THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (Toulouse INP)

Discipline ou spécialité :

Pathologie, Toxicologie, Génétique et Nutrition

Présentée et soutenue par :

M. DAVID GONZALEZ DIEGUEZ

le jeudi 22 octobre 2020

Titre :

Genomic selection accounting for non-additive genetic effects in pig and corn crossbreeding schemes

---

Ecole doctorale :

Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :

Génétique, Physiologie et Systèmes d'Elevage (GENPHYSE)

Directeur(s) de Thèse :

MME ZULMA VITEZICA

MME LLIBERTAT TUSELL

Rapporteurs :

M. AARON LORENZ, UNIVERSITY OF MINNESOTA MINNEAPOLIS

M. MARIO CALUS, WAGENINGEN UNIVERSITY

Membre(s) du jury :

M. JEAN-PIERRE BIDANEL, INRA JOUY EN JOSAS, Président

M. LEOPOLDO SANCHEZ RODRIGUEZ, INRA ORLEANS, Membre

Mme ZULMA VITEZICA, TOULOUSE INP, Membre

This thesis is dedicated to my parents.
For their endless love, inspiration and tremendous support.

Esta tesis está dedicada a mis padres.
Por su infinito amor, inspiración e inmenso apoyo.

Acknowledgements

I would like to express my heartfelt gratitude to all those who helped me to successfully complete my Ph.D. and shared their friendship and love in this life adventure. It would be impossible to cite all of you in this small piece of paper but all of you are implicitly cited in this page and explicitly included in the list of my heart forever.

First and foremost, I would like to express my gratitude to Zulma Vitezica, for her kindness, wise guidance, lots of patience, encouragement, and constant support throughout my Ph.D. and stay in Toulouse. I would also like to thank Andres Legarra, for his friendship, common sense, for the trips to the mountains and beautiful places and cities, but also for his advice, close collaboration and support during my Ph.D. I am infinitely grateful to both of you.

I would like to express my gratitude to Llibertat Tusell, for her friendship, good humor, advice, guidance, encouragement and support during my Ph.D. I remember the day you invited us for cycling towards the lake and fly kites. It was unforgettable. Endless thanks.

Thanks to Nicolas Gillodes and his beautiful family, for their kindness, discussions, wines, tapas and beers shared.

I would like to thank to…

…All those who contributed to my academic background during my Ph.D. by sharing their valuable knowledge and expertise with dedication and patience.

…All those who contributed to improve my thesis with their advice and recommendations.

…All members of the GenPhySE unit that helped me with various kind of administrative task. Thanks for your kindness and friendship.

…My colleagues, for those talks, lunches, daily greetings and fun. You made my stay at INRAE comfortable and enjoyable.

…My friends, for the special moments shared that made this experience unforgettable. You are always in my heart.

…My family in Guatemala, for always being there whenever I need you. I love you deeply.

…Thank you, Lord, for giving me life and health. You are the one who gives wisdom to man and You are the way of truth.

Finally, this thesis would not have been possible without the financial support of INRAE, RAGT 2n and France Génétique Porc. I would also like to thank the support of the Toulouse Midi-Pyrénées bioinformatics platform.

Toulouse, October 22th 2020
David

Agradecimientos

Quiero expresar mi más sincero agradecimiento a todos aquellos que me ayudaron a culminar con éxito mi doctorado y compartieron su amistad y amor en esta aventura de vida. Sería imposible citarlos a todos en esta pequeña hoja pero todos ustedes están implícitamente citados aquí y explícitamente citados en la lista de mi corazón por siempre.

En primer lugar, quiero expresar mi gratitud a Zulma Vitezica, por su gentileza, sabia guianza, mucha paciencia, ánimo y constante apoyo durante mi Ph.D. y estadía en Toulouse. También quiero agradecer a Andres Legarra, por su amistad, sentido común, por los viajes a las montañas y hermosos lugares y ciudades, y también por sus consejos, estrecha colaboración y apoyo durante mi Ph.D. Estoy infinitamente agradecido con ustedes dos.

Quiero expresar mi gratitud a Llibertat Tusell, por su amistad, buen humor, sus consejos, guianza, ánimo y apoyo durante mi Ph.D. Recuerdo el día que nos invitaste a bicicletear al lago y a volar barriletes. Fue inolvidable. Infinitas gracias por todo.

Gracias a Nicolas Gillodes y su hermosa familia, por su amabilidad, charlas, vinos, tapas y cañas compartidas.

Quiero agradecer a…

…Todos aquellos que contribuyeron a mi formación académica durante mi doctorado compartiendo sus valioso conocimiento y experiencia con dedicación y paciencia.

…Todos aquellos que contribuyeron a mejorar mi tesis con sus consejos y recomendaciones.

…Todo el personal de la unidad de GenPhySE que me ayudaron con las tareas administrativas. Gracias por su amabilidad y amistad.

...Mis compañeros, por aquellas charlas, almuerzos, saludos y diversión. Hicieron mi estadía en INRAE agradable y placentera.

…Mis amigos, por los especiales momentos compartidos que hicieron esta experiencia inolvidable. Los llevaré siempre en mi corazón.

…Mi familia en Guatemala, por siempre estar allí cuando los necesito. Los amo profundamente.

…Gracias Señor, por darme vida y salud. Tú eres el que da la sabiduría al hombre y Tú eres el camino de la verdad.

Finalmente, esta tesis no hubiera sido posible sin el apoyo financiero de INRAE, RAGT 2n y France Génétique Porc. También quiero agradecer el apoyo de la plataforma bio-informática Toulouse Midi-Pyrénées.

Toulouse, 22 de Octubre de 2020
David

Résumé

Cette thèse explore et développe des méthodes pour exploiter les effets génétiques de dominance ou/et d'épistasie sur des modèles de sélection génomique dans les schémas de sélection utilisant le croisement chez le porc et le maïs. Le Chapitre 2 a consisté à estimer et à exploiter la variance de dominance intra-race à travers des stratégies de planification des accouplements pour maximiser le progrès génétique global de l'âge à 100 kg (AGE), de l'épaisseur de lard dorsal (BD) et du poids moyen des porcelets par portée (APWL), dans une population porcine de race Landrace française. La maximisation de la valeur génétique totale au lieu de la seule partie additive dans la planification des accouplements a donné à la descendance un avantage moyen de -0,79 jour, -0,04 mm et 11,3 g pour AGE, BD et APWL, respectivement. En contrepartie, le gain génétique additif attendu a légèrement été réduit (1,8% pour AGE par exemple). Ces résultats indiquent que la planification des accouplements peut améliorer les performances des descendants sans compromettre considérablement le progrès génétique. Dans le Chapitre 3, l'efficacité de la planification des accouplements et de l'évaluation génomique, en tenant compte des effets additifs et de dominance, pour améliorer les performances des individus croisés (CB) a été étudiée par simulation dans un croisement à deux voies chez le porc. Les effets de l'utilisation de différentes sources d'informations dans l'évaluation génétique (uniquement des données de race pure (PB) ou des données PB et CB), de plusieurs valeurs d'héritabilité au sens étroit et large, et de plusieurs stratégies d'accouplement pour produire les animaux CB (accouplements au hasard, minimisant la consanguinité future ou maximisant la valeur génétique totale attendue des animaux CB) ont été évaluées. La sélection des animaux PB sur leurs performances en PB a donné un gain génétique de 0,2 écart-type génétique par génération pour le caractère « performance en croisement ». Ce gain a été doublé lorsque les animaux PB étaient sélectionnés sur leur performance en croisés. Les stratégies d'accouplement ont entraîné une légère augmentation des performances des animaux CB. Lorsque la corrélation génétique entre les performances exprimées chez les animaux PB et CB est faible, la sélection des animaux PB pour leur performance en croisés en utilisant les informations CB est une stratégie plus efficace pour exploiter l'hétérosis et augmenter les performances des animaux CB au niveau commercial. Dans le Chapitre 4, la théorie des modèles d'évaluation génétique chez des hybrides à partir du croisement de lignées pures (comme pour le maïs) a été revue dans un contexte génomique. La covariance entre les hybrides due aux effets de substitution additifs, à la dominance et à

l’épistasie a été dérivée analytiquement. En utilisant les marqueurs SNP, il est possible de décomposer l’aptitude spécifique à la combinaison (SCA) en dominance et épistasie inter-groupes, et de décomposer les aptitudes générales à la combinaison (GCA) en effets additifs intra-lignée et une épistasie additive par additive intra-ligne. Un jeu de données publiques sur des hybrides Dent × Flint a été analysé. Le modèle proposé a été comparé à d'autres modèles génomiques en termes d'estimations des composantes de variance et de capacité prédictive, y compris un modèle supposant un effet commun des gènes des lignées pures. L'étude confirme que la majeure partie de la variabilité observée chez les hybrides est expliquée par la GCA et que les variances dues à la dominance et à l'épistasie sont moins importantes et du même ordre de grandeur. Les modèles basés sur la définition d’effets différents (comme traditionnellement considérés chez le maïs), ou commun aux origines (comme considérés intra-race en sélection animale) ont abouti à des capacités prédictives similaires pour les hybrides.

