boldsymbol{u}$ is the vector of genetic values following a different modeling (see below) from that of Eq. (3.1), while other terms remain unchanged. Here, the normality of genetic values results from that of allele effects: $\boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{\phi}\sigma^2_U)$, where $(\boldsymbol{\phi})_{ij} = \phi_{ij}$ is the IBS matrix.

Note that the variance σ^2_G is different from σ^2_U. Following Gianola et al. (2009), we can show that the expected value of σ^2_G over an infinite sampling of allele effects is:

$$\begin{aligned} \mathrm{E}\left(\sigma^2_G\right) &= 2 \sum_{m=1}^{M} f_m \left(1 - f_m\right) \sigma^2_\beta \\ &= \frac{2}{M} \sum_{m=1}^{M} f_m \left(1 - f_m\right) \sigma^2_U \end{aligned}$$

These two formalisms are tightly linked and both of them are widely used in the GBLUP literature.

MAGBLUP

Let us consider a population of homozygous inbred lines divided into two genetic groups A and B, that also includes admixed lines. If we suppose a polygenic trait with bi-allelic QTLs, whose effects depend both on the allele at the QTL (0/1) and its ancestry or local admixture (A/B), and no epistatic interactions among loci, we can model the genetic value of an individual as:

$$G_i = \sum_{m=1}^{M} \left(A_{imA} \left(\beta^0_{mA} + W_{im} \left(\beta^1_{mA} - \beta^0_{mA} \right) \right) + A_{imB} \left(\beta^0_{mB} + W_{im} \left(\beta^1_{mB} - \beta^0_{mB} \right) \right) \right)$$

where G_i is the genetic value of individual i, M is the number of QTLs controlling the trait, A_{imA} is the allele ancestry at locus m, taking the value "1" if individual i inherited its allele from group A and "0" otherwise, $A_{imB} = 1 - A_{imA}$, W_{im} is the QTL genotype at locus m, taking the value "1" if individual i has the allele 1 and "0" otherwise, β^0_{mA}, β^1_{mA}, β^0_{mB} and β^1_{mB} refer to the effects of allele 0 and 1 in group A and B at locus m respectively.

MAGBLUP 1 - First formalism

According to the first statistical formalism, local ancestries are modeled as being drawn in a Bernoulli distribution: $A_{imA} \sim \mathcal{B}(\pi_i)$ where π_i is the genome proportion that individual i received from group A. QTL genotypes are modeled as being drawn in a Bernoulli distribution conditionally to allele ancestry: $(W_{im}|A_{imA} = 1) \sim \mathcal{B}(f_{mA})$ and $(W_{im}|A_{imB} = 1) \sim \mathcal{B}(f_{mB})$ where f_{mA} and f_{mB} are the frequencies of allele 1 at locus m in group A and group B respectively. An absence of LD is assumed between QTLs, which amounts to supposing independence between QTLs.

For a given trait, let $\mathrm{E}\left(G_i|\pi_i\right)$ and $\mathrm{cov}(G_i, G_j|\pi_i, \pi_j, \theta^A_{ij}, \theta^B_{ij}, \alpha^A_{ij}, \alpha^B_{ij})$ be the expected genetic value and the genetic covariance, assuming a proportion π_i of genome A for i and π_j for j, a proportion of shared

ancestry (or shared admixture) θ_{ij}^A (being formally defined as $\theta_{ij}^A = \mathrm{E}(A_{imA}A_{jmA})$ for all m) between i and j for the genome originated from group A, a proportion of shared ancestry θ_{ij}^B between i and j for the genome originated from group B, a kinship α_{ij}^A (being formally defined as $\alpha_{ij}^A = \mathrm{cor}(W_{im}, W_{jm}|A_{imA} = 1, A_{jmA} = 1)$ for all m) between i and j on their shared ancestries for the genome originated from group A and a kinship α_{ij}^B (being formally defined as $\alpha_{ij}^B = \mathrm{cor}(W_{im}, W_{jm}|A_{imB} = 1, A_{jmB} = 1)$ for all m) between i and j on their shared ancestries for the genome originated from group B (Fig. 3.1). One has (see Appendices E and F):

$$\begin{aligned}\mathrm{E}\left(G_i|\pi_i\right) &= \pi_i \sum_{m=1}^{M} \mu_{mA} + (1-\pi_i) \sum_{m=1}^{M} \mu_{mB} \\ &= \pi_i \mu_A + (1-\pi_i)\mu_B\end{aligned}$$

where $\mu_{mA} = \beta_{mA}^0 + f_{mA}(\beta_{mA}^1 - \beta_{mA}^0)$ and $\mu_{mB} = \beta_{mB}^0 + f_{mB}(\beta_{mB}^1 - \beta_{mB}^0)$ are the means at QTL m in group A and B respectively, μ_A and μ_B are the global intercepts in group A and B respectively, and:

$$\mathrm{cov}(G_i, G_j|\pi_i, \pi_j, \theta_{ij}^A, \theta_{ij}^B, \alpha_{ij}^A, \alpha_{ij}^B) = \Delta_{ij}\sigma_S^2 + \theta_{ij}^A \alpha_{ij}^A \sigma_{G_A}^2 + \theta_{ij}^B \alpha_{ij}^B \sigma_{G_B}^2$$

where $\Delta_{ij} = \theta_{ij}^A - \pi_i\pi_j$ is the covariance between the group A allele ancestries of i and j, $\sigma_S^2 = \sum_{m=1}^{M}(\mu_{mA} - \mu_{mB})^2$ is the segregation variance caused by group-specific means at QTLs, $\sigma_{G_A}^2 = \sum_{m=1}^{M} f_{mA}(1-f_{mA})(\beta_{mA}^1 - \beta_{mA}^0)^2$ and $\sigma_{G_B}^2 = \sum_{m=1}^{M} f_{mB}(1-f_{mB})(\beta_{mB}^1 - \beta_{mB}^0)^2$ are the genetic variances in groups A and B respectively. Note that $\theta_{ij}^B = 1 - \pi_i - \pi_j + \theta_{ij}^A$ and when $i = j$, the covariance simplifies to the variance $\mathrm{V}\left(G_i|\pi_i\right) = \pi_i(1-\pi_i)\sigma_S^2 + \pi_i\sigma_{G_A}^2 + (1-\pi_i)\sigma_{G_B}^2$

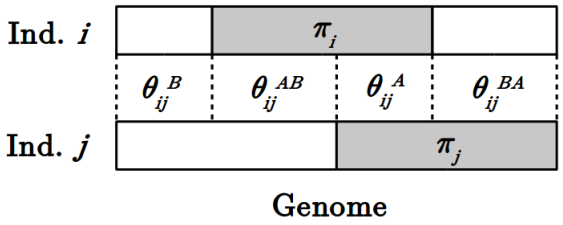


Figure 3.1: Diagram illustrating the genome-wide allele ancestry of two individuals i and j, with a proportion π_i of genome A for i and π_j for j, a proportion of shared ancestry θ_{ij}^A between i and j for the genome originated from group A, a proportion of shared ancestry $\theta_{ij}^B = 1-\pi_i-\pi_j+\theta_{ij}^A$ between i and j for the genome originated from group B, $\theta_{ij}^{AB} = \pi_i - \theta_{ij}^A$ is the proportion of not shared ancestries corresponding to the genome of i originated from group A and of j originated from group B, $\theta_{ij}^{BA} = \pi_j - \theta_{ij}^A$ is the proportion of not shared ancestries corresponding to the genome of i originated from group B and of j originated from group A

From this formalism, we can model the phenotypic value of a set of individuals as the sum of fixed group effects and four random components: an admixture component, two group-specific genetic components and an error component, the four components being independent of each other:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{a} + \boldsymbol{g}_A + \boldsymbol{g}_B + \boldsymbol{e} \tag{3.4}$$

where $\boldsymbol{X} = (\boldsymbol{q}, \boldsymbol{1} - \boldsymbol{q})$ is the incidence matrix for fixed effects with $\boldsymbol{q}$ being the vector of genome proportions for the genome originated from group A, $\boldsymbol{\beta} = (\mu_A, \mu_B)^T$ is the vector of group-specific intercepts, $\boldsymbol{a}$ is the vector of the admixture component of the genetic value, $\boldsymbol{g}_A$ is the vector of the group A component of the genetic value, $\boldsymbol{g}_B$ is the vector of the group B component of the genetic value. All other terms are identical to those described in Eq. (3.1). Assuming an infinitesimal model, the normality of $\boldsymbol{a}$, $\boldsymbol{g}_A$ and $\boldsymbol{g}_B$ is assumed: $\boldsymbol{a} \sim \mathcal{N}(0, \boldsymbol{\Delta}\sigma_S^2)$ where $\boldsymbol{\Delta}$ is the covariance matrix between allele ancestries (or local admixtures) with $(\boldsymbol{\Delta})_{ij} = \Delta_{ij}$, $\boldsymbol{g}_A \sim \mathcal{N}(0, (\boldsymbol{\theta}_A \circ \boldsymbol{K}_A)\sigma_{G_A}^2)$ where $\boldsymbol{\theta}_A$ is the matrix with proportions of shared ancestries for the genome originated from group A with $(\boldsymbol{\theta}_A)_{ij} = \theta_{ij}^A$, $\boldsymbol{K}_A$ is the kinship matrix on shared ancestries for the genome originated from group A with $(\boldsymbol{K}_A)_{ij} = \alpha_{ij}^A$, "$\circ$" refer to the Hadamard product, and $\boldsymbol{g}_B \sim \mathcal{N}(0, (\boldsymbol{\theta}_B \circ \boldsymbol{K}_B)\sigma_{G_B}^2)$ where $\boldsymbol{\theta}_B$ is the matrix with proportions of shared ancestries for the genome

originated from group B with $(\boldsymbol{\theta}_B)_{ij} = \theta^B_{ij}$ and $\boldsymbol{K}_B$ is the kinship matrix on shared ancestries for the genome originated from group B with $(\boldsymbol{K}_B)_{ij} = \alpha^B_{ij}$.

To apply this model, covariance matrices can be estimated as follows:

$$(\boldsymbol{\theta}_A)_{ij} = \frac{1}{M}\sum_{m=1}^{M} A_{imA}A_{jmA}$$

$$(\boldsymbol{\theta}_B)_{ij} = \frac{1}{M}\sum_{m=1}^{M} A_{imB}A_{jmB}$$

$$(\boldsymbol{\Delta})_{ij} = (\boldsymbol{\theta}_A)_{ij} - \widehat{\pi}_i\widehat{\pi}_j$$

$$(\boldsymbol{K}_A)_{ij} = \frac{\sum_{m=1}^{M} A_{imA}\left(W_{im} - \widehat{f}_{mA}\right)A_{jmA}\left(W_{jm} - \widehat{f}_{mA}\right)}{\sum_{m=1}^{M} A_{imA}A_{jmA}\widehat{f}_{mA}(1-\widehat{f}_{mA})}$$

$$(\boldsymbol{K}_B)_{ij} = \frac{\sum_{m=1}^{M} A_{imB}\left(W_{im} - \widehat{f}_{mB}\right)A_{jmB}\left(W_{jm} - \widehat{f}_{mB}\right)}{\sum_{m=1}^{M} A_{imB}A_{jmB}\widehat{f}_{mB}(1-\widehat{f}_{mB})}$$

where $\widehat{\pi}_i = \frac{1}{M}\sum_{m=1}^{M} A_{imA}$, $\widehat{f}_{mA} = \frac{\sum_{i=1}^{N} A_{imA}W_{im}}{\sum_{i=1}^{N} A_{imA}}$ and $\widehat{f}_{mB} = \frac{\sum_{i=1}^{N} A_{imB}W_{im}}{\sum_{i=1}^{N} A_{imB}}$ refer to the empirical estimates of π_i, f_{mA} and f_{mB} respectively. Note that the estimators of kinship matrices were proposed in analogy with Eq. (3.2), although other other estimators could have been used for all covariance matrices.

MAGBLUP 2 - Re-parametrization of allele effects and second formalism

Using the second formalism, it is possible to allow for genetic covariance between individuals from different groups by applying a re-parametrization of allele effects into a main QTL effect and group-specific deviations:

- $\beta^0_{mA} = \gamma^0_m + \delta^0_{mA}$
- $\beta^1_{mA} = \gamma^1_m + \delta^1_{mA}$
- $\beta^0_{mB} = \gamma^0_m + \delta^0_{mB}$
- $\beta^1_{mB} = \gamma^1_m + \delta^1_{mB}$

where $\gamma^0_m \sim \mathcal{N}(0, \sigma^2_\gamma)$ independent and identically distributed (IID), $\gamma^1_m \sim \mathcal{N}(0, \sigma^2_\gamma)$ IID, $\delta^0_{mA} \sim \mathcal{N}(0, \sigma^2_{\delta_A})$ IID, $\delta^1_{mA} \sim \mathcal{N}(0, \sigma^2_{\delta_A})$ IID, $\delta^0_{mB} \sim \mathcal{N}(0, \sigma^2_{\delta_B})$ IID and $\delta^1_{mB} \sim \mathcal{N}(0, \sigma^2_{\delta_B})$ IID, with σ^2_γ, $\sigma^2_{\delta_A}$ and $\sigma^2_{\delta_B}$ being the QTL effect variance for the main QTL effects (γ^0_m and γ^1_m) and for the deviation in group A (δ^0_{mA} and δ^1_{mA}) and B (δ^0_{mB} and δ^1_{mB}) respectively. All types of random effects are assumed to be independent from each other.

The genetic values of an individual can then be written as:

$$G_i = \sum_{m=1}^{M}\left(\gamma^0_m + W_{im}\left(\gamma^1_m - \gamma^0_m\right) + A_{imA}\left(\delta^0_{mA} + W_{im}\left(\delta^1_{mA} - \delta^0_{mA}\right)\right) + A_{imB}\left(\delta^0_{mB} + W_{im}\left(\delta^1_{mB} - \delta^0_{mB}\right)\right)\right) \tag{3.5}$$

Let $\mathrm{E}\left(G_i|\boldsymbol{a}_i, \boldsymbol{w}_i\right)$ and $\mathrm{cov}(G_i, G_j|\boldsymbol{a}_i, \boldsymbol{a}_j, \boldsymbol{w}_i, \boldsymbol{w}_j)$ be the expected genetic value and the covariance between genetic values of individuals i and j. Here the genotypes $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ and the allele ancestries in group A $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ are assumed to be fixed, and the expectation is computed over an infinite sampling of allele effects. One has (see Appendices G and H):

$$\mathrm{E}\left(G_i|\boldsymbol{a}_i, \boldsymbol{w}_i\right) = 0$$

and:

$$\text{cov}(G_i, G_j | \boldsymbol{a}_i, \boldsymbol{a}_j, \boldsymbol{w}_i, \boldsymbol{w}_j) = \phi_{ij}\sigma^2_U + \phi^A_{ij}\sigma^2_{U_A} + \phi^B_{ij}\sigma^2_{U_B}$$

where ϕ_{ij} is the IBS between i and j (Eq. 3.3), ϕ^A_{ij} and ϕ^A_{ij} are the IBS between i and j on shared ancestries for the genome originated from group A and B over the total number of loci respectively, $\sigma^2_U = M\sigma^2_\gamma$ is the variance component due to main QTL effects, $\sigma^2_{U_A} = M\sigma^2_{\delta_A}$ and $\sigma^2_{U_B} = M\sigma^2_{\delta_B}$ are the variance component due to QTL deviation effects in group A and B respectively. Note that when $i = j$, the covariance simplifies to the variance $\text{V}(G_i|\boldsymbol{a}_i, \boldsymbol{w}_i) = \sigma^2_U + \pi_i\sigma^2_{U_A} + (1-\pi_i)\sigma^2_{U_B}$.

The expression of the IBS are explicitly indicated in the expression of the covariance:

$$\phi^A_{ij} = \frac{1}{M}\sum_{m=1}^{M} A_{imA}A_{jmA}\left((1-W_{im})(1-W_{jm}) + W_{im}W_{jm}\right)$$

$$\phi^B_{ij} = \frac{1}{M}\sum_{m=1}^{M} A_{imB}A_{jmB}\left((1-W_{im})(1-W_{jm}) + W_{im}W_{jm}\right)$$

From this formalism, we can model the phenotypic value of a set of individuals as the sum between four components, one genetic component that is due to main QTL effects, two genetic components that are due to group-specific deviation effects, and an error component. Like with GBLUP, a fixed intercept can be assumed although not explicitly specified by the generative model:

$$\boldsymbol{y} = \boldsymbol{1}\mu + \boldsymbol{u} + \boldsymbol{u}_A + \boldsymbol{u}_B + \boldsymbol{e} \tag{3.6}$$

$\boldsymbol{u}$ is the vector of the genetic component that is due to main QTL effects, $\boldsymbol{u}_A$ is the vector of the genetic component that is due to QTL deviation effects in group A, $\boldsymbol{u}_B$ is the vector of the genetic component that is due to QTL deviation effects in group B. All other terms are identical to those described in Eq. (3.1). Here, the normality of the three genetic components results from those of allele effects: $\boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{\phi}\sigma^2_U)$ where $(\boldsymbol{\phi})_{ij} = \phi_{ij}$, $\boldsymbol{u}_A \sim \mathcal{N}(0, \boldsymbol{\phi}_A\sigma^2_{U_A})$ where $(\boldsymbol{\phi}_A)_{ij} = \phi^A_{ij}$ and $\boldsymbol{u}_B \sim \mathcal{N}(0, \boldsymbol{\phi}_B\sigma^2_{U_B})$ where $(\boldsymbol{\phi}_B)_{ij} = \phi^B_{ij}$

Material and Methods

Flint-Dent dataset

The "Flint-dent" panel was the one already presented in Chapter 2. It consists of 970 maize inbred lines including 300 pure dent, 304 pure flint and 366 admixed lines which were genotyped for 482,013 polymorphic SNPs. For all individuals, SNP alleles (coded 0/1) and alleles ancestries (dent or flint) were known. The panel was evaluated in two trials for five traits: male flowering (MF) and female flowering (FF) in calendar days after sowing, plant height (PH) in centimeters, ear leaf number (ELN) and total number of leaves (TNL). The phenotypic analysis for flowering traits was presented in Chapter 2. The same procedure was applied for the three remaining traits and the results were presented for all traits in Supplementary Table S3.1. Least-square means were computed over the whole design and were further referred to as phenotypes.

Statistical inference and genomic predictions

The genomic prediction models presented in the previous section were adapted to the Flint-Dent dataset where group A referred to the dent group (D) and group B referred to the flint group (F). The three models considered were: GBLUP as defined in Eq. (3.1) using a kinship matrix computed following Eq. (3.2), MAGBLUP 1 as defined in Eq. (3.4) using the covariances matrices described with the model, and MAGBLUP 2 as defined in Eq. (3.6) using the three IBS matrices described with the model. For all models,

the inference of parameters was done using the R-package MM4LMM (Laporte et al., 2019). Genomic predictions were computed as BLUPs (Searle et al., 2008) of the phenotypic values.

Simulated traits

Phenotypic traits were simulated to study the precision of MAGBLUP 1 and 2 in terms of variance estimates and genomic predictions. Genetic values were simulated using the generative model presented in Eq. (3.5). Three different types of genetic determinism were defined concerning QTL allele effects and were summarized in Table 3.1. Type A referred to a trait with only group-specific QTL deviations effects, type B referred to a trait with only main QTL allele effects, while type C included all types of effects. 1,000 loci were sampled among all SNPs to be used as QTLs. Allele effects were sampled independently in normal distributions with a variance defined by the type of genetic determinism. For given genotypes and allele ancestries, the genetic value of each individual was computed as a sum of allele effects according to both the allele and its ancestry. Residuals were sampled in a normal distribution $\mathcal{N}\left(0, \sigma^2_E\right)$, with σ^2_E chosen to reach a heritability of 0.8.

Table 3.1: Variances of allele effects for the three types of genetic determinism

Genetic determinism	σ^2_γ	$\sigma^2_{\delta_D}$	$\sigma^2_{\delta_F}$
A	0	1	3
B	2	0	0
C	2	1	3

Assessment of the precision of variances estimates

The precision of the genetic variance component estimates was evaluated for both MAGBLUP 1 and MAGBLUP 2. The 1,000 SNPs sampled to be used as QTLs were later considered as the only genotypic information available to build the covariance matrices for variances estimation.

The variance components of MAGBLUP 1 were defined conditionally to allele effects. To assess the precision of variance estimates, one sample of allele effects was simulated for each type of genetic determinism, for a total of three simulated traits. Admixed and composite population samples were simulated based on real genotypic data. The admixed population samples included 366 admixed lines which were simulated by generating gametes from hybrids. The hybrids were simulated by randomly sampling real dent and flint lines to be mated. The SNPs were located on a genetic map for which the genetic distance between pairs of markers was calculated as being proportional to their physical distance, and the scale parameter was determined relative to chromosome 1 (200 cM for around 300 Mbp). For each chromosome, the recombination breakpoints were sampled in a Poisson distribution with parameter λ equal to the length of the chromosome in Morgan. The composite population samples included 300 pure dent lines, 304 pure flint lines and 366 admixed lines. The dent lines were simulated by randomly sampling each chromosome from all existing versions of the given chromosome within the real dent lines. The 304 flint lines were simulated in an equivalent manner. The 366 admixed lines were simulated as described above but using the simulated dent and flint lines as parents. This procedure was repeated and led to 1,000 population samples of 366 admixed lines and 1,000 population samples of 970 lines including dent, flint and admixed lines. The phenotypes of the individuals were simulated following the procedure described in the previous section. The covariance matrices were computed using the genotypic and allele ancestry information of the simulated dataset. The variance estimates were compared to the three reference variances: σ^2_S, $\sigma^2_{G_D}$ and $\sigma^2_{G_F}$. Each reference variance was computed using the simulated allele effects and the reference allele frequencies that were estimated on the real dataset using all 970 individuals.

The variance components of MAGBLUP 2 were defined conditionally to genotypes and allele ancestries.

To assess the precision of variance estimates, 1,000 samples of alleles effects were simulated for each type of genetic determinism, forming 3,000 simulated traits which were used as replicates. Two populations samples were used to estimate variances: admixed including the 366 real admixed lines and composite including all 970 real lines. The estimates were compared to the three reference variances: σ^2_U, $\sigma^2_{U_D}$ and $\sigma^2_{U_F}$. Each variance was computed using the variances of allele effects and the number of QTLs.

The variance components of GBLUP, MAGBLUP 1 and MAGBLUP 2 were estimated for the five real traits using the whole dataset.

Assessment of the accuracy of genomic predictions

The genomic prediction accuracy of MAGBLUP 1 and 2 were compared to that of GBLUP using two different cross-validation (CV) procedures based simulated traits and on real genotypic data.

The first CV procedure was called averaged holdout (HO) and was applied to 150 simulated traits, 50 for each type of genetic determinism (50 samples of QTL allele effects), and real traits. The dataset was split with proportions $\frac{4}{5}$ and $\frac{1}{5}$ for the TS and the PS, respectively. For the simulated traits, the splitting was done 20 times and the accuracy of genomic predictions was averaged over repetitions. The accuracy was computed by correlating the predicted genetic values of the PS individuals to their true genetic value. For the real traits, the splitting was done 100 times and the predictive ability was averaged over repetitions. The predictive ability was computed by correlating the predicted genetic values of the PS individuals to their phenotypic value.

The second CV procedure was called structured holdout (SHO). It allowed us to test the impact of the composition of the TS in terms of group origin (and more particularly the interest of including admixed individuals) as well as the efficiency of the different prediction models (GBLUP, MAGBLUP 1 and 2). This procedure was applied to 50 traits simulated according to genetic determinism C (50 samples of QTL allele effects) and real traits. It considered samples of restricted sizes where 90 individuals were predicted using a model trained on 210 other individuals. Those numbers were chosen in order to fit with all the scenarios described in Table 3.2. All the scenarios were designated as TS_PS, TS and PS referring to the genetic backgrounds (dent (D), flint (F), admixed (A)) represented in the TS and the PS respectively. When there was more than one genetic background in the TS or the PS, the composition was always perfectly balanced between them. As an example, DFA_A referred to a TS equally composed of individuals from the three genetic backgrounds and a PS composed of admixed lines only. In scenarios where only a dent or flint genetic background was found in the TS (D_D, F_F, D_A and F_A), only GBLUP could be evaluated. In configurations where admixed individuals were absent of the TS (DF_D, DF_F and DF_A), the admixture term "$\boldsymbol{a}$" of MAGBLUP 1 was removed as its variance component could not be estimated. For the simulated traits, the splitting was done 20 times and the accuracy of genomic predictions was averaged over repetitions. For the real traits, the splitting was done 100 times and the predictive ability was averaged over repetitions.

Results

Variance estimates for simulated traits

MAGBLUP 1 and 2 were both evaluated for their precision in variance component estimation using simulated traits.

As the variance components of MAGBLUP 1 were defined conditionally to allele effects, the precision of their estimation was evaluated by simulating a single sample of QTL allele effects for each type of genetic

Table 3.2: Scenarios evaluated with the structure-based CV scenarios (SHO) where 90 individuals are predicted by 210 other individuals. The TS and the PS were balanced considering their composition in genetic backgrounds (dent (D), flint (F) and admixed(A))

Scenario	**TS composition**	**PS composition**
DFA_DFA	$\frac{1}{3}$D + $\frac{1}{3}$F + $\frac{1}{3}$A	$\frac{1}{3}$D + $\frac{1}{3}$F + $\frac{1}{3}$A
DFA_D	$\frac{1}{3}$D + $\frac{1}{3}$F + $\frac{1}{3}$A	D
D_D	D	D
DF_D	$\frac{1}{2}$D + $\frac{1}{2}$F	D
DA_D	$\frac{1}{2}$D + $\frac{1}{2}$A	D
FA_D	$\frac{1}{2}$F + $\frac{1}{2}$A	D
F_D	F	D
A_D	A	D
DFA_F	$\frac{1}{3}$D + $\frac{1}{3}$F + $\frac{1}{3}$A	F
F_F	F	F
DF_F	$\frac{1}{2}$D + $\frac{1}{2}$F	F
FA_F	$\frac{1}{2}$F + $\frac{1}{2}$A	F
DA_F	$\frac{1}{2}$D + $\frac{1}{2}$A	F
D_F	D	F
A_F	A	F
DFA_A	$\frac{1}{3}$D + $\frac{1}{3}$F + $\frac{1}{3}$A	A
A_A	A	A
DF_A	$\frac{1}{2}$D + $\frac{1}{2}$F	A
DA_A	$\frac{1}{2}$D + $\frac{1}{2}$A	A
FA_A	$\frac{1}{2}$F + $\frac{1}{2}$A	A
D_A	D	A
F_A	F	A

determinism, and 1,000 admixed or composite population samples used as replicates. The estimates are summarized in Table 3.3. The three variance components were generally well estimated for the three types of genetic determinism. The only exception was the segregation variance σ^2_S which was overestimated for genetic determinism A using both types of training population sample. Using a composite population sample of 970 lines generally led to more accurate estimates of variance components than using an admixed population sample of 366 lines.

As the variance components of MAGBLUP 2 were defined conditionally to genotypes and alleles ancestries, the precision of their estimation was evaluated by simulating 1,000 samples of QTL allele effects for each type of genetic determinism. Variances are summarized in Table 3.4. The variance components were well-estimated in average for the three types of genetic determinism. Here also, using all 970 lines generally led to more accurate estimates of variance components than using the 366 admixed lines.

Genomic prediction accuracy for simulated traits

The three models were first compared for their genomic prediction by CV (HO method) applied to 150 simulated traits, 50 for each type of genetic determinism, using real genotypic data. This procedure was

Table 3.3: Average of variances estimated by MAGBLUP 1 for a trait simulated according to each type of genetic determinism (see Table 3.1), compared to their reference value. 1,000 admixed and composite population samples were simulated, including 366 admixed and 970 lines respectively, and were used as replicates

Genetic determinism		Variance	Reference	Admixed	Composite
A		σ^2_S	2723.6	3604.6 (557.5)	3522.9 (451.6)
A	(deviations effects only)	$\sigma^2_{G_D}$	300.5	326.9 (175.2)	311.3 (47.4)
A		$\sigma^2_{G_F}$	1089.2	1179.2 (243.4)	1100.9 (94.7)
B		σ^2_S	164.3	237.0 (114.7)	213.1 (85.6)
B	(main effects only)	$\sigma^2_{G_D}$	627.4	631.9 (149.6)	647.9 (60.7)
B		$\sigma^2_{G_F}$	676.8	712.0 (148.9)	711.1 (64.9)
C		σ^2_S	2954.9	2810.8 (603.8)	2843.5 (581.4)
C	(all types of effects)	$\sigma^2_{G_D}$	964.1	856.4.8 (321.3)	939.4 (121.4)
C		$\sigma^2_{G_F}$	1553.5	1559.0 (369.0)	1518.9 (170.5)

Standard deviations computed over the 1,000 replicates are shown between brackets

Table 3.4: Average of variances estimated by MAGBLUP 2 using the real admixed and composite population samples, including all 366 admixed and all 970 lines respectively, and compared to their reference value. 1,000 traits were simulated according to each type of genetic determinism (see Table 3.1), and were used as replicates

Genetic determinism		Variance	Reference	Admixed	Composite
A		σ^2_U	0	176.2 (250.7)	94.7 (132.4)
A	(deviations effects only)	$\sigma^2_{U_D}$	1000	889.7 (495.4)	901.1 (216.0)
A		$\sigma^2_{U_F}$	3000	2894.1 (671.7)	2919.8 (340.2)
B		σ^2_U	2000	1946 (295.4)	1986.8 (170.2)
B	(main effects only)	$\sigma^2_{U_D}$	0	35.7 (75.6)	19.5 (40.2)
B		$\sigma^2_{U_F}$	0	38.2 (83.0)	22.0 (43.4)
C		σ^2_U	2000	1985.6 (801.3)	1991.2 (439.7)
C	(all types of effects)	$\sigma^2_{U_D}$	1000	1102.4 (825.4)	1010.8 (459.2)
C		$\sigma^2_{U_F}$	3000	2964.1 (1006.2)	3008.0 (546.6)

Standard deviations computed over the 1,000 replicates are shown between brackets

applied both to all 970 lines and to all 366 admixed lines. Boxplots of average accuracies are presented in Fig. 3.2. The average genomic prediction accuracy was higher within the 970 individuals than within the 366 admixed lines. MAGBLUP 1 and 2 outperformed GBLUP for the genetic determinism A and C, for which group specific QTL deviations effects were simulated. For instance, using all 970 individuals and considering genetic determinism A, a mean accuracy of 0.78 was obtained for MAGBLUP 1 and 2 compared to 0.68 for GBLUP. However, GBLUP and MAGBLUP 2 outperformed MAGBLUP 1 when considering genetic determinism B, for which only main QTL allele effects were simulated. For instance, using all 970 individuals, a mean accuracy of 0.69 was obtained for MAGBLUP 2 and GBLUP, compared to 0.67 for MAGBLUP 1.

The three models were then compared for their genomic prediction accuracy using the SHO CV scenarios which aimed at evaluating the impact of the composition of the TS in terms of genetic backgrounds (dent (D), flint(F) and admixed (A)) to predict a given PS. This procedure was applied to 50 simulated traits simulated according to genetic determinism C, and accuracies are summarized in Table 3.5.

Considering GBLUP, the highest average accuracy was obtained for scenario DFA_DFA, for which the PS and TS composition were balanced between the three genetic backgrounds. To predict a specific genetic background, the highest accuracies were achieved when the TS was trained on individuals from the same genetic background. On average, a higher level of accuracy was observed for predictions within the flint

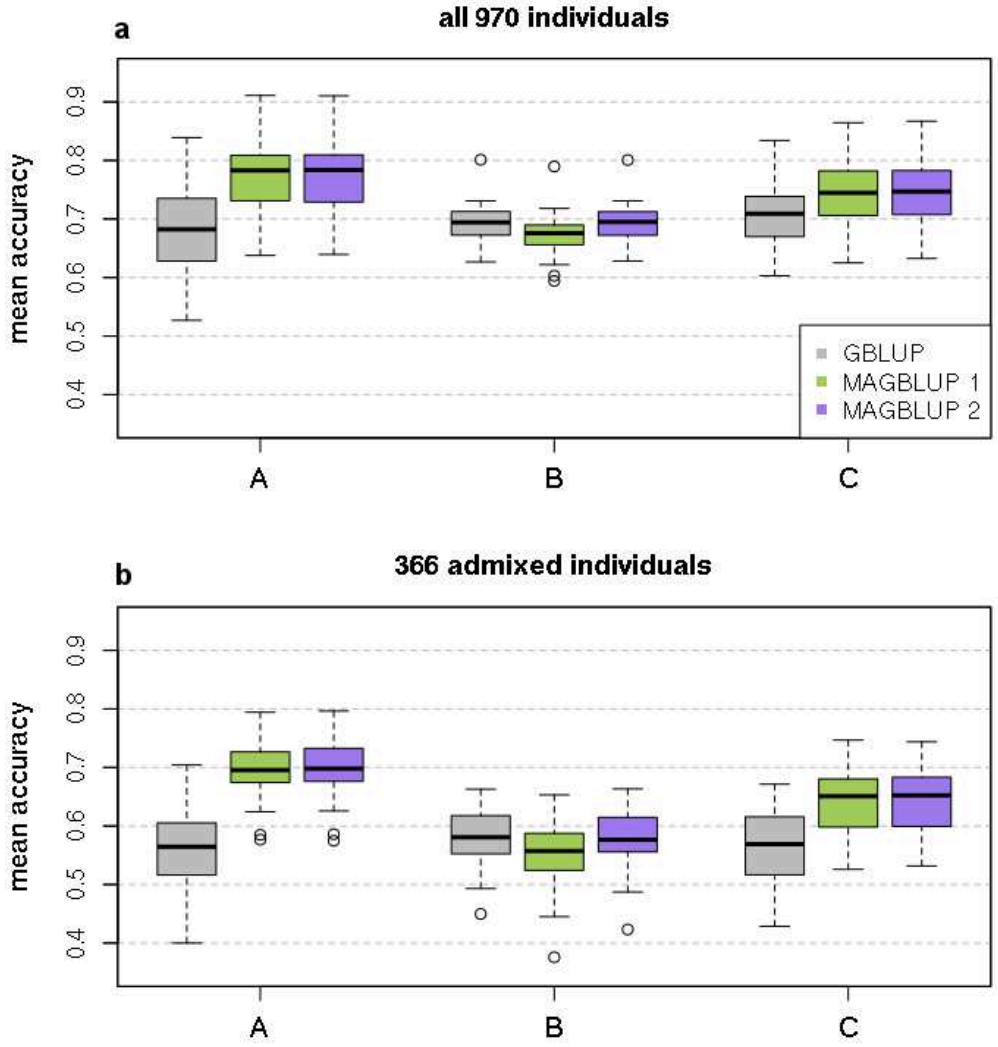


Figure 3.2: Boxplots of average accuracies over 20 CV replicates (HO method) obtained using GBLUP, MAGBLUP 1 and MAGBLUP 2 for 50 traits simulated according to each types of genetic determinism (A, B and C, see Table 3.1), using **a**. all 970 lines, or **b**. all 366 admixed lines

lines (0.54 for F_F), compared to predictions within the dent lines (0.49 for D_D), while admixed lines were intermediate (0.52 for A_A). To predict a dent PS, replacing half of the dent lines of the TS by admixed lines (DA_D) depreciated less the average accuracy (0.45) than replacing them by flint lines (0.41 for DF_D). To predict a flint PS, the same observation could be made as replacing half of the flint lines from the TS by admixed lines (0.51 for FA_F) depreciated less the accuracy than replacing them by dent lines (0.44 for DF_F). When a dent PS was predicted using a TS including only flint lines (F_D), a low average accuracy was observed (0.13) while a higher accuracy was obtained when using only admixed lines (0.41 for A_D), or both both admixed and flint lines (0.32 for FA_D). Similarly, when a flint PS was predicted using a TS including only dent lines (D_F), a low average accuracy was observed (0.09) while a higher accuracy was obtained when using only admixed lines (0.48 for A_F), or both admixed and dent lines (0.39 for DA_F). Predicting a dent PS using a TS balanced between the three genetic backgrounds (DFA_D) led to a similar accuracy (0.40) as using a TS composed of dent and flint lines (0.41 for DF_D), or with only admixed lines (0.41 for A_D). Note that the relative contribution of the dent and flint genetic groups to the TS was similar for the three scenarios mentioned. Conversely, a higher average accuracy was observed when predicting a flint PS using admixed individuals (0.48 for A_F), compared to a TS including the three genetic backgrounds (0.46 for DFA_ F) or including only flint and dent lines (0.44 for DF_F).