Summary

This thesis explores and develops methodology to exploit dominance or/and epistasis genetic effects on genomic selection models in pig and maize crossbreeding schemes. The Chapter 2 consisted of estimating and exploiting within-breed dominance variance through mate allocation strategies to maximize the overall genetic merit of the traits age at 100 Kg (AGE), backfat depth (BD) and average piglet weight per litter (APWL), in a French Landrace pig population. Maximizing total genetic values instead of breeding values in matings gave to the progeny an average advantage of − 0.79 days, − 0.04 mm, and 11.3 g for AGE, BD and APWL, respectively, but slightly reduced the expected additive genetic gain (e.g. 1.8 % for AGE). These results indicate that genomic mate allocation can improve the performance of the offspring without dramatically compromising the additive genetic gain. In Chapter 3, the effectiveness of mate allocation strategies and genomic evaluations, accounting for additive and dominance effects, to improve crossbred (CB) performance were investigated by simulation in a two-way pig crossbreeding scheme. Effects of the sources of information used in the genetic evaluation (only purebred (PB) data or PB and CB data), of several narrow and broad-sense heritability values, and of several options for mate allocation to produce the CB (mating at random, minimizing expected future inbreeding, or maximizing the expected total genetic value of crossbred animals) were evaluated. Selecting PB animals for PB performance yielded a genetic gain of 0.2 genetic standard deviations of the trait "CB performance" per generation, whereas selecting PB animals for CB performance doubled the genetic response. Mate allocation strategy resulted in a slight increase of the CB performance. When the genetic correlation between PB and CB is low, selecting PB animals for CB performance using CB information is a more efficient strategy to exploit heterosis and increase performance at the CB commercial level. In Chapter 4, the theory of hybrid genetic evaluation models from single-cross of pure lines (as in maize) was revisited in a genomic context. Covariance between hybrids due to additive substitution effects and dominance and epistatic deviations were analytically derived. Using SNP genotypes, it is possible to split specific combining ability (SCA) into dominance and across-groups epistasis, and to split general combining ability (GCA) into within-line additive effects and within-line additive by additive epistasis. A publicly available maize data set of Dent × Flint hybrids was analyzed. The proposed model was compared to other genomic models in terms of variance components estimation and predictive ability, including a model assuming a common effect of genes across origins. The study confirms that

most variation in hybrids is accounted for by GCA, and that variances due to dominance and epistasis are small and have similar magnitudes. Models based on defining effects either differently (as it is traditionally done in maize) or identically across origins (as it is done in single breeds in livestock) resulted in similar predictive abilities for hybrids.

Table of Contents

1 Chapter 1 General introduction and objective 1
1.1 Genetic selection in animal and plant breeding schemes 3
1.2 Heterosis and inbreeding 4
1.3 Crossbreeding 6
1.4 Genetic selection in crossbreeding schemes 7
1.4.1 Maize hybrid breeding scheme 8
1.4.2 Pig crossbreeding selection scheme 12
1.4.3 Comparison between pig and maize breeding schemes 15
1.5 Non-additive genetic effects in genetic evaluations 17
1.6 Genomic selection models in crossbreeding schemes 19
1.6.1 Genomic selection for crossbred performance in pigs 20
1.6.2 Genomic selection in maize hybrids 23
1.7 Mate allocation 25
1.8 Objective of this thesis 27
1.9 Outline 27
1.10 Appendices 29
Appendix 1.1 29
Appendix 1.2 31
2 Chapter 2 SNP-based mate allocation strategies to maximize total genetic value in pigs 35
2.1 Appendices 47
Appendix 2.1 47
Appendix 2.2 49
3 Chapter 3 Purebred and crossbred genomic evaluation and mate allocation strategies to exploit dominance in pig crossbreeding schemes 51
3.1 Appendices 66
Supplemental material 66
S1. Simulation of QTL effects for two parental breeds and their crossbred 66
S2. Genomic evaluation models 67
4 Chapter 4 Genomic prediction of hybrid crops allows disentangling dominance and epistasis 71
4.1 Appendices 115

Appendix A 115
Appendix B 121
5 Chapter 5 General discussion and perspectives 125
5.1 Introduction 127
5.2 Genomic selection accounting for non-additive genetic effects 129
5.2.1 Genomic selection in pig crossbreeding schemes 129
5.2.2 Genomic selection in maize crossbreeding scheme 131
5.3 Considering directional dominance in genomic evaluations 133
5.4 Inbreeding depression in crossbreds 134
5.5 Mate allocation strategies 135
5.5.1 Mate allocation in pig breeding schemes 135
5.5.2 Potential applications of mate allocation in maize crossbreeding 140
5.6 General conclusion 142
References 143
Training and education during the PhD thesis 155
Résumé/Summary 158

List of abbreviations

AGE: Age at 100 kg

APWL: Average piglet weight at birth within litter

ASGM: effects of allele the same across breeds

BD: Backfat depth

BLUP: best linear unbiased prediction

BSAM: breed-specific allele effects

BV: breeding value

CB: crossbreds

CCPS: combined crossbred purebred selection

DH: double haploid technology

DHL: double haploid line

EBV: estimated breeding value

GBLUP: genomic best linear unbiased prediction

GCA: general combining ability

GEBV: genomic estimated breeding value

GS: genomic selection

GxE: genotype-by-environment interactions

GxG: genotype-by-genotype interactions

HWE: Hardy-Weinberg Equilibrium

LD: linkage disequilibrium

PB: purebreds

QTL: quantitative trait loci

r_{PC}: genetic correlation between PB and CB performances

REML: restricted maximum likelihood

RFLP: restriction fragment length polymorphisms

RRS: reciprocal recurrent selection

SCA: specific combining ability

SD: standard deviation

SNP: single nucleotide polymorphism

Chapter 1

General introduction and objective

1.1 Genetic selection in animal and plant breeding schemes

Animal and plant breeding schemes share the common goal to improve the average performance of a population for complex traits of economic importance through genetic improvement. This genetic improvement can be achieved by selecting individuals with superior genetic merit to be parents of the next generation within a breed or population. Animal and plant breeding have Mendel's laws of inheritance as a common origin, and quantitative genetics has played an important role in the development of both fields (Hickey *et al.* 2017). Quantitative genetics has provided the theoretical basis for the development of different genetic selection evaluations for animals and plants. It allows breeders to maximize the selection response per unit of resources spent on traits that are genetically controlled by a large number of loci with small effects (quantitative trait loci, QTL) (Falconer 1981). Genetic improvement relies on the fact that part of the overall genetic merit of parents is inherited to their progeny, thus, the identification of genetically superior individuals to be parents of the next generations is essential. The heritable part of the total genetic merit of individuals is called breeding value (BV), which cannot be measured directly in the individuals, but it can be inferred from analyses of phenotype records from selection candidates and/or their close relatives (Falconer 1981). Estimated breeding values (EBV) are then used to rank and select the best individuals to be parents of the next generation. This selection is performed recurrently through several generations to increase the frequency of favorable alleles of highly polygenic traits (e.g. grain yield in maize and average daily gain in pigs), what results in an improvement in the population mean of the selected trait.

Although conceptual fundaments of breeding are the same across species, breeding schemes and selection methods in animals and plants have diverged over the decades. This is mainly due to specific features of each species, such as the number of progeny produced per year (plants are much more prolific than animals) and mode of reproduction (allogamy, autogamy or asexual propagation in plants versus sexual reproduction in animals) (Schön and Simianer 2015). Moreover, plant breeders generate a large number of genetically identical individuals as inbred lines, hybrids, or clones, which can be evaluated in well-designed experiments, and repeated across many environments and years, whereas animal breeders use unbalanced phenotypic data recorded from individual animals raised in selection nucleus or production farms (Hickey *et al.* 2017). Furthermore, animal breeders have to deal with sex-limited traits (e.g. milk yield) and traits that cannot be measured on selection candidates

themselves (e.g. meat quality), that need to be measured on their relatives (Hickey *et al.* 2017). Therefore, in animal breeding, the best linear unbiased prediction (BLUP) method, developed by Henderson (1975), has been extensively used in genetic evaluations to exploit information from relatives, i.e. pedigree-based relationships, to estimate BVs. Instead, plant breeders generally do not have to deal with sex-limited traits, thus the genetic merit of a cultivar can be often measured directly from phenotypic records obtained from trials. Nevertheless, similar approaches than in animal breeding are used in some plant breeding programs of outbreeding crops (e.g. maize and forages), where the theory of mixed-models has been adopted for the analyses of genotype-by-environment interaction and prediction of hybrid performance (Schön and Simianer 2015). For instance, in maize, the BLUP method has proved to be useful to account for genotype-by-environment interaction and for hybrid prediction (Bernardo 1996).

The large-scale genome discovery in the last few decades has reduced the gap between animal and plant breeding due to the re-emergence of quantitative genetics as a framework for the incorporation of genomic information in genetic evaluations (Schön and Simianer 2015; Hickey *et al.* 2017). The availability of high-density single nucleotide polymorphism (SNP) markers led to the development of genomic-based prediction methods for complex traits in both plant and animal breeding programs (Meuwissen *et al.* 2001; de los Campos *et al.* 2013). Genomic selection refers to the selection of individuals based on genomic predictions i.e. genomic EBV (GEBV). This approach enhanced genetic gains by increasing the prediction accuracy of EBV of difficult-to-measure, sex-limited and/or low heritable traits compared to pedigree-based methods, but especially by reducing generation intervals in both animal and plant breeding programs as it enables selecting animals at an early age (right after being genotyped). With the decreasing trend of the cost of genotyping, the use of genomic data becames standard practice in genetic evaluations in both fields (Meuwissen *et al.* 2013; Crossa *et al.* 2017).

1.2 Heterosis and inbreeding

In 1876, Charles Darwin described for the first time the phenomenon of “hybrid vigor” in crosses and the harmful effect of inbreeding in self-fertilized plants. Later, Shull (1914) coined the term “heterosis” for designating the superior performance of hybrids over the average performance of its parental inbred lines in maize. The phenomenon of heterosis has

been exploited in livestock and plant production through the use of crosses (Falconer 1981). In animal breeding, it is common to distinguish among maternal, paternal and individual heterosis, however, only individual heterosis is considered here, which depends on the genotype of the individual itself.

There are three main non-exclusive hypotheses to explain the genetic basis of heterosis. These hypotheses rely on the presence of non-additive genetic effects, which are the result of allele by allele interaction, like dominance (complete, partial or overdominance) and epistasis. Dominance is the result of interaction between alleles at the same locus while epistasis results from interactions between alleles at different loci (Falconer 1981). The *dominance hypothesis* states that, in heterozygote loci, the deleterious effects of recessive alleles inherited from one parent are suppressed by the dominance (complete/partially) action of alleles received from the other parent (Davenport 1908; Bruce 1910). Dominance explains inbreeding depression as the accumulation of homozygous deleterious recessive alleles and hybrid vigor by masking such recessive alleles. The *overdominance hypothesis* states the superiority of heterozygote over homozygote genotypes (Shull 1908; Crow 1998; Wang *et al.* 2016). However, overdominance is rarely observed (Crow 1999). Moreover, a strong association between two loci (or linkage disequilibrium (LD)) with favorable dominance alleles in repulsion-phase can result in apparent or pseudo-overdominance (Jones 1917). The *epistasis hypothesis* states that interactions between two or more loci inherited from the parents result in heterosis (Richey 1942; Powers 1944; Williams 1959). Few empirical studies suggest that contributions from epistasis to heterosis can be relevant in pigs, *Arabidopsis* and rice (Bidanel 1993; Melchinger *et al.* 2007; Shen *et al.* 2014).

Although in reality the three hypotheses can act simultaneously, there is little statistical evidence of overdominance and epistasis contributing to heterosis in practice. Therefore, in animal and plant breeding, dominance has been widely considered as the major cause of heterosis (Davenport 1908; Bruce 1910; Crow 1998; Visscher *et al.* 2000).

Empirically, heterosis can be measured as the deviation from additivity i.e. the mid-parent value (Falconer 1981). In theory, in the absence of epistasis, the amount of heterosis in a cross considering one locus is proportional to the magnitude of the dominance effect and the squared difference in allele frequency between parental populations (Falconer 1981). This means that with positive dominance in one locus, an increase in the difference in allele frequencies between parental populations will increase the heterosis. Then, with many loci, the overall heterosis in

one individual will be the combination of the heterosis effect of each locus affecting the trait (Falconer 1981). Therefore, the manifestation of heterosis or inbreeding depression depends on the existence of directional dominance, which states that dominance effects on average across all loci tend to be favorable for the trait (Falconer 1981; Varona *et al.* 2018a). This can be written as $E(d) = \mu_d$, where the dominance effects d across all loci controlling the trait have a positive favorable mean μ_d.

The opposite and complementary to heterosis is inbreeding, which increases both the homozygosity and the frequency of detrimental alleles in the population resulting in a reduction of fitness, a phenomenon known as "inbreeding depression" (Falconer 1981). In plant breeding, there are many cultivars constituted by fully inbred lines (e.g. wheat, barley) and inbreeding is also used in cross-pollinated species to develop inbred lines to produce hybrid crops (e.g. maize). In livestock populations, animal breeders prevent matings between close relatives to avoid the negative effects of inbreeding. For example, many important traits in dairy cattle, such as yield, fertility and udder health, show inbreeding depression (González-Recio *et al.* 2007; Pryce *et al.* 2014; Doekes *et al.* 2019).

Prior to genomic information, the classical inbreeding coefficient computed from pedigree-based relationships was used (VanRaden and Smith 1999), which is an expectation of the proportion of the genome that is identical by descent between two individuals (Howard *et al.* 2017). With the availability of genomic information, the realized proportion of the genome shared by two individuals can be measured more accurately with SNP markers than with pedigree-based information (Hill and Weir 2011). Inbreeding can be characterized SNP-by-SNP or at specific regions. The first approach measures the genomic inbreeding coefficient (f) of an individual as the proportion of homozygous SNP genotypes across loci per individual (e.g. Silió *et al.* 2013). Conversely, its individual heterozygosity can be computed as $(1 - f)$. The second approach considers long uninterrupted segments of homozygous genotypes rather than individual SNP, which are called runs of homozygosity (ROH) (McQuillan *et al.* 2008).

1.3 Crossbreeding

Crossbreeding or hybridization (the latter term is mostly used in plant breeding) is a common practice consisting of crossing two (or more) breeds or lines genetically different that

often results in a descendant with superior performance compared to its parents (Falconer 1981). This superiority of the crossbred animals (or hybrids in plants) is attributed to heterosis effect or hybrid vigor, which, as it has been explained above, is mainly due to dominance. Crossbreeding also gives breeders the opportunity of combining the desirable characteristics of each breed/line to complement the strengths and weaknesses of each other (Smith 1964; Sellier 1976). These two features, heterosis and complementarity between breeds/lines, are the main motivations for using crossbreeding in animal and plant breeding schemes. Since the heterosis effect is lost in subsequent generations, hybrids offer a built-in plant/animal protection system that confers an additional commercial benefit to breeding companies (Schulthess *et al.* 2017).

Maize pioneered the use of crossbreeding for the development of F_1 hybrids (Crow 1998), which was adopted later in other cultivars as well as in some livestock species (Visscher *et al.* 2000). In plant production, hybrids have been extensively used in the field in a variety of economically important crops such as maize, sorghum, sunflower and rice (Duvick 1999). In addition to high grain yield, hybrid crops have other desirable traits such as uniformity and yield stability. In animal production, crossbreeding can be used in beef and dairy cattle (Ahlborn-Breier and Hohenboken 1991; Gregory et al. 1992) and is intensively used in monogastric species, such as pigs, poultry and rabbits (Kosba 1978; Visscher *et al.* 2000; Baselga 2004). This thesis deals with maize and pigs.

1.4 Genetic selection in crossbreeding schemes

The main goal in crossbreeding selection schemes is to improve the performance of crossbred animals or hybrid plants (CB). Please, note that the term “crossbred” or “hybrid” will be used interchangeably throughout this thesis. Given that CB individuals are not suitable for breeding because heterosis is not fully retained in the F_2 descendants, genetic improvement of CB individuals is made through selection of purebred (PB) individuals (or inbred lines) in parental populations (Wei and van der Werf 1994; Technow *et al.* 2012). Nevertheless, the source of information (PB and/or CB performance) used to select the PB animals or inbred lines in parental populations and the genetic correlation between PB and CB performances (r_{PC}) determines how much selection response will be obtained at the CB commercial level as a genetic correlated response. Selection of parental individuals can be based on PB or CB performance only, or combining both PB and CB sources of information.

The genetic improvement achieved in the CB population when PB individuals are selected using only PB performance is only made through a genetic correlated response. This means that if the r_{PC} differs markedly from unity, the genetic gain obtained in PB parental populations will not be efficiently transmitted to the CB progeny (Wei *et al.* 1991) as a genetic correlated response. In most of the economically important traits in pig and maize, the r_{PC} has been estimated lower than one, (e.g. grain yield in maize or average daily gain in pigs) (Richey 1924; Smith 1986; Wientjes and Calus 2017). A r_{PC} can be lower than one due to genotype-by-environment interactions (GxE), but also due to genotype-by-genotype interactions (GxG) (i.e. dominance and/or epistasis) (Bernardo 2010; Wientjes and Calus 2017; Duenk *et al.* 2020). As a consequence, crossbreeding schemes aiming to achieve improved CB performance may be suboptimum if they rely only on selection based on PB performance (Wei *et al.* 1991).

To overcome that, selection methods to explicitly improve CB performance by using hybrid progeny performance rather than PB performance were proposed by plant breeders (Hull 1945; Comstock *et al.* 1949). In 1949, Comstock *et al.* proposed the reciprocal recurrent selection (RRS) method to increase genetic response on hybrid performance by improving the two parental populations simultaneously (see details in the next section). This selection strategy was efficient to improve hybrid performance in maize and was immediately adopted in commercial maize breeding programs with success (Hallauer *et al.* 2010). These methods were also introduced in animal crossbreeding and extended to include both PB and CB performances in pedigree-based genetic evaluations, but up to date, their use is limited due to practical constraints and costs of routine collection of phenotypic and pedigree data from CB (Wei and van der Werf 1994; Dekkers 2007).

In the next sections, a description of how commercial maize and pig crossbreeding selection schemes are organized to exploit heterosis and their main limitations is presented.

1.4.1 Maize hybrid breeding scheme

Maize (*Zea mays* L.) is a cross-pollinated species that exhibits high heterosis for important agronomic traits like grain yield and biomass yield. This phenomenon of hybrid vigor in maize was reported independently by Shull and East in 1908. Later, Shull (1909, 1910) proposed the “pure line method of corn breeding” that provided the basis of modern hybrid

breeding. The method involved the creation of inbred lines through repeated generations of self-pollination, followed by their evaluation in hybrid combination when crossed with other inbred lines. Thus, the combination of inbreeding and hybridization constitutes the basis of maize improvement. Since then, commercial maize breeding programs are focused on exploiting heterosis through the development of F_1 hybrids instead of open-pollinated varieties (Crow 1998).

In the last 100 years, maize breeders did several modifications to Shull's method to optimize the genetic response on hybrid performance. (1) Selection of inbred lines based on the performance of their hybrid progeny or "testcross" was proposed by Jenkins and Brunson (1932) to overcome the limitation of low r_{PC}. (2) Then, Sprague and Tatum (1942) partitioned the genotypic value of a hybrid into a general (GCA) and a specific (SCA) combining ability. They empirically defined the GCA as the average performance of the progeny from a given inbred line when crossed with other inbred lines, while the SCA is the deviation of the hybrid's genotypic value from the parental average based on GCA. Hence, the GCA effect is an indicator of the additive effects for hybrid breeding of a line, while the SCA refers to non-additive genetic effects (Lv *et al.* 2012). Based on the latter concepts, new approaches for the use of progeny test or testcross were suggested. (3) Maize breeders adopted the RRS method to improve GCA and SCA in both parental populations simultaneously (Comstock *et al.* 1949). The method consists of crossing inbred lines from one population with inbred lines from the opposite parental population. Then, the inbred lines with superior GCA are often recycled as parents to develop source populations for line development within each group. (4) The organization of inbred lines into heterotic groups for structuring the genetic diversity. Heterotic groups can be defined as groups of inbred lines that show similar combining ability and heterotic response when crossed to inbred lines from other genetically distinct germplasm groups (Melchinger and Gumber 1998; Smith *et al.* 1999; Reif *et al.* 2005). In modern maize breeding, hybrids are exclusively made by crossing lines between heterotic groups to increase the probability of obtaining superior hybrids. For instance, European flints and US dent (usually called Flint and Dent) are the most used heterotic groups in hybrid breeding programs in Central Europe (Stich *et al.* 2007). (5) Adoption of the early-generation selection strategy, which relies on the fact that the hybrid performance of a line in an early generation (e.g. F_2 and F_4) is a good predictor of its hybrid performance in later generations (e.g. F_6) (Jenkins 1940; Sprague 1946; Bernardo 1991). This strategy is crucial in hybrid breeding, in terms of cost and time savings. (6) More recently, with the development of double haploid (DH) technology, a great number of fully

homozygous lines can be generated rapidly (e.g. three cycles), shortening the generation interval at low cost (Smith *et al.* 2008; Chaikam *et al.* 2019).