Table 3.5: Average of accuracies over 50 traits simulated according to genetic determinism C (see Table 3.1) and 20 CV replicates (SHO method), obtained using GBLUP, MAGBLUP 1 and MAGBLUP 2

	GBLUP	MAGBLUP 1	MAGBLUP 2
DFA_DFA	0.57 (0.11)	0.60 (0.12)	0.60 (0.12)
DFA_D	0.40 (0.08)	0.42 (0.09)	0.42 (0.09)
D_D	0.49 (0.07)	-	-
DF_D	0.41 (0.08)	0.42 (0.09)	0.42 (0.08)
DA_D	0.45 (0.07)	0.47 (0.08)	0.47 (0.08)
FA_D	0.32 (0.08)	0.33 (0.09)	0.34 (0.09)
F_D	0.13 (0.13)	-	-
A_D	0.41 (0.07)	0.41 (0.08)	0.42 (0.08)
DFA_F	0.46 (0.10)	0.48 (0.08)	0.48 (0.09)
F_F	0.54 (0.08)	-	-
DF_F	0.44 (0.11)	0.46 (0.09)	0.47 (0.09)
FA_F	0.51 (0.09)	0.52 (0.08)	0.53 (0.08)
DA_F	0.39 (0.11)	0.41 (0.09)	0.42 (0.09)
D_F	0.09 (0.13)	-	-
A_F	0.48 (0.09)	0.49 (0.08)	0.50 (0.08)
DFA_A	0.48 (0.04)	0.54 (0.06)	0.55 (0.05)
A_A	0.52 (0.06)	0.59 (0.07)	0.60 (0.06)
DF_A	0.40 (0.05)	0.41 (0.05)	0.41 (0.05)
DA_A	0.47 (0.05)	0.54 (0.06)	0.55 (0.05)
FA_A	0.49 (0.04)	0.56 (0.05)	0.57 (0.05)
D_A	0.29 (0.07)	-	-
F_A	0.39 (0.08)	-	-

Standard deviations over the 50 average accuracies (computed over 20 CV replicates) are shown between brackets

"-" indicated that a model could not be applied for the given configuration

When predicting admixed lines, using an admixed TS (A_A) led to higher accuracies (0.52) than using all genetic background in the TS (0.48 for DFA_A), or replacing half of the admixed lines by dent (0.47 for DA_A) or flint lines (0.49 for FA_A). When no admixed lines was present in the TS, the average accuracy was depreciated down to 0.29 by using a dent TS (D_A). For a given size of TS, training the GBLUP model on admixed individuals allowed high accuracies no matter the target PS, and was a better option than just combining flint and dent lines into a multi-group TS.

MAGBLUP 1 and 2 were considered as an alternative to GBLUP with the exception of the scenarios for which only dent or flint lines were included in the TS. Both models led to small increases in average accuracies when predicting dent or flint lines. For instance, in scenario DFA_D, the average accuracy was 0.40 for GBLUP and 0.42 for MAGBLUP 1 and 2. As expected, the gain in accuracy was higher when MAGBLUP 1 and 2 were used to predict admixed lines. For instance, in scenario DFA_A, the average accuracy was 0.48 for GBLUP, 0.54 for MAGBLUP 1 and 0.55 MAGBLUP 2. The average accuracy of MAGBLUP 2 was often slightly higher to that of MAGBLUP 1.

Application to real traits

Variance components were estimated using the three models for five traits and are summarized in Table 3.6. The global genetic variance σ^2_G estimated using GBLUP could be compared to the group specific genetic variances $\sigma^2_{G_D}$ and $\sigma^2_{G_F}$ estimated using MAGBLUP 1. For all traits but MF, σ^2_G was larger than $\sigma^2_{G_D}$ and $\sigma^2_{G_F}$. For instance, σ^2_G was estimated at 19.51 for FF while $\sigma^2_{G_D}$ and $\sigma^2_{G_F}$ were estimated at 17.69 and 15.99 respectively. The segregation variance estimates σ^2_S were always smaller than group-specific genetic variances for all traits, but were substantial especially for PH. Considering MAGBLUP 2, the variance component due to main QTL effects σ^2_U was always larger than the variances due to group-specific deviations effects $\sigma^2_{U_D}$ and $\sigma^2_{U_F}$, which suggested a minor contribution of group-specific deviations effects within this dataset. For instance, σ^2_U was estimated as being equal to 37.90 for MF, while it was estimated as being equal to 2.64 and 0.00 for $\sigma^2_{U_D}$ and $\sigma^2_{U_F}$ respectively. The variance component that is due to flint deviations effects was lower than the component due to dent deviations effects for all traits but TNL. Residual variance estimates were comparable between models for all traits.

Table 3.6: Variance of real traits estimated by GBLUP, MAGBLUP 1 and MAGBLUP 2 using all 970 lines

	Type	MF	FF	PH	ELN	TNL
GBLUP	σ^2_G	13.95	19.51	640.35	1.22	1.74
	σ^2_E	3.77	2.89	114.97	0.32	0.46
MAGBLUP 1	σ^2_S	3.56	4.60	312.28	0.49	0.43
	$\sigma^2_{G_D}$	14.36	17.69	583.61	1.16	1.52
	$\sigma^2_{G_F}$	12.10	15.99	487.25	1.11	1.71
	σ^2_E	3.63	3.46	131.00	0.32	0.46
MAGBLUP 2	σ^2_U	37.90	48.89	1567.70	3.29	4.60
	$\sigma^2_{U_D}$	2.64	2.38	246.96	0.27	0.12
	$\sigma^2_{U_F}$	0.00	0.01	0.07	0.02	0.33
	σ^2_E	3.81	3.65	127.53	0.32	0.47

The three models were compared for their predictive ability using the HO CV scenarios applied to the five real traits. Boxplots of predictive abilities are presented in Fig. 3.3. Consistent with simulated traits, the average predictive ability was higher within the 970 individuals than within the 366 admixed individuals. Lower predictive abilities were obtained for PH compared to the four other traits. The three models led to very similar predictive abilities no matter the trait and the population evaluated. Considering the 970 individuals, MAGBLUP 2 always led to slightly higher predictive abilities compared to GBLUP, itself leading to higher predictive abilities than MAGBLUP 1. For instance, with FF, the average predictive ability was equal to 0.769 for MAGBLUP 2, to 0.759 for GBLUP and 0.758 for MAGBLUP 1. Considering the 366 admixed individuals, MAGBLUP 2 and GBLUP led to very similar predictive abilities while MAGBLUP 1 was always slighly lower. For instance, with FF, the average predictive ability was equal to 0.438 for MAGBLUP 2, to 0.440 for GBLUP and 0.429 for MAGBLUP 1. In conclusion, the three prediction models showed similar performances for these five traits.

The three models were then compared for their predictive ability using the SHO CV scenarios applied to the five traits. The predictive abilities obtained using GBLUP are summarized in Table 3.7. The predictive abilities obtained using MAGBLUP 1 and 2 are summarized in Supplementary Table S3.2 and S3.3 respectively. For most traits and scenarios, GBLUP and MAGBLUP 2 reached very similar levels of predictive abilities and were superior to those obtained using MAGBLUP 1.

When focusing on GBLUP, the highest predictive abilities were obtained for scenario DFA_DFA which was consistent with the SHO results on simulated traits. For all traits but PH, higher predictive abilities were observed when predicting within the dent (D_D) or the flint lines (F_F) than within the admixed lines (A_A). For instance, average predictive abilities of 0.70 and 0.69 were obtained for MF using D_D

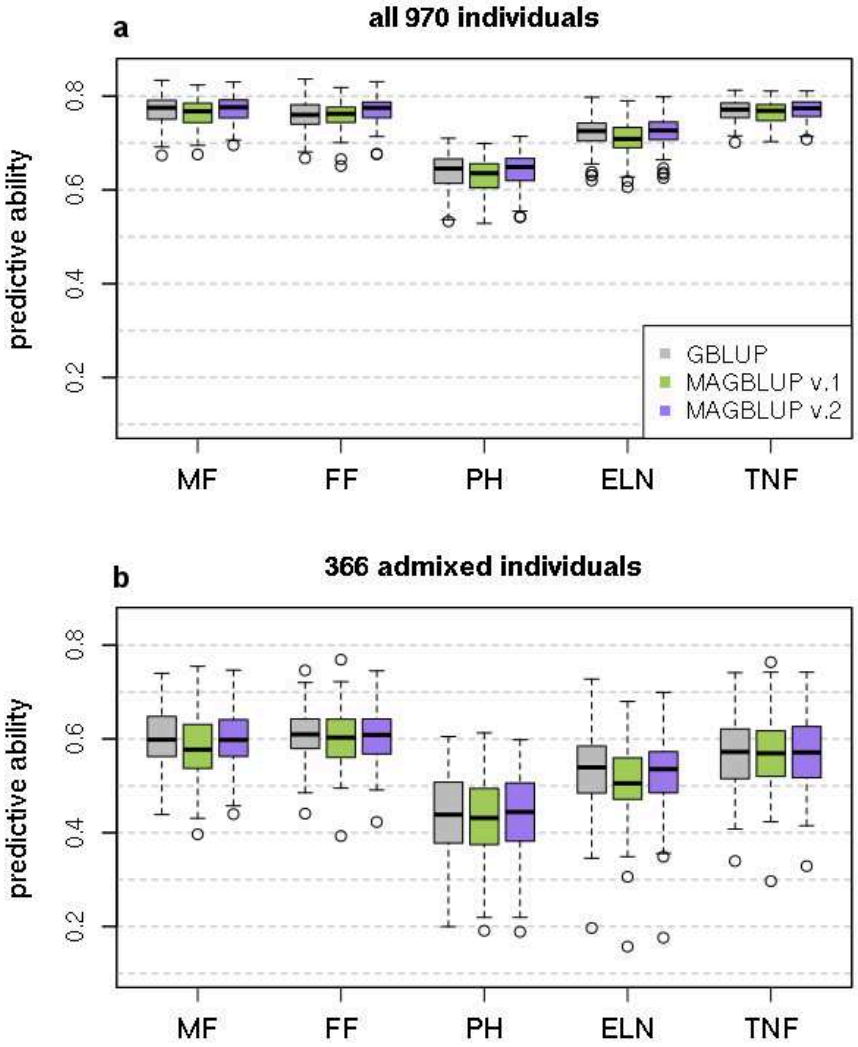


Figure 3.3: Boxplots of predictive abilities obtained by CV (HO method) for GBLUP, MAGBLUP 1 and MAGBLUP 2 on real traits and by considering **a**. all 970 lines, or **b**, all 366 admixed lines

and F_F respectively, compared to 0.55 using A_A. Contrary to what was observed on simulated traits, applying genomic predictions within a given genetic background did not always lead to the highest predictive abilities. For instance, when a flint PS was predicted using flint lines for PH (F_F), the average predictive ability was lower (0.37) than when using both flint and admixed lines (FA_F with 0.41). Like for simulated traits, replacing half of the dent lines of the TS by admixed lines to predict a dent PS (DA_D) generally depreciated less the average predictive ability than replacing them by flint lines (DF_D). For instance, an average predictive ability of 0.68 was obtained using DA_D for MF compared to 0.66 using DF_D. Similar trends could be observed when predicting a flint PS, as replacing half of the flint lines of the TS by admixed lines (FA_F) generally depreciated less the predictive ability compared to using dent lines (DF_F). To predict a dent PS, the lowest predictive abilities were achieved using a flint TS (F_D), as observed for MF (0.33). Similar trends were observed when predicting flint lines using a dent TS (D_F) but with a higher level of predictive ability, like with MF (0.60). A strong dissymmetry is thus observed between the dent and flint lines, as flint lines are well predicted by dent lines whereas the opposite is less true. Unlike with simulated traits, predicting a dent PS using a TS including dent and flint lines (DF_D) led to a higher average accuracy than using an admixed TS (A_D) or a TS including all genetic backgrounds (DFA_D). For instance, when considering MF, an average predictive ability of 0.66 was observed using DF_D compared to 0.60 using A_D and 0.63 using DFA_D. The same trend was not always observed when predicting a flint PS as a higher average predictive ability was observed for PH using an admixed TS (0.41 for A_F) compared to using a TS

Table 3.7: Average of predictive abilities over 100 CV replicates (SHO method) for the five traits using GBLUP

	MF	FF	PH	ELN	TNL
DFA_DFA	0.76 (0.04)	0.75 (0.04)	0.57 (0.06)	0.70 (0.05)	0.76 (0.04)
DFA_D	0.63 (0.06)	0.65 (0.05)	0.41 (0.08)	0.49 (0.06)	0.56 (0.06)
D_D	0.70 (0.04)	0.72 (0.04)	0.52 (0.06)	0.54 (0.05)	0.59 (0.05)
DF_D	0.66 (0.05)	0.68 (0.05)	0.44 (0.08)	0.52 (0.06)	0.57 (0.06)
DA_D	0.68 (0.05)	0.70 (0.05)	0.48 (0.06)	0.52 (0.07	0.59 (0.06)
FA_D	0.53 (0.07)	0.55 (0.07)	0.27 (0.11)	0.40 (0.09)	0.49 (0.09)
F_D	0.33 (0.11)	0.43 (0.10)	0.07 (0.10)	0.31 (0.09)	0.39 (0.08)
A_D	0.60 (0.06)	0.61 (0.05)	0.34 (0.09)	0.44 (0.07)	0.53 (0.07)
DFA_F	0.71 (0.05)	0.70 (0.05)	0.37 (0.08)	0.66 (0.05)	0.67 (0.05)
F_F	0.69 (0.05)	0.68 (0.05)	0.37 (0.07)	0.67 (0.05)	0.67 (0.05)
DF_F	0.70 (0.05)	0.69 (0.05)	0.35 (0.08)	0.67 (0.05)	0.68 (0.05)
FA_F	0.70 (0.05)	0.70 (0.05)	0.41 (0.08)	0.66 (0.05)	0.67 (0.05)
DA_F	0.66 (0.06)	0.66 (0.06)	0.37 (0.08)	0.61 (0.05)	0.65 (0.05)
D_F	0.60 (0.07)	0.58 (0.08)	0.15 (0.11)	0.58 (0.07)	0.56 (0.07)
A_F	0.67 (0.05)	0.68 (0.05)	0.41 (0.08)	0.61 (0.06)	0.64 (0.05)
DFA_A	0.56 (0.08)	0.57 (0.08)	0.37 (0.08)	0.48 (0.08)	0.52 (0.07)
A_A	0.55 (0.06)	0.56 (0.07)	0.39 (0.08)	0.48 (0.07)	0.53 (0.07)
DF_A	0.57 (0.08)	0.58 (0.07)	0.39 (0.08)	0.48 (0.07)	0.53 (0.08)
DA_A	0.56 (0.06)	0.57 (0.06)	0.36 (0.09)	0.49 (0.07)	0.55 (0.06)
FA_A	0.54 (0.08)	0.56 (0.07)	0.39 (0.10)	0.48 (0.08)	0.51 (0.07)
D_A	0.53 (0.07)	0.54 (0.07)	0.36 (0.08)	0.46 (0.07)	0.49 (0.07)
F_A	0.52 (0.08)	0.54 (0.08)	0.38 (0.06)	0.44 (0.08)	0.45 (0.08)

Standard deviations over the predictive abilities of the 100 CV replicates are shown between brackets

including both flint and dent lines (0.35 for DF_F).

When predicting an admixed PS, using admixed lines (A_A) was not necessarily the best option. For instance a higher predictive ability was observed when using a TS including both dent and flint lines (DF_A with 0.57) compared to using A_A (0.55). In general, the level of predictive ability was similar for all scenarios when predicting admixed individuals.

Discussion

Modeling group-specific allele in admixed populations

When a population is stratified into genetic groups, a heterogeneity of marker allele effects may be observed across groups. To evaluate such effects and take them into account in genomic prediction, we developed two genomic prediction models adapted to the prediction of admixed individuals, called MAGBLUP 1 and MAGBLUP 2, implemented using linear mixed model. Both are based on an additive genetic model with SNP-QTL allele effects that depend on both SNP alleles and ancestries (or local admixtures).

MAGBLUP 1 was derived using a formalism in which the genotypic information at SNPs is random, and is thus in line with the "animal model" (Henderson, 1984) and the decomposition of variance in admixed populations proposed by Lo et al. (1993) and García-Cortés and Toro (2006). We proposed estimators of the covariance matrices that take advantage of both genomic data and local admixtures, unlike Strandén and Mäntysaari (2013) and Makgahlela et al. (2013) who adapted these models using global admixture proportions and a standard kinship matrix estimated with SNPs. For given genetic groups A and B, the model is expressed as a variance component model including a segregation variance σ^2_S and two group-specific genetic variances $\sigma^2_{G_A}$ and $\sigma^2_{G_B}$. The segregation variance σ^2_S was presented by Lande (1981), Lo et al. (1993) or Lynch and Walsh (1998) and corresponds to the additional variance observed in an admixed population that is due to group specific means at SNPs. It depends on two factors: the differentiation allele frequencies between groups and the existence of group-specific allele effects. From a breeding perspective, it highlights the possibility to generate genetic variance from differentiated genetic groups whose genetic diversity is low. This variance was also recently used in the metafounders theory, which is dedicated to the connection of pedigrees with partial genomic information that are used in single-step evaluation of structured populations (Legarra et al., 2015).

MAGBLUP 2 was derived using a formalism in which SNP allele effects are random, and is thus is line with a bayesian conception of genomic prediction models (Meuwissen et al., 2001; Gianola et al., 2009). Using this formalism, it is possible to allow for genetic covariances between individuals from different groups, assuming that SNP allele effects are at least partly conserved between groups. Different genomic prediction models were proposed which explicitly accounted for covariances between effects of different groups, as proposed by Karoui et al. (2012) and Lehermeier et al. (2015). This same formalism was also used to derive a priori indicators of accuracy or to find relevant estimators of relatedness in structured populations (Wientjes et al., 2015a, 2017). Rather than modeling directly covariances between effects across groups, we re-parametrized QTL allele effects into a main effect and group-specific deviations, as proposed by Schulz-Streeck et al. (2012), de los Campos et al. (2015) or Technow and Totir (2015). Here also, the main innovation of our model lies in the valorization of genomic data and local admixtures, whereas other methods based on the second formalism did not account for admixture. MAGBLUP 2 could be expressed as a variance component model including a component that is due to main SNP allele effects σ^2_U and two components that are due to group-specific deviations effects $\sigma^2_{U_A}$ and $\sigma^2_{U_B}$. These components can be used to better understand the genetic determinism of a given trait as they provide insights concerning the conservation of SNP allele effects across genetic groups. This information is tightly linked to the concept of genetic correlation between genetic groups, which is an important parameter to consider when applying GS in a structured population (**?**Wientjes et al., 2017).

Both genomic prediction models MAGBLUP 1 and 2 can theoretically be interpreted in terms of the origin of genetic covariance between individuals. According to MAGBLUP 1, a pure individual from a given group A can be predicted by a multi-group TS using the intercept of group A and genetic information from both admixed and other group A individuals, through the covariance of $\boldsymbol{g}_A$. However, no information can be borrowed from a pure individual of the alternative group B, as a null kinship is assumed between individuals coming from different groups. Admixed individuals can be predicted using an average of group-specific intercepts weighted by admixture proportions, and using genetic information from all types of individuals. Interestingly, genetic information can be borrowed from another admixed individual, even though they do not share any allele ancestry, through the segregation covariance of $\boldsymbol{a}$. According to MAGBLUP 2, a pure individual from a given group A can be predicted by a multi-group TS using genetic information from all types of individuals through the covariance of $\boldsymbol{u}$, and from both admixed and group A individuals through the covariance of $\boldsymbol{u}_A$. Interestingly, MAGBLUP 2 allows for genetic information to be borrowed across genetic groups if the SNP allele effects are at least partially conserved between groups (i.e. $\sigma^2_U > 0$). In conclusion, these two models underline the main sources of genetic covariance between individuals: (i) their kinship, which tend to be null between individuals from different groups, (ii) the conservation of QTL alleles effects between groups, and (iii) the segregation of allele ancestries within admixed individuals.

Variance components and genomic predictions

Based on simulated traits, the models were evaluated concerning their precision in variance component estimation and genomic prediction accuracy, using a "Flint-Dent" maize inbred panel including admixed individuals. The precision in variance component estimation was evaluated for traits simulated using various contributions of group-specific deviations effects at QTLs. Both models generally estimated their variance components accurately, although a bias has been observed when estimating the segregation variance σ^2_S using MAGBLUP 1 for the genetic determinism with highly differentiated allele effects across groups. Note that the variance components of MAGBLUP 1 and 2 are not directly comparable as the origin of the variation either comes from the genotypes for MAGBLUP 1 or from the allele effects for MAGBLUP 2.

For genomic prediction, both models were then compared to GBLUP for the same three types of genetic determinisms using standard CV scenarios. Both MAGBLUP 1 and 2 led to higher accuracies than GBLUP when group-specific QTL allele effects were simulated, and the gain was the highest for the genetic determinism with QTL allele effects drawn independently within each group. When evaluated for the genetic determinism with conserved QTL allele effects between groups, MAGBLUP 2 led to accuracies similar as GBLUP, while MAGBLUP 1 resulted in slightly lower accuracies. These results indicated that MAGBLUP 2 is more robust than MAGBLUP 1 against a wide variety of genetic determinisms. This can be explained by the possibility for genetic information to be borrowed from one group to another, which gives a substantial advantage when QTL allele effects are conserved between groups, as discussed by Lehermeier et al. (2015). These simulations also show evidence of the high robustness of GBLUP with respect to the heterogeneity of SNP allele effects across groups, as the gain in accuracy did not exceed around 0.15 in scenario A.

Using five real traits, MAGBPLUP 1 and 2 were compared to GBLUP for variance component estimates and genomic prediction accuracy. The genetic variance σ^2_G estimated with GBLUP was comparable and generally higher than the group-specific variances $\sigma^2_{G_D}$ and $\sigma^2_{G_F}$ estimated with MAGBLUP 1. The segregation variance estimates were relatively low for flowering traits compared to group-specific genetic variances, but was substantial for PH. These results suggest an additional variance generated by admixture for PH, which is consistent with the higher phenotypic variance estimated in the phenotypic analyses for admixed lines compared to pure lines (Supplementary Table S3.1). Using MAGBLUP 2, the proportion of variance estimated to be due to main SNP allele effects was much higher than those due to group-specific deviations effects for all traits. These results suggested that the genetic determinism of these five traits consisted of a polygenic background whose QTL allele effects are mainly conserved between the dent and the flint groups. As expected on the basis of this statement, the two MAGBLUP models did not lead to a substantial gain in accuracy, even though MAGBLUP 2 allowed limited gains using all the individuals.

Based on the conclusion of Chapter 2, we could have expected that the proportion of variance due to group-specific QTL deviations effects would be higher for MF and FF. Previous QTL mapping and GWAS studies had also shown differences in terms of genetic determinism between dent and flint groups for flowering traits (**???**). These results with the ones observed in the present study may suggest the existence of group differences at main QTLs, while the polygenic background is mainly conserved between groups. Alternatively, group differences in SNP allele effects may be mainly due to interactions between QTLs and the genetic background so that the SNP allele effects would not be conserved for a given ancestry between the pure and the admixed genetic backgrounds. In such a configuration, modeling the genetic value of an individual, as being a sum of QTL allele effects which depend on both the allele at the QTL and its ancestry, would not be relevant. Several results support this hypothesis concerning flowering time: (i) QTLs interacting with the genetic background were detected in Chapter 2, (ii) directional epistasis was detected in Chapter 2, and (iii) variance components estimated using MAGBLUP 2 while excluding admixed individuals suggested a higher contribution of group-specific deviations effects (Supplementary Table S3.4). It such a situation, it could be more appropriate to perform genomic predictions using models that account directly for epistatic interactions between QTLs (Vitezica et al., 2017), or other methods accounting for various types of heterogeneity between

genetic groups, such as computing an alternative covariance matrix based on specific kernel functions (Heslot et al., 2015).

In conclusion, MAGBLUP 1 and 2 showed their complementarity as genomic prediction and variance component models in the context of a structured population including admixed individuals. Even though, we did not observe substantial gains in terms of genomic prediction accuracy when applied to a "Flint-Dent" panel evaluated for traits, these models would find applications if the existence of group-specific SNP allele effects were identified or highly suspected for a given species. As discussed by Ibánez-Escriche et al. (2009) or Technow et al. (2012), the modeling of group-specific allele effects would probably be more beneficial compared to standard GBLUP for datasets genotyped as low to medium density. This statement is based on the hypothesis that LD between SNPs and QTLs is more likely to differ between groups at these densities. Finally, extension to more than two groups is straightforward for MAGBLUP 2 but not for MAGBLUP 1 as it would require to divide the segregation variance into several components, as shown by Lo et al. (1993) and (García-Cortés and Toro, 2006).

Benefits from admixed individuals in multi-group training sets

Genetic structure may impact the accuracy of genomic predictions, particularly if a given genetic group to be predicted is not represented in the TS, as shown by Olson et al. (2012), Chen et al. (2013), Technow et al. (2013) or Lehermeier et al. (2014). The use of multi-group TS was proposed for several applications including the possibility to apply predictions to a wide range of genetic diversity, the improvement of genomic selection efficiency in genetic groups with limited size or the optimization of resources for traits that are expensive to evaluate. One could question whether including admixed individuals, instead of assembling pure individuals, would help to create connections between genetic groups and allow for more genetic information to be borrowed.

The interest of admixed individuals was first evaluated using the "Flint-Dent" maize inbred panel and simulated traits for which QTL allele effects were partially conserved between dent and flint genetic groups. Different CV scenarios were defined, for a given size of TS and PS, by leveraging the contribution of each genetic background (dent, flint or admixed) to the TS and the PS. As previously shown in Chapter 1 (Rio et al., 2019), a given group-specific PS was best predicted with GBLUP using a TS including only individuals from the same genetic group. Conversely, applying across-group predictions could highly depreciate genomic prediction accuracy, while multi-group TSs showed a relatively high accuracy no matter the target PS. Using MAGBLUP 1 or 2 instead of using GBLUP led to limited gains when predicting dent or flint lines but greatly improved the accuracy when predicting admixed lines. Along with the simulations made by Toosi et al. (2013), these results support the use of multi-group TSs with admixed individuals, particularly when the target PS also includes admixed individuals.

When the same procedure was applied to the five real traits using GBLUP, differences could be observed compared to the simulated traits. First the level of accuracy, obtained when applying genomic prediction within a given genetic background was generally lower for the admixed lines while it was intermediate for simulated traits. Flint lines were generally well predicted by dent lines while the opposite was not true. These results may suggest more QTLs being only polymorphic in the dent group for MF, FF, ELN and TNL, although genome wide more SNPs are only polymorphic in the flint group (results not shown). Another hypothesis lies in the higher contribution of dent- compared to flint-specific deviations effects (see Table 3.6 and Supplementary Table S3.4), making dent lines more difficult to predict by the flint lines than the opposite. Considering flowering traits, the level of accuracy obtained for across-group predictions is higher than those observed by Lehermeier et al. (2014), but may be explained by the use of a diversity panel compared to the bi-parental progenies used in their study. The consistence between the results observed on flowering traits and the traits related to plant architecture (ELN and TNL) is probably due to the proximity of these traits in terms of genetic determinism (Li et al., 2016a; Bouchet et al., 2016). Replacing pure lines by admixed lines

in a multi-group TS generally depreciated the accuracy to predict pure lines, and possibly admixed lines. A noticeable exception was PH for which FA_F lead to a better accuracy than F_F. Using MAGBLUP 1 or 2 instead of GBLUP did not improve the genomic prediction accuracy, as expected from the results based on standard CV scenarios. Surprisingly, these results are not really supporting the idea of using a multi-group TSs with admixed individuals rather than using pure individuals. Here also, a hypothesis to explain these results would be the existence of multiple epistatic interactions between QTLs and the genetic background. Such interactions would be shuffled within admixed individuals and would limit the amount of genetic information to be borrowed. In such context, the best source of genetic information to predict a given individual would consist in other individuals from the same genetic group.

Conclusion

In conclusion, MAGBLUP 1 and 2 showed their complementarity as genomic prediction models in the context of a structured population with admixed individuals. While MAGBLUP 1 can be used to identify the segregation variance generated by admixture, MAGBLUP 2 can be used to disentangle the variance that is due to main SNP allele effects from that due to group-specific deviations. The benefits of using admixed individuals in multi-group TSs were not systematic across traits and genetic groups for this panel but remains to be tested for other genetic groups, hybrid performance evaluation, or other species.

Acknowledgments

This research was supported by the "Investissement d'Avenir" project "Amaizing". S. Rio is jointly funded by the program AdmixSel of the INRA metaprogram SelGen and by the breeding companies partners of the Amaizing project: Caussade-Semences, Euralis, KWS, Limagrain, Maisadour, RAGT and Syngenta. We thank Valerie Combes, Delphine Madur and Stéphane Nicolas (GQE - Le Moulon) for DNA extraction, analysis and assembly of genotypic data. We thank Cyril Bauland (GQE - Le Moulon), Carine Palaffre, Bernard Lagardère, Jean-René Loustalot (INRA Saint-Martin de Hinx) for the panel assembly and the coordination of seed production, all the breeding companies partners of the Amaizing project for the production of admixed lines and the company Limagrain for the genotyping of admixed lines.

Appendix A

For a given trait, let $\mathrm{E}\left(G_i\right)$ be the expected genetic value:

$$
\begin{aligned}
\mathrm{E}(G_i) &= \mathrm{E}\left(\sum_{m=1}^{M}\left(\beta_m^0 + W_{im}\left(\beta_m^1 - \beta_m^0\right)\right)\right) \\
\mathrm{E}(G_i) &= \sum_{m=1}^{M}\mathrm{E}\left(\beta_m^0 + W_{im}\left(\beta_m^1 - \beta_m^0\right)\right) \\
\mathrm{E}(G_i) &= \sum_{m=1}^{M}\left(\beta_m^0 + \mathrm{E}(W_{im})(\beta_m^1 - \beta_m^0)\right) \\
\mathrm{E}(G_i) &= \sum_{m=1}^{M}\left(\beta_m^0 + f_m(\beta_m^1 - \beta_m^0)\right) \stackrel{\text{def}}{=} \mu
\end{aligned}
$$

Appendix B

For a given trait, let $\mathrm{cov}\left(G_i, G_j|\alpha_{ij}\right)$, later referred to as $\mathrm{cov}\left(G_i, G_j\right)$, be the covariance between the genetic values of individuals assuming a kinship α_{ij}:

$$
\begin{aligned}
\mathrm{cov}(G_i, G_j) &= \mathrm{cov}\left(\sum_{m=1}^{M}\left(\beta_m^0 + W_{im}\left(\beta_m^1 - \beta_m^0\right)\right), \sum_{m'=1}^{M}\left(\beta_{m'}^0 + W_{jm'}\left(\beta_{m'}^1 - \beta_{m'}^0\right)\right)\right) \\
\mathrm{cov}(G_i, G_j) &= \sum_{m=1}^{M}\mathrm{cov}\left(\beta_m^0 + W_{im}\left(\beta_m^1 - \beta_m^0\right), \beta_m^0 + W_{jm}\left(\beta_m^1 - \beta_m^0\right)\right) \\
\mathrm{cov}(G_i, G_j) &= \sum_{m=1}^{M}\mathrm{cov}\left(W_{im}, W_{jm}\right)\left(\beta_m^1 - \beta_m^0\right)^2 \\
\mathrm{cov}(G_i, G_j) &= \sum_{m=1}^{M}\left(\mathrm{E}\left(W_{im}, W_{jm}\right) - \mathrm{E}\left(W_{im}\right)\mathrm{E}\left(W_{jm}\right)\right)\left(\beta_m^1 - \beta_m^0\right)^2 \\
\mathrm{cov}(G_i, G_j) &= \sum_{m=1}^{M}\left(\alpha_{ij} f_m + (1-\alpha_{ij}) f_m^2 - f_m^2\right)\left(\beta_m^1 - \beta_m^0\right)^2 \\
\mathrm{cov}(G_i, G_j) &= \alpha_{ij}\sum_{m=1}^{M} f_m\left(1 - f_m\right)\left(\beta_m^1 - \beta_m^0\right)^2 \stackrel{\text{def}}{=} \alpha_{ij}\sigma_G^2
\end{aligned}
$$

Note that an absence of LD was assumed between QTLs

Appendix C

Let $\mathrm{E}\left(G_i|\boldsymbol{w}_i\right)$, later referred to as $\mathrm{E}\left(G_i\right)$, be the expected genetic value of an individual, with given alleles at QTLs, over an infinite sampling of allele effects:

$$\begin{aligned}
\mathrm{E}(G_i) &= \mathrm{E}\left(\sum_{m=1}^{M}\left(\beta_m^0 + W_{im}\left(\beta_m^1 - \beta_m^0\right)\right)\right) \\
\mathrm{E}(G_i) &= \sum_{m=1}^{M}\mathrm{E}\left(\beta_m^0 + W_{im}\left(\beta_m^1 - \beta_m^0\right)\right) \\
\mathrm{E}(G_i) &= \sum_{m=1}^{M}\left(\mathrm{E}(\beta_m^0) + \mathrm{E}(\beta_m^1)W_{im} - \mathrm{E}(\beta_m^0)W_{im}\right) \\
\mathrm{E}(G_i) &= 0
\end{aligned}$$

Appendix D

Let $\mathrm{cov}(G_i, G_j|\boldsymbol{w}_i, \boldsymbol{w}_j)$, later referred to as $\mathrm{cov}(G_i, G_j)$, be the covariance between the genetic values of individuals, with given alleles at QTLs, over an infinite sampling of allele effects:

$$\begin{aligned}
\mathrm{cov}(G_i, G_j) &= \mathrm{cov}\left(\sum_{m=1}^{M}\left(\beta_m^0 + W_{im}\left(\beta_m^1 - \beta_m^0\right)\right), \sum_{m'=1}^{M}\left(\beta_{m'}^0 + W_{jm'}\left(\beta_{m'}^1 - \beta_{m'}^0\right)\right)\right) \\
\mathrm{cov}(G_i, G_j) &= \sum_{m=1}^{M}\mathrm{cov}\left(\beta_m^0 + W_{im}\left(\beta_m^1 - \beta_m^0\right), \beta_m^0 + W_{jm}\left(\beta_m^1 - \beta_m^0\right)\right) \\
\mathrm{cov}(G_i, G_j) &= \sum_{m=1}^{M}\left(\mathrm{V}\left(\beta_m^0\right) - \mathrm{V}\left(\beta_m^0\right)W_{im} + W_{im}W_{jm}\mathrm{V}\left(\beta_m^1\right) - W_{im}\mathrm{V}\left(\beta_m^0\right) + W_{im}W_{jm}\mathrm{V}\left(\beta_m^0\right)\right) \\
\mathrm{cov}(G_i, G_j) &= \sum_{m=1}^{M}\left(\sigma_\beta^2 - \sigma_\beta^2 W_{im} + W_{im}W_{jm}\sigma_\beta^2 - W_{im}\sigma_\beta^2 + W_{im}W_{jm}\sigma_\beta^2\right) \\
\mathrm{cov}(G_i, G_j) &= \sum_{m=1}^{M}\left(\left(1 - W_{im}\right)\left(1 - W_{jm}\right) + W_{im}W_{jm}\right)\sigma_\beta^2 \\
\mathrm{cov}(G_i, G_j) &= \frac{1}{M}\sum_{m=1}^{M}\left(\left(1 - W_{im}\right)\left(1 - W_{jm}\right) + W_{im}W_{jm}\right)M\sigma_\beta^2 \overset{\mathrm{def}}{=} \phi_{ij}\sigma_U^2
\end{aligned}$$

Appendix E

For a given trait, let $\mathrm{E}\left(G_i|\pi_i\right)$, later referred to as $\mathrm{E}(G_i)$, be the expected genetic value of an with a proportion of genome A π_i:

$$\mathrm{E}(G_i) = \mathrm{E}\left(\sum_{m=1}^{M}\left(A_{imA}\left(\beta^0_{mA} + W_{im}(\beta^1_{mA} - \beta^0_{mA})\right) + A_{imB}\left(\beta^0_{mB} + W_{im}(\beta^1_{mB} - \beta^0_{mB})\right)\right)\right)$$

$$\mathrm{E}(G_i) = \sum_{m=1}^{M}\mathrm{E}\left(A_{imA}\left(\beta^0_{mA} + W_{im}(\beta^1_{mA} - \beta^0_{mA})\right) + A_{imB}\left(\beta^0_{mB} + W_{im}(\beta^1_{mB} - \beta^0_{mB})\right)\right)$$

$$\mathrm{E}(G_i) = \sum_{m=1}^{M}\left(\mathrm{E}(A_{imA})\beta^0_{mA} + \mathrm{E}(A_{imA}W_{im})(\beta^1_{mA} - \beta^0_{mA}) + \mathrm{E}(A_{imB})\beta^0_{mB} + \mathrm{E}(A_{imB}W_{im})(\beta^1_{mB} - \beta^0_{mB})\right)$$

$$\mathrm{E}(G_i) = \sum_{m=1}^{M}\left(\pi_i\beta^0_{mA} + \pi_i f_{mA}(\beta^1_{mA} - \beta^0_{mA}) + (1-\pi_i)\beta^0_{mB} + (1-\pi_i)f_{mB}(\beta^1_{mB} - \beta^0_{mB})\right)$$

$$\mathrm{E}(G_i) = \sum_{m=1}^{M}\left(\pi_i\left(\beta^0_{mA} + f_{mA}(\beta^1_{mA} - \beta^0_{mA})\right) + (1-\pi_i)\left(\beta^0_{mB} + f_{mB}(\beta^1_{mB} - \beta^0_{mB})\right)\right)$$

$$\mathrm{E}(G_i) = \pi_i\sum_{m=1}^{M}\left(\beta^0_{mA} + f_{mA}(\beta^1_{mA} - \beta^0_{mA})\right) + (1-\pi_i)\sum_{m=1}^{M}\left(\beta^0_{mB} + f_{mB}(\beta^1_{mB} - \beta^0_{mB})\right)$$

$$\mathrm{E}(G_i) = \pi_i\sum_{m=1}^{M}\mu_{mA} + (1-\pi_i)\sum_{m=1}^{M}\mu_{mB} \overset{\text{def}}{=} \pi_i\mu_A + (1-\pi_i)\mu_B$$