A commercial maize crossbreeding program can be seen as a pyramidal structure. The breeding nucleus where the genetic improvement and development of hybrids occurs is on the top. In the middle, there are the multiplier farms where the basic seed of parental inbred lines and certified hybrid seeds are increased in number and sold to the commercial farms. Finally, on the base, the commercial farms cultivate the hybrids on a large scale to meet market demand.

The conventional phenotypic hybrid maize breeding program consists basically of two stages: line development and hybrid evaluation (Kadam and Lorenz 2018). Figure 1.1 is an illustration of the maize breeding scheme. First, maize breeders develop new inbred lines within each heterotic group from an initial "breeding cross" between two elite inbred lines (often parents of successful hybrids), followed by several generations of inbreeding through self-pollination to create new fully homozygous lines. In early-generations (e.g. F_2 and F_4), candidate lines are evaluated based on their testcross progeny performance in few environments (e.g. five), by crossing them with an elite tester line (or 3 testers in F_4) from the opposite heterotic group (Bernardo 2010). In addition, lines are selected based on per se performance for traits having high heritability (e.g. plant morphology, cob traits, quality traits, resistance to diseases and insects), but lines with very poor performance are discarded (Hallauer *et al.* 2010). In case of generating double haploid lines (DHLs), only those DHLs having suitable *per se* performance are evaluated in two rounds of testcross (Bernardo 2010). Second, in an advanced stage, selected inbred lines (or DHLs) in one heterotic group are crossed with those selected in the opposite heterotic group to evaluate their GCA and SCA in hybrid combinations across many environments (e.g. 300-1500). In this stage, special attention is given to identify the best performing single-cross hybrids. Then, superior inbred lines that produce the best performing hybrids are recycled as parents in the next cycle to develop source populations for line development within each heterotic group (Comstock *et al.* 1949).

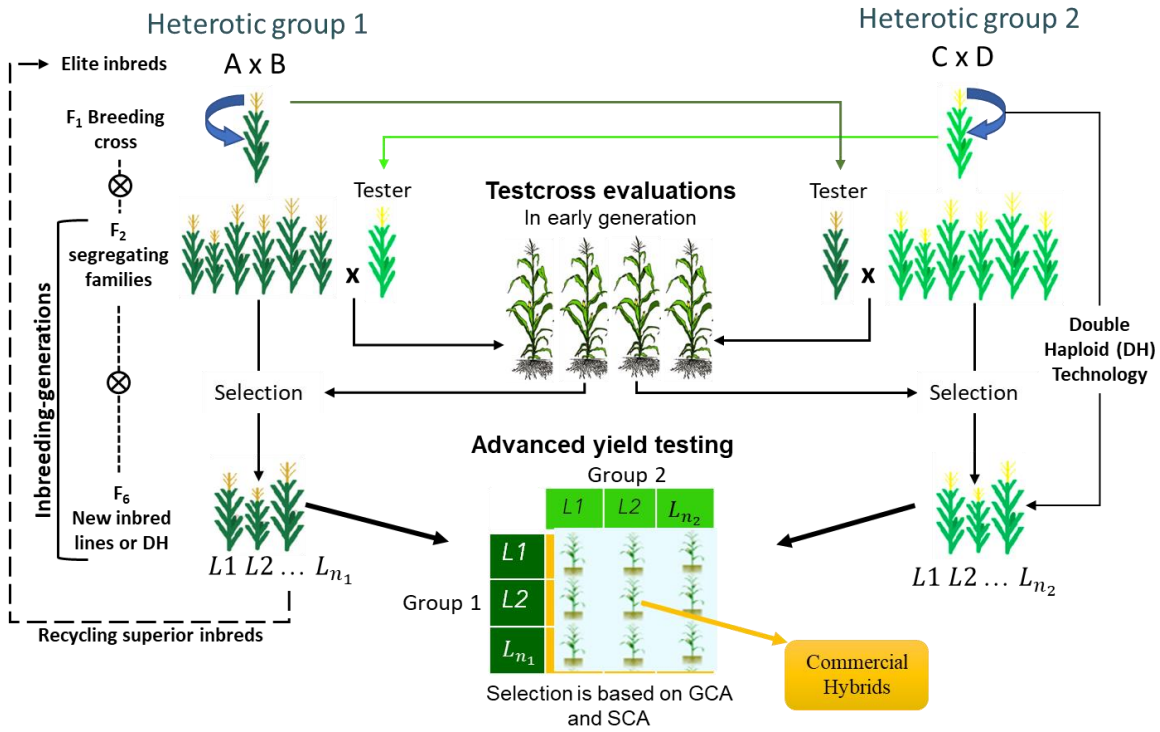

Figure 1.1 Schematic representation of a breeding nucleus in a maize crossbreeding scheme.

The identification of promising pairs of inbred lines that produce the best performing hybrid is still challenging and is the most expensive and critical operation in hybrid breeding schemes (Kadam and Lorenz 2018). For instance, the full assessment of GCA and SCA would be a complete factorial design, where each line from one heterotic group is crossed with each line from the opposite group (Hallauer *et al.* 2010). Thus, even if the breeding program has just 100 inbreds from each group, the number of potential hybrids to evaluate is 10,000. The adoption of DH technology has further exacerbated this situation (e.g. with 1000 DHLs from each pool, the number of potential hybrids reaches one million). In practice, only a small fraction of all potential hybrids is evaluated in field trials (e.g. in an incomplete factorial design). Typically, <1% of the tested hybrids eventually become commercial hybrids (Bernardo 1996). Therefore, hybrid breeders are interested in predictive approaches to predict the genetic potential of untested lines and to identify the most promising untested hybrids to be evaluated (Albrecht *et al.* 2011; Kadam and Lorenz 2018). In the current maize crossbreeding scheme, only hybrid performance records are used for the prediction of untested hybrids and estimation of GCA effects of inbred lines. The performance of inbred lines is often ignored.

Bernardo (1994) introduced to plant breeding the BLUP methodology developed by animal breeders (Henderson 1975). He used the BLUP methodology for hybrid prediction by exploiting hybrid phenotypic data together with genetic (co)variances between tested and untested hybrids. He used either pedigree-based (Bernardo 1996) or marker-based (Bernardo 1994) relationship approaches to predict hybrid performance. The marked-based approach, based on restriction fragment length polymorphisms (RFLPs) markers, resulted in slightly higher prediction accuracy than the pedigree-based approach (Bernardo 1994). With the development of high-throughput genotyping technologies, several studies have been carried out with larger population sizes and high-density SNP markers in plant breeding (Bernardo and Yu 2007; Crossa *et al.* 2010; de los Campos *et al.* 2013) and in maize hybrid prediction (Maenhout *et al.* 2010; Technow *et al.* 2014; Kadam *et al.* 2016). In section 1.6.2 genomic prediction for hybrid performance in maize is introduced, emphasizing how SCA effects are modeled.

1.4.2 Pig crossbreeding selection scheme

In pigs (*Sus scrofa*), the benefit of implementing crossbreeding was demonstrated by Smith and King (1964), who reported higher performance in CB animals for reproductive and growth traits. Since then, crossbreeding schemes became a standard practice in pig breeding programs to produce commercial crossbred animals for production farms.

To exploit heterosis and breed complementarity, pig breeders have developed sire and dam breeds specialized for different objectives in order to answer to different production and market needs. Sire breeds are selected mainly to improve production traits (e.g., growth, carcass quality, feed efficiency), whereas dam breeds are selected for reproduction traits and maternal abilities (e.g. litter size, number of piglets weaned, average piglet weight per litter) while still keeping appropriate production performances (Smith 1964; Visscher *et al.* 2000).

Traditional pig breeding schemes use crossbreeding to produce three-way (sometimes four-way) CB commercial animals, in which, two-way CB females (F_1 sows) from two maternal breeds are mated with boars from a third sire breed (or a two-way CB sire). Figure 1.2 illustrates how the pig industry is organized in a three-way pyramidal structure (Dekkers *et al.* 2011). The nucleus breeding farms are at the top of the pyramid, where specialized pure breeds are bred and selected in high-health environments. Multiplier farms are in the middle, where a large

number of two-way F_1 sows are produced and raised in a medium-health environment to be sold to production farms. Finally, at the base of the pyramid, F1 sows are crossed with PB sires (or two-way CB sires) to produce and raise the crossbred animals (growing-finisher pigs) in production farms in field conditions that are sold to slaughterhouses to produce the pork. The recurrent selection based on PB performance normally takes one or two generations depending if it is a production (e.g. backfat depth) or reproduction (e.g. litter size) trait, respectively. The generation interval is around 1.5 to 2 years in the nucleus. The time lag between selection at the nucleus and slaughter of commercial CB pigs is around 4-5 years depending on the type of trait and the type of crossbreeding scheme (Visscher *et al.* 2000; Dekkers *et al.* 2011; IFIP 2013). In France, French Landrace and Large White are the most popular maternal breeds used to produce F1 sows, while Pietrain is the most commonly used paternal breed (IFIP 2013).

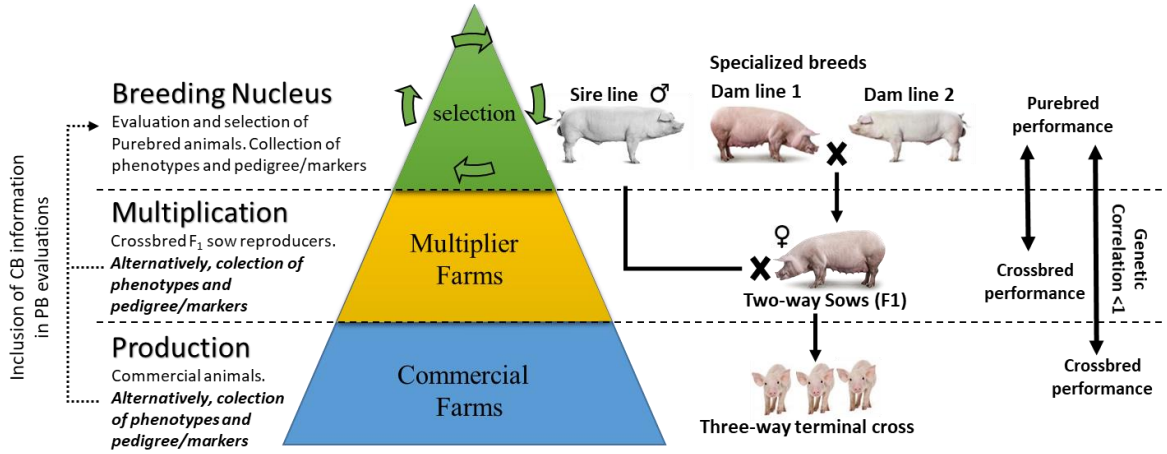

Figure 1.2 Schematic representation of a pig crossbreeding scheme.

In current pig crossbreeding schemes, the BVs of PB animals are commonly estimated from PB information (phenotypes, pedigree and/or SNP markers) recorded at the nucleus level, which may result in a correlated response in CB performance depending on the r_{PC} (Wei *et al.* 1991). Differences between the PB nucleus and commercial farm environments, and the presence of non-additive genetic effects (dominance as a main cause of heterosis) between PB and CB, result in a low r_{PC} (Wientjes and Calus 2017; Duenk *et al.* 2020). Estimates of r_{PC} in pigs, based on different trait categories (growth, meat amount, meat quality, and feed), are on average 0.63, with 50% of the estimates between 0.45 and 0.87 (Wientjes and Calus 2017). As a result, the genetic response in CB performance may be suboptimal if it relies only on PB

information. The impact of r_{PC} can be seen in the equation for the genetic gain in CB performance (ΔG_{CB}) as a correlated response by selecting on PB performance derived by Wientjes and Calus (2017): $\Delta G_{CB} = r_{PC} \frac{i \times \rho_{PB} \times \sigma_{a_{CB}}}{L}$, where i is the intensity of selection, ρ_{PB} is the accuracy of selecting purebreds for PB performance, $\sigma_{a_{CB}}$ is the genetic standard deviation in purebreds for CB performance and L is the generation interval. This formula shows that the ΔG_{CB} is proportional to the r_{PC} value and allows breeders to predict the genetic progress in CB due to selection in PB animals.

To overcome the limitation of r_{PC} lower than unit, animal breeders introduced the RRS method (presented in the previous section) developed by plant breeders (Bell *et al.* 1950). The RRS is a kind of progeny-testing system where each selection cycle takes two generations. However, the practical value of RRS was not as encouraging as expected (Wei and van der Steen 1991). Wei and van der Werf (1994) proposed to use both PB and CB information (phenotypes and pedigree records) in a combined crossbred and purebred selection (CCPS) model, to maximize genetic response for CB performance. The model considered the PB and CB performance as two genetically different traits and provided EBV for both. Later, Lo *et al.* (1993) derived the theory for an additive model to implement the CCPS approach in a BLUP framework. This model was extended to include non-additive genetic effects by Lo *et al.* (1995). However, Lo *et al.*'s model (1995) was too complex for practical applications. Later, Lo *et al.* (1997) proposed an alternative model for a two-breed terminal-crosses where PB animals have different BV depending on whether they are mated to individuals from the same or different breed. The approach of selecting PB animals when they are mated to a different breed is called "selection for CB performance". Lutaaya *et al.* (2001) used the Lo *et al.*'s model (1997) in pigs and showed that it was optimal for obtaining a better genetic response on CB performance. However, it has not been extensively used in pig crossbreeding schemes mainly due to the costs and practical constraints of routine collection of performance and pedigree data of CB animals in field conditions (Ibánẽz-Escriche *et al.* 2009). Further, CCPS increases inbreeding in PB animals because it increases the probability of co-selection of family members (Bijma *et al.* 2001).

With the availability of high-density SNP, genomic evaluations renewed the interest in selection for CB performance. Dekkers (2007) pointed out that genotyping could overcome the need for recording pedigree information from CB by using SNP markers while reducing rates of inbreeding. In section 1.6.1 genomic selection for crossbred performance in pigs is presented.

1.4.3 Comparison between pig and maize breeding schemes

When comparing pig and maize crossbreeding schemes, it can be seen that both breeding schemes share some similarities but also present substantial differences. Both breeding programs rely on the exploitation of heterosis in crossbreeding systems and recurrent selection for improving the parental populations. They are both organized in separated breeds or heterotic groups to maintain or increase genetic divergence in allele frequencies in parental populations to better exploit heterosis. A major difference is that current pig breeding emphasizes on achieving genetic progress in parental populations by selecting PB animals based on PB performance, with the expectation that part of this genetic progress achieved will be transmitted to the commercial crossbred pigs as a genetic correlated response. Instead, in maize breeding inbred lines are selected directly to improve hybrids by using CB performance as selection criterion.

Furthermore, breeding schemes can be split in two basic components according to Hickey *et al.* (2017): (i) A recurrent selection component for genetic improvement and (ii) a product development component (commercial crossbreds or hybrids) for farmers to produce. Figure 1.3 shows a comparison of pig and maize breeding schemes. The main difference is that maize breeding includes both population improvement and product development (hybrids as the marketable product) at the breeding nucleus. Instead, in pig breeding, the nucleus is exclusively for population improvement, and the product development occurs in the subsequent levels in the pyramidal structure. Recurrent selection in pig breeding is faster (around 1.5 to 2 years) than in maize breeding (a slow RRS process, between 7 to 10 years including the product development). Further, the time needed to develop the product in a pig breeding is almost half of that one in maize breeding (4 -5 years).

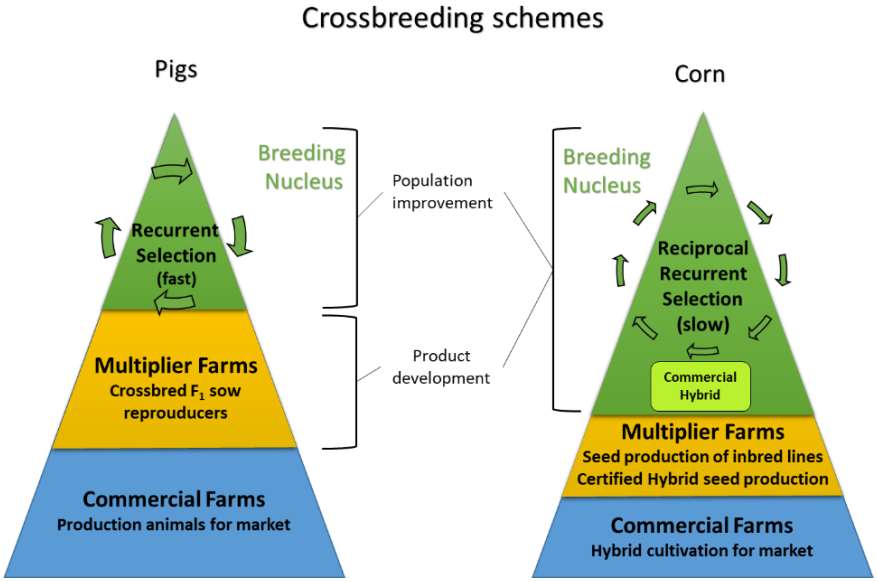

Figure 1.3 Comparison between pig and maize crossbreeding schemes. Adapted from Hickey *et al.* (2017).

Pig and maize crossbreeding schemes have the common goal of improving hybrid performance at the field level by genetically selecting the best pure animals/lines. Despite the substantial differences, similar evaluation methods and tools for prediction of genetic values can be used to achieve this goal. Plant and animal breeders can benefit from exchanging knowledge and expertise to address problems that are common to both fields (Hickey *et al.* 2017). This thesis has intended to establish a bridge between both domains.

Furthermore, the sources of information more often available for genetic evaluations in pig and maize crossbreeding schemes can be compared (Figure 1.4). They both have genotypes, phenotypes and pedigree information from PB animals or inbred lines. However, the performance of inbred lines is often ignored for analysis of hybrid prediction in maize whereas pig selection programs rely on PB performances. Maize breeders have genotypes (derived from parental homozygous genotypes) and pedigree (at least individual and its parents) of all potential hybrids, but also phenotypes of hybrid performance from a subset of hybrids tested across several environments. Instead, in pig breeding, such information (CB performance) either is not always available (or it is not reliable due to PB-CB pedigree bad traceability) or it is available for a small proportion of individuals.