Appendix F

For a given trait, let $\mathrm{cov}(G_i, G_j|\pi_i, \pi_j, \theta^A_{ij}, \alpha^A_{ij}, \alpha^B_{ij})$, later referred to as $\mathrm{cov}(G_i, G_j)$, be the genetic covariance between two individuals i and j, assuming a proportion π_i of genome A for i and π_j for j, a proportion of shared ancestry θ^A_{ij} between i and j for the genome originated from group A, a kinship α^A_{ij} between i and j on their shared ancestries for the genome originated from group A and a kinship α^B_{ij} between i and j on their shared ancestries for the genome originated from group B. We have also $\theta^B_{ij} = 1 - \pi_i - \pi_j + \theta^A_{ij}$ being the proportion of shared ancestries of i and j for the genome originated from group B, $\theta^{AB}_{ij} = \pi_i - \theta^A_{ij}$ is the proportion of not shared ancestries corresponding to the genome of i originated from group A and of j originated from group B, and $\theta^{BA}_{ij} = \pi_j - \theta^A_{ij}$ is the proportion of not shared ancestries corresponding to the genome of i originated from group B and of j originated from group A:

$$\begin{aligned}\mathrm{cov}(G_i, G_j) = \mathrm{cov}\Bigg(&\sum_{m=1}^{M}\left(\gamma^0_m + W_{im}\left(\gamma^1_m - \gamma^0_m\right) + A_{imA}\left(\delta^0_{mA} + W_{im}\left(\delta^1_{mA} - \delta^0_{mA}\right)\right)\right.\\ &\left.+A_{imB}\left(\delta^0_{mB} + W_{im}\left(\delta^1_{mB} - \delta^0_{mB}\right)\right)\right), \sum_{m'=1}^{M}\left(\gamma^0_{m'} + W_{jm'}\left(\gamma^1_{m'} - \gamma^0_{m'}\right)\right.\\ &\left.+A_{jm'A}\left(\delta^0_{m'A} + W_{jm'}\left(\delta^1_{m'A} - \delta^0_{m'A}\right)\right) + A_{jm'B}\left(\delta^0_{m'B} + W_{jm'}\left(\delta^1_{m'B} - \delta^0_{m'B}\right)\right)\right)\Bigg)\end{aligned}$$

$$
\begin{aligned}
\operatorname{cov}(G_i, G_j) = \sum_{m=1}^{M} & \operatorname{cov}\left(\gamma_m^0 + W_{im}\left(\gamma_m^1 - \gamma_m^0\right) + A_{imA}\left(\delta_{mA}^0 + W_{im}\left(\delta_{mA}^1 - \delta_{mA}^0\right)\right)\right. \\
& + A_{imB}\left(\delta_{mB}^0 + W_{im}\left(\delta_{mB}^1 - \delta_{mB}^0\right)\right), \gamma_m^0 + W_{jm}\left(\gamma_m^1 - \gamma_m^0\right) \\
& \left. + A_{jmA}\left(\delta_{mA}^0 + W_{jm}\left(\delta_{mA}^1 - \delta_{mA}^0\right)\right) + A_{jmB}\left(\delta_{mB}^0 + W_{jm}\left(\delta_{mB}^1 - \delta_{mB}^0\right)\right)\right) \\
\operatorname{cov}(G_i, G_j) = \sum_{m=1}^{M} & \left(\operatorname{cov}\left(A_{imA}, A_{jmA}\right)\left(\beta_{mA}^0\right)^2 + \operatorname{cov}\left(A_{imB}, A_{jmB}\right)\left(\beta_{mB}^0\right)^2\right. \\
& + \operatorname{cov}\left(A_{imA}, A_{jmB}\right)\beta_{mA}^0\beta_{mB}^0 + \operatorname{cov}\left(A_{imB}, A_{jmA}\right)\beta_{mA}^0\beta_{mB}^0 \\
& + \operatorname{cov}\left(A_{imA}, A_{jmA}W_{jm}\right)\beta_{mA}^0\left(\beta_{mA}^1 - \beta_{mA}^0\right) + \operatorname{cov}\left(A_{jmA}, A_{imA}W_{im}\right)\beta_{mA}^0\left(\beta_{mA}^1 - \beta_{mA}^0\right) \\
& + \operatorname{cov}\left(A_{imA}, A_{jmB}W_{jm}\right)\beta_{mA}^0\left(\beta_{mB}^1 - \beta_{mB}^0\right) + \operatorname{cov}\left(A_{jmA}, A_{imB}W_{im}\right)\beta_{mA}^0\left(\beta_{mB}^1 - \beta_{mB}^0\right) \\
& + \operatorname{cov}\left(A_{imB}, A_{jmA}W_{jm}\right)\beta_{mB}^0\left(\beta_{mA}^1 - \beta_{mA}^0\right) + \operatorname{cov}\left(A_{jmB}, A_{imA}W_{im}\right)\beta_{mB}^0\left(\beta_{mA}^1 - \beta_{mA}^0\right) \\
& + \operatorname{cov}\left(A_{imB}, A_{jmB}W_{jm}\right)\beta_{mB}^0\left(\beta_{mB}^1 - \beta_{mB}^0\right) + \operatorname{cov}\left(A_{jmB}, A_{imB}W_{im}\right)\beta_{mB}^0\left(\beta_{mB}^1 - \beta_{mB}^0\right) \\
& + \operatorname{cov}\left(A_{imA}W_{im}, A_{jmB}W_{jm}\right)\left(\beta_{mA}^1 - \beta_{mA}^0\right)\left(\beta_{mB}^1 - \beta_{mB}^0\right) \\
& + \operatorname{cov}\left(A_{jmA}W_{jm}, A_{imB}W_{im}\right)\left(\beta_{mA}^1 - \beta_{mA}^0\right)\left(\beta_{mB}^1 - \beta_{mB}^0\right) \\
& \left. + \operatorname{cov}\left(A_{imA}W_{im}, A_{jmA}W_{jm}\right)\left(\beta_{mA}^1 - \beta_{mA}^0\right)^2 + \operatorname{cov}\left(A_{jmB}W_{jm}, A_{imB}W_{im}\right)\left(\beta_{mB}^1 - \beta_{mB}^0\right)^2\right) \\
\operatorname{cov}(G_i, G_j) = \sum_{m=1}^{M} & \left(\Delta_{ij}\left(\beta_{mA}^0\right)^2 + \Delta_{ij}\left(\beta_{mB}^0\right)^2 - 2\Delta_{ij}\beta_{mA}^0\beta_{mB}^1 + 2\Delta_{ij}f_{mA}\beta_{mA}^0\left(\beta_{mA}^1 - \beta_{mA}^0\right)\right. \\
& - 2\Delta_{ij}f_{mB}\beta_{mA}^0\left(\beta_{mB}^1 - \beta_{mB}^0\right) - 2\Delta_{ij}f_{mA}\beta_{mB}^0\left(\beta_{mA}^1 - \beta_{mA}^0\right) + 2\Delta_{ij}f_{mB}\beta_{mB}^0\left(\beta_{mB}^1 - \beta_{mB}^0\right) \\
& - 2\Delta_{ij}f_{mA}f_{mB}\left(\beta_{mA}^1 - \beta_{mA}^0\right)\left(\beta_{mB}^1 - \beta_{mB}^0\right) \\
& + \theta_{ij}^A\alpha_{ij}^A f_{mA}\left(1 - f_{mA}\right)\left(\beta_{mA}^1 - \beta_{mA}^0\right)^2 + \Delta_{ij}f_{mA}^2\left(\beta_{mA}^1 - \beta_{mA}^0\right)^2 \\
& \left. + \theta_{ij}^B\alpha_{ij}^B f_{mB}\left(1 - f_{mB}\right)\left(\beta_{mB}^1 - \beta_{mB}^0\right)^2 + \Delta_{ij}f_{mB}^2\left(\beta_{mB}^1 - \beta_{mB}^0\right)^2\right) \\
\operatorname{cov}(G_i, G_j) = \sum_{m=1}^{M} & \left(\Delta_{ij}\left(\left(\beta_{mA}^0 + f_{mA}\left(\beta_{mA}^1 - \beta_{mA}^0\right)\right) - \left(\beta_{mB}^0 + f_{mB}\left(\beta_{mB}^1 - \beta_{mB}^0\right)\right)\right)^2\right. \\
& \left. \theta_{ij}^A\alpha_{ij}^A f_{mA}\left(1 - f_{mA}\right)\left(\beta_{mA}^1 - \beta_{mA}^0\right)^2 + \theta_{ij}^B\alpha_{ij}^B f_{mB}\left(1 - f_{mB}\right)\left(\beta_{mB}^1 - \beta_{mB}^0\right)^2\right) \\
\operatorname{cov}(G_i, G_j) = \Delta_{ij}\sum_{m=1}^{M}\left(\mu_{mA} - \mu_{mB}\right)^2 & + \theta_{ij}^A\alpha_{ij}^A\sum_{m=1}^{M} f_{mA}\left(1 - f_{mA}\right)\left(\beta_{mA}^1 - \beta_{mA}^0\right)^2 \\
& + \theta_{ij}^B\alpha_{ij}^B\sum_{m=1}^{M} f_{mB}\left(1 - f_{mB}\right)\left(\beta_{mB}^1 - \beta_{mB}^0\right)^2 \overset{\text{def}}{=} \Delta_{ij}\sigma_S^2 + \theta_{ij}^A\alpha_{ij}^A\sigma_{G_A}^2 + \theta_{ij}^B\alpha_{ij}^B\sigma_{G_B}^2
\end{aligned}
$$

Note that an absence of LD was assumed between QTLs and:

$$
\begin{aligned}
&\operatorname{cov}\left(A_{imA}, A_{jmA}\right) = \theta_{ij}^A - \pi_i\pi_j = \Delta_{ij} \\
&\operatorname{cov}\left(A_{imB}, A_{jmB}\right) = \theta_{ij}^B - \left(1 - \pi_i\right)\left(1 - \pi_j\right) = \Delta_{ij} \\
&\operatorname{cov}\left(A_{imA}, A_{jmB}\right) = \theta_{ij}^{AB} - \pi_i\left(1 - \pi_j\right) = -\Delta_{ij} \\
&\operatorname{cov}\left(A_{imB}, A_{jmA}\right) = \theta_{ij}^{BA} - \left(1 - \pi_i\right)\pi_j = -\Delta_{ij} \\
&\operatorname{cov}\left(A_{imA}W_{im}, A_{jmA}W_{jm}\right) = \theta_{ij}^A\left(\alpha_{ij}^A f_{mA} + (1 - \alpha_{ij}^A)f_{mA}^2\right) - \pi_i\pi_j f_{mA}^2 = \theta_{ij}^A\alpha_{ij}^A f_{mA}\left(1 - f_{mA}\right) + \Delta_{ij}f_{mA}^2
\end{aligned}
$$

Appendix G

Let $\mathrm{E}\left(G_i|\boldsymbol{a}_i, \boldsymbol{w}_i\right)$, later referred to as $\mathrm{E}(G_i)$, be the expected genetic value of an individual, with given local admixtures and alleles at QTLs, over an infinite number of samples of allele effects:

$$\mathrm{E}(G_i) = \mathrm{E}\left(\sum_{m=1}^{M}\left(\gamma_m^0 + W_{im}\left(\gamma_m^1 - \gamma_m^0\right) + A_{imA}\left(\delta_{mA}^0 + W_{im}\left(\delta_{mA}^1 - \delta_{mA}^0\right)\right) + A_{imB}\left(\delta_{mB}^0 + W_{im}\left(\delta_{mB}^1 - \delta_{mB}^0\right)\right)\right)\right)$$

$$\mathrm{E}(G_i) = \sum_{m=1}^{M}\mathrm{E}\left(\gamma_m^0 + W_{im}\left(\gamma_m^1 - \gamma_m^0\right) + A_{imA}\left(\delta_{mA}^0 + W_{im}\left(\delta_{mA}^1 - \delta_{mA}^0\right)\right) + A_{imB}\left(\delta_{mB}^0 + W_{im}\left(\delta_{mB}^1 - \delta_{mB}^0\right)\right)\right)$$

$$\begin{aligned}\mathrm{E}(G_i) = \sum_{m=1}^{M}\Big(&\mathrm{E}\left(\gamma_m^0\right) + W_{im}\mathrm{E}\left(\gamma_m^1\right) - W_{im}\mathrm{E}\left(\gamma_m^0\right) + A_{imA}\mathrm{E}\left(\delta_{mA}^0\right) + A_{imA}W_{im}\mathrm{E}\left(\delta_{mA}^1\right) - A_{imA}W_{im}\mathrm{E}\left(\delta_{mA}^0\right)\\ &+ A_{imB}\mathrm{E}\left(\delta_{mB}^0\right) + A_{imB}W_{im}\mathrm{E}\left(\delta_{mB}^1\right) - A_{imB}\mathrm{E}\left(\delta_{mB}^0\right)\Big)\end{aligned}$$

$$\mathrm{E}(G_i) = 0$$

Appendix H

Let $\text{cov}(G_i, G_j|\boldsymbol{a}_i, \boldsymbol{a}_j, \boldsymbol{w}_i, \boldsymbol{w}_j)$ be the covariance between the genetic values of pairs of individuals, with given alleles at QTLs, over an infinite number of samples of allele effects:

$$
\begin{aligned}
\text{cov}(G_i, G_j) = \text{cov}\Bigg(&\sum_{m=1}^{M}\left(\gamma_m^0 + W_{im}\left(\gamma_m^1 - \gamma_m^0\right) + A_{imA}\left(\delta_{mA}^0 + W_{im}\left(\delta_{mA}^1 - \delta_{mA}^0\right)\right)\right.\\
&\left. + A_{imB}\left(\delta_{mB}^0 + W_{im}\left(\delta_{mB}^1 - \delta_{mB}^0\right)\right)\right), \sum_{m'=1}^{M}\left(\gamma_{m'}^0 + W_{jm'}\left(\gamma_{m'}^1 - \gamma_{m'}^0\right)\right.\\
&\left. + A_{jm'A}\left(\delta_{m'A}^0 + W_{jm'}\left(\delta_{m'A}^1 - \delta_{m'A}^0\right)\right) + A_{jm'B}\left(\delta_{m'B}^0 + W_{jm'}\left(\delta_{m'B}^1 - \delta_{m'B}^0\right)\right)\right)\Bigg)\\
\text{cov}(G_i, G_j) = &\sum_{m=1}^{M}\text{cov}\left(\gamma_m^0 + W_{im}\left(\gamma_m^1 - \gamma_m^0\right) + A_{imA}\left(\delta_{mA}^0 + W_{im}\left(\delta_{mA}^1 - \delta_{mA}^0\right)\right)\right.\\
&\left. + A_{imB}\left(\delta_{mB}^0 + W_{im}\left(\delta_{mB}^1 - \delta_{mB}^0\right)\right), \gamma_m^0 + W_{jm}\left(\gamma_m^1 - \gamma_m^0\right)\right.\\
&\left. + A_{jmA}\left(\delta_{mA}^0 + W_{jm}\left(\delta_{mA}^1 - \delta_{mA}^0\right)\right) + A_{jmB}\left(\delta_{mB}^0 + W_{jm}\left(\delta_{mB}^1 - \delta_{mB}^0\right)\right)\right.\\
\text{cov}(G_i, G_j) = &\sum_{m=1}^{M}\left(\text{V}\left(\gamma_m^0\right) - W_{jm}\text{V}\left(\gamma_m^0\right) + W_{im}W_{jm}\text{V}\left(\gamma_m^1\right) - W_{im}\text{V}\left(\gamma_m^0\right) + W_{im}W_{jm}\text{V}\left(\gamma_m^0\right)\right.\\
&\left. + A_{imA}A_{jmA}\text{V}\left(\delta_{mA}^0\right) - A_{imA}A_{jmA}W_{jm}\text{V}\left(\delta_{mA}^0\right) + A_{imA}W_{im}A_{jmA}W_{jm}\text{V}\left(\delta_{mA}^1\right)\right.\\
&\left. - A_{imA}W_{im}A_{jmA}\text{V}\left(\delta_{mA}^0\right) + A_{imA}W_{im}A_{jmA}W_{jm}\text{V}\left(\delta_{mA}^0\right)\right.\\
&\left. + A_{imB}A_{jmB}\text{V}\left(\delta_{mB}^0\right) - A_{imB}A_{jmB}W_{jm}\text{V}\left(\delta_{mB}^0\right) + A_{imB}W_{im}A_{jmB}W_{jm}\text{V}\left(\delta_{mB}^1\right)\right.\\
&\left. - A_{imB}W_{im}A_{jmB}\text{V}\left(\delta_{mB}^0\right) + A_{imB}W_{im}A_{jmB}W_{jm}\text{V}\left(\delta_{mB}^0\right)\right)\\
\text{cov}(G_i, G_j) = &\sum_{m=1}^{M}\left(\sigma_\gamma^2 - W_{jm}\sigma_\gamma^2 + W_{im}W_{jm}\sigma_\gamma^2 - W_{im}\sigma_\gamma^2 + W_{im}W_{jm}\sigma_\gamma^2\right.\\
&\left. + A_{imA}A_{jmA}\sigma_{\delta_A}^2 - A_{imA}A_{jmA}W_{jm}\sigma_{\delta_A}^2 + A_{imA}W_{im}A_{jmA}W_{jm}\sigma_{\delta_A}^2 - A_{imA}W_{im}A_{jmA}\sigma_{\delta_A}^2\right.\\
&\left. + A_{imA}W_{im}A_{jmA}W_{jm}\sigma_{\delta_A}^2\right.\\
&\left. + A_{imB}A_{jmB}\sigma_{\delta_B}^2 - A_{imB}A_{jmB}W_{jm}\sigma_{\delta_B}^2 + A_{imB}W_{im}A_{jmB}W_{jm}\sigma_{\delta_B}^2 - A_{imB}W_{im}A_{jmB}\sigma_{\delta_B}^2\right.\\
&\left. + A_{imB}W_{im}A_{jmB}W_{jm}\sigma_{\delta_B}^2\right)\\
\text{cov}(G_i, G_j) = &\sum_{m=1}^{M}\left(\left(1 - W_{im}\right)\left(1 - W_{jm}\right) + W_{im}W_{jm}\right)\sigma_\gamma^2\\
&+ \sum_{m=1}^{M}A_{imA}A_{jmA}\left(\left(1 - W_{im}\right)\left(1 - W_{jm}\right) + W_{im}W_{jm}\right)\sigma_{\delta_A}^2\\
&+ \sum_{m=1}^{M}A_{imB}A_{jmB}\left(\left(1 - W_{im}\right)\left(1 - W_{jm}\right) + W_{im}W_{jm}\right)\sigma_{\delta_B}^2\\
\text{cov}(G_i, G_j) = &\,\phi_{ij}M\sigma_\gamma^2 + \phi_{ij}^A M\sigma_{\delta_A}^2 + \phi_{ij}^B M\sigma_{\delta_B}^2 \overset{\text{def}}{=} \phi_{ij}\sigma_U^2 + \phi_{ij}^A\sigma_{U_A}^2 + \phi_{ij}^B\sigma_{U_B}^2
\end{aligned}
$$

General discussion

The advent of high density genotyping opened new perspectives in quantitative genetics studies including the identification of QTLs in GWAS, and the prediction of breeding values in GS to improve the efficiency of selection. The stratification of breeding populations into genetic groups challenges these methods, notably through the existence of group-specific allele effects at SNPs. Generating admixed individuals, to be included in GWAS and GS studies, is an appealing solution to allow for a better connection between genetic groups and to better understand the genetic determinism of traits. This thesis had several objectives concerning the identification and the modeling of group-specific QTL allele effects, as well as the evaluation of the interest of admixed individuals, from both GWAS and GS perspectives. Two maize diversity panels were studied, with different levels of genetic diversity: the "Amaizing Dent" panel which could be subdivided in three genetic sub-groups, and the "Flint-Dent" panel including flint lines, most dent lines of the "Amaizing Dent" panel, as well as admixed lines generated by crossing pure dent and flint lines.

The first objective of this thesis was to study the impact of genetic structure on genomic selection efficiency within the "Amaizing-Dent" panel. Assembling genetic groups into multi-group TSs was proposed by de Roos et al. (2009) to (i) apply genomic predictions to a broad range of genetic diversity, (ii) improve the accuracy for genetic groups with limited size or (iii) optimize resources for traits that are expensive to evaluate. In this panel, assembling genetic groups was an effective solution for achieving good levels of predictive abilities, regardless of the target PS. To predict a group-specific PS, increasing the size of the TS by adding extra-group individuals generally improved the predictive ability. The benefits were especially important when targeting the Iodents (group C), which is consistent with their recent origin and their proximity to certain group A individuals. It illustrated that genetic information can be borrowed from one group to another in the "Amaizing Dent" panel, even when close relatives of the PS are already included in the TS. These results are in favor of the development of generic TSs that could be evaluated on high-throughput genotyping platforms, or through extensive field trials, for traits like drought tolerance in maize (Millet et al., 2016). Using the multivariate genomic prediction models proposed by Lehermeier et al. (2015) called Multi-group GBLUP (MGBLUP), the genetic correlations between groups can be estimated to get insights concerning the conservation of SNP allele effects across genetic groups. The genetic correlation estimates were close to 1 for all traits but grain moisture (around 0.7), suggesting a high conservation of SNP allele effects. Depending on the genetic correlation, different prediction strategies strategies can be considered as discussed by Lehermeier et al. (2015): null genetic correlations suggest to train the model separately within each group using GBLUP, genetic correlations close to 1 suggest to train the model on a multi-group TS using GBLUP, and intermediate correlations suggest to train the model on a multi-group TS using MGBLUP. No gain was observed in terms of genomic prediction accuracy when comparing GBLUP to MGBLUP, even for grain moisture. These results suggested that the impact of genetic structure on the accuracy is probably not resulting from a heterogeneity of SNP allele effects across groups in this dataset. We also evaluated the efficiency of *a priori* indicators to forecast genomic prediction accuracy in this panel. The stantard CD indicator was used to get estimates of accuracy that were compared to those empirically obtained with the structure-based scenarios. Standard CD was not always efficient to forecast the best scenario. We also evaluated new indicators proposed by Wientjes et al. (2015a) which use the genetic correlations estimated using MGBLUP. These led to no improvement, suggesting that the erratic performances of standard CD observed in this dataset were not caused by its

underlying hypothesis, being the conservation of SNP allele effects between groups. Our hypothesis is that the poor performances of CD is rather caused by group differences in allele frequencies at QTLs.

Another objective of this thesis was to develop a GWAS methodology to identify QTLs with group-specific allele effects. While many genomic prediction models have been developed to account for the heterogeneity of SNP and QTL allele effects across genetic groups (Karoui et al., 2012; Schulz-Streeck et al., 2012; Lehermeier et al., 2015; de los Campos et al., 2015; Technow and Totir, 2015), to our knowledge no GWAS model was proposed to identify QTLs with contrasted effects across groups. We developed a GWAS methodology to identify such regions and showed how including admixed individuals can help to disentangle the factors causing the heterogeneity of allele effects across groups: local genomic differences (group differences in LD or group specific mutations) or epistatic interactions between QTLs and the genetic background. The methodology was applied to the "Flint-Dent" panel which assembles a broader diversity than that observed in the "Amaizing-Dent" panel. Several new flowering time QTLs were identified compared to applying the standard GWAS model proposed by Yu et al. (2006) separately within each group. Applying different GWAS strategies was shown as being complementary to maximize the number of QTLs detected. The high number of QTLs detected for maize flowering in our study is in accordance with the results of Chardon et al. (2004) and Buckler et al. (2009) concerning maize flowering as being highly polygenic, with at least several 10s of QTLs involved. Various configurations were observed, especially at known flowering QTLs such as Vgt1 (Salvi et al., 2007; Ducrocq et al., 2008), Vgt2 (Bouchet et al., 2013) or Vgt3 (Salvi et al., 2011, 2017). In our study, group-specific allele effect resulted from both local genomic differences (group differences in LD or group-specific mutations) and/or interactions with the genetic background depending on the QTL. Our results also suggest the existence of QTLs with either additive or interacting profiles. The importance of epistatic interactions with the genetic background on group-specific allele effects was supported by evidence of directional epistasis revealed by admixed individuals.

Knowing the possible existence of group-specific SNP allele effects, we developed two genomic prediction models, dedicated to admixed individuals, that would account for this heterogeneity. The first model, Multi-Group Admixed GBLUP (MAGBLUP) 1, was derived according to the "animal model" and decomposed the genetic variance into group-specific variance components and a segregation variance component due to admixture, as proposed by Lo et al. (1993) and García-Cortés and Toro (2006). The second model, MAGBLUP 2, was derived assuming that the SNP alleles are drown from random distributions and decomposed these effects into a main SNP allele effect and group-specific deviations, along with Schulz-Streeck et al. (2012), de los Campos et al. (2015) or Technow and Totir (2015). The originality of these models lies in their ability to take advantage of genomic data, local admixtures and group specific allele effects. While MAGBLUP 1 can be used to identify the additional genetic variance generated by admixture, MAGBLUP 2 can be used to analyze the conservation of SNP allele effects across genetic groups. Using the "Flint-Dent" panel and simulated traits, both models were effective to estimate their respective variance components and prove their efficiency in terms of genomic prediction accuracy compared to GBLUP. As expected , the gain was particularly important when both TS and PS included admixed individuals. However, no gain was observed for the real traits. The estimates of variances obtained with MAGBLUP 2 suggested a very limited contribution of group-specific deviation effects compared to main SNP effects. This could either suggest very conserved polygenic background effects, which is apparently contradictory with the conclusions of Chapter 2, or a lack of consistency between the SNP allele effects for a given allele ancestry. The underlying assumption of MAGBLUP 1 and 2 is indeed that group-specific allele effects only result from local genomic differences and not from epistatic interactions with the genetic background. Using the same panel, we also evaluated the benefits of including admixed individuals in multi-group TSs, in order to allow for more connection between genetic groups. Their interest was clearly shown using simulated traits, and was consistent with the results obtained by Toosi et al. (2013) using a simulated dataset. The simulation results were not confirmed for most of the real traits, for which assembling pure dent and flint lines was generally the most efficient. Here also, one hypothesis to explain the differences between simulated and real traits was the role of epistatic interactions between QTLs and the genetic background.

More generally, this thesis aimed at investigating the importance of the heterogeneity of SNP allele effects in structured maize diversity panels, resulting either from group-specific QTL allele effects or from group differences in LD between SNPs and QTLs. One the one hand, applying GWAS within the "Flint-Dent" panel highlighted a certain heterogeneity of allele effects for the detected QTLs. But on the other hand, applying MGBLUP in the "Amaizing-Dent" panel or MAGBLUP 1 and 2 in the "Flint-Dent" panel rather suggested a good conservation of the polygenic background across groups. As discussed by Ibánez-Escriche et al. (2009) and Technow et al. (2012), the divergence between the observed SNP effects across genetic groups should be more pronounced using low and medium density genotyping. This statement is based on the hypothesis that LD between SNPs and QTLs is more likely to differ between groups at these densities. Within the "Flint-Dent" panel, LD extent and linkage phase appear to be highly conserved over short distances between physically linked markers (Supplementary Fig. S2.4 and S2.5). Using high density genotyping, group-specific SNPs allele effects are likely to result mainly from interactions between QTLs and the genetic background, or group-specific mutations in the region of QTLs. With the advent of whole genome sequencing for all individuals, modeling group-specific allele effects for SNPs or structural variations should not be discarded as it should provide a good understanding of the genetic determinism of traits across genetic groups.

This thesis also aimed at studying the interest of using admixed individuals in quantitative genetics studies. From a GWAS perspective, they allowed to observe an allele with given group ancestry in different genetic backgrounds. Observing differences in allele effects within admixed individuals versus pure individuals suggests the existence of QTLs interacting with the genetic background. It might be questioned whether the detection of QTLs interacting with the genetic background would be studied more efficiently by directly modeling the interactions between QTLs, following Jannink (2007) or Crawford et al. (2017). From a GS perspective, admixed individuals were suggested as a solution to allow for more connections between genetic groups in multi-group TSs. Their efficiency was proven by Toosi et al. (2013) using simulations and was consistent with our observations on simulated traits. However, the efficiency of admixed individuals was more variable for the real traits tested which suggested complex genetic determinisms. More research needs to be conducted to decide on their interest in multi-group TSs. A simple question arises from this thesis concerning admixed individuals: does the cost of their production justify generating them? For academic purposes, the necessity to further evaluate the pioneer approach of this thesis totally justifies their production for other maize genetic groups or other plant and animal species. For breeding purposes, it remains unclear whether the expected results justify their cost. Beyond the specific case of the Dents and Flints in maize, the introduction of new sources of diversity in breeding programs generate admixed individuals. When introgressing genetic diversity, the expectation of a breeder is to generate additional genetic variance. Such variance can be finely modeled using MAGBLUP 1, as discussed in the perspectives.

Perspectives

The different results of this thesis confirmed a number of hypotheses related to the impact of genetic structure on quantitative genetics studies, such as the relevance of using multi-group TSs for genomic prediction. We found also more surprising results such as the importance of directional epistasis for maize flowering time, or the variable efficiency of admixed individuals to improve genomic prediction accuracy in multi-group TSs. Altogether, this thesis suggested interesting perspectives to get a further understanding of the impact of genetic structure in quantitative genetics studies.

Regarding the *a priori* estimation of accuracy, a new indicator called EthAcc was proposed recently by Mangin et al. (2019). Based on the results of Chapter 1, group differences in QTL allele frequencies are suspected to be as the main factor affecting the efficiency of CD-based *a priori* estimation of accuracy. The EthAcc indicator is based on the estimated effect of the main QTLs identified by GWAS and thus accounts for differences in allele frequencies at these same QTLs. It would therefore be very interesting to forecast the genomic prediction accuracy using the EhtAcc indicator in the "Amaizing-Dent" and the "Flint-Dent" panels, and compare its estimates to those based on CD. Further improvements could be obtained by using group-specific estimates of QTL allele effects for traits that show evidence of a heterogeneity of allele effects across genetic groups at main QTLs, such as flowering time in maize.

According to Chapter 2, there are interactions between QTLs and the genetic background for flowering time in maize, and these interactions helped to interpret the GS results of Chapter 3. It may be beneficial to perform genomic predictions using models that account directly for epistatic interactions between QTLs as proposed by Vitezica et al. (2017). However, this modeling may not be suited to the evidence of directional epistasis for flowering time. The problem of directional epistasis could be tackled by adding a covariate amounting the level of admixture of the individual, based on the analogy of adding a covariate to account for inbreeding depression in presence of directional dominance (Xiang et al., 2016).

In GWAS, the genetic background is classically modeled using a random effect of the polygenic background, as suggested by Yu et al. (2006). This polygenic effect is used to control for spurious associations due to the stratification of the population into groups of related individuals, possibly creating long-range LD between loci. Although no excess of false positives was apparent in Chapter 2 based on QQ-plot observation, a better modeling of the genetic background, by accounting for group-specific SNP allele effects, may reduce the residual error and increase statistical power. This would amount to modeling the polygenic background according to MAGBLUP 1 or 2. Very little gain is to be expected for flowering time in maize, based on the variance component estimates of Chapter 3, but this solution could be considered for traits showing evidence of a polygenic background with group-specific allele effects.

MAGBLUP 1 and 2 prove their complementarity when applied to admixed populations to get insights concerning the impact of genetic structure on given traits. In the short term, it would be interesting to apply these models to other ear-architecture traits, that were also evaluated in 2015 for the "Flint-Dent" panel. These traits included the length of the ear, the width of the ear, the number of rows, the number or kernel per row, the mean size of a kernel, and many others. Using these traits, we would have a new opportunity to evaluate the benefits of applying MAGBLUP models to predict breeding values. These traits could also be

of interest to reveal new evidence of group-specific SNP allele effects, possibly interacting with the genetic background, using our new GWAS methodology. In addition to studying complementary traits using the "Flint-Dent" panel, it would be interesting to apply these methods to other maize admixed populations and to other species. As both GWAS and GS methods were developed in the context of homozygous inbred lines, minor adjustments should be considered to account for heterozygosity, especially to be applied to animal species. Accounting for group-specific dominance effects could then be considered.

Using the "Flint-Dent" panel, we studied the interest of admixed individuals for traits evaluated *per se*. However, inbred lines are never used directly as cultivars in maize, but are further crossed to generate across-group hybrids. In order to study the interest of admixed individuals in a hybrid context, the 366 admixed lines have recently been crossed according to a highly incomplete diallel design to generate hybrids. These admixed hybrids were evaluated for both phenology and productivity traits. Using such genetic material, it should be possible to better understand the genetic determinism of heterosis. The observation of homozygous and heterozygous loci being locally "dent-dent", "flint-flint" or "dent-flint", should enable the identification of regions that require (i) the complementarity between flint and dent alleles versus both alleles with the same group ancestry, and (ii) the heterozygosity versus the homozygosity at the QTL. One could imagine a re-definition of heterotic groups according to the results. The benefit of using these hybrids in a generic TS for hybrid predictions could also be evaluated, possibly along with an adaptation of MAGBLUP to heterogeneous dominance effects depending on the combination of group ancestries of the two parental lines at the locus.

Both MAGBLUP models should also find interesting applications in the context of the exploitation of genetic resources. When crossing a diversity donor to a pool of elite genetic material, a breeder could get an estimate of the additional genetic variance generated by the donor within the progeny using MAGBLUP 1 (with the segregation variance). This additional variance is of great interest for a breeder because the genetic variance is involved in the calculation of the usefulness criterion which determines the interest of a cross. The conservation of QTL allele effects between the diversity donor and the elite material could be investigated using MAGBPLUP 2, and may lead to a gain in genomic prediction accuracy if the donor is highly distant of the elite material. Finally, applying our GWAS methodology could give a further insight concerning the cause of the heterogeneity of QTL allele effects, whether it comes from differences in LD or epistatic interactions with the genetic background.

Bibliography

Albrecht, T., Auinger, H.-J., Wimmer, V., Ogutu, J. O., Knaak, C., Ouzunova, M., Piepho, H.-P., and Schön, C.-C. (2014). Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. *Theoretical and Applied Genetics*, 127(6):1375–1386.

Alexander, D., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 19:1655–1664.

Astle, W. and Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471.

Badke, Y. M., Bates, R. O., Ernst, C. W., Schwab, C., and Steibel, J. P. (2012). Estimation of linkage disequilibrium in four us pig breeds. *BMC Genomics*, 13(1):24.

Barroso, I., Luan, J., Wheeler, E., Whittaker, P., Wasson, J., Zeggini, E., Weedon, M. N., Hunt, S., Venkatesh, R., Frayling, T. M., Delgado, M., Neuman, R. J., Zhao, J., Sherva, R., Glaser, B., Walker, M., Hitman, G., McCarthy, M. I., Hattersley, A. T., Permutt, M. A., Wareham, N. J., and Deloukas, P. (2008). Population-specific risk of type 2 diabetes conferred by hnf4a p2 promoter variants. *Diabetes*, 57(11):3161–3165.

Beavis, W. D. (1994). The power and deceit of qtl experiments: Lessons from comparative qtl studies. *Proceedings of the Forty-Ninth Annual Corn and Sorghum Industry Research Conference*, pages 250–266.

Beavis, W. D. (1998). *Molecular Dissection of Complex Traits*, chapter QTL analyses: Power, Precision and Accuracy, pages 145–162. CRC Press, New York.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300.

Bordes, J., Dumas de Vaulx, R., Lapierre, A., and Pollacsek, M. (1997). Haplodiploidization of maize (zea mays l) through induced gynogenesis assisted by glossy markers and its use in breeding. *Agronomie*, 17:291–297.

Bouché, F., Lobet, G., Tocquin, P., and Périlleux, C. (2015). FLOR-ID: an interactive database of flowering-time gene networks in Arabidopsis thaliana. *Nucleic Acids Research*, 44(D1):D1167–D1171.

Bouchet, S., Bertin, P., Presterl, T., Jamin, P., Coubriche, D., Gouesnard, B., Laborde, J., and Charcosset, A. (2016). Association mapping for phenology and plant architecture in maize shows higher power for developmental traits compared with growth influenced traits. *Heredity*, 118:249 EP –. Original Article.

Bouchet, S., Servin, B., Bertin, P., Madur, D., Combes, V., Dumas, F., Brunel, D., Laborde, J., Charcosset, A., and Nicolas, S. (2013). Adaptation of maize to temperate climates: Mid-density genome-wide association genetics and diversity patterns reveal key genomic regions, with a major contribution of the vgt2 (zcn8) locus. *PLOS ONE*, 8(8):1–17.

Brandenburg, J.-T., Mary-Huard, T., Rigaill, G., Hearne, S. J., Corti, H., Joets, J., Vitte, C., Charcosset, A., Nicolas, S. D., and Tenaillon, M. I. (2017). Independent introductions and admixtures have contributed to adaptation of european maize and its american counterparts. *PLOS Genetics*, 13(3):1–30.

Brard, S. and Ricard, A. (2015). Is the use of formulae a reliable way to predict the accuracy of genomic selection? *Journal of Animal Breeding and Genetics*, 132(3):207–217.