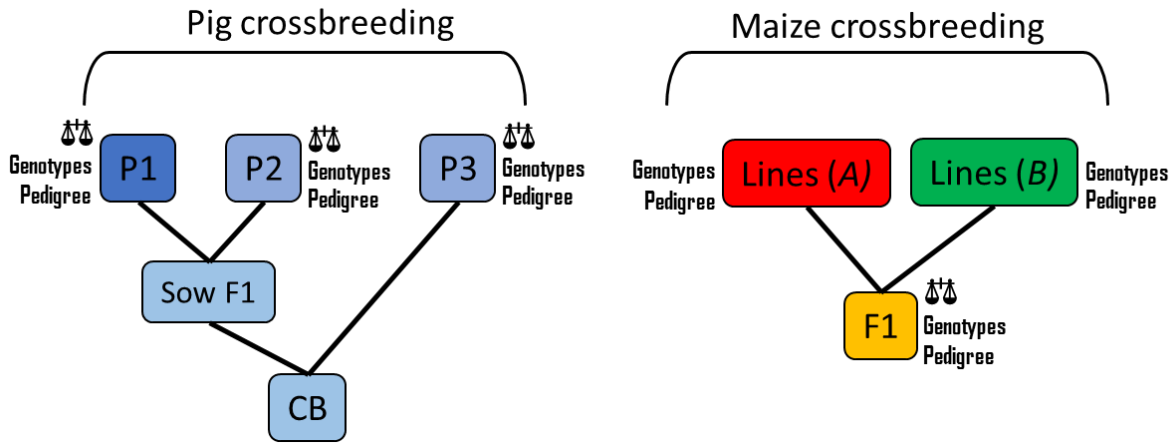

Figure 1.4 Comparison of the sources of information available for genetic evaluations in pig and maize crossbreeding schemes. CB, P1, P2 and P3 stands for crossbreds and purebred animals in breeds 1 to 3. F1 stands for single-cross hybrids between lines in complementary heterotic groups A and B. The "⚖" symbol stands for recorded phenotypes.

Besides, there are other differences between pig and maize crossbreeding due to specific features of each species. All gametes produced from fully homozygous inbred lines are identical. Thus, all F1 hybrids derived from a pair of inbred lines are genetically identical (there is not Mendelian sampling). All hybrids with the same parental inbred line on the paternal or maternal side, share identical copies of the parental gametes. Moreover, because the entire gamete (without recombination) is transmitted to the hybrid progenies, the epistatic combinations in the parental inbred lines are also transmitted. There is recombination from F1 to F2 (segregating population). In contrast, in pigs recombination during meiosis results in a variation of gametes produced by the same animal, which causes that full sibs are not genetically identical (there is Mendelian sampling). In addition, recombination always breaks down epistasis combinations and they are not transmitted to the progeny.

1.5 Non-additive genetic effects in genetic evaluations

The total genetic merit of an individual is the result of both additive and non-additive components. The part of the genetic merit of an individual attributable to the non-additive genetic effects is due to the interaction of the alleles inherited from parents. The value of this interaction or combination of genes is what is called the SCA. This concept is used more often

in plant breeding than in animal breeding probably because animal breeders are more focused on the use of additive effects for genetic improvement and plant breeders on hybrid improvement. Although non-additive effects are not inherited from parents to the offspring because only one allele of each pair is transmitted to the progeny, it might be advantageous to include them in genetic evaluations to improve the prediction of genetic values of individuals. This is of special interest in crossbreeding, where the main goal is to improve the overall genetic merit of CB individuals by exploiting heterosis. Moreover, as the r_{PC} is also influenced by non-additive effects and allele frequencies, it is useful to take into account additive and dominance (co)variances in genetic evaluation models of purebred for crossbred performances (Wei and van der Steen 1991).

The estimation of non-additive effects in a pedigree-based context is difficult. In inbred populations, the dominance relationship matrix is complex to obtain because it involves covariances between BV and dominance deviations (de Boer and Hoeschele 1993). In addition, its inversion is computationally demanding (Misztal 1997). Moreover, to obtain accurate pedigree-based estimates of dominance effects requires large groups of individuals with dominance relationships (the probability of identical genotypes of two individuals at a locus), such as full-sibs (de Boer and Hoeschele 1993; Fernández *et al.* 2017). Furthermore, the use of dominance effects was difficult in practice (e.g. in mate allocation). These are the reasons why non-additive effects were ignored in most of the classical pedigree-based genetic evaluations.

With genomic information, the estimation of dominance effects becomes much easier in a genomic BLUP framework than in pedigree-based methods. This is because instead of probabilities of identical genotypes, the heterozygous states at SNPs across the genome are known (Vitezica *et al.* 2013). Thus, with the availability of high-density SNP markers and the increasing amount of genotyped animals, dominance has been investigated in several livestock populations, including dairy cattle, pigs, sheep, and layer chickens (Varona *et al.* 2018b). In general, estimates of dominance variance account for around 10% of the phenotypic variance in PB livestock populations, but it varies depending on the species and traits (Varona *et al.* 2018b; Tusell *et al.* 2019). In general, the inclusion of dominance does not increase the accuracy of EBV, as reviewed by Varona *et al.* (2018b). In maize, SCA effects are crucial for predicting the total genetic merit of hybrids. Estimates of SCA variance using genomic models, expressed as the ratio SCA/GCA, range from 0.07 to 0.6, depending on the trait (Technow *et al.* 2014; Alves *et al.* 2019). Moreover, estimates of SCA variance are expected to be lower in hybrids

derived from two divergent heterotic groups than in hybrids derived from a homogeneous population (e.g. convergent populations) (Reif *et al.* 2007; Duvick *et al.* 2010). Further, there is little evidence that including SCA effects leads to an increase in genomic prediction accuracy (Kadam and Lorenz 2018).

In presence of directional dominance, genomic models accounting for non-additive genetic effects must consider inbreeding depression to avoid biased estimates of dominance variance, as was proved analytically by Boer and Hoeschele (1993) in a pedigree context and Xiang *et al.* (2016) in a genomic context (see Appendix 1.1). Directional dominance can be accounted for by including a regression on genomic inbreeding (as a covariate) (Xiang *et al.* 2016a; Aliloo *et al.* 2017; Varona *et al.* 2018a). Otherwise, estimates of non-additive effects will be overestimated.

1.6 Genomic selection models in crossbreeding schemes

Genomic selection (GS) is an approach in which dense molecular markers across the genome are used to estimate BV of selection candidates for a trait. GS models fit marker effects either explicitly, as was originally proposed by Meuwissen *et al.* (2001) or implicitly, by using an equivalent model that uses the marker data to build a "genomic" relationship matrix (VanRaden 2008) in a genomic BLUP (GBLUP) framework to predict individual GEBV.

GS relies on the LD between markers and QTL to capture the effects of QTL. Thus, a strong LD between markers and QTL will tend to increase the accuracy of genomic prediction. Other critical factors affecting the genomic prediction that are under breeder's control are the statistical model of choice, marker density, size of the reference population and relationship between individuals in the reference and target populations (Kadam and Lorenz 2018; VanRaden 2020). It has been proved that the accuracy of genomic prediction is higher than pedigree-based prediction in livestock and plant breeding (Bernardo and Yu 2007; Hayes *et al.* 2009; de los Campos *et al.* 2013; Tusell *et al.* 2013; Crossa *et al.* 2017). Furthermore, as mentioned before, GS is especially beneficial for traits that are difficult to measure, including traits that are measured or expressed only on relatives in different environments i.e. crossbred progeny. This thesis is focused on statistical models for genomic selection of PB animals or inbred lines for crossbred performance.

Crossbreeding schemes are widely used in animal and plant breeding, and GS offers new tools to improve CB performance. The next section describes GS in crossbreeding schemes with more details, but before, some general concepts are presented in Appendix 1.2.

1.6.1 Genomic selection for crossbred performance in pigs

GS has been successfully applied in selection based on PB performance for several traits in pigs (Ibáñez-Escriche *et al.* 2014; Knol *et al.* 2016; Samorè and Fontanesi 2016). With the decreasing trend in genotyping cost, the need of accurate recording of pedigree at the crossbred level can be alleviated. Thus, there is a renewed interest in selection for improving CB performance, which is the ultimate goal of a crossbreeding scheme. In addition, CB individuals can be used as training animals (Bijma and Bastiaansen 2014). Thus, several genomic models have been proposed to account for CB information (genotypes and/or phenotypes) in genetic evaluations in pig crossbreeding schemes.

1.6.1.1 Crossbred models with additive effects

As an alternative to CCPS (presented in section 1.4.2), Dekkers (2007) proposed an additive model to use marker data and phenotypes from CB animals to select PB animals for CB performance. The selection index theory was used to evaluate different scenarios (e.g. different types of PB and CB performances) according to their response in CB performance and inbreeding rate. This approach resulted in higher genetic response in CB performance and lower rates of inbreeding than the CCPS method based on simulation. Later, Ibánẽz-Escriche *et al.* (2009) investigated a model to evaluate PB for CB performance considering breed-specific allele effects (BSAM) and compared it to a model that assumed the same allele effects across breeds (ASGM). In their model, CB genotypes and phenotypes were used for training. The SNP effects are expected to be breed-specific due to different LD patterns between markers and QTL and differences in allele frequencies in the parental breeds (Roos *et al.* 2009). The BSAM performed better than the ASGM only when the number of markers was small, the size of the CB training population was large, and parental breeds were distantly related. This result was confirmed later with real pig data (Lopes *et al.* 2017). Besides, Vandenplas *et al.* (2016) and Sevillano *et al.* (2016) evaluated several approaches to assign breed-of-origin of alleles on

simulated and real data. These approaches were efficient for the assignment of breed-of-origin, which allow the implementation of BSAM models in genomic prediction. Later, Vandenplas *et al.* (2017) derived equations for predicting the reliability of GEBV for CB performance using ASGM or BSAM models, which are useful to optimize the design of crossbreeding programs. Sevillano *et al.* (2019) extended the BSAM to distinguish between SNPs strongly associated with CB performance and all remaining SNP. The superiority of their model was only observed when the selected SNPs explained 10% of the total CB genetic variance and when the r_{PC} for the selected and non-selected SNP differed around 20%.

Besides, Hidalgo *et al.* (2016) evaluated the predictive ability of PB and CB performance in pigs using CB and PB genomic data. Using only CB data in the training resulted in lower accuracy of BV for CB performance than using PB in the training. The authors stated that prediction based on PB performance was favored by a high r_{PC} (>0.9) and a small CB training population size. In contrast to these results, Lopes *et al.* (2017) found that training in CB performance produced higher prediction accuracy than training only using PB performance. Overall, these previous studies promote the use of CB information in the genetic evaluation of PB candidates for CB performance. However, all these approaches were single-trait models and considered only CB animals with genotypes and genotypes (excluding all the phenotyped CB animals without genotypes).