Brøndum, R., Rius-Vilarrasa, E., Strandén, I., Su, G., Guldbrandtsen, B., Fikse, W., and Lund, M. (2011). Reliabilities of genomic prediction using combined reference data of the nordic red dairy cattle populations. *Journal of Dairy Science*.

Browning, B. and Browning, S. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84:210–223.

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J. C., Goodman, M. M., Harjes, C., Guill, K., Kroon, D. E., Larsson, S., Lepak, N. K., Li, H., Mitchell, S. E., Pressoir, G., Peiffer, J. A., Rosas, M. O., Rocheford, T. R., Romay, M. C., Romero, S., Salvo, S., Villeda, H. S., Sofia da Silva, H., Sun, Q., Tian, F., Upadyayula, N., Ware, D., Yates, H., Yu, J., Zhang, Z., Kresovich, S., and McMullen, M. D. (2009). The genetic architecture of maize flowering time. *Science*, 325(5941):714–718.

Buitenhuis, B., Janss, L. L., Poulsen, N. A., Larsen, L. B., Larsen, M. K., and Sørensen, P. (2014). Genome-wide association and biological pathway analysis for milk-fat composition in danish holstein and danish jersey cattle. *BMC Genomics*, 15(1):1112.

Buitenhuis, B., Poulsen, N. A., Larsen, L. B., and Sehested, J. (2015). Estimation of genetic parameters and detection of quantitative trait loci for minerals in danish holstein and danish jersey milk. *BMC Genetics*, 16(1):52.

Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822–e1002822.

Butler, D. G., Cullis, B. R., Gilmour, A. R., and Gogel, B. J. (2009). *ASReml-R reference manual*.

Camus-Kulandaivelu, L., Veyrieras, J.-B., Madur, D., Combes, V., Fourmann, M., Barraud, S., Dubreuil, P., Gouesnard, B., Manicacci, D., and Charcosset, A. (2006). Maize adaptation to temperate climate: Relationship between population structure and polymorphism in the dwarf8 gene. *Genetics*, 172(4):2449–2463.

Carillier, C., Larroque, H., and Robert-Granié, C. (2014). Comparison of joint versus purebred genomic evaluation in the french multi-breed dairy goat population. *Genetics Selection Evolution*, 46(1):67.

Cavanagh, C., Morell, M., Mackay, I., and Powell, W. (2008). From mutations to magic: resources for gene discovery, validation and delivery in crop plants. *Current Opinion in Plant Biology*, 11(2):215 – 221. Genome studies and Molecular Genetics, edited by Juliette de Meaux and Maarten Koornneef / Plant Biotechnology, edited by Andy Greenland and Jan Leach.

Chardon, F., Hourcade, D., Combes, V., and Charcosset, A. (2005). Mapping of a spontaneous mutation for early flowering time in maize highlights contrasting allelic series at two-linked qtl on chromosome 8. *Theoretical and Applied Genetics*, 112(1):1–11.

Chardon, F., Virlon, B., Moreau, L., Falque, M., Joets, J., Decousset, L., Murigneux, A., and Charcosset, A. (2004). Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. *Genetics*, 168(4):2169–2185. 15611184[pmid].

Chen, L., Schenkel, F., Vinsky, M., Crews, D. H., and Li, C. (2013). Accuracy of predicting genomic breeding values for residual feed intake in angus and charolais beef cattle. *Journal of Animal Science*, 91:4669–4678.

Cole, J. B., Wiggans, G. R., Ma, L., Sonstegard, T. S., Lawlor, T. J., Crooker, B. A., Van Tassell, C. P., Yang, J., Wang, S., Matukumalli, L. K., and Da, Y. (2011). Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary u.s. holstein cows. *BMC Genomics*, 12(1):408.

Cole, J.B.and VanRaden, P., O'Connell, J., Van Tassell, C., Sonstegard, T., Schnabel, R., Taylor, J., and Wiggans, G. (2009). 'racial' differences in genetic effects for complex diseases. *Journal of Dairy Science*, 92(6):2931–2946.

Crawford, L., Zeng, P., Mukherjee, S., and Zhou, X. (2017). Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLOS Genetics*, 13(7):1–37.

Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLOS ONE*, 3(10):1–8.

de los Campos, G., Veturi, Y., Vazquez, A. I., Lehermeier, C., and Pérez-Rodríguez, P. (2015). Incorporating genetic heterogeneity in whole-genome regressions using interactions. *Journal of Agricultural, Biological, and Environmental Statistics*, 20(4):467–490.

de Roos, A. P. W., Hayes, B. J., and Goddard, M. E. (2009). Reliability of genomic predictions across multiple populations. *Genetics*, 183(4):1545–1553.

de Roos, A. P. W. M., Hayes, B. J., Spelman, R. J., and Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in holstein-friesian, jersey and angus cattle. *Genetics*, 179:1503–1512.

Ducrocq, S., Madur, D., Veyrieras, J.-B., Camus-Kulandaivelu, L., Kloiber-Maitz, M., Presterl, T., Ouzunova, M., Manicacci, D., and Charcosset, A. (2008). Key impact of vgt1 on flowering time adaptation in maize: Evidence from association mapping and ecogeographical information. *Genetics*, 178(4):2433–2437.

Duhnen, A., Gras, A., Teyssèdre, S., Romestant, M., Claustres, B., Daydé, J., and Mangin, B. (2017). Genomic selection for yield and seed protein content in soybean: A study of breeding program data and assessment of prediction accuracy. *Crop Sci.*, 57:1–13.

Durand, E., Bouchet, S., Bertin, P., Ressayre, A., Jamin, P., Charcosset, A., Dillmann, C., and Tenaillon, M. I. (2012). Flowering time in maize: Linkage and epistasis at a major effect locus. *Genetics*, 190(4):1547–1562.

Duvick, D. N. (2005). The contribution of breeding to yield advances in maize (zea mays l.). *Advances in Agronomy*, 86:83–145.

Elsen, J.-M. (2016). Approximated prediction of genomic selection accuracy when reference and candidate populations are related. *Genetics Selection Evolution*, 48(1):18.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. *PLOS ONE*, 6(5):1–10.

Endelman, J. B., Atlin, G. N., Beyene, Y., Semagn, K., Zhang, X., Sorrells, M. E., and Jannink, J. (2014). Optimal design of preliminary yield trials with genome-wide markers. *Crop Sci*, 54:48–59.

Erbe, M., Gredler, B., Seefried, F. R., Bapst, B., and Simianer, H. (2013). A function accounting for training set size and marker density to model the average accuracy of genomic prediction. *PLOS ONE*, 8(12):1–11.

Esfandyari, H., Sørensen, A. C., and Bijma, P. (2015). A crossbred reference population can improve the response to genomic selection for crossbred performance. *Genetics Selection Evolution*, 47(1):76.

Evangelou, E. and Ioannidis, J. P. A. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nature Genetics*, 14:379–389.

Evans, D. M. and Cardon, L. R. (2005). A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *The American Journal of Human Genetics*, 76:681–687.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.

FAO (2019). Maize production and area harvested in the world in 2017. `http://www.fao.org/faostat/en/#data/QC`. Accessed: 2019-02-24.

Fisher, R. A. (1918). Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433.

Flint-Garcia, S. A., Thuillet, A.-C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., Doebley, J., Kresovich, S., Goodman, M. M., and Buckler, E. S. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *The Plant Journal*, 44(6):1054–1064.

Foll, M. and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A bayesian perspective. *Genetics*, 180(2):977–993.

Ganal, M. W., Durstewitz, G., Polley, A., érard, A., Buckler, E. S., Charcosset, A., Clarke, J. D., Graner, E.-M., Hansen, M., Joets, J., Le Paslier, M.-C., McMullen, M. D., Montalent, P., Rose, M., Schön, C.-C., Sun, Q., Walter, H., Martin, O. C., and Falque, M. (2011). A large maize (zea mays l.) snp genotyping array: Development and germplasm genotyping, and genetic mapping to compare with the b73 reference genome. *PLOS ONE*, 6(12):1–15.

García-Cortés, L. A. and Toro, M. A. (2006). Multibreed analysis by splitting the breeding values. *Genet Sel Evol*, 38(6):601–615.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the bayesian alphabet. *Genetics*, 183(1):347–363.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., and Buckler, E. S. (2014). Tassel-gbs: A high capacity genotyping by sequencing analysis pipeline. *PLOS ONE*, 9(2):1–11.

Goddard, M., Hayes, B., and Meuwissen, T. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics*, 128(6):409–421.

Guo, Z., Tucker, D. M., Basten, C. J., Gandhi, H., Ersoz, E., Guo, B., Xu, Z., Wang, D., and Gay, G. (2014). The impact of population structure on genomic prediction in stratified populations. *Theoretical and Applied Genetics*, 127(3):749–762.

Hallauer, A. R., Russell, W. A., and Lamkey, K. R. (1988). chapter Corn Breeding1, pages 463–564. Agronomy Monograph. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI.

Hao, C., Wang, L., Ge, H., Dong, Y., and Zhang, X. (2011). Genetic diversity and linkage disequilibrium in chinese bread wheat (triticum aestivum l.) revealed by ssr markers. *PLOS ONE*, 6(2):1–13.

Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K., and Goddard, M. E. (2009). Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution*, 41(1):51.

Hayes, B. J., Corbet, N. J., Allen, J. M., Laing, A. R., Fordyce, G., McGowan, M. R., Lyons, R., and Burns, B. M. (2018). Towards multi-breed genomic evaluations for female fertility of tropical beef cattle1. *Journal of Animal Science*, 97(1):55–62.

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., and Goddard, M. E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in holstein cattle as contrasting model traits. *PLOS Genetics*, 6(9):1–11.

Heffner, E. L., Sorrells, M., and Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Science*, 49:1–12.

Helgadottir, A., Manolescu, A., Helgason, A., Thorleifsson, G., Thorsteinsdottir, U., Gudbjartsson, D. F., Gretarsdottir, S., Magnusson, K. P., Gudmundsson, G., Hicks, A., Jonsson, T., Grant, S. F. A., Sainz, J., O'Brien, S. J., Sveinbjornsdottir, S., Valdimarsson, E. M., Matthiasson, S. E., Levey, A. I., Abramson, J. L., Reilly, M. P., Vaccarino, V., Wolfe, M. L., Gudnason, V., Quyyumi, A. A., Topol, E. J., Rader, D. J., Thorgeirsson, G., Gulcher, J. R., Hakonarson, H., Kong, A., and Stefansson, K. (2006). A variant of the gene encoding leukotriene a4 hydrolase confers ethnicity-specific risk of myocardial infarction. *Nature Genetics*, 38(1):68–74.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2):423–447.

Henderson, C. R. (1984). *Applications of linear models in animal breeding*. Guelph : University of Guelph.

Heslot, N., Jannink, J.-L., and Sorrells, M. (2015). Perspectives for genomic selection applications and research in plants. *Crop Science*, 55:1–12.

Heslot, N., Yang, H., Sorrells, M. E., and Jannink, J. (2012). Genomic selection in plant breeding: A comparison of models. *Crop Sci.*, 52:146–160.

Holsinger, K. E. and Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting fst. *Nature Reviews Genetics*, 10:639 EP –. Review Article.

Huang, X. and Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annual Review of Plant Biology*, 65(1):531–551.

Ibánez-Escriche, N., Fernando, R. L., Toosi, A., and Dekkers, J. C. (2009). Genomic selection of purebreds for crossbred performance. *Genetics Selection Evolution*, 41(1):12.

Ioannidis, J. P. A., Ntzani, E. E., and Trikalinos, T. A. (2004). 'racial' differences in genetic effects for complex diseases. *Nature Genetics*, 36(12):1312–1318.

Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics*, 128(1):145–158.

Jannink, J.-L. (2007). Identifying quantitative trait locus by genetic background interactions in association studies. *Genetics*, 176(1):553–561.

Karoui, S., Carabaño, M. J., Díaz, C., and Legarra, A. (2012). Joint genomic evaluation of french dairy cattle breeds using multiple-trait models. *Genetics Selection Evolution*, 44(1):39.

Kruuk, L. E. B. (2004). Estimating genetic parameters in natural populations using the "animal model". *Philos Trans R Soc Lond B Biol Sci*, 359(1446):873–890. 15306404[pmid].

Lande, R. (1981). The minimum number of genes contributing to quantitative variation between and within populations. *Genetics*, 99(3-4):541–553.

Laporte, F., Charcosset, A., and Mary-Huard, T. (2019). Efficient reml inference in variance component mixed models using min-max algorithms. Manuscript submitted for publication. https://CRAN.R-project.org/package=MM4LMM.

Larsson, S. J., Lipka, A. E., and Buckler, E. S. (2013). Lessons from dwarf8 on the strengths and weaknesses of structured association mapping. *PLOS Genetics*, 9(2):1–11.

Le Rouzic, A. (2014). Estimating directional epistasis. *Frontiers in Genetics*, 5:198.

Legarra, A., Christensen, O. F., Vitezica, Z. G., Aguilar, I., and Misztal, I. (2015). Ancestral relationships using metafounders: Finite ancestral populations and across population relationships. *Genetics*, 200(2):455–468.

Lehermeier, C., Krämer, N., Bauer, E., Bauland, C., Camisan, C., Campo, L., Flament, P., Melchinger, A. E., Menz, M., Meyer, N., Moreau, L., Moreno-González, J., Ouzunova, M., Pausch, H., Ranc, N., Schipprack, W., Schönleben, M., Walter, H., Charcosset, A., and Schön, C.-C. (2014). Usefulness of multiparental populations of maize (zea mays l.) for genome-based prediction. *Genetics*, 198(1):3–16.

Lehermeier, C., Schön, C.-C., and de los Campos, G. (2015). Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. *Genetics*, 201(1):323–337.

Li, D., Wang, X., Zhang, X., Chen, Q., Xu, G., Xu, D., Wang, C., Liang, Y., Wu, L., Huang, C., Tian, J., Wu, Y., and Tian, F. (2016a). The genetic architecture of leaf number and its genetic relationship to flowering time in maize. *New Phytologist*, 210(1):256–268.

Li, Y. R. and Keating, B. J. (2014). Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Medicine*, 6(10):91.

Li, Y.-x., Li, C., Bradbury, P. J., Liu, X., Lu, F., Romay, C. M., Glaubitz, J. C., Wu, X., Peng, B., Shi, Y., Song, Y., Zhang, D., Buckler, E. S., Zhang, Z., Li, Y., and Wang, T. (2016b). Identification of genetic variants associated with maize flowering time using an extremely large multi-genetic background population. *The Plant Journal*, 86(5):391–402.

Liang, Y., Liu, Q., Wang, X., Huang, C., Xu, G., Hey, S., Lin, H.-Y., Li, C., Xu, D., Wu, L., Wang, C., Wu, W., Xia, J., Han, X., Lu, S., Lai, J., Song, W., Schnable, P. S., and Tian, F. (2019). Zmmads69 functions as a flowering activator through the zmrap2.7-zcn8 regulatory module and contributes to maize flowering time adaptation. *New Phytologist*, 221(4):2335–2347.

Liu, Y., Nyunoya, T., Leng, S., Belinsky, S. A., Tesfaigzi, Y., and Bruse, S. (2013). Softwares and methods for estimating genetic ancestry in human populations. *Hum Genomics*, 7(1):1–1. 23289408[pmid].

Lo, L. L., Fernando, R. L., and Grossman, M. (1993). Covariance between relatives in multibreed populations: additive model. *Theoretical and Applied Genetics*, 87(4):423–430.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

Makgahlela, M., Mäntysaari, E., Strandén, I., Koivula, M., Nielsen, U., Sillanpää, M., and Juga, J. (2013). Across breed multi-trait random regression genomic predictions in the nordic red dairy cattle. *Journal of Animal Breeding and Genetics*, 130(1):10–19.

Mangin, B., Rincent, R., Rabier, C.-E., Moreau, L., and Goudemand-Dugue, E. (2019). Training set optimization of genomic prediction by means of ethacc. *PLOS ONE*, 14(2):1–21.

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., and Cierco-Ayrolles, C. (2011). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity*, 108:285 – 291.

Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). Rfmix: A discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, 93(2):278–288.

Marigorta, U. M. and Navarro, A. (2013). High trans-ethnic replicability of gwas results implies common causal variants. *PLOS Genetics*, 9(6):1–13.

Mather, K. (1967). Complementary and duplicate gene interactions in biometrical genetics. *Heredity*, 22:97 EP –. Original Article.

Matsuoka, Y., Vigouroux, Y., Goodman, M., Sanchez G, J., Buckler, E., and Doebley, J. (2002). A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences of the United States of America*, 99:6080–4.

Mauricio, R. (2001). Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. *Nature Reviews Genetics*, 2:370 EP –. Review Article.

McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S. E., Peterson, B., Pressoir, G., Romero, S., Rosas, M. O., Salvo, S., Yates, H., Hanson, M., Jones, E., Smith, S., Glaubitz, J. C., Goodman, M., Ware, D., Holland, J. B., and Buckler, E. S. (2009). Genetic properties of the maize nested association mapping population. *Science*, 325(5941):737–740.

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

Millet, E. J., Welcker, C., Kruijer, W., Negro, S., Coupel-Ledru, A., Nicolas, S. D., Laborde, J., Bauland, C., Praud, S., Ranc, N., Presterl, T., Tuberosa, R., Bedo, Z., Draye, X., Usadel, B., Charcosset, A., Van Eeuwijk, F., and Tardieu, F. (2016). Genome-wide analysis of yield in europe: Allelic effects vary with drought and heat scenarios. *Plant Physiology*, 172(2):749–764.

Moreau, L., Charcosset, A., and Gallais, A. (2004). Experimental evaluation of several cycles of marker-assisted selection in maize. *Euphytica*, 137(1):111–118.

Neuman, R. J., Wasson, J., Atzmon, G., Wainstein, J., Yerushalmi, Y., Cohen, J., Barzilai, N., Blech, I., Glaser, B., and Permutt, M. A. (2010). Gene-gene interactions lead to higher risk for development of type 2 diabetes in an ashkenazi jewish population. *PLOS ONE*, 5(3):1–6.

Ntzani, E. E., Liberopoulos, G., Manolio, T. A., and Ioannidis, J. P. A. (2012). Consistency of genome-wide associations across major ancestral groups. *Human Genetics*, 131(7):1057–1071.

Olson, K. M., Van Raden, P. M., and Tooker, M. E. (2012). Multibreed genomic evaluations using purebred holsteins, jerseys, and brown swiss. *Journal of Dairy Science*, 95(9):5378–5383.

Padhukasahasram, B. (2014). Inferring ancestry from population genomic data and its applications. *Frontiers in Genetics*, 5:204.

Pilet-Nayel, M.-L., Moury, B., Caffier, V., Montarry, J., Kerlan, M.-C., Fournet, S., Durel, C.-E., and Delourme, R. (2017). Quantitative resistance to plant pathogens in pyramiding strategies for durable crop protection. *Frontiers in Plant Science*, 8:1838.

Plieschke, L., Edel, C., Pimentel, E. C., Emmerling, R., Bennewitz, J., and Götz, K.-U. (2015). A simple method to separate base population and segregation effects in genomic relationship matrices. *Genetics Selection Evolution*, 47(1):53.

Porto-Neto, L. R., Kijas, J. W., and Reverter, A. (2014). The extent of linkage disequilibrium in beef cattle breeds using high-density snp genotypes. *Genetics Selection Evolution*, 46(1):22.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:904–909.

Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data. *The American Journal of Human Genetics*, 69(1):1–14.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Pryce, J. E., Gredler, B., Bolormaa, S., Bowman, P. J., Egger-Danner, C., Fuerst, C., Emmerling, R., Solkner, J., Goddard, M. E., and Hayes, B. J. (2011). Short communication: Genomic selection using a multi-breed, across-country reference population. *Journal of Dairy Science*, 94(5):2625–2630.

Pszczola, M., Strabel, T., Mulder, H., and Calus, M. (2012). Reliability of direct genomic values for animals with different relationships within and to the reference population. *Journal of Dairy Science*, 95(1):389 – 400.

Rabier, C.-E., Barre, P., Asp, T., Charmet, G., and Mangin, B. (2016). On the accuracy of genomic selection. *PLOS ONE*, 11(6):1–23.

Raven, L.-A., Cocks, B. G., and Hayes, B. J. (2014). Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. *BMC Genomics*, 15(1):62.

Revilla, P., Rodríguez, V. M., Ordás, A., Rincent, R., Charcosset, A., Giauffret, C., Melchinger, A. E., Schön, C.-C., Bauer, E., Altmann, T., Brunel, D., Moreno-González, J., Campo, L., Ouzunova, M., Álvarez, Á., Ruíz de Galarreta, J. I., Laborde, J., and Malvar, R. A. (2016). Association mapping for cold tolerance in two large maize inbred panels. *BMC Plant Biology*, 16(1):127.

Rincent, R., Charcosset, A., and Moreau, L. (2017). Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. *Theoretical and Applied Genetics*, 130(11):2231–2247.

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodríguez, V., Moreno-Gonzalez, J., Melchinger, A., Bauer, E., Schoen, C.-C., Meyer, N., Giauffret, C., Bauland, C., Jamin, P., Laborde, J., Monod, H., Flament, P., Charcosset, A., and Moreau, L. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (zea mays l.). *Genetics*, 192(2):715–728.

Rincent, R., Moreau, L., Monod, H., Kuhn, E., Melchinger, A. E., Malvar, R. A., Moreno-Gonzalez, J., Nicolas, S., Madur, D., Combes, V., Dumas, F., Altmann, T., Brunel, D., Ouzunova, M., Flament, P., Dubreuil, P., Charcosset, A., and Mary-Huard, T. (2014a). Recovering power in association mapping panels with variable levels of linkage disequilibrium. *Genetics*, 197(1):375–387.

Rincent, R., Nicolas, S., Bouchet, S., Altmann, T., Brunel, D., Revilla, P., Malvar, R. A., Moreno-Gonzalez, J., Campo, L., Melchinger, A. E., Schipprack, W., Bauer, E., Schoen, C.-C., Meyer, N., Ouzunova, M., Dubreuil, P., Giauffret, C., Madur, D., Combes, V., Dumas, F., Bauland, C., Jamin, P., Laborde, J., Flament, P., Moreau, L., and Charcosset, A. (2014b). Dent and flint maize diversity panels reveal important genetic potential for increasing biomass production. *Theoretical and Applied Genetics*, 127(11):2313–2331.

Rio, S., Mary-Huard, T., Moreau, L., and Charcosset, A. (2019). Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. *Theoretical and Applied Genetics*, 132(1):81–96.

Rius, M. and Darling, J. A. (2014). How important is intraspecific genetic admixture to the success of colonising populations? *Trends in Ecology & Evolution*, 29(4):233 – 242.

Rogers, A. R. (2014). How population growth affects linkage disequilibrium. *Genetics*, 197(4):1329–1341.

Romay, M. C., Millard, M. J., Glaubitz, J. C., Peiffer, J. A., Swarts, K. L., Casstevens, T. M., Elshire, R. J., Acharya, C. B., Mitchell, S. E., Flint-Garcia, S. A., McMullen, M. D., Holland, J. B., Buckler, E. S., and Gardner, C. A. (2013). Comprehensive genotyping of the usa national maize inbred seed bank. *Genome Biology*, 14(6):R55.

Salvi, S., Corneti, S., Bellotti, M., Carraro, N., Sanguineti, M. C., Castelletti, S., and Tuberosa, R. (2011). Genetic dissection of maize phenology using an intraspecific introgression library. *BMC plant biology*, 11:4.

Salvi, S., Emanuelli, F., Soriano, J. M., Zamariola, L., Giuliani, S., Bovina, R., Ormanbekova, D., Koumproglou, R., Burdo, B., Rouster, J., Wyatt, P., Tuberosa, R., Jahrmann, T., Kaeppler, S., Praud, S., and Salvi, S. (2017). Cloning of vgt3, a major qtl for flowering time in maize. In *59th Annual Maize Genetics Conference*.

Salvi, S., Sponza, G., Morgante, M., Tomes, D., Niu, X., Fengler, K. A., Meeley, R., Ananiev, E. V., Svitashev, S., Bruggemann, E., Li, B., Hainey, C. F., Radovic, S., Zaina, G., Rafalski, J.-A., Tingey, S. V., Miao, G.-H., Phillips, R. L., and Tuberosa, R. (2007). Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proceedings of the National Academy of Sciences of the United States of America*, 104(27):11376–11381.

Sanchez, M.-P., Govignon-Gion, A., Croiseau, P., Fritz, S., Hozé, C., Miranda, G., Martin, P., Barbat-Leterrier, A., Letaïef, R., Rocha, D., Brochard, M., Boussaha, M., and Boichard, D. (2017). Within-breed and multi-breed gwas on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. *Genetics Selection Evolution*, 49(1):68.

Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008). Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, 82(2):290 – 303.

Sawyer, S. L., Mukherjee, N., Pakstis, A. J., Feuk, L., Kidd, J. R., Brookes, A. J., and Kidd, K. K. (2005). Linkage disequilibrium patterns vary substantially among populations. *European Journal Of Human Genetics*, 13:677–686.

Schopp, P., Müller, D., Wientjes, Y. C. J., and Melchinger, A. E. (2017). Genomic prediction within and across biparental families: Means and variances of prediction accuracy and usefulness of deterministic equations. *G3: Genes, Genomes, Genetics*.

Schulz-Streeck, T., Ogutu, J. O., Karaman, Z., Knaak, C., and Piepho, H. P. (2012). Genomic selection using multiple populations. *Crop Science*, 52:2453–2461.

Searle, S. R., Casella, G., and McCulloch, C. E. (2008). *Prediction of Random Variables*, chapter 7, pages 258–289. John Wiley & Sons, Ltd.

Spitze, K. (1993). Population structure in daphnia obtusa: quantitative genetic and allozymic variation. *Genetics*, 135(2):367–374.

Strandén, I. and Mäntysaari, E. A. (2013). Use of random regression model as an alternative for multibreed relationship matrix. *Journal of Animal Breeding and Genetics*, 130(1):4–9.

Stryjecki, C., Alyass, A., and Meyre, D. (2018). Ethnic and population differences in the genetic predisposition to human obesity. *Obesity Reviews*, 19(1):62–80.

Tang, H. (2006). Confronting ethnicity-specific disease risk. *Nature Genetics*, 38(1):12–15.

Technow, F., Burger, A., and Melchinger, A. E. (2013). Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. *G3: Genes/Genomes/Genetics*, 3(2):197–203.

Technow, F., Riedelsheimer, C., Schrag, T. A., and Melchinger, A. E. (2012). Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theoretical and Applied Genetics*, 125(6):1181–1194.

Technow, F. and Totir, L. R. (2015). Using bayesian multilevel whole genome regression models for partial pooling of training sets in genomic prediction. *G3: Genes, Genomes, Genetics*, 5(8):1603–1612.

Tenaillon, M. I. and Charcosset, A. (2011). A european perspective on maize history. *Comptes Rendus Biologies*, 334(3):221 – 228. On the trail of domestications, migrations and invasions in agriculture.

Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D., and Buckler IV, E. S. (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics*, 28:286 EP –.

Thornton, T., Tang, H., Thomas, J., Heather, M., Bette, J., and Risch, N. (2012). Estimating kinship in admixed populations. *The American Journal of Human Genetics*, 91:122–138.

Toosi, A., Fernando, R., and Dekkers, J. (2013). Genomic selection in admixed and crossbred populations. *Journal of Animal Scien,ce*, 130(1):10–19.

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., Meitinger, T., Strom, T. M., Fries, R., Pausch, H., Bertani, C., Davassi, A., Mayer, K. F., and Schön, C.-C. (2014). A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k snp genotyping array. *BMC Genomics*, 15(1):823.

van den Berg, I., Boichard, D., and Lund, M. S. (2016). Comparing power and precision of within-breed and multibreed genome-wide association studies of production traits using whole-genome sequence data for 5 french and danish dairy cattle breeds. *Journal of Dairy Science*, 99(11):8932–8945.

Van Inghelandt, D., Reif, J. C., Dhillon, B. S., Flament, P., and Melchinger, A. E. (2011). Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm. *Theoretical and Applied Genetics*, 123(1):11–20.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423.

Vitezica, Z. G., Legarra, A., Toro, M. A., and Varona, L. (2017). Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. *Genetics*, 206(3):1297–1307.

Whittaker, J. C., Thompson, R., and Denham, M. C. (2000). Marker-assisted selection using ridge regression. *Genetical Research*, 75(2):249–252.

Wientjes, Y. C., Veerkamp, R. F., Bijma, P., Bovenhuis, H., Schrooten, C., and Calus, M. P. (2015a). Empirical and deterministic accuracies of across-population genomic prediction. *Genetics Selection Evolution*, 47(1):5.

Wientjes, Y. C., Veerkamp, R. F., and Calus, M. P. (2015b). Using selection index theory to estimate consistency of multi-locus linkage disequilibrium across populations. *BMC Genetics*, 16(1):87.

Wientjes, Y. C. J., Bijma, P., Vandenplas, J., and Calus, M. P. L. (2017). Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations. *Genetics*, 207(2):503–515.

Wientjes, Y. C. J., Bijma, P., Veerkamp, R. F., and Calus, M. P. L. (2016). An equation to predict the accuracy of genomic values by combining data from multiple traits, populations, or environments. *Genetics*, 202(2):799–823.

Wientjes, Y. C. J., Calus, M. P. L., Hayes, B. J., Goddard, M. E., and Hayes, B. J. (2015c). Impact of qtl properties on the accuracy of multi-breed genomic prediction. *Genetics Selection Evolution*, 47(1):42.

Wright, S. (1943). Isolation by distance. *Genetics*, 28(2):114–138.

Wright, S. (1949). The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354.

Xiang, T., Christensen, O. F., Vitezica, Z. G., and Legarra, A. (2016). Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. *Genetics Selection Evolution*, 48(1):92.

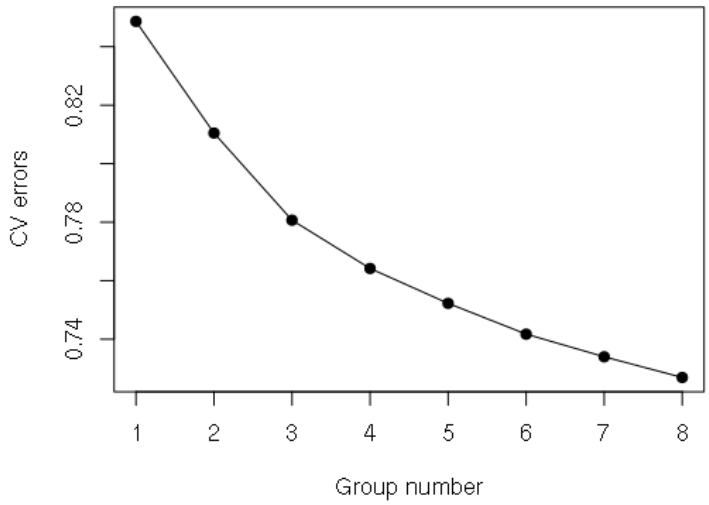
Young, N. D. (1996). Young nd. qtl mapping and quantitative disease resistance in plants. annu rev phytopathol 34: 479-501. *Annual review of phytopathology*, 34:479–501.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38:203–208.

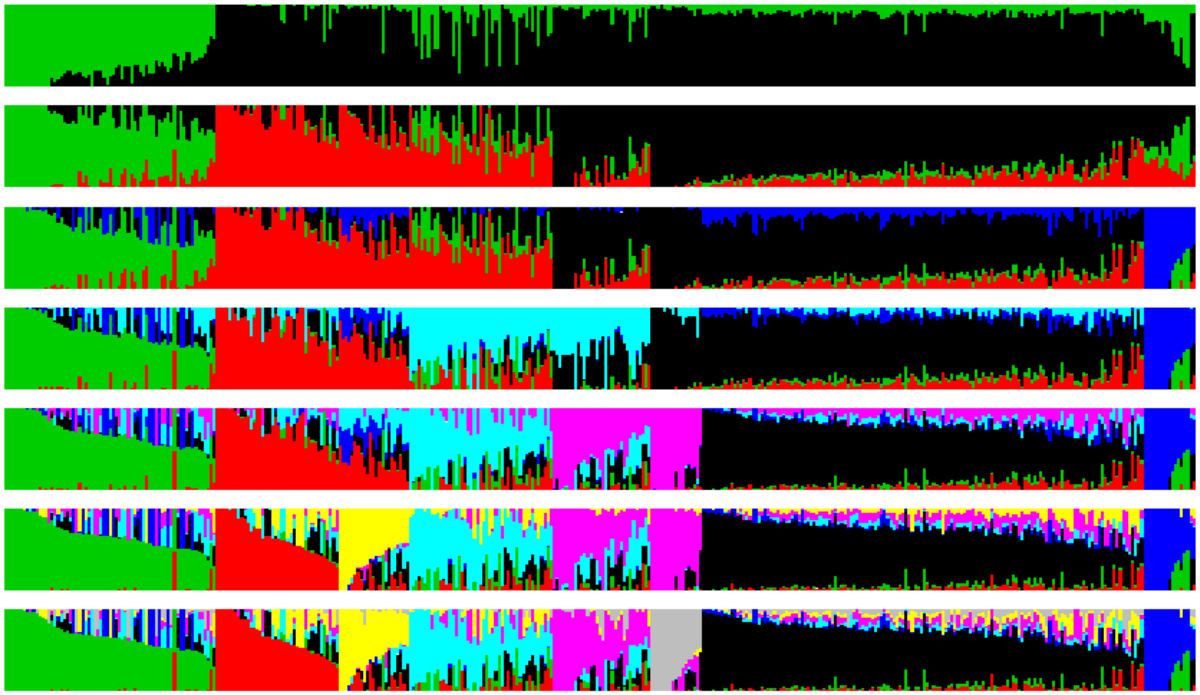
Zhou, L., Ding, X., Zhang, Q., Wang, Y., Lund, M. S., and Su, G. (2013). Consistency of linkage disequilibrium between chinese and nordic holsteins and genomic prediction for chinese holsteins using a joint reference population. *Genetics Selection Evolution*, 45(1):7.

Zhu, C., Gore, M., Buckler, E., and Yu, J. (2008). Status and prospects of association mapping in plants. *The Plant Genome*, 1:5–20.

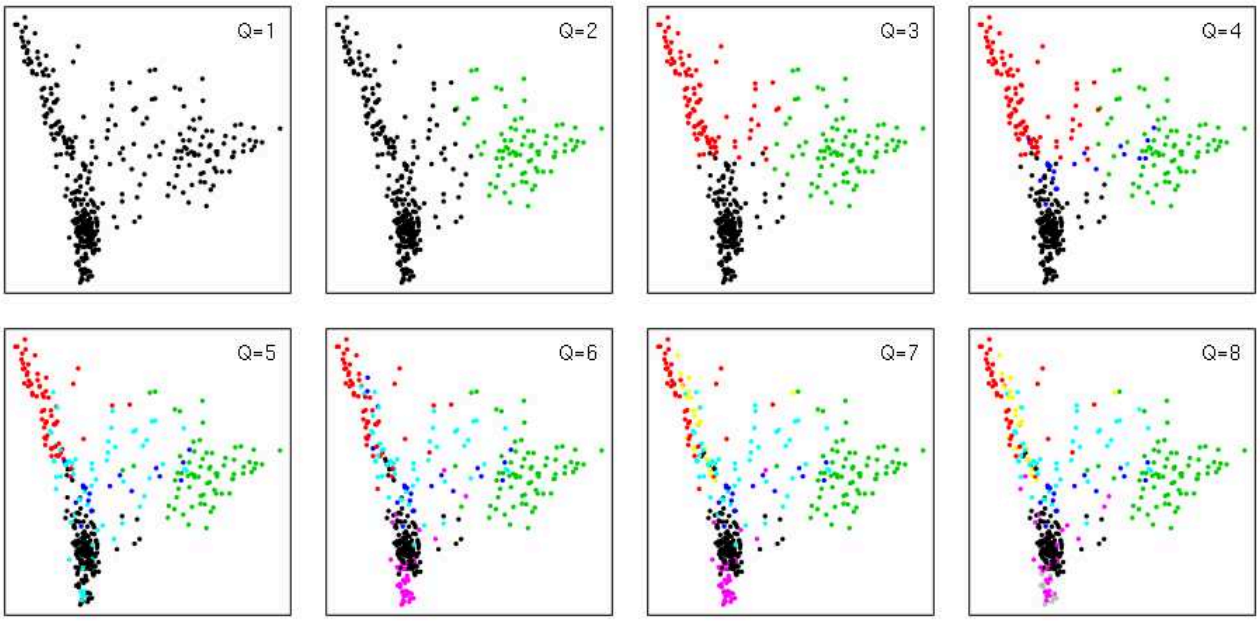
Supplementary Material - Chapitre 1



Supplementary Fig. S1.1: Evolution of the CV error criterion while increasing the number of groups from 2 to 8 using ADMIXTURE software on SNP data



Supplementary Fig. S1.2: Bar-plot featuring individual admixture proportion for a number of groups ranging from 2 to 8. Individuals are ordered using their maximal admixture coefficient at $Q = 8$. Each step from Q to $Q + 1$ groups can either be characterized by the subdivision of one group into two others (ex from $Q = 2$ to $Q = 3$) or the union of individuals of several groups to form a new one (from $Q = 3$ to $Q = 4$)



Supplementary Fig. S1.3: PCoA on genetic distances with coloration of individuals using their maximal admixture coefficient for different number of groups K

Trait	Model	Location	Plot h^2	Repetitions	h^2	Mean	Genotypic σ^2	Residual σ^2
Grain Yield	Block AR1	Graneros	0.59	1.91	0.74	87.44	131.1	89.9
Grain Yield	Block AR1	Bernburg	0.66	1.94	0.79	87.72	101.7	52
Grain Yield	Block AR1	Niederhergheim	0.68	1.86	0.79	77.89	106.1	51
Grain Yield	Block AR1	Villampuy	0.62	1.93	0.76	77.1	88.1	55
Grain Yield	replicate	Blois	0.7	1.79	0.81	93.51	100.3	42.1
Grain Yield	Rep row col	Mons	0.55	1.94	0.7	81.42	118.4	98.8
Grain Yield	Rep row col	Souprosse	0.82	1.97	0.9	88.15	102	22.3
Grain Moisture	Bloc AR1	Graneros	0.81	1.99	0.89	19.67	5.3349	1.251
Grain Moisture	Block AR1	Bernburg	0.92	2	0.96	30.61	4.7025	0.3837
Grain Moisture	Block AR1	Blois	0.81	1.82	0.89	30.36	3.8288	0.8748
Grain Moisture	Block AR1	Niederhergheim	0.87	1.95	0.93	22.68	5.332	0.8321
Grain Moisture	Block AR1	Souprosse	0.87	2	0.93	26.8	4.6982	0.6743
Grain Moisture	Block AR1	Villampuy	0.67	2.01	0.81	29.87	3.5885	1.7298
Grain Moisture	Rep row col	Mons	0.88	2	0.94	32.38	3.2301	0.4349
Yield Index	Block AR1	Graneros	0.57	1.91	0.71	38.32	118.3	90.4
Yield Index	Block AR1	Niederhergheim	0.62	1.86	0.76	21.27	98	59.2
Yield Index	Rep row col	Bernburg	0.7	1.94	0.82	11.17	103.2	45.1
Yield Index	Rep row col	Blois	0.68	1.79	0.79	17.68	95.9	46
Yield Index	Rep row col	Mons	0.56	1.94	0.71	0.55	133.7	105
Yield Index	Rep row col	Souprosse	0.77	1.97	0.87	21.18	85.9	26.3
Yield Index	Rep row col	Villampuy	0.56	1.93	0.71	2.47	84.3	65
Male Flowering	Bloc AR1	Graneros	0.28	2	0.44	914.67	340.2	859.8
Male Flowering	Block sub-block	Mons	0.81	2.01	0.89	931.84	1450.1	351.2
Male Flowering	Block sub-block	Niederhergheim	0.74	2.01	0.85	922.77	1339.4	471.9
Male Flowering	Rep row col	Bernburg	0.78	2.01	0.87	805.13	692.6	199.6
Male Flowering	Rep row col	Blois	0.86	2.01	0.93	853.37	1200.9	190.2
Male Flowering	Rep row col	Souprosse	0.82	2.01	0.9	887.41	1023.8	226.3
Male Flowering	Rep row col	Villampuy	0.77	2	0.87	927.38	1046.8	306

Supplementary Table S1.1: Information on phenotypic data in each field trials where Plot h^2 is the plot heritability, Repetition is the average number of repetitions of each genotype, h^2 is the heritability of the genotypic mean, Genotypic and Residual σ^2 are the genotypic and residual variances respectively. The "Model" column define the model chose to correct for within trials spatial effects using AIC criterion:

Replicate model: $Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}$ where Y_{ij} is the phenotype of genotype i in block j, μ is the intercept, α_i is the fixed effect of genotype i, β_j is the fixed effect of block j, E_{ij} is the error where $E_{ij} \sim \mathcal{N}(0, \sigma^2_E)$ i.i.d. and σ^2_E is the error variance.

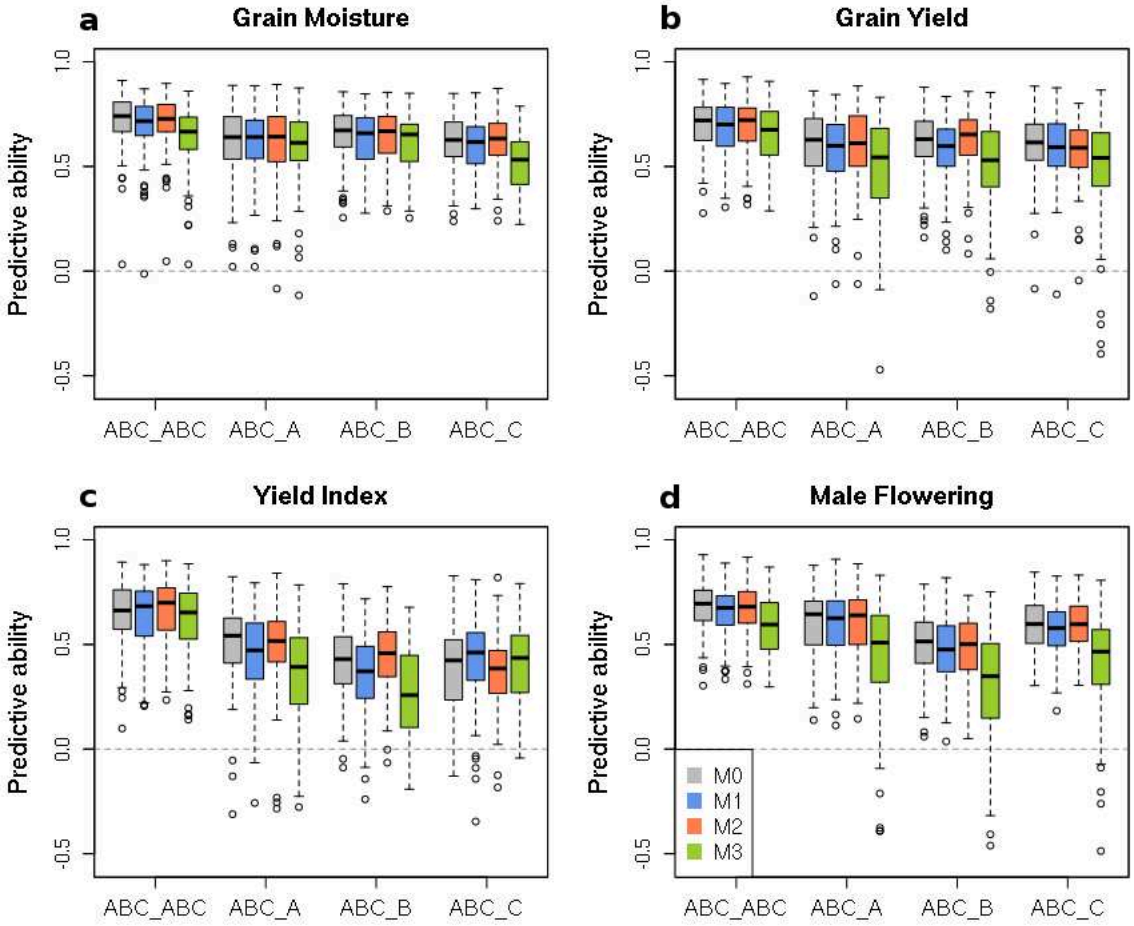
Block AR1: $Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}$ is identical to Replicate model but $\boldsymbol{e} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_r \otimes \boldsymbol{\Sigma}_c \sigma^2_E)$ where $\boldsymbol{e}$ is the vector of errors, $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_c$ are the autoregressive correlation matrices for row and columns respectively.

Block sub-block: $Y_{ijk} = \mu + \alpha_i + \beta_j + B_{k(j)} + E_{ijk}$ is identical to Replicate model but $B_{k(j)}$ is the random effect of sub-block k with $B_{k(j)} \sim \mathcal{N}(0, \sigma^2_B)$ i.i.d.

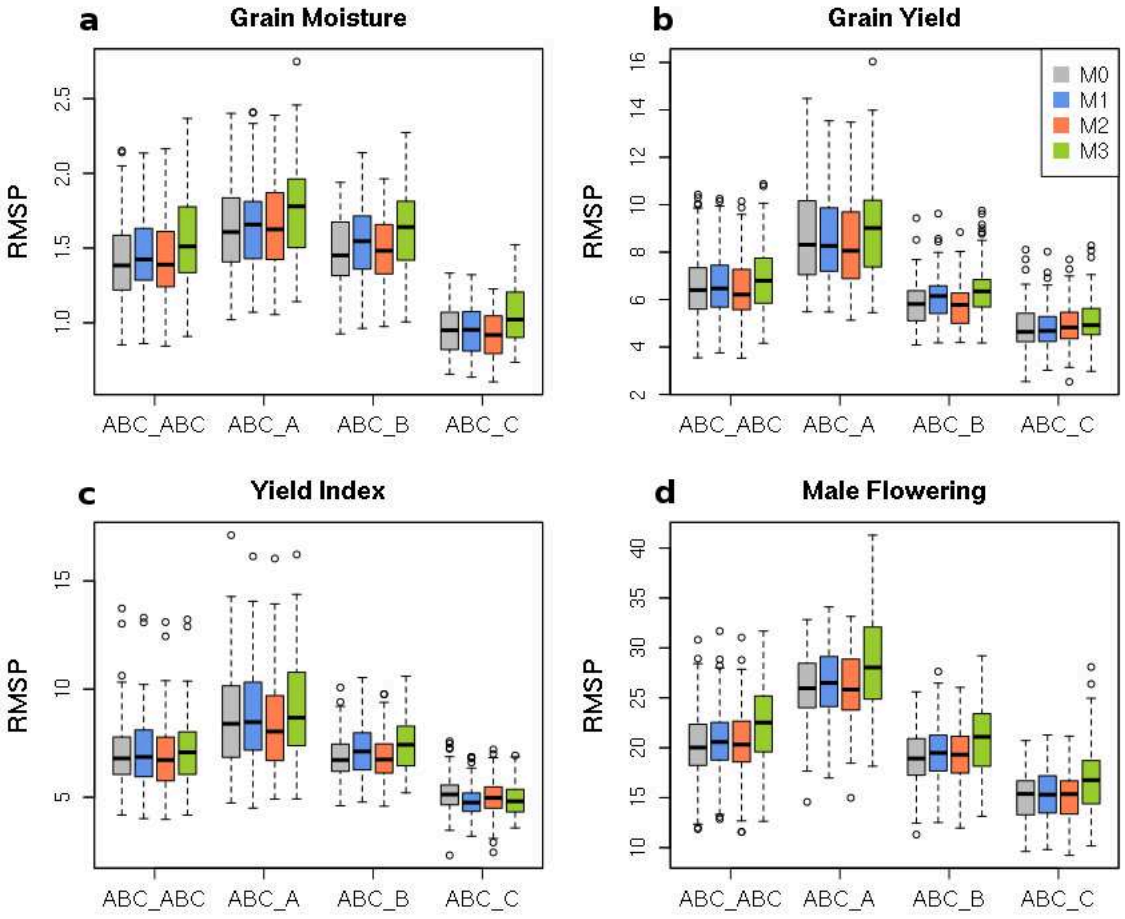
Rep row col: $Y_{ijrl} = \mu + \alpha_i + \beta_j + X_r + Y_c + E_{ijrl}$ is identical to Replicate model but X_r and X_c are the row and column random effects respectively where $X_r \sim \mathcal{N}(0, \sigma^2_X)$ i.i.d. and $Y_c \sim \mathcal{N}(0, \sigma^2_Y)$ i.i.d.

Scenario	Grain Moisture	Grain Yield	Yield Index	Male Flowering
ABC_ABC	0.72 (0.12)	0.70 (0.13)	0.65 (0.15)	0.68 (0.12)
ABC_A	0.61 (0.16)	0.60 (0.17)	0.51 (0.19)	0.60 (0.15)
A_A	0.68 (0.12)	0.65 (0.13)	0.57 (0.15)	0.68 (0.12)
BC_A	0.54 (0.17)	0.47 (0.17)	0.23 (0.23)	0.54 (0.16)
B_A	0.49 (0.20)	0.51 (0.17)	0.28 (0.20)	0.53 (0.16)
C_A	0.57 (0.16)	0.45 (0.17)	0.14 (0.22)	0.51 (0.16)
ABC_B	0.66 (0.13)	0.61 (0.14)	0.42 (0.17)	0.50 (0.15)
B_B	0.69 (0.10)	0.65 (0.11)	0.48 (0.15)	0.56 (0.13)
AC_B	0.58 (0.16)	0.50 (0.18)	0.40 (0.18)	0.50 (0.15)
A_B	0.49 (0.18)	0.48 (0.17)	0.27 (0.21)	0.45 (0.18)
C_B	0.60 (0.13)	0.53 (0.15)	0.31 (0.19)	0.45 (0.16)
ABC_C	0.61 (0.14)	0.60 (0.14)	0.38 (0.19)	0.59 (0.12)
C_C	0.52 (0.12)	0.59 (0.13)	0.51 (0.13)	0.53 (0.14)
AB_C	0.61 (0.15)	0.57 (0.13)	0.22 (0.23)	0.60 (0.15)
A_C	0.56 (0.15)	0.50 (0.16)	0.05 (0.26)	0.57 (0.13)
B_C	0.55 (0.17)	0.43 (0.17)	0.21 (0.19)	0.40 (0.20)

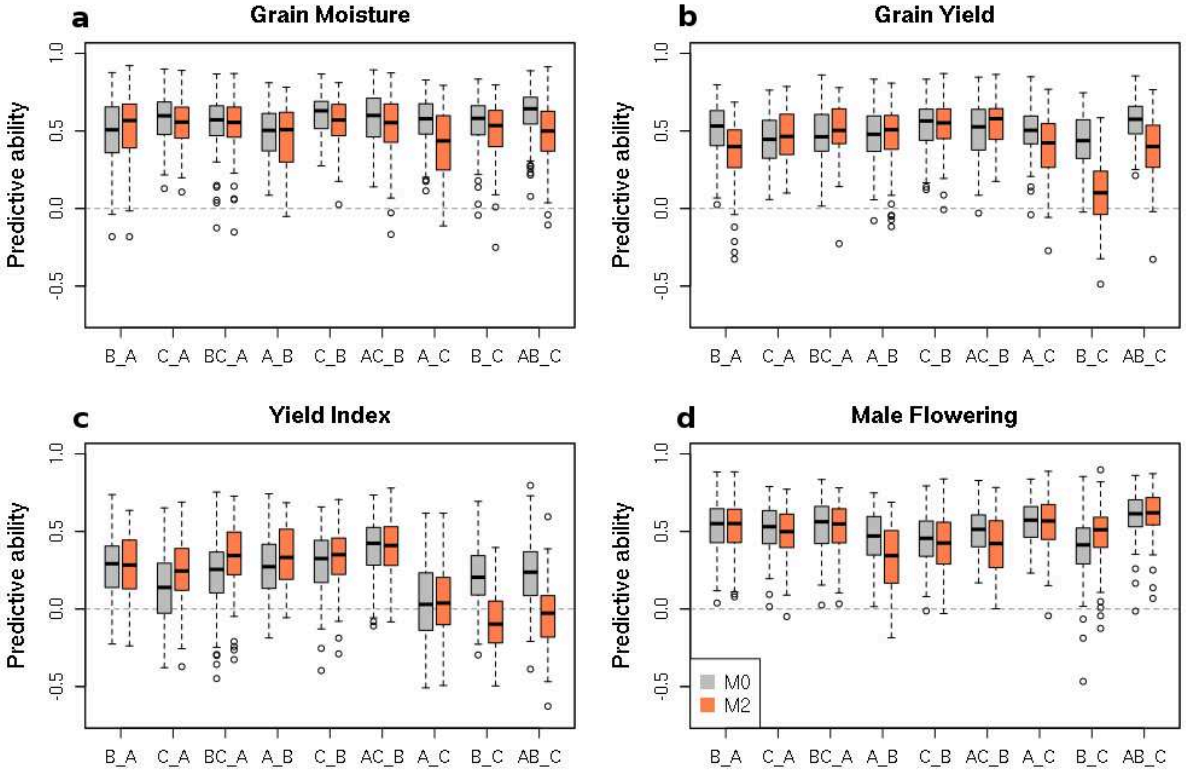
Supplementary Table S1.2: Average of predictive abilities obtained with the the structure-based cross-validations (SHO) using $\mathbf{M_0}$. Standard deviations are shown between brackets



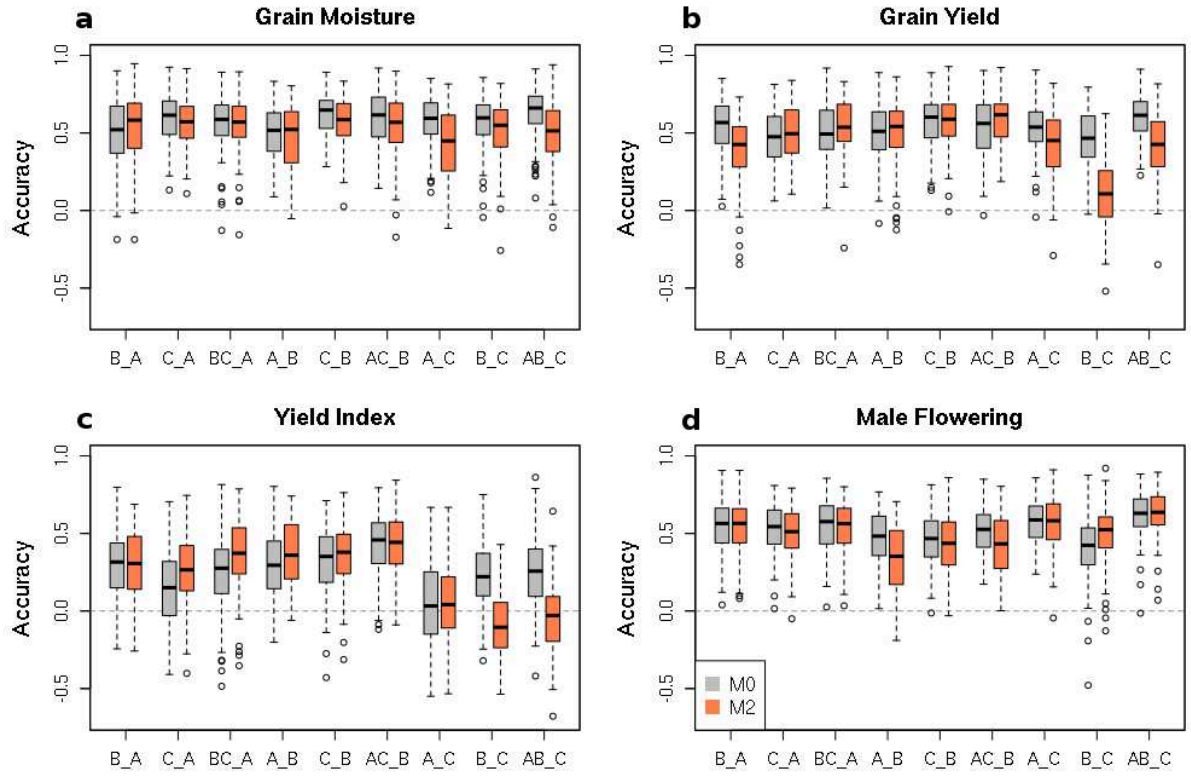
Supplementary Fig. S1.4: Box-plots of predictive abilities obtained with the structure-based cross-validations (SHO) for scenarios ABC_ABC and ABC_X using different models of prediction for **a**. grain Moisture, **b**. Grain Yield, **c**. Yield Index and **d**. Male Flowering



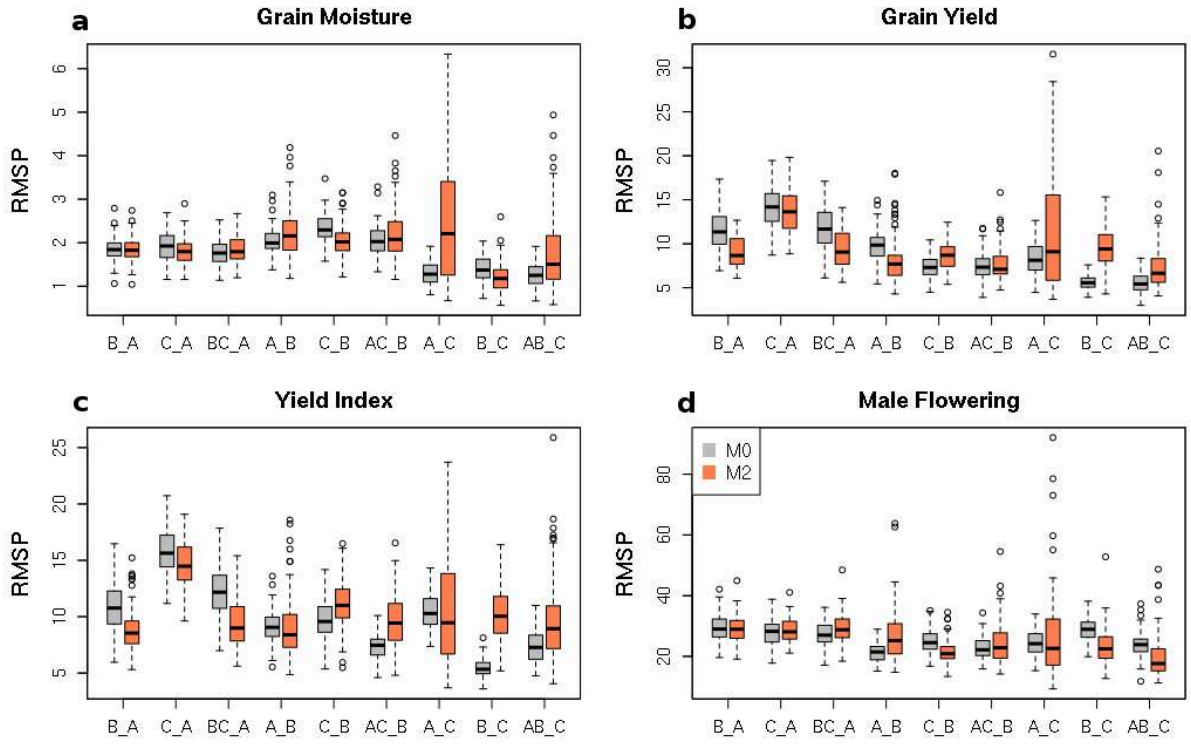
Supplementary Fig. S1.5: Box-plots of RMSP obtained with the structure-based cross-validations (SHO) for scenarios ABC_ABC and ABC_X using different models of prediction for **a**. grain Moisture, **b**. Grain Yield, **c**. Yield Index and **d**. Male Flowering



Supplementary Fig. S1.6: Box-plots of predictive abilities obtained with the structure-based cross-validations (SHO) for across-group scenarios using different models of prediction for **a**. grain Moisture, **b**. Grain Yield, **c**. Yield Index and **d**. Male Flowering



Supplementary Fig. S1.7: Box-plots of accuracies obtained with the structure-based cross-validations (SHO) for across-group scenarios using different models of prediction for **a**. grain Moisture, **b**. Grain Yield, **c**. Yield Index and **d**. Male Flowering



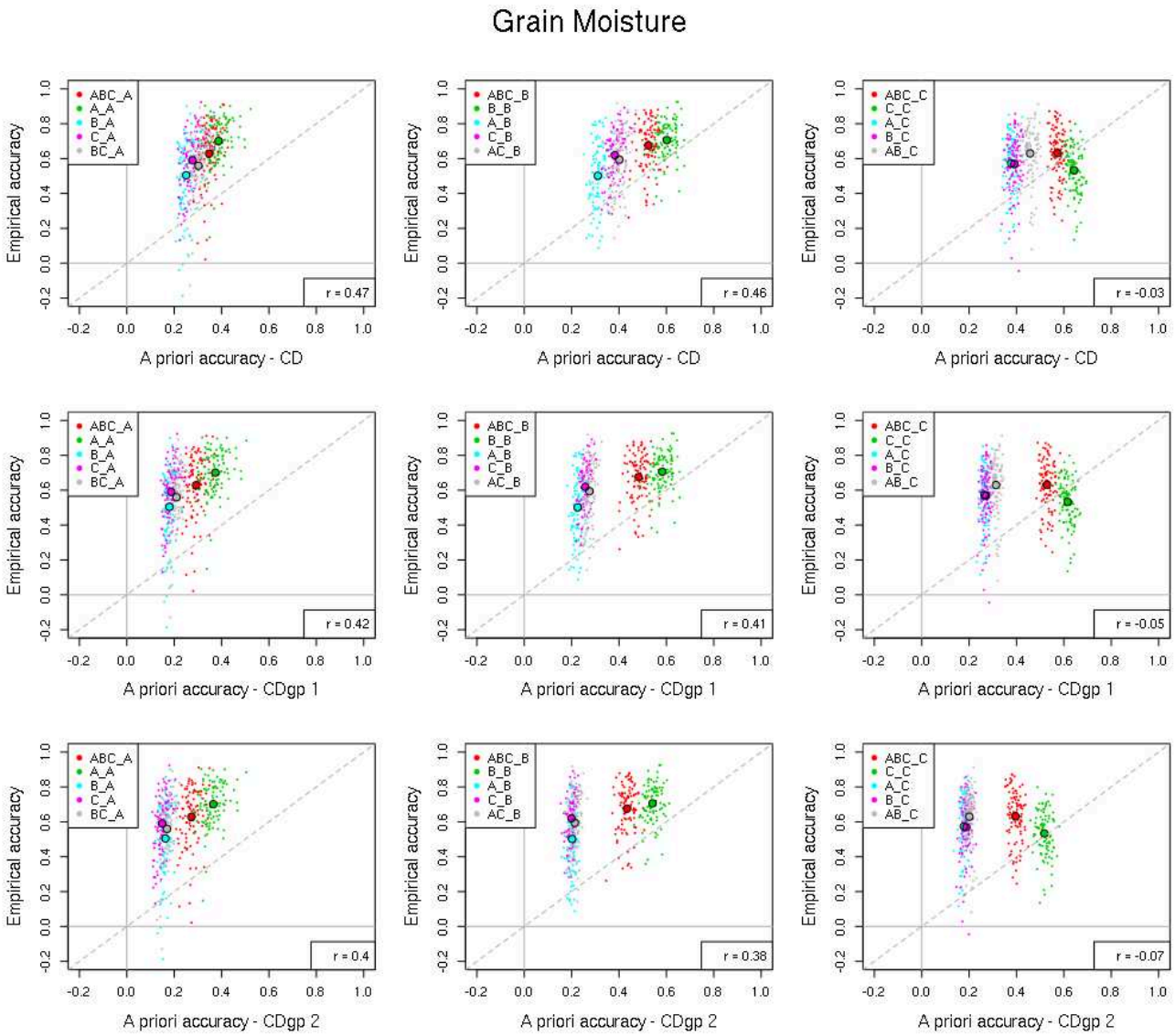
Supplementary Fig. S1.8: Box-plots of RMSP obtained with the structure-based cross-validations (SHO) for across-group scenarios using different models of prediction for **a**. grain Moisture, **b**. Grain Yield, **c**. Yield Index and **d**. Male Flowering

	Grain Moisture	**Grain Yield**	**Yield Index**	**Male Flowering**
$\sigma^2_{G_A}$	2.35 (0.28)	59.02 (6.89)	66.93 (7.78)	559.87 (61.62)
$\sigma^2_{G_B}$	1.94 (0.37)	28.60 (6.33)	34.43 (9.66)	335.19 (62.82)
$\sigma^2_{G_C}$	0.70 (0.14)	20.19 (4.80)	14.12 (4.88)	219.61 (34.94)
r_{AB}	0.72 (0.12)	0.96 (0.04)	0.95 (0.04)	0.99 (0.01)
r_{AC}	0.65 (0.11)	0.92 (0.08)	0.88 (0.14)	0.99 (0.01)
r_{BC}	0.62 (0.12)	0.92 (0.07)	0.88 (0.14)	0.99 (0.01)
$\sigma^2_{E_A}$	0.36 (0.11)	5.31 (2.35)	4.57 (2.13)	46.85 (17.68)
$\sigma^2_{E_B}$	0.51 (0.19)	7.12 (3.46)	14.17 (6.40)	32.37 (22.89)
$\sigma^2_{E_C}$	0.29 (0.09)	4.24 (2.67)	8.15 (3.73)	9.65 (6.22)

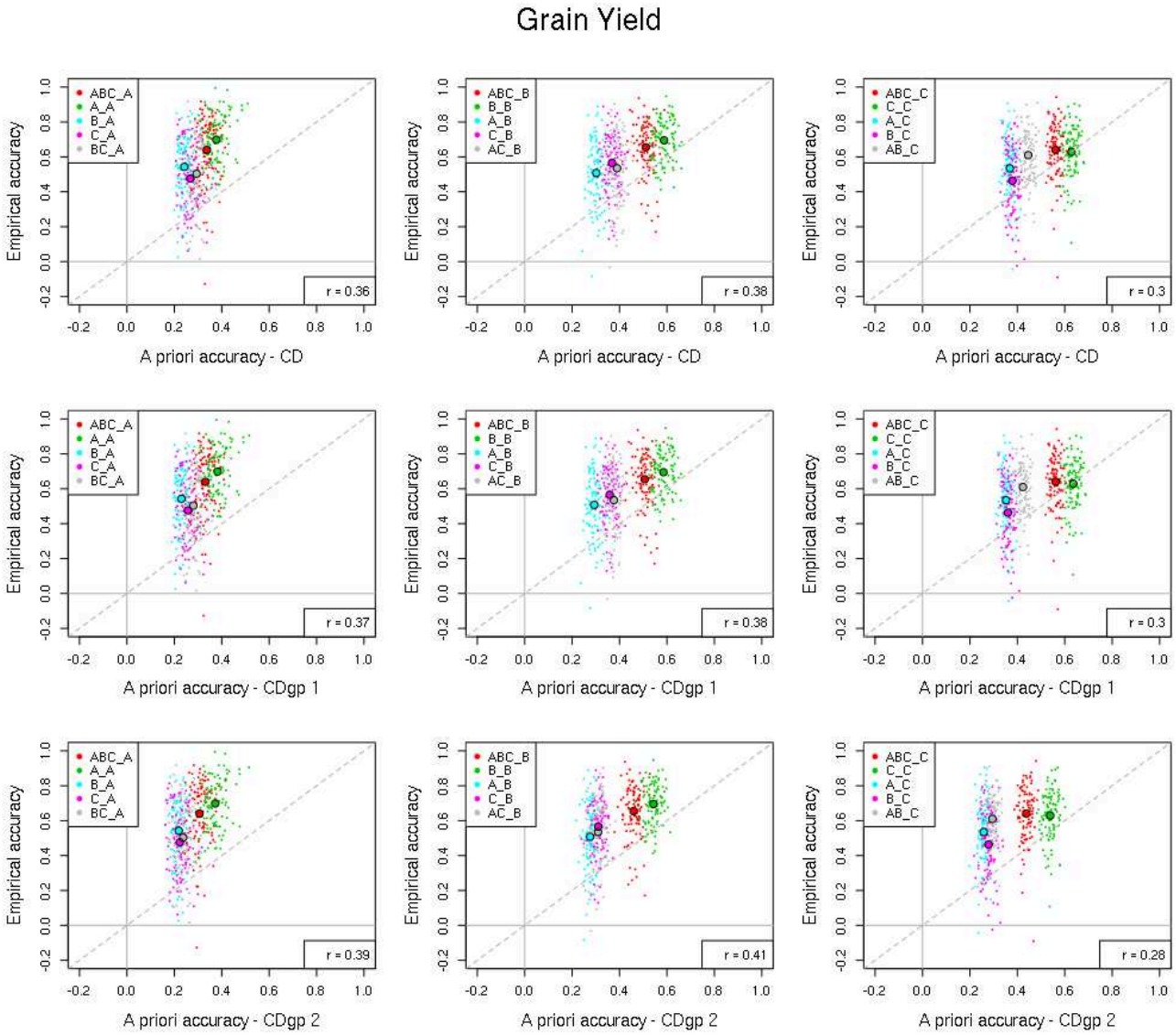
Supplementary Table S1.3: Posterior mean of group-specific genetic variances, genetic correlations and environmental variances estimated with $\mathbf{M_3}$ and $\boldsymbol{K^3}$ on all the data. Posterior standard deviations, obtained on Gibbs samples, are shown between brackets

	Grain Moisture			**Grain Yield**			**Yield Index**			**Male Flowering**		
VS	A	B	C	A	B	C	A	B	C	A	B	C
CD	0.33	0.23	0.21	0.32	0.23	0.20	0.27	0.21	0.29	0.31	0.18	0.20
$CDgp^1$	0.39	0.30	0.29	0.33	0.24	0.21	0.26	0.21	0.28	0.31	0.18	0.20
$CDgp^2$	0.41	0.35	0.36	0.35	0.27	0.28	0.27	0.21	0.24	0.33	0.20	0.26

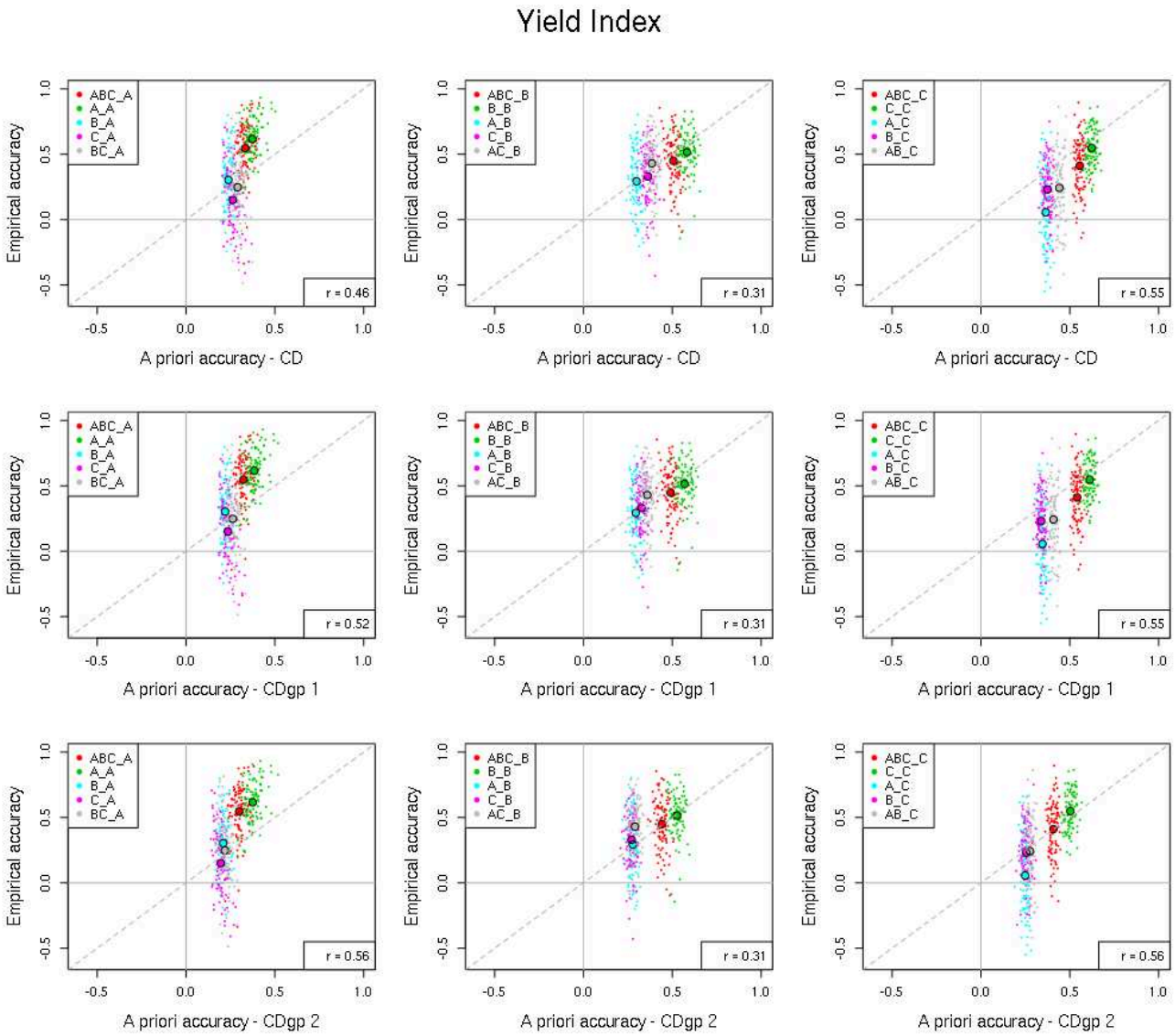
Supplementary Table S1.4: RMSE between a priori estimates of accuracy and empirical accuracies using $\mathbf{M_0}$ and structure-based cross-validations (SHO) for group-specific VS (e.g. A includes ABC_A, A_A, BC_A, B_A and C_A)