To overcome some of these limitations, Christensen *et al.* (2014) proposed an additive model that combines the benefit of the multiple-trait model from Wei and van der Werf (1994) with the single-step GBLUP approach (Legarra *et al.* 2009; Christensen and Lund 2010). Their model can evaluate both PB and CB performances simultaneously in a two-breed crossbreeding system and include genotyped and non-genotyped animals. This approach uses two breed-specific covariance matrices that combine pedigree and genomic information (partial relationship matrices), one for each parental breed. This model makes full use of CB genotypes that need to be phased and therefore accounts for the exact contribution of alleles of the sire and the dam to a given CB performance. This model was also extended for a three-way crossbreeding system by Christensen *et al.* (2015). Xiang *et al.* (2016b) validated the model from Christensen *et al.* (2014) in real data in a pig two-way crossbreeding scheme. They reported that the inclusion of CB genomic information improved the accuracy of prediction for CB performance. Later, Xiang *et al.* (2017) used one relationship matrix with metafounders in a single-step GBLUP model. This matrix related the animals of the three populations and the

breed origin of CB alleles was not needed. They used data from a two-way pig crossbreeding scheme (Xiang *et al.* 2016a). Single-step GBLUP with metafounders performed as well as the single-step GBLUP requiring phasing in terms of prediction accuracy, but with a lower complexity.

Furthermore, based on the model of Wei and van der Werf (1994), Tusell *et al.* (2016) proposed a bivariate single-step model to evaluate two-way cross pigs with their respective sire line. In this model, the additive genetic effect of the CB individuals is decomposed into its sire and dam additive genetic effects. This allows estimating the genetic correlation between the CB sire genetic effect and the PB genetic animal and does not requires CB genotypes (which can be cost-saving and easier to implement). They reported higher accuracy of BV for CB performance compared to univariate single-step models considering either PB or CB animals as training, and also higher than those obtained in pedigree-based models.

1.6.1.2 Crossbred models with additive and non-additive genetic effects

Dominance is considered the main cause of heterosis and has special importance in CB populations. Thereby, different genomic models for crossbred performance have been proposed to include dominance effects. The inclusion of dominance in a genomic model led to an extra response in CB performance (Zeng *et al.* 2013; Esfandyari *et al.* 2015). However, these models are univariate models and preclude the inclusion of PB and CB performances, to account for r_{PC}. Further, it is not possible to account for different patterns of LD between SNP and QTL in parental breeds. To overcome these issues, a trivariate GBLUP model that includes PB and CB data and dominance genetic effects was proposed by Vitezica *et al.* (2016). In their model, PB and CB performance were treated as different but genetically correlated traits and additive and dominance effects were correlated across PB and CB populations. Xiang *et al.* (2016a) used Vitezica *et al.*'s model, but including genomic inbreeding (observed homozygosity) as a covariate in the model for PB and CB animals to account for potential directional dominance. More recently, Christensen *et al.* (2019) extended the model of Vitezica *et al.* (2016) and Xiang *et al.* (2016a) for the case of a purebred sire line and three-way crossbred pigs. They showed that the proposed genomic model may be useful for the implementation of selection in a three-way crossbreeding system.

Although the methodology is available for accounting simultaneously for all these aspects in genomic selection, the benefit of its implementation in practice in a crossbreeding scheme has not been evaluated yet.

1.6.2 Genomic selection in maize hybrids

After Meuwissen *et al.* (2001) defined the foundations of GS, plant breeders were interested in using this approach for the prediction of hybrid performance. Recently, several studies explored the prospects of genomic prediction of maize hybrids derived from inbred lines from two different heterotic groups (e.g. Maenhout *et al.* 2010; Technow *et al.* 2014; Kadam *et al.* 2016). In all these studies, the authors estimated the covariance (or kinship) for GCA and SCA effects based on the genetic covariance between hybrids presented by Stuber and Cockerham (1966) assuming allele effects according to the parental origin, as was done by Bernardo (1994).

Stuber and Cockerham (1966) defined the covariance between hybrids under infinitesimal assumptions for a two-locus genetic model, as follows: assume i and i' are two inbred lines from heterotic group 1, j and j' are two inbred lines from heterotic group 2. Lines between heterotic groups are unrelated by pedigree (Figure 1.5). The genetic covariance between single-cross hybrids (ij) and ($i'j'$) is:

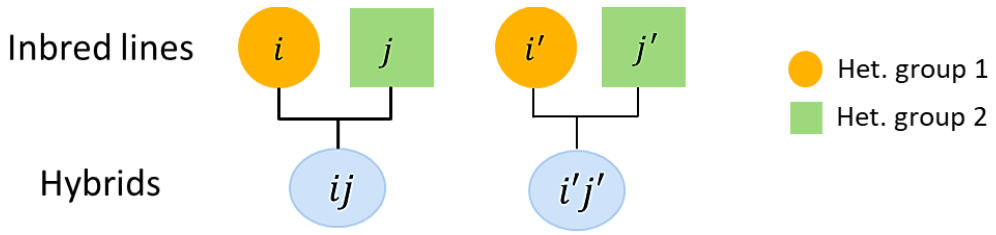

Figure 1.5 Pedigree relationship between hybrids.

$$Cov[(ij),(i'j')] = f_{ii'}V_{A^{(1)}} + f_{jj'}V_{A^{(2)}} + f_{ii'}f_{jj'}V_D + f_{ii'}{}^2V_{AA^{(1)}} + f_{jj'}{}^2V_{AA^{(2)}} + f_{ii'}f_{jj'}V_{AA^{(12)}}$$

where $f_{ii'}$ is the coefficient of coancestry between inbred lines i and i' and $f_{jj'}$ the coefficient of coancestry between inbred lines j and j'. $V_{A^{(1)}}$ and $V_{A^{(2)}}$ are the respective additive variance

in heterotic groups 1 and 2, based on alleles effects of lines from one heterotic group crossed to lines from the opposite heterotic group. V_D is the dominance deviation variance of pairs of parental inbred lines. $V_{AA^{(1)}}$ and $V_{AA^{(2)}}$ are the within-line additive-by-additive epistatic variance across loci for heterotic groups 1 and 2, respectively. $V_{AA^{(12)}}$ is the additive-by-additive epistatic variance across loci in alleles coming from different heterotic groups or "across population". However, in a pedigree context, the V_D and $V_{AA^{(12)}}$ have the same coefficient $f_{ii'}f_{jj'}$, thereby both variances are confounded into SCA variance, $V_{SCA} = V_D + V_{AA^{(12)}}$ and cannot be estimated separately (Bernardo 1995, 1996). Notice that all the epistatic components are involved in the concepts of GCA and SCA. Genetically, the GCA variance is due to the additive genetic effects and additive x additive gene interaction, whereas the SCA variance is due to dominance deviation effects, additive x dominance, dominance x dominance gene interactions (Griffing 1956). However, epistatic effects are often ignored because they are difficult to be estimated and its inclusion in the models does not produce a significant improvement in prediction accuracy (Bernardo 1995). Thus, the covariance between hybrids becomes more simple if epistasis is ignored (Bernardo 2010):

$$Cov[(ij),(i'j')] = f_{ii'}V_{A^{(1)}} + f_{jj'}V_{A^{(2)}} + f_{ii'}f_{jj'}V_{SCA}$$

Under an infinitesimal model, the covariance of SCA effects is the product $f_{ii'}f_{jj'}$, which are obtained directly by multiplying the corresponding elements in the additive relationship matrix for specific hybrids. This idea was transposed to marker-based models and has been widely spread in the literature of genomic-based prediction of hybrid performance (Massman *et al.* 2013; Technow *et al.* 2014; Bouvet *et al.* 2016; Kadam *et al.* 2016; Acosta-Pech *et al.* 2017). With SNP genotypes, it is not clear if it is possible to distinguish between relationships within locus (i.e. dominance) and across loci (i.e. epistasis). So far, how to model the non-additive genetic effects in a genomic hybrid prediction model is still not clear.

The previous model assumes allele effects according to origin to estimate GCAs within groups for the selection of inbred lines. Stuber and Cockerham (1966) also defined a model where allele effects are defined "uniquely" within hybrids, regardless of the origin, like a single population (i.e. a convergent population in maize). Several genomic models have been used following this approach based on models previously used in animal breeding. For instance, Fritsche-Neto *et al.* (2018) and Alves *et al.* (2019) used the genotypic model from Su *et al.* (2012) to model dominance effects. Other authors have implemented a genomic model with a

dominance relationship matrix as in Vitezica *et al.* (2013) in hybrid breeding crops like maize, sorghum and rice (e.g. Dias *et al.* 2018; Cui *et al.* 2020; Hunt *et al.* 2020). However, in all these studies, genomic models have been transposed from animal breeding without considering the particularities of hybrid crops, especially the fact that a hybrid population is not in Hardy-Weinberg Equilibrium (HWE).

Based on this short review, it can be noticed that there is a need to revisit the theory of quantitative genetics for genomic selection in hybrids. This could allow the development of models adapted to the particularities of hybrid crops and the physical nature of the genome (finite and not infinitesimal), better understood, and potentially more accurate.

1.7 Mate allocation

The non-additive part of the genetic merit of an individual is due to the interaction of the genes (dominance and epistasis) inherited from its parents. Non-additive effects can be used to produce specific matings or mate allocation. In animal breeding, mate allocation (or planed matings) is an important component in every breeding program. It has been used mainly to control inbreeding by using pedigree information to avoid matings between animals with a common ancestor (VanRaden and Smith 1999; Kinghorn 2011). Mate allocation can also be used to increase connectedness among herds, make culling decisions, preserve genetic diversity and to profit from non-additive genetic effects (Toro and Varona 2010; Kinghorn 2011). The latter is of interest because mate allocation can be a potential strategy to maximize the total genetic merit of future progeny by exploiting dominance variation either within-breed or/and across-breeds (DeStefano and Hoeschele 1992; Hayes and Miller 2000; Toro and Varona 2010). A limitation is that the dominance-specific advantage obtained from a specific mating cannot be accumulated in subsequent generations because it is not inherited (Toro and Varona 2010). Therefore, mate allocation has to be applied in each generation to profit from dominance. Its interest is to boost performance for sex-limited traits in PB lines, for example increasing litter size in pigs to have more selection candidates, or in CB to enhance the overall performance of CB females or final marketed products. Mate allocation implementation could change the organization of crossbreeding schemes.