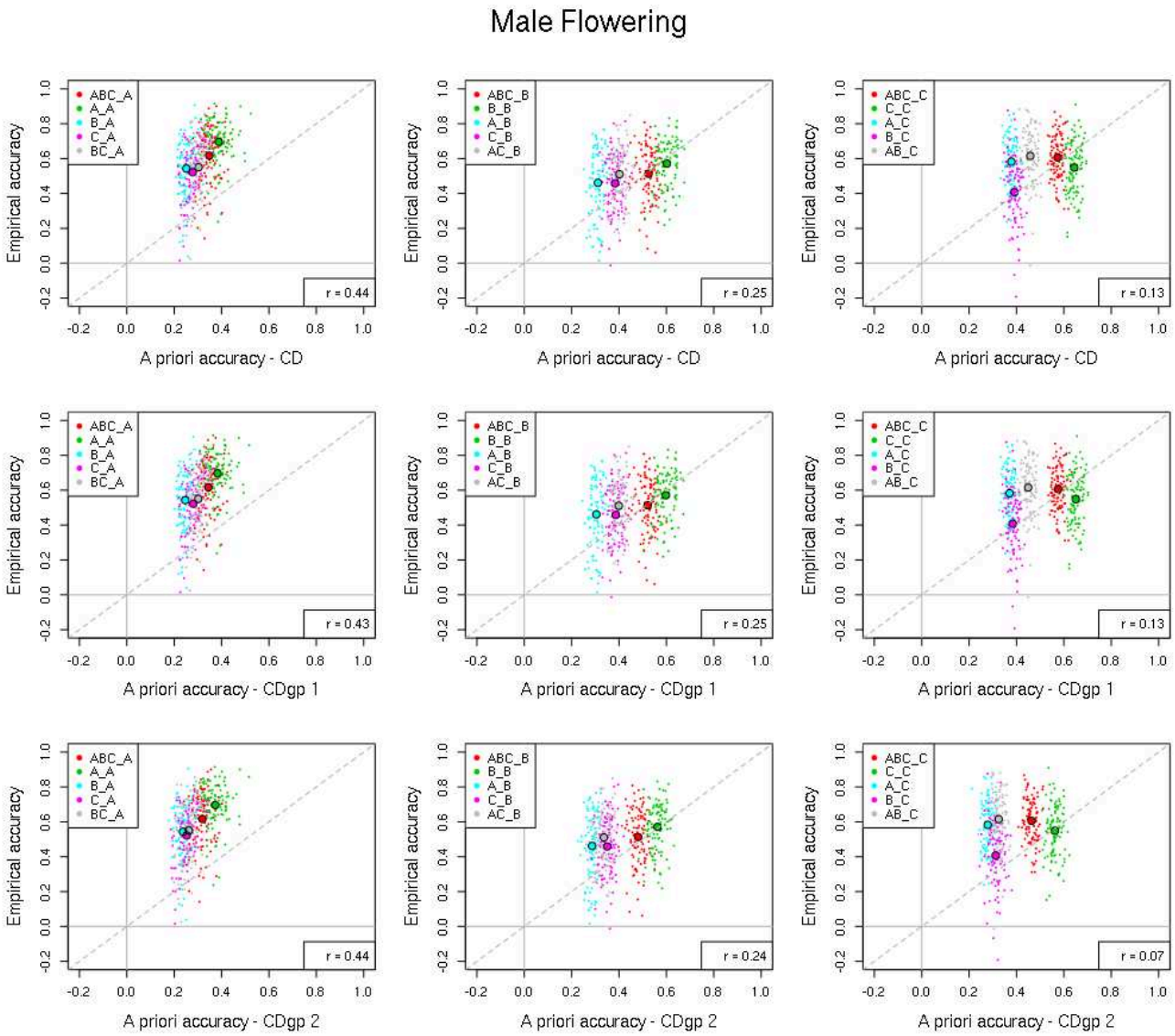
Supplementary Fig. S1.9: A priori estimates of accuracy, using three different CD indicators, plotted against empirical accuracies obtained on the SHO replicates using $\mathbf{M_0}$, for Grain Moisture



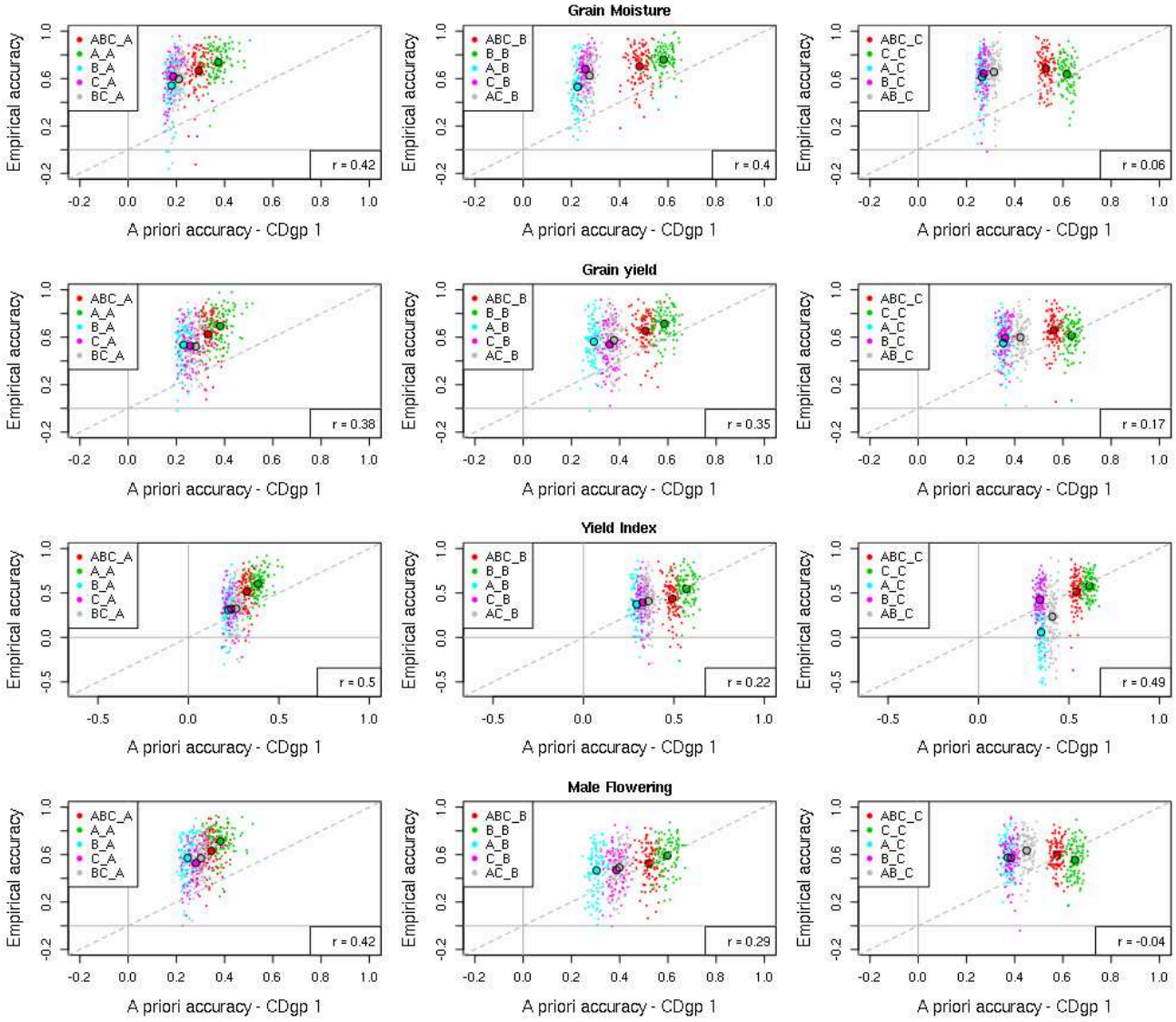
Supplementary Fig. S1.10: A priori estimates of accuracy, using three different CD indicators, plotted against empirical accuracies obtained on the SHO replicates using $\mathbf{M_0}$, for Grain Yield



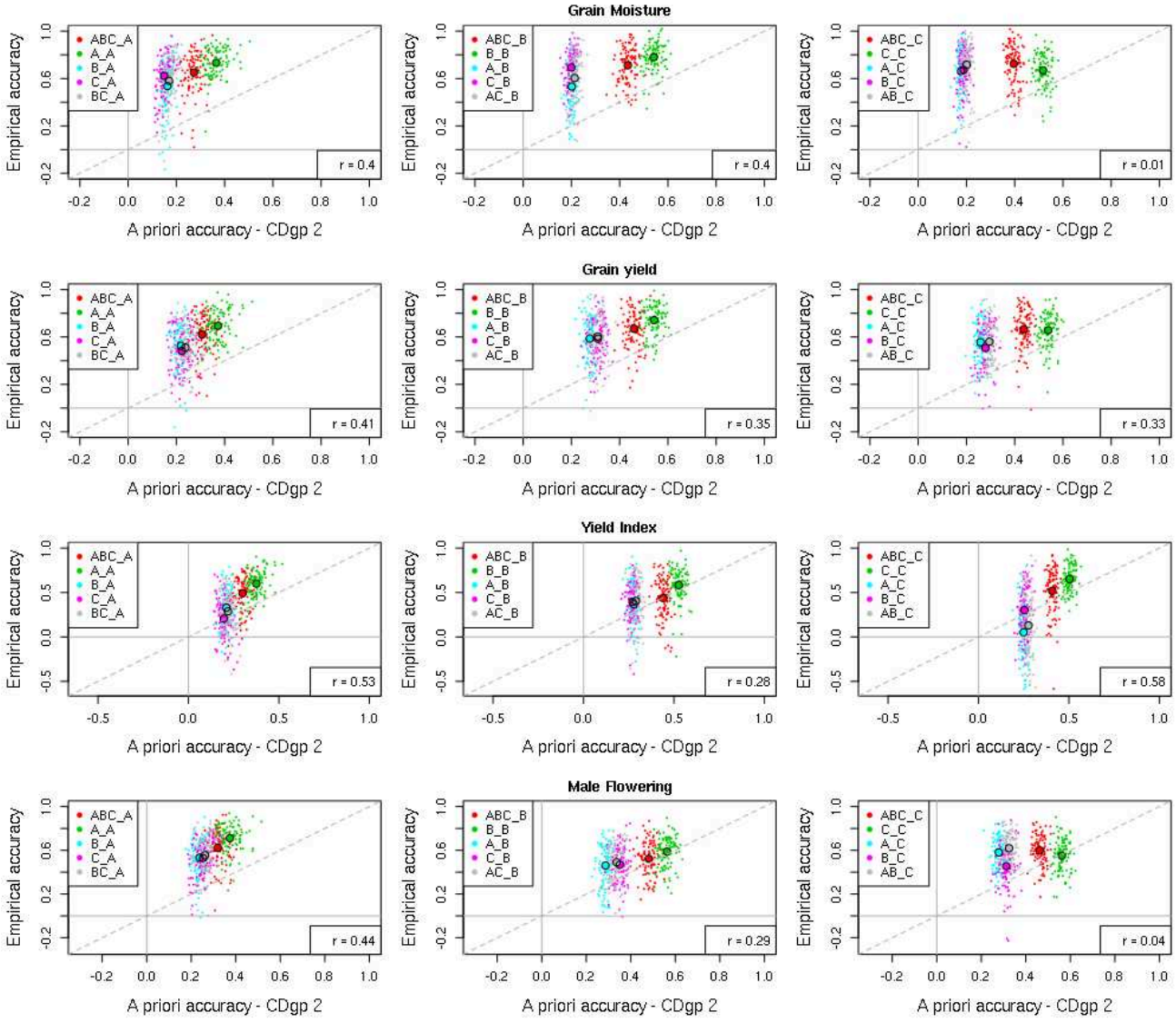
Supplementary Fig. S1.11: A priori estimates of accuracy, using three different CD indicators, plotted against empirical accuracies obtained on the SHO replicates using $\mathbf{M_0}$, for Yield Index



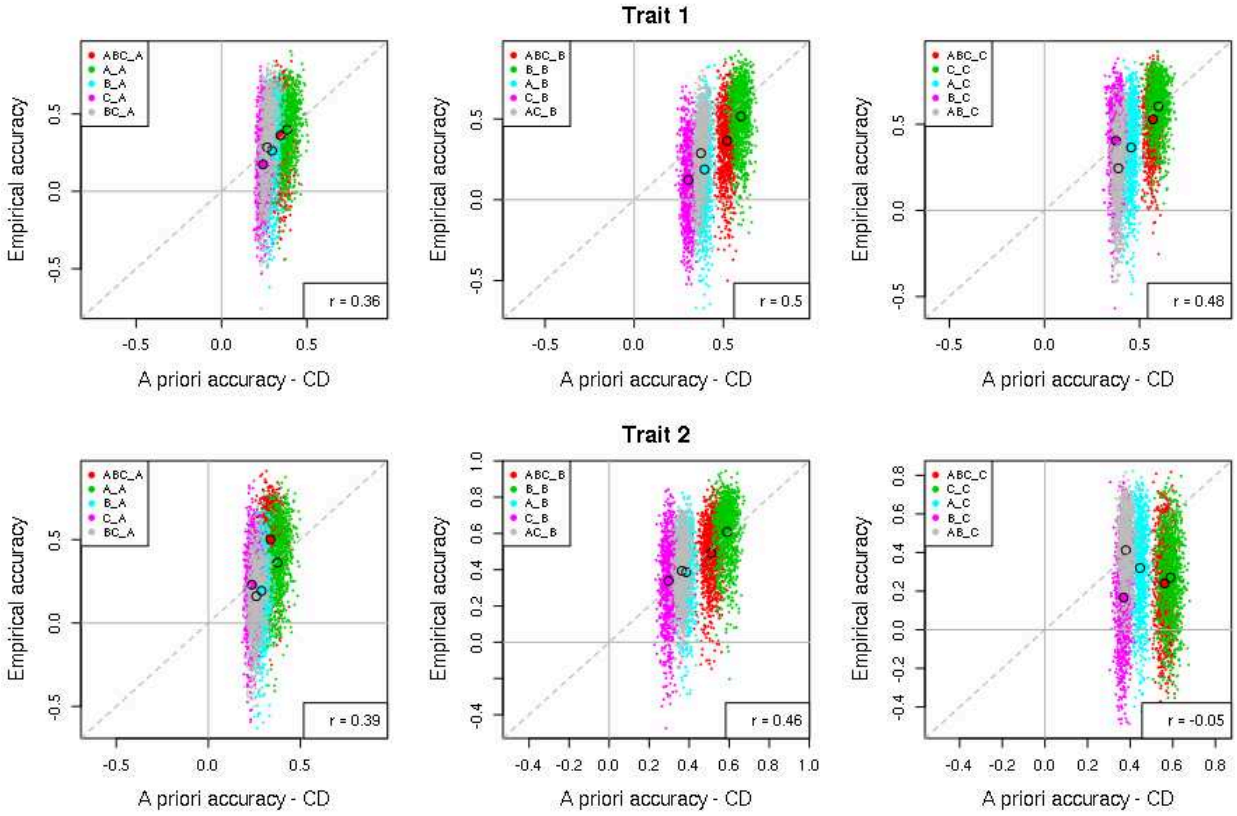
Supplementary Fig. S1.12: A priori estimates of accuracy, using three different CD indicators, plotted against empirical accuracies obtained on the SHO replicates using $\mathbf{M_0}$, for Male Flowering



Supplementary Fig. S1.13: A priori estimates of accuracy ($CDgp^1$), plotted against empirical accuracies obtained on the SHO replicates with predictions of breeding values computed using $\mathbf{M_3}$ and $\boldsymbol{K^0}$. To explore all scenarios, including across group predictions, variances parameters were estimated on the whole dataset (using $\mathbf{M_3}$, $\boldsymbol{K^0}$) presented in Table 8



Supplementary Fig. S1.14: A priori estimates of accuracy ($CDgp^2$), plotted against empirical accuracies obtained on the SHO replicates with predictions of breeding values computed using $\mathbf{M_3}$ and $\boldsymbol{K^3}$. To explore all scenarios, including across group predictions, variances parameters were estimated on the whole dataset (using $\mathbf{M_3}$, $\boldsymbol{K^3}$) presented in Supplementary Table S4



Supplementary Fig. S1.15: A priori estimates of accuracy, using standard CD indicator, plotted against empirical accuracies obtained with 1,000 SHO replicates using $\mathbf{M_0}$, for two simulated traits. A priori accuracy using CD is a good proxy of empirical accuracies for trait 1 but is less efficient for trait 2

Simulation Procedure:

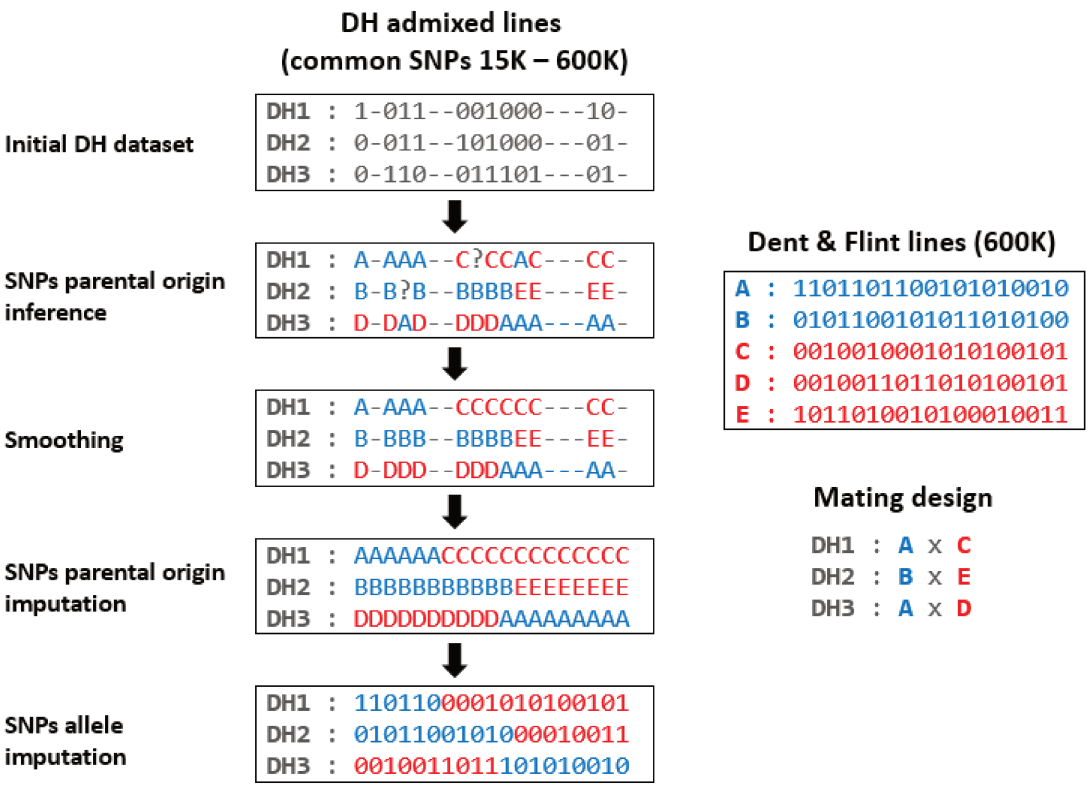
1,000 SNPs were selected among the 986,045 SNPs to be declared as QTLs.

Two vectors of allele effects $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_0}$ were sampled in a normal distribution of variance $\sigma^2_\beta = 1$.

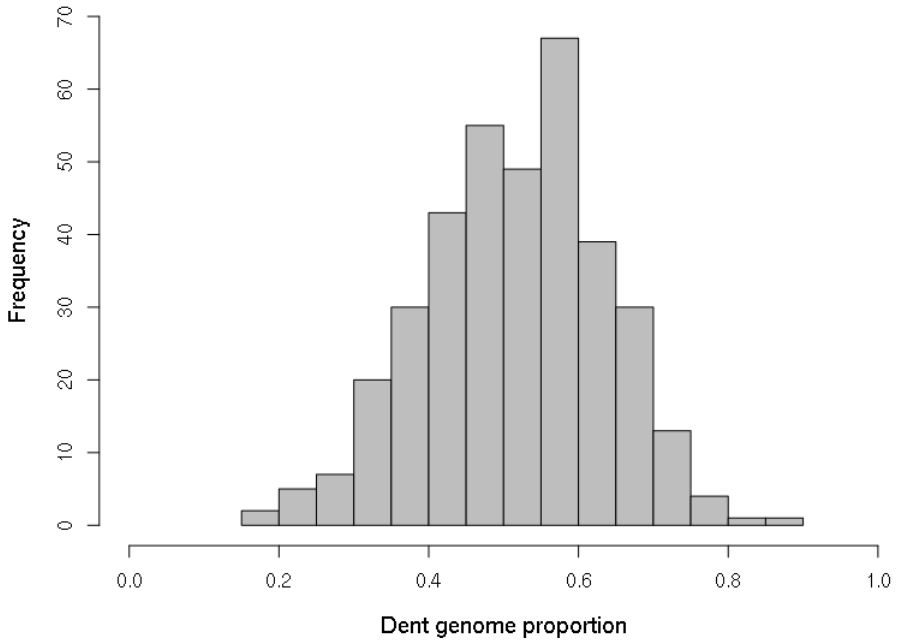
Breeding values were computed as $\boldsymbol{g} = (\boldsymbol{J} - \boldsymbol{W})\boldsymbol{\beta_0} + \boldsymbol{W}\boldsymbol{\beta_1}$ where $\boldsymbol{g}$ is the vector of breedingvalues, $\boldsymbol{J}$ is a matrix of 1 with N lines and 1,000 columns and $\boldsymbol{W}$ is the genotypic matrix at QTLs (coded 0, 0.5, 1).

A residual noise was sampled in a normal distribution of variance σ^2_E to obtain simulated phenotypes with a $h^2 = 0.9$ using: $\boldsymbol{y} = \boldsymbol{g} + \boldsymbol{e}$ where $\boldsymbol{y}$ is the vector of phenotypes and $\boldsymbol{e}$ is the vector of residuals.

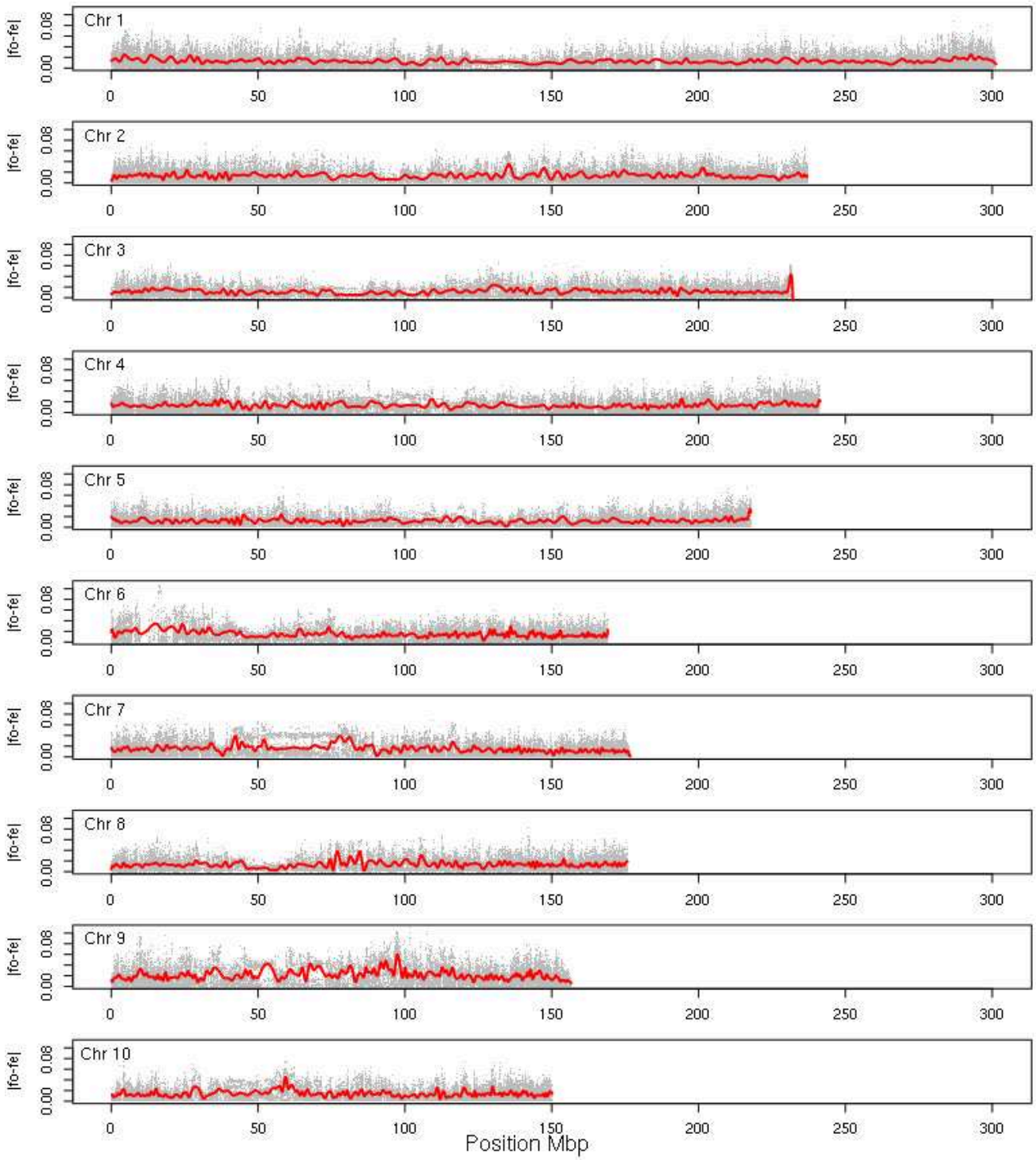
Supplementary Material - Chapitre 2



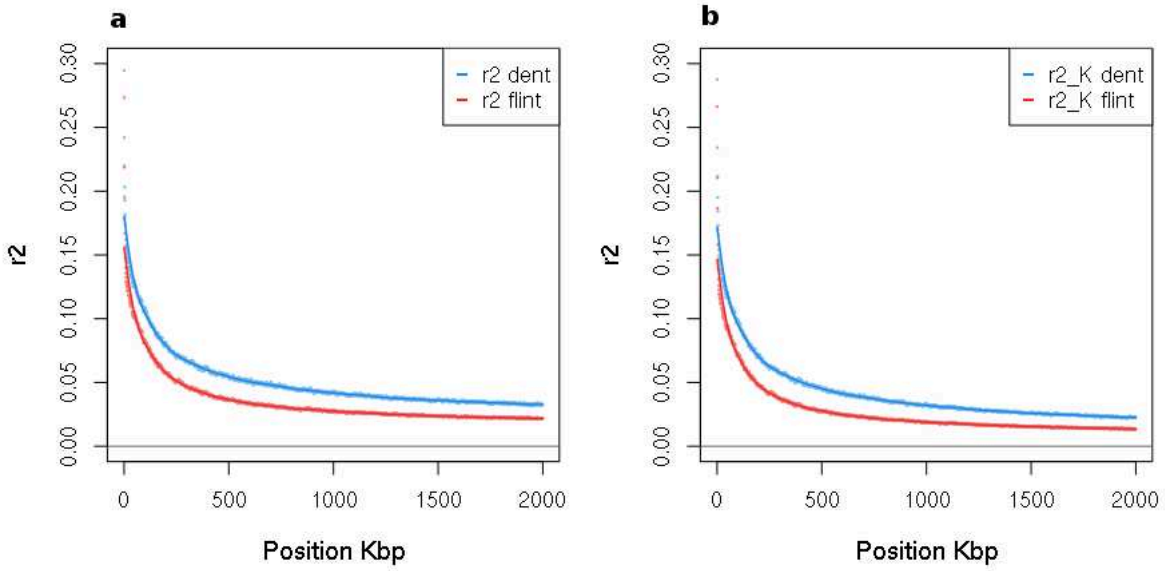
Supplementary Figure S2.1: Diagram illustrating the procedure applied to impute admixed DH lines to 600K genotyping based on the parental origins of alleles



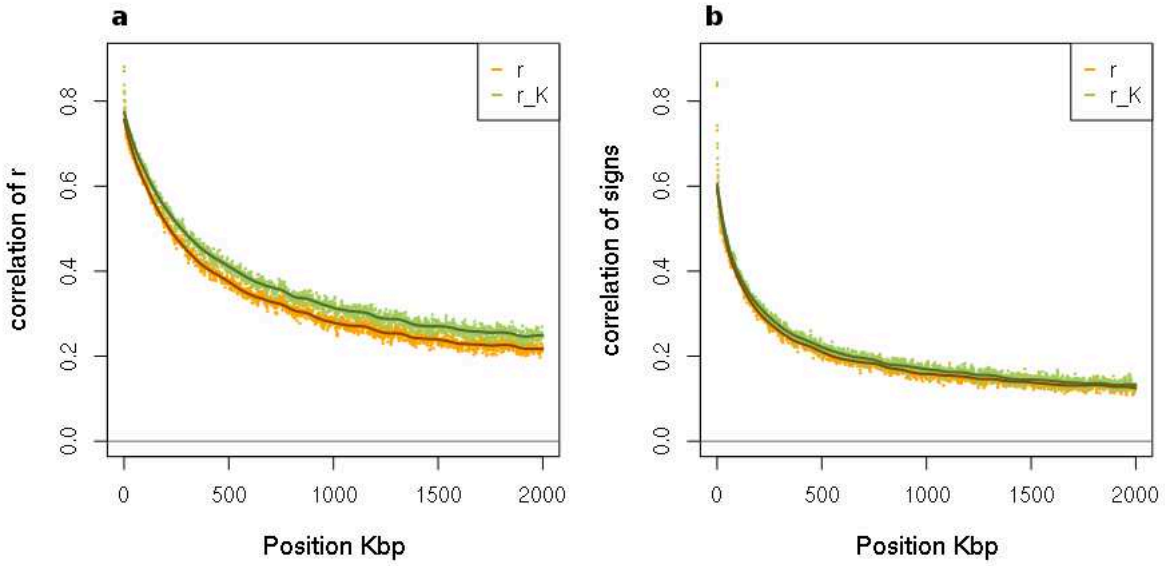
Supplementary Figure S2.2: Distribution of dent genome proportion among the admixed lines.



Supplementary Figure S2.3: Absolute difference between observed allele frequency of the reference allele f_o estimated on the admixed lines and their expected value f_e along each chromosome. The expected allele frequencies were computed as the mean of flint and dent allele frequencies estimated on the parental lines by taking into account the contribution of each parent. A cubic smoothing spline was adjusted using the R function "smooth.spline", and plotted in red



Supplementary Figure S2.4: LD extent estimated, with a sliding window of physical distances between pairs of loci, in dent and flint genetic group using the average of **a**. the standard r^2 or **b**. the r^2_K accounting for relatedness between individuals. A cubic smooth spline was adjusted for each group, using the R function "smooth.spline"



Supplementary Figure S2.5: Conservation of LD phases estimated, with a sliding window of physical distances between pairs of loci, using the correlation **a**. between the r of dent and flint groups (or the r_K accounting for relatedness between individuals), and **b**. between the signs of the r dent and flint groups (or the signs of the r_K). A cubic smooth spline was adjusted for method, using the R function "smooth.spline"

Supplementary Table S2.1: Phenotypic analysis

	MF	FF
Row-Column 2015	AR1	AR1
Row-Column 2016	IID	IID
μ_{2015}	64.22	65.86
μ_{2016}	71.20	72.40
μ_{D}	70.81	72.28
μ_{F}	64.95	66.39
μ_{A}	67.66	69.11
$\sigma^2_{G_{\mathrm{D}}}$	24.15	27.65
$\sigma^2_{G_{\mathrm{F}}}$	23.06	27.03
$\sigma^2_{G_{\mathrm{A}}}$	16.89	20.07
$\sigma^2_{(G\times\beta)_{2015,\mathrm{D}}}$	1.31	0.00
$\sigma^2_{(G\times\beta)_{2015,\mathrm{F}}}$	1.35	1.99
$\sigma^2_{(G\times\beta)_{2015,\mathrm{A}}}$	4.64	4.84
$\sigma^2_{(G\times\beta)_{2016,\mathrm{D}}}$	1.09	2.67
$\sigma^2_{(G\times\beta)_{2016,\mathrm{F}}}$	0.00	0.00
$\sigma^2_{(G\times\beta)_{2016,\mathrm{A}}}$	2.11	3.51
$\sigma^2_{E_{2015}}$	2.46	2.56
$\sigma^2_{E_{2016}}$	1.54	1.94
h^2_{D}	0.96	0.96
h^2_{F}	0.96	0.96
h^2_{A}	0.88	0.88
$\bar{r}_{2015}$	1.94	1.94
$\bar{r}_{2016}$	1.99	1.99

The lines "Row-Column" refers to the modeling of row and columns. AR1 refers to the modeling of row and column effects, as defined by the experimental design, following an autoregressive model AR1, while IID refers to the modeling of row and column as being independent and identically distributed among rows and among columns for a given trial. For more information, see the ASReml-R reference manual by Butler et al. (2009).

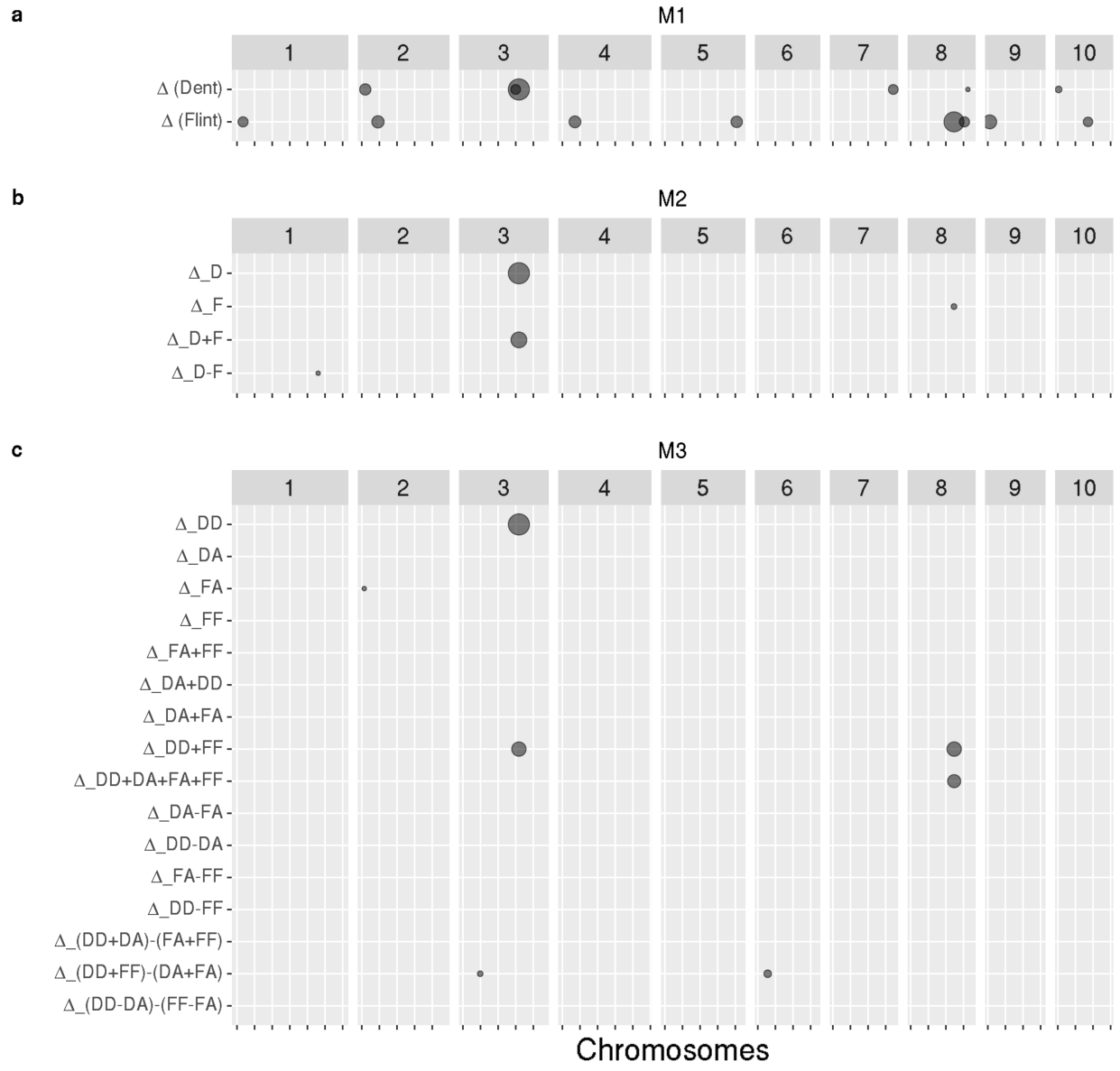
The mean of each trial was computed following: $\mu_j = \mu + \beta_j + \sum_{k=1}^{3} \frac{N_k}{N} \alpha_k$ where N_k is the number of individuals (genotypes) in genetic background k and N is the total number of individuals.

The mean of each genetic background was computed following: $\mu_k = \mu + \alpha_k + \frac{1}{2} \sum_{j=1}^{2} \beta_j$

The heritabilities of each genetic background k were computed as:

$$h^2_k = \frac{\sigma^2_{G_k}}{\sigma^2_{G_k} + \frac{1}{4}\sum_{j=1}^{2} \sigma^2_{(G\times\beta)_{jk}} + \frac{1}{4}\sum_{j=1}^{2} \frac{1}{\bar{r}_j}\sigma^2_{E_j}}$$

where $\bar{r}_i$ is the mean number of genotype replicates in trial j



Supplementary Figure S2.6: Position of QTLs detected for FF with a FDR of 20% using **a**. $\mathbf{M_1}$, **b**. $\mathbf{M_2}$ and **c**. $\mathbf{M_3}$. The size of the grey dots is proportional to the -$\log_{10}$(pval) of the test at the most significant SNP of the region

Supplementary Table S2.2: Information regarding significant SNPs for MF using all GWAS strategies: the name of the SNP, the chromosome on which it is located, its position in bp along the chromosome, the frequency of the allelic state observed in the dataset in which it was tested, the GWAS model applied, the hypothesis tested, the -$\log_{10}$(pval) of the test and the FDR for which it was declared significant

Trait	SNP	Chromosome	Position	0FF	1FF	0FA	1FA	0DA	1DA	0DD	1DD	Model	Test	-$\log_{10}$(pval)	FDR
MF	AX-91341754	3	6,649,723	-	-	-	-	-	-	201	99	M1	Δ^m (Dent)	5.41	20%
MF	AX-90645849	3	8,265,448	-	-	-	-	-	-	123	177	M1	Δ^m (Dent)	4.65	20%
MF	AX-90795999	3	8,266,018	-	-	-	-	-	-	139	161	M1	Δ^m (Dent)	5.20	20%
MF	AX-91557112	3	19,841,904	-	-	-	-	-	-	67	233	M1	Δ^m (Dent)	4.73	20%
MF	AX-90834909	3	158,880,237	-	-	-	-	-	-	107	193	M1	Δ^m (Dent)	4.61	20%
MF	AX-90566336	3	158,889,565	-	-	-	-	-	-	109	191	M1	Δ^m (Dent)	5.17	20%
MF	AX-91583291	3	158,891,367	-	-	-	-	-	-	108	192	M1	Δ^m (Dent)	4.96	20%
MF	AX-90834898	3	158,896,163	-	-	-	-	-	-	106	194	M1	Δ^m (Dent)	5.10	20%
MF	AX-91583310	3	158,974,594	-	-	-	-	-	-	97	203	M1	Δ^m (Dent)	10.99	5%
MF	AX-90590040	3	158,974,646	-	-	-	-	-	-	96	204	M1	Δ^m (Dent)	9.49	5%
MF	AX-90834934	3	158,974,756	-	-	-	-	-	-	102	198	M1	Δ^m (Dent)	9.21	5%
MF	AX-91408371	3	158,975,082	-	-	-	-	-	-	101	199	M1	Δ^m (Dent)	8.96	5%
MF	AX-91583371	3	159,390,464	-	-	-	-	-	-	120	180	M1	Δ^m (Dent)	5.07	20%
MF	AX-91583384	3	159,391,016	-	-	-	-	-	-	119	181	M1	Δ^m (Dent)	5.45	20%
MF	AX-91583382	3	159,392,021	-	-	-	-	-	-	121	179	M1	Δ^m (Dent)	5.17	20%
MF	AX-91583403	3	159,405,200	-	-	-	-	-	-	119	181	M1	Δ^m (Dent)	5.85	20%
MF	AX-90835029	3	159,413,980	-	-	-	-	-	-	119	181	M1	Δ^m (Dent)	5.33	20%
MF	AX-90835045	3	159,415,239	-	-	-	-	-	-	120	180	M1	Δ^m (Dent)	5.57	20%
MF	AX-91583404	3	159,487,904	-	-	-	-	-	-	119	181	M1	Δ^m (Dent)	4.84	20%
MF	AX-90835056	3	159,506,081	-	-	-	-	-	-	119	181	M1	Δ^m (Dent)	4.89	20%
MF	AX-90835061	3	159,514,077	-	-	-	-	-	-	118	182	M1	Δ^m (Dent)	5.26	20%
MF	AX-90846668	3	201,723,987	-	-	-	-	-	-	220	80	M1	Δ^m (Dent)	4.61	20%
MF	AX-90859038	4	12,912,108	-	-	-	-	-	-	12	288	M1	Δ^m (Dent)	4.96	20%
MF	AX-90919240	4	237,451,186	-	-	-	-	-	-	256	44	M1	Δ^m (Dent)	4.84	20%
MF	AX-90552295	7	130,495,196	-	-	-	-	-	-	279	21	M1	Δ^m (Dent)	4.77	20%
MF	AX-91736802	7	130,499,284	-	-	-	-	-	-	288	12	M1	Δ^m (Dent)	4.94	20%
MF	AX-91055771	7	130,500,226	-	-	-	-	-	-	289	11	M1	Δ^m (Dent)	5.38	20%
MF	AX-91737837	7	136,148,844	-	-	-	-	-	-	218	82	M1	Δ^m (Dent)	4.78	20%
MF	AX-91058427	7	140,576,170	-	-	-	-	-	-	271	29	M1	Δ^m (Dent)	4.71	20%
MF	AX-90633821	7	140,674,983	-	-	-	-	-	-	270	30	M1	Δ^m (Dent)	4.99	20%
MF	AX-91404598	7	140,675,794	-	-	-	-	-	-	238	62	M1	Δ^m (Dent)	4.86	20%
MF	AX-91744205	7	171,556,516	-	-	-	-	-	-	216	84	M1	Δ^m (Dent)	5.45	20%
MF	AX-91382048	7	171,572,202	-	-	-	-	-	-	214	86	M1	Δ^m (Dent)	4.69	20%
MF	AX-91100608	8	123,504,353	-	-	-	-	-	-	232	68	M1	Δ^m (Dent)	4.64	20%
MF	AX-91112607	8	165,932,930	-	-	-	-	-	-	161	139	M1	Δ^m (Dent)	4.63	20%
MF	AX-91456671	1	17,179,645	277	27	-	-	-	-	-	-	M1	Δ^m (Flint)	6.10	20%
MF	AX-90588172	1	231,185,911	278	26	-	-	-	-	-	-	M1	Δ^m (Flint)	5.41	20%
MF	AX-90592743	1	278,511,074	226	78	-	-	-	-	-	-	M1	Δ^m (Flint)	5.02	20%
MF	AX-90567285	2	59,761,720	283	21	-	-	-	-	-	-	M1	Δ^m (Flint)	4.99	20%
MF	AX-90591872	4	35,758,347	293	11	-	-	-	-	-	-	M1	Δ^m (Flint)	5.74	20%
MF	AX-90544623	5	12,227,383	287	17	-	-	-	-	-	-	M1	Δ^m (Flint)	5.55	20%
MF	AX-90974105	5	203,928,025	63	241	-	-	-	-	-	-	M1	Δ^m (Flint)	5.76	20%
MF	AX-91685722	6	24,606,589	272	32	-	-	-	-	-	-	M1	Δ^m (Flint)	5.03	20%
MF	AX-91355042	6	148,530,168	43	261	-	-	-	-	-	-	M1	Δ^m (Flint)	5.06	20%
MF	AX-91717426	7	25,909,531	269	35	-	-	-	-	-	-	M1	Δ^m (Flint)	4.89	20%

Trait	SNP	Chromosome	Position	0FF	1FF	0FA	1FA	0DA	1DA	0DD	1DD	Model	Test	-log$_{10}$(pval))	FDR
MF	AX-91100149	8	121,884,992	282	22	-	-	-	-	-	-	M1	Δ^m (Flint)	4.84	20%
MF	AX-91100415	8	122,949,646	284	20	-	-	-	-	-	-	M1	Δ^m (Flint)	10.19	5%
MF	AX-91359941	8	122,950,264	286	18	-	-	-	-	-	-	M1	Δ^m (Flint)	5.36	20%
MF	AX-91100407	8	122,950,963	279	25	-	-	-	-	-	-	M1	Δ^m (Flint)	7.89	5%
MF	AX-91428720	8	122,952,245	291	13	-	-	-	-	-	-	M1	Δ^m (Flint)	6.76	5%
MF	AX-91768145	8	122,952,961	289	15	-	-	-	-	-	-	M1	Δ^m (Flint)	5.32	20%
MF	AX-90596641	8	122,952,991	284	20	-	-	-	-	-	-	M1	Δ^m (Flint)	5.52	20%
MF	AX-91768204	8	123,372,565	293	11	-	-	-	-	-	-	M1	Δ^m (Flint)	5.99	20%
MF	AX-91100612	8	123,509,765	23	281	-	-	-	-	-	-	M1	Δ^m (Flint)	5.22	20%
MF	AX-90557598	9	103,752,251	11	293	-	-	-	-	-	-	M1	Δ^m (Flint)	5.95	20%
MF	AX-91364734	10	85,834,561	279	25	-	-	-	-	-	-	M1	Δ^m (Flint)	5.65	20%
MF	AX-91341754	3	6,649,723	232	72	-	-	-	-	201	99	M2	Δ^m_D	5.83	20%
MF	AX-90645849	3	8,265,448	164	140	-	-	-	-	123	177	M2	Δ^m_D	5.99	20%
MF	AX-90795999	3	8,266,018	170	134	-	-	-	-	139	161	M2	Δ^m_D	6.53	20%
MF	AX-90796019	3	8,268,719	218	86	-	-	-	-	187	113	M2	Δ^m_D	4.95	20%
MF	AX-90566336	3	158,889,565	238	66	-	-	-	-	109	191	M2	Δ^m_D	5.07	20%
MF	AX-91439319	3	158,898,522	234	70	-	-	-	-	106	194	M2	Δ^m_D	4.92	20%
MF	AX-90834949	3	158,900,592	237	67	-	-	-	-	106	194	M2	Δ^m_D	4.93	20%
MF	AX-91583310	3	158,974,594	158	146	-	-	-	-	97	203	M2	Δ^m_D	9.65	5%
MF	AX-90590040	3	158,974,646	150	154	-	-	-	-	96	204	M2	Δ^m_D	8.21	5%
MF	AX-90834934	3	158,974,756	246	58	-	-	-	-	102	198	M2	Δ^m_D	8.14	5%
MF	AX-91408371	3	158,975,082	246	58	-	-	-	-	101	199	M2	Δ^m_D	7.97	5%
MF	AX-91583371	3	159,390,464	261	43	-	-	-	-	120	180	M2	Δ^m_D	4.97	20%
MF	AX-91583384	3	159,391,016	261	43	-	-	-	-	119	181	M2	Δ^m_D	5.32	20%
MF	AX-91583382	3	159,392,021	261	43	-	-	-	-	121	179	M2	Δ^m_D	5.09	20%
MF	AX-90835020	3	159,398,400	262	42	-	-	-	-	119	181	M2	Δ^m_D	5.26	20%
MF	AX-91583403	3	159,405,200	262	42	-	-	-	-	119	181	M2	Δ^m_D	5.57	20%
MF	AX-90835029	3	159,413,980	262	42	-	-	-	-	119	181	M2	Δ^m_D	5.14	20%
MF	AX-90835045	3	159,415,239	262	42	-	-	-	-	120	180	M2	Δ^m_D	5.30	20%
MF	AX-91583388	3	159,448,068	260	44	-	-	-	-	119	181	M2	Δ^m_D	5.37	20%
MF	AX-90835061	3	159,514,077	261	43	-	-	-	-	118	182	M2	Δ^m_D	5.18	20%
MF	AX-91837147	3	159,556,555	259	45	-	-	-	-	119	181	M2	Δ^m_D	5.27	20%
MF	AX-90859038	4	12,912,108	73	231	-	-	-	-	12	288	M2	Δ^m_D	5.12	20%
MF	AX-91641186	4	237,449,027	232	72	-	-	-	-	258	42	M2	Δ^m_D	5.66	20%
MF	AX-90919240	4	237,451,186	233	71	-	-	-	-	256	44	M2	Δ^m_D	6.15	20%
MF	AX-91398141	4	237,452,916	233	71	-	-	-	-	258	42	M2	Δ^m_D	5.67	20%
MF	AX-90572541	4	237,453,761	232	72	-	-	-	-	252	48	M2	Δ^m_D	5.08	20%
MF	AX-90919578	4	238,793,343	31	273	-	-	-	-	46	254	M2	Δ^m_D	5.28	20%
MF	AX-90552295	7	130,495,196	235	69	-	-	-	-	279	21	M2	Δ^m_D	5.44	20%
MF	AX-91736802	7	130,499,284	215	89	-	-	-	-	288	12	M2	Δ^m_D	5.83	20%
MF	AX-91736834	7	130,500,002	216	88	-	-	-	-	288	12	M2	Δ^m_D	5.81	20%
MF	AX-91055771	7	130,500,226	213	91	-	-	-	-	289	11	M2	Δ^m_D	6.39	20%
MF	AX-91055769	7	130,500,394	215	89	-	-	-	-	288	12	M2	Δ^m_D	5.82	20%
MF	AX-91736821	7	130,500,692	216	88	-	-	-	-	289	11	M2	Δ^m_D	5.10	20%
MF	AX-91737837	7	136,148,844	222	82	-	-	-	-	218	82	M2	Δ^m_D	5.19	20%

Trait	SNP	Chromosome	Position	0FF	1FF	0FA	1FA	0DA	1DA	0DD	1DD	Model	Test	-log$_{10}$(pval)	FDR
MF	AX-90633821	7	140,674,983	273	31	-	-	-	-	270	30	M2	Δ^m_D	4.96	20%
MF	AX-91404598	7	140,675,794	254	50	-	-	-	-	238	62	M2	Δ^m_D	5.71	20%
MF	AX-91744205	7	171,556,516	216	84	-	-	-	-	251	53	M2	Δ^m_D	4.93	20%
MF	AX-90974105	5	203,928,025	63	241	-	-	-	-	108	192	M2	Δ^m_F	5.23	20%
MF	AX-91100415	8	122,949,646	284	20	-	-	-	-	274	26	M2	Δ^m_F	5.85	20%
MF	AX-91100407	8	122,950,963	279	25	-	-	-	-	266	34	M2	Δ^m_F	5.75	20%
MF	AX-90605790	1	278,649,743	189	115	-	-	-	-	250	50	M2	Δ^m_{D+F}	5.27	20%
MF	AX-90633029	2	11,366,954	290	10	-	-	-	-	239	65	M2	Δ^m_{D+F}	5.02	20%
MF	AX-91583310	3	158,974,594	158	146	-	-	-	-	97	203	M2	Δ^m_{D+F}	7.47	5%
MF	AX-90590040	3	158,974,646	150	154	-	-	-	-	96	204	M2	Δ^m_{D+F}	6.10	20%
MF	AX-91641186	4	237,449,027	232	72	-	-	-	-	258	42	M2	Δ^m_{D+F}	5.95	20%
MF	AX-90919240	4	237,451,186	233	71	-	-	-	-	256	44	M2	Δ^m_{D+F}	6.32	20%
MF	AX-91398141	4	237,452,916	233	71	-	-	-	-	258	42	M2	Δ^m_{D+F}	5.96	20%
MF	AX-90572541	4	237,453,761	232	72	-	-	-	-	252	48	M2	Δ^m_{D+F}	5.47	20%
MF	AX-91059083	7	142,908,218	215	89	-	-	-	-	129	171	M2	Δ^m_{D+F}	5.00	20%
MF	AX-91100620	8	123,504,889	230	70	-	-	-	-	162	142	M2	Δ^m_{D+F}	5.25	20%
MF	AX-90555247	8	133,562,520	180	124	-	-	-	-	205	95	M2	Δ^m_{D+F}	5.12	20%
MF	AX-90710320	1	230,470,022	169	135	-	-	-	-	222	78	M2	Δ^m_{D-F}	5.30	20%
MF	AX-90843571	3	191,134,946	277	27	-	-	-	-	219	81	M2	Δ^m_{D-F}	4.93	20%
MF	AX-91345590	4	10,216,379	198	106	-	-	-	-	174	126	M2	Δ^m_{D-F}	4.93	20%
MF	AX-90859038	4	12,912,108	73	231	-	-	-	-	12	288	M2	Δ^m_{D-F}	5.01	20%
MF	AX-90863948	4	31,102,452	160	144	-	-	-	-	115	185	M2	Δ^m_{D-F}	5.47	20%
MF	AX-91218190	4	31,102,486	161	143	-	-	-	-	115	185	M2	Δ^m_{D-F}	5.76	20%
MF	AX-90863950	4	31,102,555	161	143	-	-	-	-	115	185	M2	Δ^m_{D-F}	5.14	20%
MF	AX-91602775	4	31,102,595	161	143	-	-	-	-	116	184	M2	Δ^m_{D-F}	5.18	20%
MF	AX-90863961	4	31,102,637	160	144	-	-	-	-	116	184	M2	Δ^m_{D-F}	5.51	20%
MF	AX-90863956	4	31,106,881	169	135	-	-	-	-	146	154	M2	Δ^m_{D-F}	5.35	20%
MF	AX-91624224	4	145,102,674	291	13	-	-	-	-	247	53	M2	Δ^m_{D-F}	4.99	20%
MF	AX-90645628	4	165,993,620	166	138	-	-	-	-	146	154	M2	Δ^m_{D-F}	4.92	20%
MF	AX-90967192	5	178,524,725	282	22	-	-	-	-	277	23	M2	Δ^m_{D-F}	5.11	20%
MF	AX-91736802	7	130,499,284	215	89	-	-	-	-	288	12	M2	Δ^m_{D-F}	5.48	20%
MF	AX-91736834	7	130,500,002	216	88	-	-	-	-	288	12	M2	Δ^m_{D-F}	5.74	20%
MF	AX-91055771	7	130,500,226	213	91	-	-	-	-	289	11	M2	Δ^m_{D-F}	6.34	20%
MF	AX-91055769	7	130,500,394	215	89	-	-	-	-	288	12	M2	Δ^m_{D-F}	5.57	20%
MF	AX-91736821	7	130,500,692	216	88	-	-	-	-	289	11	M2	Δ^m_{D-F}	5.12	20%
MF	AX-91454277	1	4,450,457	185	119	99	72	73	122	132	168	M3	Δ_{DD}	5.22	20%
MF	AX-91341754	3	6,649,723	232	72	129	47	139	51	201	99	M3	Δ_{DD}	6.42	20%
MF	AX-90645849	3	8,265,448	164	140	80	91	92	103	123	177	M3	Δ_{DD}	6.05	20%
MF	AX-90795999	3	8,266,018	170	134	82	89	102	93	139	161	M3	Δ_{DD}	6.04	20%
MF	AX-90566336	3	158,889,565	238	66	135	42	56	133	109	191	M3	Δ_{DD}	5.30	20%
MF	AX-91583291	3	158,891,367	232	72	131	46	56	133	108	192	M3	Δ_{DD}	5.25	20%
MF	AX-90834897	3	158,895,171	231	73	130	47	56	133	108	192	M3	Δ_{DD}	5.24	20%
MF	AX-90834898	3	158,896,163	231	73	129	48	55	134	106	194	M3	Δ_{DD}	5.13	20%

Trait	SNP	Chromosome	Position	0FF	1FF	0FA	1FA	0DA	1DA	0DD	1DD	Model	Test	-log$_{10}$(pval)	FDR
MF	AX-91439319	3	158,898,522	234	70	133	44	54	135	106	194	M3	Δ_{DD}	5.32	20%
MF	AX-90834949	3	158,900,592	237	67	133	44	54	135	106	194	M3	Δ_{DD}	5.33	20%
MF	AX-91583310	3	158,974,594	158	146	92	85	48	141	97	203	M3	Δ_{DD}	10.53	5%
MF	AX-90590040	3	158,974,646	150	154	87	90	48	141	96	204	M3	Δ_{DD}	8.97	5%
MF	AX-90834934	3	158,974,756	246	58	139	38	50	139	102	198	M3	Δ_{DD}	8.78	5%
MF	AX-91408371	3	158,975,082	246	58	139	38	50	139	101	199	M3	Δ_{DD}	8.69	5%
MF	AX-90838756	3	173,800,721	173	131	110	74	63	119	82	218	M3	Δ_{DD}	5.68	20%
MF	AX-90887576	4	121,595,414	216	88	139	47	143	37	234	66	M3	Δ_{DD}	5.27	20%
MF	AX-91620133	4	123,620,457	153	151	91	95	85	95	141	159	M3	Δ_{DD}	5.14	20%
MF	AX-91624125	4	144,753,436	215	89	125	62	140	39	226	74	M3	Δ_{DD}	5.87	20%
MF	AX-91624156	4	144,823,042	231	73	129	58	142	37	237	63	M3	Δ_{DD}	5.95	20%
MF	AX-90893690	4	144,826,673	228	76	129	58	142	37	236	64	M3	Δ_{DD}	6.17	20%
MF	AX-91624155	4	144,833,742	231	73	129	58	142	37	236	64	M3	Δ_{DD}	5.13	20%
MF	AX-91851677	4	165,971,204	169	135	95	85	163	23	248	52	M3	Δ_{DD}	5.71	20%
MF	AX-90645628	4	165,993,620	166	138	102	78	85	101	146	154	M3	Δ_{DD}	5.35	20%
MF	AX-90904393	4	185,184,378	192	112	97	98	152	19	261	39	M3	Δ_{DD}	5.28	20%
MF	AX-90578187	8	14,591,468	249	55	165	32	127	42	226	74	M3	Δ_{DD}	5.81	20%
MF	AX-91100502	8	123,272,484	168	136	84	105	119	58	233	67	M3	Δ_{DD}	5.21	20%
MF	AX-91100608	8	123,504,353	155	149	78	111	119	58	232	68	M3	Δ_{DD}	5.40	20%
MF	AX-91100620	8	123,504,889	162	142	81	108	119	58	230	70	M3	Δ_{DD}	5.21	20%
MF	AX-90620050	8	161,789,135	66	238	28	149	45	144	93	207	M3	Δ_{DD}	5.74	20%
MF	AX-91424173	8	161,789,150	71	233	32	145	52	137	100	200	M3	Δ_{DD}	5.86	20%
MF	AX-91451275	8	162,519,150	216	88	127	49	82	108	122	178	M3	Δ_{DD}	5.21	20%
MF	AX-91112607	8	165,932,930	180	124	89	87	119	71	161	139	M3	Δ_{DD}	5.64	20%
MF	AX-91020683	6	164,143,140	190	114	103	74	176	13	280	20	M3	Δ_{DA}	6.55	20%
MF	AX-90731948	2	6,860,236	242	62	136	46	120	64	217	83	M3	Δ_{FA}	5.64	20%
MF	AX-91510472	2	6,861,005	240	64	134	48	120	64	216	84	M3	Δ_{FA}	6.16	20%
MF	AX-90731977	2	6,992,374	101	203	70	112	59	125	112	188	M3	Δ_{FA}	5.14	20%
MF	AX-90617761	2	6,993,631	161	143	104	78	94	90	175	125	M3	Δ_{FA}	6.10	20%
MF	AX-90731985	2	6,994,751	151	153	96	86	84	100	147	153	M3	Δ_{FA}	5.49	20%
MF	AX-91844146	2	6,995,423	160	144	104	78	89	95	153	147	M3	Δ_{FA}	5.96	20%
MF	AX-91449028	2	6,995,647	116	188	79	103	58	126	95	205	M3	Δ_{FA}	5.24	20%
MF	AX-90601996	2	7,041,069	102	202	74	108	50	134	75	225	M3	Δ_{FA}	8.24	5%
MF	AX-91397720	2	7,044,549	121	183	77	105	50	134	78	222	M3	Δ_{FA}	7.38	5%
MF	AX-90780177	2	188,534,511	255	49	124	44	188	10	283	17	M3	Δ_{FA}	5.63	20%
MF	AX-91456671	1	17,179,645	277	27	159	10	180	17	247	53	M3	Δ_{FF}	6.09	20%
MF	AX-90601996	2	7,041,069	102	202	74	108	50	134	75	225	M3	Δ_{FA+FF}	5.91	20%
MF	AX-90824812	3	119,685,903	71	233	37	143	106	80	159	141	M3	Δ_{FA+FF}	5.19	20%
MF	AX-90826563	3	126,522,366	218	86	141	39	135	51	185	115	M3	Δ_{FA+FF}	5.19	20%
MF	AX-91120582	9	18,615,185	34	270	23	127	156	60	182	118	M3	Δ_{FA+FF}	5.95	20%
MF	AX-91583310	3	158,974,594	158	146	92	85	48	141	97	203	M3	Δ_{DD+DA}	6.01	20%
MF	AX-91619685	4	121,506,569	170	134	110	76	139	41	215	85	M3	Δ_{DD+DA}	5.27	20%
MF	AX-91410184	6	87,412,104	83	221	31	133	61	141	114	186	M3	Δ_{DD+DA}	5.87	20%
MF	AX-91380030	6	87,416,360	84	220	31	132	61	142	114	186	M3	Δ_{DD+DA}	6.03	20%
MF	AX-90549146	6	87,416,883	89	215	33	130	61	142	114	186	M3	Δ_{DD+DA}	5.99	20%

Trait	SNP	Chromosome	Position	0FF	1FF	0FA	1FA	0DA	1DA	0DD	1DD	Model	Test	-log$_{10}$(pval)	FDR
MF	AX-91100444	8	123,079,618	138	166	69	120	79	98	170	130	M3	Δ_{DD+DA}	5.64	20%
MF	AX-91100441	8	123,080,508	137	167	69	120	79	98	170	130	M3	Δ_{DD+DA}	5.65	20%
MF	AX-91100502	8	123,272,484	168	136	84	105	119	58	233	67	M3	Δ_{DD+DA}	5.41	20%
MF	AX-91100608	8	123,504,353	155	149	78	111	119	58	232	68	M3	Δ_{DD+DA}	5.47	20%
MF	AX-91100620	8	123,504,889	162	142	81	108	119	58	230	70	M3	Δ_{DD+DA}	5.35	20%
MF	AX-90651064	1	4,980,734	235	69	147	22	155	42	237	63	M3	Δ_{DA+FA}	5.30	20%
MF	AX-91547985	2	206,217,985	261	43	161	11	154	40	241	59	M3	Δ_{DA+FA}	5.12	20%
MF	AX-91412553	3	126,510,640	216	88	139	41	135	51	185	115	M3	Δ_{DA+FA}	5.20	20%
MF	AX-91577365	3	126,511,574	217	87	139	41	135	51	185	115	M3	Δ_{DA+FA}	5.40	20%
MF	AX-90826563	3	126,522,366	218	86	141	39	135	51	185	115	M3	Δ_{DA+FA}	5.98	20%
MF	AX-90608260	3	126,629,929	221	83	142	38	134	52	189	111	M3	Δ_{DA+FA}	6.39	20%
MF	AX-90628418	5	12,229,516	165	139	109	80	162	15	260	40	M3	Δ_{DA+FA}	5.15	20%
MF	AX-90614466	6	163,890,118	222	82	109	66	178	13	284	16	M3	Δ_{DA+FA}	5.52	20%
MF	AX-91711651	6	163,917,087	236	68	123	52	178	13	284	16	M3	Δ_{DA+FA}	5.18	20%
MF	AX-91711655	6	163,918,261	237	67	126	49	178	13	284	16	M3	Δ_{DA+FA}	5.29	20%
MF	AX-91020644	6	163,918,306	236	68	123	52	178	13	285	15	M3	Δ_{DA+FA}	5.17	20%
MF	AX-90577717	1	290,958,790	243	61	137	20	150	59	211	89	M3	Δ_{DD+FF}	5.15	20%
MF	AX-90777295	2	178,185,474	250	54	164	19	171	12	274	26	M3	Δ_{DD+FF}	5.46	20%
MF	AX-91341754	3	6,649,723	232	72	129	47	139	51	201	99	M3	Δ_{DD+FF}	5.74	20%
MF	AX-91583310	3	158,974,594	158	146	92	85	48	141	97	203	M3	Δ_{DD+FF}	7.86	5%
MF	AX-90590040	3	158,974,646	150	154	87	90	48	141	96	204	M3	Δ_{DD+FF}	6.48	20%
MF	AX-90846101	3	199,827,365	46	258	10	176	20	160	45	255	M3	Δ_{DD+FF}	5.29	20%
MF	AX-91642098	5	805,345	92	212	54	150	77	85	152	148	M3	Δ_{DD+FF}	5.30	20%
MF	AX-91221941	5	896,677	223	81	164	40	101	61	182	118	M3	Δ_{DD+FF}	5.26	20%
MF	AX-90962762	5	161,466,615	15	289	10	177	28	151	46	254	M3	Δ_{DD+FF}	5.25	20%
MF	AX-91351897	5	161,467,164	16	288	10	177	28	151	47	253	M3	Δ_{DD+FF}	5.19	20%
MF	AX-90962856	5	161,633,865	16	288	10	177	18	161	35	265	M3	Δ_{DD+FF}	5.58	20%
MF	AX-91766335	8	113,022,360	52	252	16	170	55	125	118	182	M3	Δ_{DD+FF}	5.69	20%
MF	AX-91099737	8	120,244,941	285	19	180	11	150	25	234	66	M3	Δ_{DD+FF}	5.74	20%
MF	AX-91767760	8	120,457,839	285	19	180	11	150	25	235	65	M3	Δ_{DD+FF}	5.23	20%
MF	AX-91099775	8	120,467,292	285	19	180	11	150	25	237	63	M3	Δ_{DD+FF}	5.35	20%
MF	AX-91768187	8	123,200,729	167	137	83	106	118	59	228	72	M3	Δ_{DD+FF}	5.65	20%
MF	AX-91100502	8	123,272,484	168	136	84	105	119	58	233	67	M3	Δ_{DD+FF}	6.17	20%
MF	AX-91100608	8	123,504,353	155	149	78	111	119	58	232	68	M3	Δ_{DD+FF}	6.86	20%
MF	AX-91100620	8	123,504,889	162	142	81	108	119	58	230	70	M3	Δ_{DD+FF}	7.15	5%
MF	AX-90555039	8	123,506,141	27	277	10	179	30	147	81	219	M3	Δ_{DD+FF}	6.27	20%
MF	AX-91100596	8	123,510,186	28	276	10	179	30	147	81	219	M3	Δ_{DD+FF}	5.85	20%
MF	AX-90588565	8	123,510,776	27	277	10	179	30	147	81	219	M3	Δ_{DD+FF}	6.22	20%
MF	AX-91430137	8	123,511,566	23	281	10	179	30	147	79	221	M3	Δ_{DD+FF}	5.31	20%
MF	AX-91427084	8	123,511,765	23	281	10	178	30	148	79	221	M3	Δ_{DD+FF}	5.30	20%
MF	AX-90555041	8	123,511,872	23	281	10	179	30	147	79	221	M3	Δ_{DD+FF}	5.30	20%
MF	AX-90635840	8	123,512,089	23	281	10	179	24	153	61	239	M3	Δ_{DD+FF}	5.47	20%
MF	AX-91439741	8	123,512,155	28	276	10	179	24	153	64	236	M3	Δ_{DD+FF}	6.04	20%
MF	AX-91770045	8	133,560,452	182	122	95	93	116	62	217	83	M3	Δ_{DD+FF}	5.31	20%
MF	AX-90648672	8	133,560,831	182	122	95	93	116	62	216	84	M3	Δ_{DD+FF}	5.33	20%

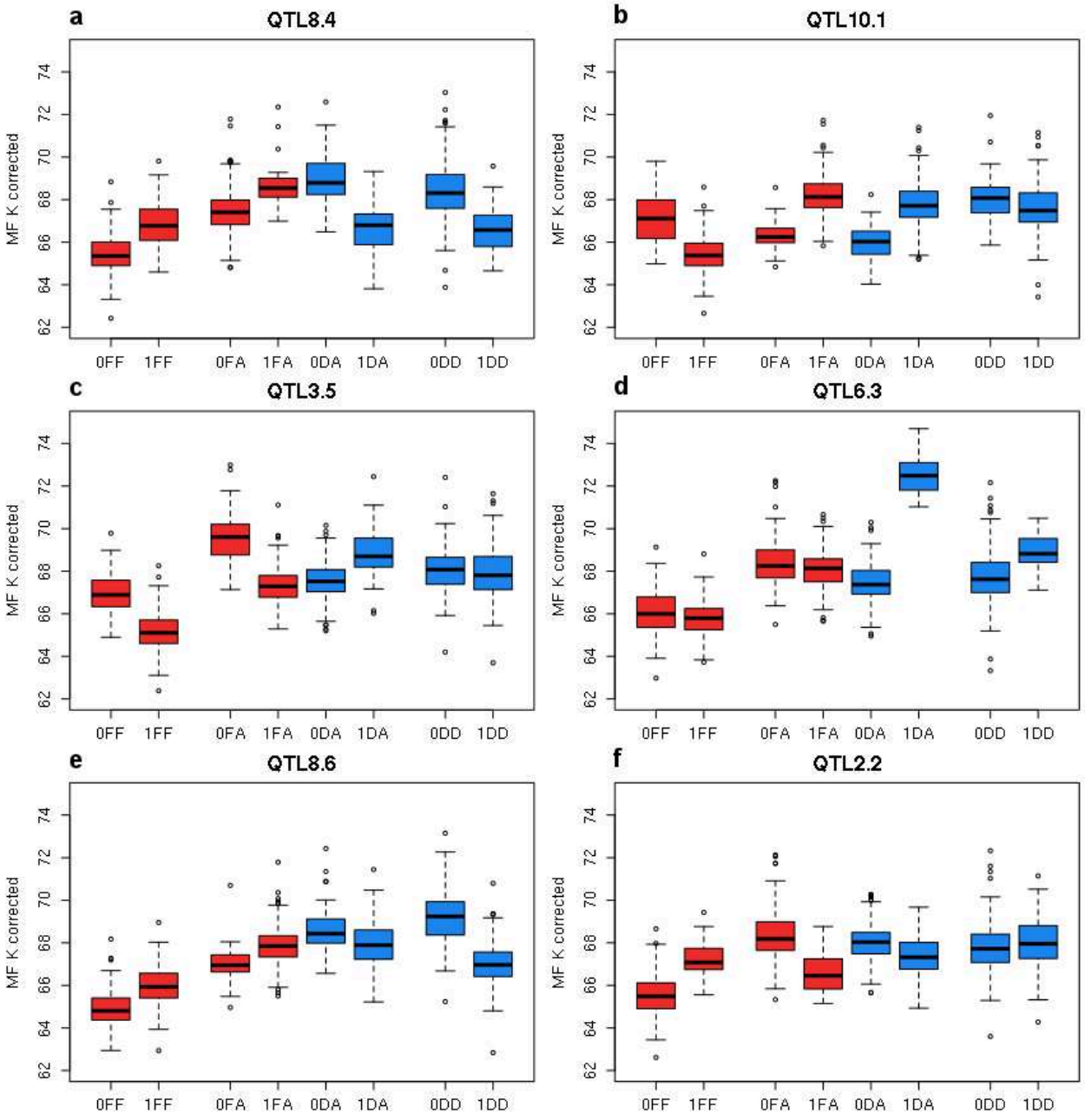
Trait	SNP	Chromosome	Position	0FF	1FF	0FA	1FA	0DA	1DA	0DD	1DD	Model	Test	-log$_{10}$(pval)	FDR
MF	AX-90555247	8	133,562,520	180	124	94	93	112	67	205	95	M3	Δ_{DD+FF}	5.42	20%
MF	AX-91111564	8	162,223,261	42	262	10	166	44	146	80	220	M3	Δ_{DD+FF}	5.49	20%
MF	AX-91775479	8	162,226,185	65	239	20	156	58	132	96	204	M3	Δ_{DD+FF}	5.18	20%
MF	AX-90560230	10	109,485,617	203	101	130	52	168	16	260	40	M3	Δ_{DD+FF}	5.70	20%
MF	AX-91827513	10	109,519,556	197	107	130	52	168	16	260	40	M3	Δ_{DD+FF}	6.19	20%
MF	AX-90651064	1	4,980,734	235	69	147	22	155	42	237	63	M3	$\Delta_{DD+DA+FA+FF}$	5.21	20%
MF	AX-91547985	2	206,217,985	261	43	161	11	154	40	241	59	M3	$\Delta_{DD+DA+FA+FF}$	5.46	20%
MF	AX-90608260	3	126,629,929	221	83	142	38	134	52	189	111	M3	$\Delta_{DD+DA+FA+FF}$	5.51	20%
MF	AX-91583310	3	158,974,594	158	146	92	85	48	141	97	203	M3	$\Delta_{DD+DA+FA+FF}$	6.59	20%
MF	AX-90590040	3	158,974,646	150	154	87	90	48	141	96	204	M3	$\Delta_{DD+DA+FA+FF}$	5.80	20%
MF	AX-91410184	6	87,412,104	83	221	31	133	61	141	114	186	M3	$\Delta_{DD+DA+FA+FF}$	5.96	20%
MF	AX-91380030	6	87,416,360	84	220	31	132	61	142	114	186	M3	$\Delta_{DD+DA+FA+FF}$	6.22	20%
MF	AX-91100441	8	123,080,508	137	167	69	120	79	98	170	130	M3	$\Delta_{DD+DA+FA+FF}$	5.27	20%
MF	AX-91768187	8	123,200,729	167	137	83	106	118	59	228	72	M3	$\Delta_{DD+DA+FA+FF}$	5.13	20%
MF	AX-91100502	8	123,272,484	168	136	84	105	119	58	233	67	M3	$\Delta_{DD+DA+FA+FF}$	5.69	20%
MF	AX-91100608	8	123,504,353	155	149	78	111	119	58	232	68	M3	$\Delta_{DD+DA+FA+FF}$	6.45	20%
MF	AX-91100620	8	123,504,889	162	142	81	108	119	58	230	70	M3	$\Delta_{DD+DA+FA+FF}$	6.46	20%
MF	AX-90555039	8	123,506,141	27	277	10	179	30	147	81	219	M3	$\Delta_{DD+DA+FA+FF}$	5.99	20%
MF	AX-91100596	8	123,510,186	28	276	10	179	30	147	81	219	M3	$\Delta_{DD+DA+FA+FF}$	5.69	20%
MF	AX-90588565	8	123,510,776	27	277	10	179	30	147	81	219	M3	$\Delta_{DD+DA+FA+FF}$	5.98	20%
MF	AX-91430137	8	123,511,566	23	281	10	179	30	147	79	221	M3	$\Delta_{DD+DA+FA+FF}$	5.23	20%
MF	AX-91427084	8	123,511,765	23	281	10	178	30	148	79	221	M3	$\Delta_{DD+DA+FA+FF}$	5.22	20%
MF	AX-90555041	8	123,511,872	23	281	10	179	30	147	79	221	M3	$\Delta_{DD+DA+FA+FF}$	5.23	20%
MF	AX-91439741	8	123,512,155	28	276	10	179	24	153	64	236	M3	$\Delta_{DD+DA+FA+FF}$	5.31	20%
MF	AX-90824812	3	119,685,903	71	233	37	143	106	80	159	141	M3	Δ_{DA-FA}	5.32	20%
MF	AX-91020683	6	164,143,140	190	114	103	74	176	13	280	20	M3	Δ_{DA-FA}	6.21	20%
MF	AX-90801998	3	30,698,113	33	271	17	159	41	149	64	236	M3	Δ_{DD-DA}	5.30	20%
MF	AX-91636629	4	213,775,578	95	209	53	148	38	127	87	213	M3	Δ_{DD-DA}	5.16	20%
MF	AX-90925388	5	17,633,689	260	44	167	14	113	72	187	113	M3	Δ_{DD-DA}	5.65	20%
MF	AX-91744673	7	173,737,798	210	94	113	62	161	30	243	57	M3	Δ_{DD-DA}	5.58	20%
MF	AX-90734998	2	17,677,961	263	41	154	28	125	59	203	97	M3	Δ_{FA-FF}	5.15	20%
MF	AX-90800545	3	25,561,024	137	167	73	102	160	31	272	28	M3	Δ_{DD-FF}	5.67	20%
MF	AX-90843586	3	191,173,981	283	21	188	11	121	46	232	68	M3	Δ_{DD-FF}	5.25	20%
MF	AX-90863948	4	31,102,452	160	144	107	79	53	127	115	185	M3	Δ_{DD-FF}	6.59	20%
MF	AX-91218190	4	31,102,486	161	143	107	79	53	127	115	185	M3	Δ_{DD-FF}	6.93	20%
MF	AX-90863950	4	31,102,555	161	143	107	79	53	127	115	185	M3	Δ_{DD-FF}	6.27	20%
MF	AX-91602775	4	31,102,595	161	143	107	79	53	127	116	184	M3	Δ_{DD-FF}	6.32	20%
MF	AX-90863961	4	31,102,637	160	144	107	79	53	127	116	184	M3	Δ_{DD-FF}	6.65	20%
MF	AX-90863956	4	31,106,881	169	135	108	78	75	105	146	154	M3	Δ_{DD-FF}	6.10	20%
MF	AX-91424173	8	161,789,150	71	233	32	145	52	137	100	200	M3	Δ_{DD-FF}	5.74	20%
MF	AX-91803234	9	134,451,420	200	104	112	44	133	77	180	120	M3	Δ_{DD-FF}	5.43	20%
MF	AX-91150818	9	134,451,487	199	105	112	44	133	77	179	121	M3	Δ_{DD-FF}	5.54	20%
MF	AX-91150814	9	134,461,729	200	104	112	44	133	77	182	118	M3	Δ_{DD-FF}	5.29	20%
MF	AX-91150847	9	134,501,960	199	105	112	44	133	77	180	120	M3	Δ_{DD-FF}	5.55	20%
MF	AX-91803264	9	134,503,559	199	105	112	44	133	77	178	122	M3	Δ_{DD-FF}	5.28	20%

Trait	SNP	Chromosome	Position	0FF	1FF	0FA	1FA	0DA	1DA	0DD	1DD	Model	Test	-log$_{10}$(pval)	FDR
MF	AX-91150842	9	134,550,712	199	105	113	44	132	77	182	118	M3	Δ_{DD-FF}	5.35	20%
MF	AX-90824812	3	119,685,903	71	233	37	143	106	80	159	141	M3	$\Delta_{(DD+DA)-(FA+FF)}$	5.17	20%
MF	AX-90863948	4	31,102,452	160	144	107	79	53	127	115	185	M3	$\Delta_{(DD+DA)-(FA+FF)}$	5.16	20%
MF	AX-91218190	4	31,102,486	161	143	107	79	53	127	115	185	M3	$\Delta_{(DD+DA)-(FA+FF)}$	5.39	20%
MF	AX-90863961	4	31,102,637	160	144	107	79	53	127	116	184	M3	$\Delta_{(DD+DA)-(FA+FF)}$	5.20	20%
MF	AX-91103145	8	132,533,242	238	66	164	23	138	41	242	58	M3	$\Delta_{(DD+DA)-(FA+FF)}$	5.52	20%
MF	AX-90723036	1	276,990,838	236	68	135	39	176	16	271	29	M3	$\Delta_{(DD+FF)-(DA+FA)}$	5.45	20%
MF	AX-90734998	2	17,677,961	263	41	154	28	125	59	203	97	M3	$\Delta_{(DD+FF)-(DA+FA)}$	5.31	20%
MF	AX-91562588	3	49,073,554	102	202	59	122	47	138	75	225	M3	$\Delta_{(DD+FF)-(DA+FA)}$	5.53	20%
MF	AX-91818269	10	60,084,694	26	278	14	167	26	159	35	265	M3	$\Delta_{(DD+FF)-(DA+FA)}$	5.19	20%
MF	AX-91172703	10	60,153,876	34	270	20	161	25	160	37	263	M3	$\Delta_{(DD+FF)-(DA+FA)}$	6.11	20%
MF	AX-90583023	7	142,783,061	175	129	104	79	167	16	253	47	M3	$\Delta_{(DD-DA)-(FF-FA)}$	5.17	20%
MF	AX-91744673	7	173,737,798	210	94	113	62	161	30	243	57	M3	$\Delta_{(DD-DA)-(FF-FA)}$	6.21	20%

Supplementary Table S2.3: Information regarding significant SNPs for FF using all GWAS strategies: the name of the SNP, the chromosome on which it is located, its position in bp along the chromosome, the frequency of the allelic state observed in the dataset in which it was tested, the GWAS model applied, the hypothesis tested, the -$\log_{10}$(pval) of the test and the FDR for which it was declared significant

Trait	SNP	Chromosome	Position	0FF	1FF	0FA	1FA	0DA	1DA	0DD	1DD	Model	Test	-$\log_{10}$(pval)	FDR
FF	AX-91447509	2	9,973,231	-	-	-	-	-	-	169	131	M1	Δ^m (Dent)	5.10	20%
FF	AX-90732788	2	9,973,666	-	-	-	-	-	-	171	129	M1	Δ^m (Dent)	5.73	20%
FF	AX-91511057	2	9,973,932	-	-	-	-	-	-	154	146	M1	Δ^m (Dent)	5.09	20%
FF	AX-91338134	2	9,980,899	-	-	-	-	-	-	87	213	M1	Δ^m (Dent)	5.16	20%
FF	AX-91581634	3	149,790,500	-	-	-	-	-	-	227	73	M1	Δ^m (Dent)	5.26	20%
FF	AX-90832635	3	149,799,411	-	-	-	-	-	-	226	74	M1	Δ^m (Dent)	5.39	20%
FF	AX-90834909	3	158,880,237	-	-	-	-	-	-	107	193	M1	Δ^m (Dent)	5.09	20%
FF	AX-90566336	3	158,889,565	-	-	-	-	-	-	109	191	M1	Δ^m (Dent)	5.68	20%
FF	AX-91583291	3	158,891,367	-	-	-	-	-	-	108	192	M1	Δ^m (Dent)	5.57	20%
FF	AX-90566337	3	158,893,642	-	-	-	-	-	-	107	193	M1	Δ^m (Dent)	5.16	20%
FF	AX-90834898	3	158,896,163	-	-	-	-	-	-	106	194	M1	Δ^m (Dent)	5.70	20%
FF	AX-90638102	3	158,897,644	-	-	-	-	-	-	105	195	M1	Δ^m (Dent)	5.10	20%
FF	AX-91583310	3	158,974,594	-	-	-	-	-	-	97	203	M1	Δ^m (Dent)	11.61	5%
FF	AX-90590040	3	158,974,646	-	-	-	-	-	-	96	204	M1	Δ^m (Dent)	10.42	5%
FF	AX-90834934	3	158,974,756	-	-	-	-	-	-	102	198	M1	Δ^m (Dent)	10.05	5%
FF	AX-91408371	3	158,975,082	-	-	-	-	-	-	101	199	M1	Δ^m (Dent)	9.65	5%
FF	AX-91583384	3	159,391,016	-	-	-	-	-	-	119	181	M1	Δ^m (Dent)	4.78	20%
FF	AX-91583403	3	159,405,200	-	-	-	-	-	-	119	181	M1	Δ^m (Dent)	5.06	20%
FF	AX-90835045	3	159,415,239	-	-	-	-	-	-	120	180	M1	Δ^m (Dent)	4.78	20%
FF	AX-91744205	7	171,556,516	-	-	-	-	-	-	216	84	M1	Δ^m (Dent)	5.39	20%
FF	AX-91451275	8	162,519,150	-	-	-	-	-	-	122	178	M1	Δ^m (Dent)	4.84	20%
FF	AX-90563249	10	2,257,467	-	-	-	-	-	-	156	144	M1	Δ^m (Dent)	4.94	20%
FF	AX-91456671	1	17,179,645	277	27	-	-	-	-	-	-	M1	Δ^m (Flint)	5.45	20%
FF	AX-90602128	2	45,989,148	198	106	-	-	-	-	-	-	M1	Δ^m (Flint)	5.99	20%
FF	AX-90591872	4	35,758,347	293	11	-	-	-	-	-	-	M1	Δ^m (Flint)	5.87	20%
FF	AX-90974105	5	203,928,025	63	241	-	-	-	-	-	-	M1	Δ^m (Flint)	5.76	20%
FF	AX-91100415	8	122,949,646	284	20	-	-	-	-	-	-	M1	Δ^m (Flint)	9.35	5%
FF	AX-91100407	8	122,950,963	279	25	-	-	-	-	-	-	M1	Δ^m (Flint)	7.64	5%
FF	AX-91428720	8	122,952,245	291	13	-	-	-	-	-	-	M1	Δ^m (Flint)	5.79	20%
FF	AX-90596641	8	122,952,991	284	20	-	-	-	-	-	-	M1	Δ^m (Flint)	5.36	20%
FF	AX-91100471	8	123,151,156	286	18	-	-	-	-	-	-	M1	Δ^m (Flint)	4.98	20%
FF	AX-91768204	8	123,372,565	293	11	-	-	-	-	-	-	M1	Δ^m (Flint)	5.09	20%
FF	AX-91108827	8	152,591,911	269	35	-	-	-	-	-	-	M1	Δ^m (Flint)	5.48	20%
FF	AX-91116977	9	5,415,277	200	104	-	-	-	-	-	-	M1	Δ^m (Flint)	6.52	5%
FF	AX-91116965	9	5,417,992	112	192	-	-	-	-	-	-	M1	Δ^m (Flint)	6.15	5%
FF	AX-91779203	9	5,420,956	207	97	-	-	-	-	-	-	M1	Δ^m (Flint)	5.28	20%
FF	AX-91779204	9	5,445,523	210	94	-	-	-	-	-	-	M1	Δ^m (Flint)	5.42	20%
FF	AX-91364734	10	85,834,561	279	25	-	-	-	-	-	-	M1	Δ^m (Flint)	5.33	20%
FF	AX-91583310	3	158,974,594	158	146	-	-	-	-	97	203	M2	Δ^m_D	10.05	5%
FF	AX-90590040	3	158,974,646	150	154	-	-	-	-	96	204	M2	Δ^m_D	9.04	5%
FF	AX-90834934	3	158,974,756	246	58	-	-	-	-	102	198	M2	Δ^m_D	8.78	5%
FF	AX-91408371	3	158,975,082	246	58	-	-	-	-	101	199	M2	Δ^m_D	8.48	5%
FF	AX-91100415	8	122,949,646	284	20	-	-	-	-	274	26	M2	Δ^m_F	6.02	20%
FF	AX-91100407	8	122,950,963	279	25	-	-	-	-	266	34	M2	Δ^m_F	6.01	20%
FF	AX-91583310	3	158,974,594	158	146	-	-	-	-	97	203	M2	Δ^m_{D+F}	7.88	5%

Trait	SNP	Chromosome	Position	0FF	1FF	0FA	1FA	0DA	1DA	0DD	1DD	Model	Test	-log$_{10}$(pval)	FDR
FF	AX-90590040	3	158,974,646	150	154	-	-	-	-	96	204	M2	Δ^m_{D+F}	7.05	5%
FF	AX-90710320	1	230,470,022	169	135	-	-	-	-	222	78	M2	Δ^m_{D-F}	5.99	20%
FF	AX-91583310	3	158,974,594	158	146	92	85	48	141	97	203	M3	Δ_{DD}	9.74	5%
FF	AX-90590040	3	158,974,646	150	154	87	90	48	141	96	204	M3	Δ_{DD}	8.82	5%
FF	AX-90834934	3	158,974,756	246	58	139	38	50	139	102	198	M3	Δ_{DD}	8.16	5%
FF	AX-91408371	3	158,975,082	246	58	139	38	50	139	101	199	M3	Δ_{DD}	7.93	5%
FF	AX-90601996	2	7,041,069	102	202	74	108	50	134	75	225	M3	Δ_{FA}	6.28	20%
FF	AX-91583310	3	158,974,594	158	146	92	85	48	141	97	203	M3	Δ_{DD+FF}	7.49	5%
FF	AX-90590040	3	158,974,646	150	154	87	90	48	141	96	204	M3	Δ_{DD+FF}	6.86	20%
FF	AX-91768187	8	123,200,729	167	137	83	106	118	59	228	72	M3	Δ_{DD+FF}	6.18	20%
FF	AX-91100502	8	123,272,484	168	136	84	105	119	58	233	67	M3	Δ_{DD+FF}	6.33	20%
FF	AX-91100608	8	123,504,353	155	149	78	111	119	58	232	68	M3	Δ_{DD+FF}	6.76	20%
FF	AX-91100620	8	123,504,889	162	142	81	108	119	58	230	70	M3	Δ_{DD+FF}	7.52	5%
FF	AX-91100608	8	123,504,353	155	149	78	111	119	58	232	68	M3	$\Delta_{DD+DA+FA+FF}$	6.86	20%
FF	AX-91100620	8	123,504,889	162	142	81	108	119	58	230	70	M3	$\Delta_{DD+DA+FA+FF}$	7.22	5%
FF	AX-91562588	3	49,073,554	102	202	59	122	47	138	75	225	M3	$\Delta_{(DD+FF)-(FA+FF)}$	6.30	20%
FF	AX-91686325	6	28,128,237	202	102	119	53	116	78	153	147	M3	$\Delta_{(DD+FF)-(FA+FF)}$	6.41	20%



Supplementary Figure S2.7: Boxplots of phenotypes for the different alleles of six other highlighted QTLs: **a**. *QTL8.4*, **b**. *QTL10.1*, **c**. *QTL3.5*, **d**. *QTL6.3*, **e**. *QTL8.6* and **f**. *QTL2.2*, after correcting for relatedness using $\mathbf{M_3}$. The denomination of the allelic states on the x-axis include the SNP allele (0/1), its ancestry (D/F) and the genetic background in which it is observed (D/A/F), as presented in Table 2.2

Supplementary Table S2.4: Information regarding the six other highlighted QTLs: *QTL8.4*, *QTL10.1*, *QTL3.5*, *QTL6.3*, *QTL8.6* and *QTL2.2*

	QTL8.4	***QTL10.1***	***QTL3.5***	***QTL6.3***	***QTL8.6***	***QTL2.2***
Trait	MF	MF	MF	MF	MF	MF
SNP	AX-91103145	AX-91172703	AX-90824812	AX-91020683	AX-91424173	AX-90734998
Chromosome	8	10	3	6	8	2
Position (Mbp)	132.53	60.15	119.69	164.14	161.79	17.68
Allele frequency						
- 0DD	242	37	159	280	100	203
- 1DD	58	263	141	20	200	97
- 0DA	138	25	106	176	52	125
- 1DA	41	160	80	13	137	59
- 0FA	164	20	37	103	32	154
- 1FA	23	161	143	74	145	28
- 0FF	238	34	71	190	71	263
- 1FF	66	270	233	114	233	41
-$\log_{10}$(pval)						
M_1						
- Δ^m (Dent)	1.85	0.49	0.43	0.32	3.51 *	0.26
- Δ^m (Flint)	2.36 .	1.83	2.24 .	0.12	1.42	3.11 *
M_2						
- Δ^m_D	2.07 .	0.57	0.49	0.61	3.56 *	0.47
- Δ^m_F	1.67	1.05	2.31 .	0.21	1.46	2.54 .
- Δ^m_{D+F}	0.02	1.27	2.16 .	0.37	0.35	2.43 .
- Δ^m_{D-F}	3.39 *	0.20	0.90	0.69	4.32 *	1.20
M_3						
- Δ^m_{DD}	2.80 .	0.40	0.16	0.82	5.86 **	0.14
- Δ^m_{DA}	3.47 *	1.73	1.97	6.55 **	0.64	0.71
- Δ^m_{FA}	1.04	1.57	3.72 *	0.29	0.71	1.80
- Δ^m_{FF}	1.74	1.42	3.52 *	0.26	1.35	2.29 .
- Δ^m_{FA+FF}	1.82	0.09	5.19 **	0.36	1.33	0.01
- Δ^m_{DD+DA}	4.33 *	0.52	0.85	4.49 *	3.25 *	0.29
- Δ^m_{DA+FA}	0.54	2.82 .	0.62	4.52 *	0.08	2.11 .
- Δ^m_{DD+FF}	0.19	1.38	2.46 .	0.50	0.99	1.82
- $\Delta^m_{DD+DA+FA+FF}$	0.46	0.41	1.76	2.99 .	0.35	0.16
- Δ^m_{DA-FA}	3.60 *	0.00	5.32 **	6.21 **	1.11	0.61
- Δ^m_{DD-DA}	0.39	2.68 .	2.22 .	3.44 *	2.21 .	0.90
- Δ^m_{FA-FF}	0.06	4.10 *	0.34	0.05	0.12	5.15 **
- Δ^m_{DD-FF}	4.15*	0.48	1.95	0.93	5.74 **	1.42
- $\Delta^m_{(DD+DA)-(FA+FF)}$	5.52 **	0.23	5.17 **	4.54 *	3.84 *	0.16
- $\Delta^m_{(DD+FF)-(DA+FA)}$	0.30	6.11 **	0.62	2.72 .	0.86	5.31 **
- $\Delta^m_{(DD-DA)-(FF-FA)}$	0.17	0.39	1.71	2.90 .	1.32	2.13 .

*** : -$\log_{10}$(pval) > 7 ; ** : 7 > -$\log_{10}$(pval) > 5 ; * : 3 > -$\log_{10}$(pval) > 5 ; . : 2 > -$\log_{10}$(pval) > 3

Supplementary Table S2.5: Additive, epistatic and residual variance components for each trait with the p-value (pval) of the epistatic component using a likelihood-ratio LR test

The existence of epistasis can be investigated using a test based on variance components. The epistatic variance component between pairs of loci was estimated on the joint dent, flint and admixed dataset using the model:

$$Y_l = \mu + G_l + (G \times G)_l + E_l$$

where $(G \times G)_l$ is the global epistatic deviation of line l where $\boldsymbol{g}_e^T = ((G \times G)_1, .., (G \times G)_N)$ is the vector of epistatic deviations with $\boldsymbol{g}_e \sim \mathcal{N}(0, \boldsymbol{K} \circ \boldsymbol{K}\sigma^2_{(G\times G)})$ where $\boldsymbol{K} \circ \boldsymbol{K}$ is the Hadamard product of the kinship matrix (Eq. 2.2) with itself and $\sigma^2_{(G\times G)}$ is the epistatic genetic variance between pairs of loci. Note that other terms are identical to those described in $\mathbf{M_1}$ (Eq. 2.1). This model can be seen as a simplified version of the one proposed by Vitezica et al. (2017), as purely homozygous lines were used. The epistatic variance component was tested using a LR test between this model and the same model without the term $(G \times G)_l$.

	σ^2_G	$\sigma^2_{(G\times G)}$	σ^2_E	pval
MF	12.50	2.91	2.13	7.55 10^{-3} **
FF	15.55	5.66	0.92	6.80 10^{-5} ***

*** : pval $< 10^{-3}$; ** : $< 10^{-3}$ pval $< 10^{-2}$; * : $< 10^{-2}$ pval $< 5 \times 10^{-2}$

Supplementary Material - Chapitre 3

Supplementary Table S3.1: Phenotypic analysis of MF, FF, PH, ELN and TNL following following the phenotypic analysis presented in Chapter 2 (see Material and Methods)

	MF	FF	PH	ELN	TNL
Row-Column 2015	AR1	AR1	AR1	AR1	AR1
Row-Column 2016	IID	IID	IID	IID	IID
μ_{2015}	64.22	65.86	190.72	11.23	15.95
μ_{2016}	71.20	72.40	182.62	11.19	16.09
μ_{D}	70.81	72.28	198.77	11.96	17.29
μ_{F}	64.95	66.39	172.77	10.48	14.90
μ_{A}	67.66	69.11	183.98	11.12	15.83
$\sigma^2_{G_{\mathrm{D}}}$	24.15	27.65	555.17	1.48	2.28
$\sigma^2_{G_{\mathrm{F}}}$	23.06	27.03	558.80	1.98	3.05
$\sigma^2_{G_{\mathrm{A}}}$	16.89	20.07	695.10	1.26	1.87
$\sigma^2_{(G\times\beta)_{2015,\mathrm{D}}}$	1.31	0.00	0.00	0.33	0.29
$\sigma^2_{(G\times\beta)_{2015,\mathrm{F}}}$	1.35	1.99	35.00	0.10	0.04
$\sigma^2_{(G\times\beta)_{2015,\mathrm{A}}}$	4.64	4.84	118.23	0.42	0.44
$\sigma^2_{(G\times\beta)_{2016,\mathrm{D}}}$	1.09	2.67	88.71	0.00	0.00
$\sigma^2_{(G\times\beta)_{2016,\mathrm{F}}}$	0.00	0.00	19.41	0.00	0.23
$\sigma^2_{(G\times\beta)_{2016,\mathrm{A}}}$	2.11	3.51	83.21	0.00	0.13
$\sigma^2_{E_{2015}}$	2.46	2.56	101.04	0.59	0.63
$\sigma^2_{E_{2016}}$	1.54	1.94	63.76	0.24	0.34
h^2_{D}	0.96	0.96	0.93	0.89	0.92
h^2_{F}	0.96	0.96	0.94	0.94	0.94
h^2_{A}	0.88	0.88	0.91	0.86	0.88
$\bar{r}_{2015}$	1.94	1.94	1.99	1.95	1.95
$\bar{r}_{2016}$	1.99	1.99	2.00	2.00	2.00

The lines "Row-Column" refers to the modeling of row and columns. AR1 refers to the modeling of row and column effects, as defined by the experimental design, following an autoregressive model AR1, while IID refers to the modeling of row and column as being independent and identically distributed for a given trial. For more information, see the ASReml-R reference manual by Butler et al. (2009).

The mean of each trial was computed following: $\mu_j = \mu + \beta_j + \sum_{k=1}^{3} \frac{N_k}{N} \alpha_k$ where N_k is the number of individuals (genotypes) in genetic background k and N is the total number of individuals.

The mean of each genetic background was computed following: $\mu_k = \mu + \alpha_k + \frac{1}{2} \sum_{j=1}^{2} \beta_j$

The heritabilities of each genetic background k were computed as:

$$h^2_k = \frac{\sigma^2_{G_k}}{\sigma^2_{G_k} + \frac{1}{4}\sum_{j=1}^{2} \sigma^2_{(G\times\beta)_{jk}} + \frac{1}{4}\sum_{j=1}^{2} \frac{1}{\bar{r}_j}\sigma^2_{E_j}}$$

where $\bar{r}_i$ is the mean number of genotype replicates in trial j

Supplementary Table S3.2: Average of predictive abilities over 100 CV replicates (SHO method) for the five traits using MAGBLUP 1

	MF	FF	PH	ELN	TNL
DFA_DFA	0.75 (0.04)	0.73 (0.04)	0.57 (0.06)	0.69 (0.05)	0.75 (0.04)
DFA_D	0.62 (0.05)	0.64 (0.05)	0.40 (0.08)	0.45 (0.06)	0.54 (0.06)
D_D	-	-	-	-	-
DF_D	0.65 (0.05)	0.67 (0.05)	0.43 (0.08)	0.48 (0.06)	0.56 (0.06)
DA_D	0.66 (0.05)	0.68 (0.05)	0.47 (0.06)	0.50 (0.07)	0.58 (0.06)
FA_D	0.52 (0.08)	0.54 (0.07)	0.24 (0.11)	0.34 (0.10)	0.46 (0.09)
D_F	-	-	-	-	-
A_D	0.59 (0.06)	0.60 (0.05)	0.32 (0.09)	0.42 (0.08)	0.51 (0.07)
DFA_F	0.69 (0.05)	0.67 (0.05)	0.33 (0.09)	0.65 (0.06)	0.65 (0.05)
F_F	-	-	-	-	-
DF_F	0.68 (0.05)	0.67 (0.05)	0.33 (.09)	0.65 (0.05)	0.66 (0.05)
FA_F	0.70 (0.05)	0.69 (0.05)	0.39 (0.08)	0.66 (0.05)	0.66 (0.05)
DA_F	0.58 (0.08)	0.57 (0.08)	0.32 (0.08)	0.54 (0.08)	0.58 (0.08)
D_F	-	-	-	-	-
A_F	0.65 (0.05)	0.64 (0.05)	0.39 (0.09)	0.59 (0.07)	0.62 (0.05)
DFA_A	0.53 (0.09)	0.55 (0.08)	0.35 (0.08)	0.45 (0.08)	0.51 (0.07)
A_A	0.53 (0.07)	0.55 (0.07)	0.37 (0.08)	0.45 (0.08)	0.51 (0.07)
DF_A	0.55 (0.08)	0.56 (0.07)	0.38 (0.08)	0.46 (0.07)	0.51 (0.08)
DA_A	0.53 (0.07)	0.54 (0.07)	0.34 (0.10)	0.43 (0.08)	0.49 (0.07)
FA_A	0.53 (0.08)	0.55 (0.08)	0.36 (0.09)	0.46 (0.08)	0.50 (0.07)
D_A	-	-	-	-	-
F_A	-	-	-	-	-

Standard deviations over the predictive abilities of the 100 CV replicates are shown between brackets

"-" indicated that the model could not be applied for the given configuration

Supplementary Table S3.3: Average of predictive abilities over 100 CV replicates (SHO method) for the five traits using MAGBLUP 2

	MF	FF	PH	ELN	TNL
DFA_DFA	0.76 (0.04)	0.75 (0.04)	0.57 (0.06)	0.70 (0.05)	0.76 (0.04)
DFA_D	0.64 (0.05)	0.65 (0.05)	0.41 (0.08)	0.49 (0.06)	0.56 (0.06)
D_D	-	-	-	-	-
DF_D	0.67 (0.05)	0.68 (0.05)	0.43 (0.08)	0.52 (0.06)	0.58 (0.06)
DA_D	0.68 (0.05)	0.69 (0.05)	0.48 (0.06)	0.52 (0.07)	0.59 (0.06)
FA_D	0.56 (0.07)	0.57 (0.07)	0.27 (0.11)	0.41 (0.09)	0.51 (0.08)
D_F	-	-	-	-	-
A_D	0.60 (0.06)	0.61 (0.05)	0.34 (0.09)	0.44 (0.08)	0.54 (0.07)
DFA_F	0.70 (0.05)	0.69 (0.05)	0.35 (0.09)	0.66 (0.05)	0.67 (0.05)
F_F	-	-	-	-	-
DF_F	0.69 (0.05)	0.68 (0.05)	0.34 (0.09)	0.67 (0.05)	0.67 (0.05)
FA_F	0.71 (0.05)	0.70 (0.05)	0.41 (0.08)	0.66 (0.05)	0.67 (0.05)
DA_F	0.65 (0.06)	0.65 (0.06)	0.35 (0.08)	0.60 (0.06)	0.63 (0.06)
D_F	-	-	-	-	-
A_F	0.67 (0.05)	0.68 (0.05)	0.41 (0.08)	0.61 (0.06)	0.64 (0.05)
DFA_A	0.55 (0.08)	0.57 (0.07)	0.36 (0.08)	0.47 (0.08)	0.52 (0.07)
A_A	0.55 (0.07)	0.56 (0.07)	0.38 (0.08)	0.47 (0.07)	0.52 (0.07)
DF_A	0.56 (0.08)	0.58 (0.07)	0.38 (0.08)	0.48 (0.07)	0.52 (0.08)
DA_A	0.55 (0.07)	0.56 (0.07)	0.35 (0.10)	0.47 (0.08)	0.52 (0.07)
FA_A	0.54 (0.08)	0.56 (0.08)	0.37 (0.09)	0.48 (0.08)	0.51 (0.07)
D_A	-	-	-	-	-
F_A	-	-	-	-	-

Standard deviations over the predictive abilities of the 100 CV replicates are shown between brackets

"-" indicated that the model could not be applied for the given configuration

Supplementary Table S3.4: Variance of real traits estimated by GBLUP, MAGBLUP 1, MAGBLUP 2 using all 300 pure dent and 304 pure flint lines. Note that σ^2_S could not be estimated in absence of admixed individuals

	Type	MF	FF	PH	ELN	TNL
GBLUP	σ^2_G	20.90	21.83	641.07	1.637	2.75
	σ^2_E	0.35	1.12	63.52	0.09	0.02
MAGBLUP 1	$\sigma^2_{G_D}$	17.86	21.61	576.96	1.44	1.84
	$\sigma^2_{G_F}$	13.81	18.10	486.98	1.36	2.02
	σ^2_E	2.05	1.48	94.23	0.18	0.32
MAGBLUP 2	σ^2_U	40.14	53.11	1528.25	4.19	5.53
	$\sigma^2_{U_D}$	13.42	12.32	311.04	0.27	0.25
	$\sigma^2_{U_F}$	1.37	1.72	0.04	0.00	0.44
	σ^2_E	2.15	1.51	88.24	0.16	0.31

université PARIS-SACLAY
ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Titre : Contributions à la sélection génomique et à la génétique d'association en populations structurées et admixées: application au maïs

Mots clés : Admixture, Coefficient de Détermination, Prédiction génomique, Structure génétique, GWAS, Epistasie

Résumé :

L'essor des marqueurs moléculaires (SNPs) a révolutionné les méthodes de génétique quantitative en permettant l'identification de régions impliquées dans le déterminisme génétique des caractères (QTLs) via la génétique d'association (GWAS), ou encore la prédiction des performances d'individus sur la base de leur information génomique (GS). La stratification des populations en groupes génétiques est courante en sélection animale et végétale. Cette structure peut impacter les méthodes de GWAS et de GS via des différences de fréquence et d'effets des allèles des QTL, ainsi que par des différences de déséquilibre de liaison (LD) entre SNP et QTL selon les groupes.

Pendant cette thèse, deux panels de diversité de maïs ont été utilisés, présentant des niveaux différents de structuration: le panel "Amaizing Dent" représentant les lignées dentées utilisées en Europe et le panel "Flint-Dent" incluant des lignées dentées, cornées européennes, ainsi que des lignées admixées entre ces deux groupes.

En GS, l'impact de la structure génétique sur la qualité des prédictions a été évalué au sein du premier panel pour des caractères de productivité et de phénologie. Cette étude a mis en évidence l'intérêt d'une population d'entraînement (TS) dont la constitution en matière de groupes génétiques est similaire à celle de la population à prédire. Assembler les différents groupes au sein d'un TS multi-groupe apparaît comme une solution efficace pour prédire un large spectre de diversité génétique. Des indicateurs a priori de la précision des prédictions génomiques, basés sur le coefficient de détermination, ont également été évalués, mettant en évidence une efficacité variable selon le groupe et le caractère étudié.

Une nouvelle méthodologie GWAS a ensuite été développée pour étudier l'hétérogénéité des effets capturés par les SNPs selon les groupes. L'intégration des individus admixés à l'analyse permet de séparer les effets des facteurs responsables de l'hétérogénéité des effets alléliques: différence génomique locale (liée au LD ou à une mutation spécifique d'un groupe) ou interactions épistatiques entre le QTL et le fonds génétique. Cette méthodologie a été appliquée au panel "Flint-Dent" pour la précocité de floraison. Des QTL ont été détéctés comme présentant des effets groupe-spécifiques interagissant ou non avec le fonds génétique. De nombreux QTL présentant un profil original ont pu être mis en évidence, incluant des locus connus tels que Vgt1, Vgt2 ou Vgt3. Une importante épistasie directionnelle a aussi été mise en évidence grâce aux individus admixés, confortant l'existence d'interactions épistatiques avec le fonds génétique pour ce caractère.

Sachant l'existence de cette hétérogénéité d'effets alléliques, nous avons développé deux modèles de prédictions génomiques nommées Multi-group Admixed GBLUP (MAGBLUP). Ceux-ci modélisent des effets groupe-spécifiques aux QTLs et sont adaptés à la prédiction d'individus admixés. Le premier permet d'identifier la variance génétique additionnelle créée par l'admixture (variance de ségrégation), alors que le second permet d'évaluer le degré de conservation des effets alléliques entre groupes. Ces deux modèles ont montré un intérêt certain par rapport à des modèles standards pour prédire des caractères simulés, mais plus limité sur des caractères réels.

Enfin, l'intérêt des individus admixés dans la constitution de TS multi-groupes a été évalué à l'aide du second panel. Si leur intérêt a clairement été mis en évidence pour des caractères simulés, des résultats plus variables ont été observés avec les caractères réels, pouvant s'expliquer par la présence d'interactions avec le fonds génétique.

Les nouvelles méthodes et l'utilisation d'individus admixés ouvrent des pistes de recherches intéressantes pour les études de génétique quantitative en population structurée.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

université PARIS-SACLAY
ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Title : Contributions to Genomic Selection and Association Mapping in Structured and Admixed Populations: Application to Maize

Keywords : Admixture , Coefficient of Determination , Genomic Prediction , Genetic Structure , GWAS , Epistasis

Abstract :

The advent of molecular markers (SNPs) has revolutionized quantitative genetics methods by enabling the identification of regions involved in the genetic determinism of traits (QTLs) thanks to association studies (GWAS), or the prediction of the performance of individuals using genomic information (GS). The stratification of populations into genetic groups is common in animal and plant breeding. This structure can impact GWAS and GS methods through group differences in QTL allele frequencies and effects, as well as in linkage disequilibrium (LD) between SNP and QTL.

During this thesis, two maize diversity panels were used, presenting different levels of structuration: the ”Amaizing Dent” panel representing the diversity of dent lines used in Europe and the ”Flint-Dent” panel including dent, flint and admixed lines between these two groups.

In GS, the impact of genetic structure on genomic prediction accuracy was evaluated in the first panel for productivity and phenology traits. This study highlighted the interest of a training population (TS) whose constitution in terms of genetic groups is similar to that of the population to be predicted. Assembling the different groups within a multi-group TS appears as an effective solution to predict a broad spectrum of genetic diversity. A priori indicators of genomic prediction accuracy, based on the coefficient of determination, were also evaluated and highlighted a variable efficiency depending on the group and the trait.

A new GWAS methodology was then developed to study the heterogeneity of the allele effects captured by SNPs depending on the group. The integration of admixed individuals to such analyses allows to disentangle the factors causing the heterogeneity of allele effects across groups: local genomic difference (related to LD or group-specific mutation) or epistatic interactions between the QTL and the genetic background. This methodology was applied to the ”Flint-Dent” panel for flowering time. QTLs have been detected as presenting group-specific effects interacting or not with the genetic background. QTLs with an original profile have been highlighted, including known loci such as Vgt1, Vgt2 or Vgt3. Significant directional epistasis has also been demonstrated using admixed individuals and supported the existence of epistatic interactions with the genetic background for this trait.

Based on the existence of such heterogeneity of allele effects, we have developed two genomic prediction models named Multi-group Admixed GBLUP (MAGBLUP). Both model group-specific QTL effects and are suited to the prediction of admixed individuals. The first allows the identification the additional genetic variance created by the admixture (segregation variance), while the second allows the evaluations of the degree of conservation of SNP allele effects across groups. These two models showed a certain interest compared to standard models to predict simulated traits, but it was more limited on real traits.

Finally, the interest of admixed individuals in multi-group TS was evaluated using the second panel. Although their interest has been clearly demonstrated for simulated traits, more variable results have been observed with the real traits, which can be explained by the presence of interactions with the genetic background.

The new methods and the use of admixed individuals open interesting lines of research for quantitative genetics studies in structured population.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l’Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France