A few simulation studies demonstrated the effectiveness of mate allocation to maximize the total genetic merit of future progeny by exploiting dominance in a single population (DeStefano and Hoeschele 1992; Toro 1993; Kinghorn 1998) and across-breed populations (Hayes and Miller 2000). However, pedigree-based models were used to predict BV and accurate estimates of dominance effects were not straightforward to obtain.

With the advent of high-density SNP markers, estimates of additive and dominance SNP effects can be obtained and easily used in mate allocation. Mate allocation strategies accounting for non-additive genetic effects have been developed by Toro and Varona (2010). A simulation study has provided encouraging results to maximize the expected total genetic merit of future progeny in within-breed populations (Toro and Varona 2010). Similar results were obtained in real data in dairy cattle (Sun *et al.* 2013; Ertl *et al.* 2014; Aliloo *et al.* 2017). They reported an additional increase in the performance of expected progeny by implementing mate allocation ranging from 10 to 28% compared to random mating. In pigs, genomic estimates of dominance variance in purebred and crossbred populations, expressed as proportion of phenotypic variance, ranged from 2 to 18%, depending on the trait (Vitezica *et al.* 2016; Xiang *et al.* 2016a; Tusell *et al.* 2019). Thus, mate allocation could take advantage of non-additive genetic effects either within-breed (e.g. in a purebred population) and/or across-breed (e.g. crosses in a 2-way crossbreeding scheme). However, the benefits of genomic mate allocation strategies accounting for non-additive genetic effects have not been evaluated in a pig crossbreeding scheme to maximize either PB or CB performances. In this thesis, the effectiveness of genomic mate allocation strategies exploiting non-additive genetic effects to maximize the overall genetic merit of future progeny in a Landrace pig population is addressed in Chapter 2; and in a two-way crossbreeding scheme in the long term in Chapter 3.

Besides, in hybrid breeding programs, mate allocation can play an important role in the definition of breeding crosses between parental lines to develop source populations for line development within each heterotic group (Bernardo 2003; Wegenast *et al.* 2008). Recently, a few studies have implemented mate allocation strategies considering only additive genetic effects to define parental-crosses to maximize the performance in progeny and limit diversity reduction for long-term genetic gain in an inbred line breeding program (Lehermeier *et al.* 2017; Allier *et al.* 2019). This mate allocation strategy can potentially be extended to hybrid breeding by including dominance effects to jointly improve the two heterotic groups. Another potential application of mate allocation in maize is the optimization of the training set composition to

maximize prediction accuracy of hybrids with minimum resources spent on phenotyping and genotyping (Rincent *et al.* 2012; Kadam and Lorenz 2018). Dominance may also be included in the optimization of the training set to improve prediction accuracy. Mate allocation for hybrid improvement in maize is not addressed in this thesis, but the theory to estimate non-additive genetic effects in hybrids was revisited and is presented in Chapter 4.

1.8 Objective of this thesis

The overall objective of this thesis is to explore and develop methodology to account for non-additive genetic effects on genomic selection models and mate allocation strategies to boost performance of pig and maize crossbreeding schemes.

1.9 Outline

This thesis is organized in 6 chapters. A general introduction and the objectives of the thesis are presented first (Chapter 1). The corps of the thesis is a compilation of three scientific papers (Chapter 2 to Chapter 4), followed by a general discussion, perspectives and general conclusions (Chapter 5).

The first research chapter of this thesis, Chapter 2, is focused on exploiting within-breed dominance variance through mate allocation strategies to maximize the overall genetic merit in a Landrace pig population. First, genetic variance components are estimated for the traits age at 100 kg, backfat depth at 140 days, and average piglet weight at birth within litter. Then, the efficiency of mate allocation strategies that account for dominance and inbreeding depression to maximize the total genetic value of future offspring are explored.

Chapter 3 investigates, by simulation, the effectiveness of mate allocation strategies and genomic evaluations, accounting for additive and dominance effects, to improve CB performance in a pig two-way crossbreeding scheme in the long term. Breed-specific QTL and individual genomic heterosis effects are explicitly simulated in a two-way pig crossbreeding scheme. Effects of the sources of information used in the genetic evaluation (only PB data or both PB and CB data), of several narrow- and broad-sense heritability values, and of several options for mate allocation to produce the crossbred are evaluated.

In Chapter 4, the theory of hybrid genetic evaluation models from single-cross of pure lines (e.g. in maize) is revisited in a genomic context. Covariances between hybrids due to additive substitution effects and dominance and epistatic deviations are analytically derived. In addition, the expressions of genomic relationship matrices for additive, dominance and epistasis effects in hybrid crops are developed. Then, the developed genomic relationship matrices are tested using a publicly available data set of maize hybrids, both for estimation of variance components and predictive ability.

1.10 Appendices

Appendix 1.1

Considering genomic inbreeding

If there is directional dominance (i.e. a higher percentage of positive than negative dominance effects), inbreeding depression (or heterosis) can occur and it should be considered in the genetic evaluation model (de Boer and Hoeschele 1993; Xiang *et al.* 2016a). Otherwise, estimates of dominance variance are inflated (de Boer and Hoeschele 1993; Aliloo *et al.* 2017).

A model including (biological) additive and dominant effects of the SNPs can be written in matrix form for a set of individuals as (Toro and Varona 2010):

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

where $\boldsymbol{y}$ is the phenotypic value of the individuals, μ is the population mean and $\boldsymbol{a}$ and $\boldsymbol{d}$ are the biological additive and dominant effects of the SNP markers. Typically, genetic models assume that $\boldsymbol{a}$ and $\boldsymbol{d}$ have zero means, which is not true for $\boldsymbol{d}$ when directional dominance exists. As,

$$E(\boldsymbol{d}) \neq 0$$

We can define $\boldsymbol{d}^* = \boldsymbol{d} - E(\boldsymbol{d}) = \boldsymbol{d} - \mu_d$ and write an equivalent genomic model as:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Za} + \boldsymbol{Wd}^* + \boldsymbol{W1}\mu_d + \boldsymbol{e} \qquad (1)$$

Xiang *et al.* (2016a) show that inclusion of inbreeding (e.g. based on SNPs) as a covariate accounts for directional dominance and inbreeding depression. The term $\boldsymbol{W1}$ defined as $\boldsymbol{h} = \boldsymbol{W1}$ contains the row sums of $\boldsymbol{W}$, i.e. individual heterozygosities (note that $\boldsymbol{W}$ has a value of 1 at heterozygous loci for an individual).

The genomic inbreeding coefficient $\boldsymbol{f}$ can be calculated as: $\boldsymbol{f} = \mathbf{1} - \boldsymbol{h}/N$, where N is the number of SNP loci. For instance, $\boldsymbol{f}$ is a vector that contains the percentage of homozygous loci for each individual. Then,

$$\boldsymbol{h} = (\mathbf{1} - \boldsymbol{f})N = \mathbf{1}N + \boldsymbol{f}(-N)$$

and with the mean μ_d

$$\boldsymbol{h}\mu_d = (\mathbf{1} - \boldsymbol{f})N\mu_d = \mathbf{1}N\mu_d + \boldsymbol{f}(-N\mu_d)$$

Thus, the model (1) can be rewritten as

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Za} + \boldsymbol{Wd}^* + \mathbf{1}N\mu_d + \boldsymbol{f}(-N\mu_d) + \boldsymbol{e}$$

and

$$\boldsymbol{y} = \mathbf{1}\mu^* + \boldsymbol{Za} + \boldsymbol{Wd}^* + \boldsymbol{f}b + \boldsymbol{e} \qquad (2)$$

where the term $\mathbf{1}N\mu_d$ is confounded with the overall mean of the model (μ^*), while the $\boldsymbol{f}(-N\mu_d)$ models the inbreeding depression and $b = (-N\mu_d)$ is the inbreeding depression summed over the SNP loci, which has to be estimated.

Appendix 1.2

Additive substitution effects for crossbred performance

F1 crossbred individuals involve gametes from two parental populations (1 and 2). In the presence of dominance gene action and because allelic frequencies differ in each parental population, the (additive) substitution effects within a breed are not equivalent to the substitution effects across the F1 population. Therefore, different BVs can be derived for PB individuals: one in the PB scale to be used when mated to individuals from the same population (i.e. to improve PB performance), and one in the CB scale when mated to a different population to generate CB offspring (to improve CB performance). The BV on the PB scale is based on the classical additive substitution effect (Falconer 1981). Instead, the BV on the CB scale is based on additive substitution effects defined in terms of the allele frequencies from the other parental breed. As a result, the allele substitution effects are breed-specific. The genetic variance for an F1 CB population from two non-inbred populations within a genomic framework was presented (Reif *et al.* 2007; Vitezica *et al.* 2016). The effects of alleles were assumed to be the same in the two parental populations and the CB population. The derivations for the allele substitution effects for BV on the CB scale are presented below.

Consider two non-inbred parental populations, 1 and 2, that are in Hardy-Weinberg equilibrium. Then, consider a biallelic single locus/gene with alleles *B/b* and the allele origin denoted as *i*=1 and *j*=2; thus population 1 has B_1 and b_1 with respective frequencies p_1 and $q_1 = 1 - p_1$ and population 2 has B_2 and b_2 with frequencies p_2 and q_2. The F1 CB population has genotypes (frequencies) B_1B_2 (p_1p_2), B_1b_2 (p_1q_2), b_1B_2 (q_1p_2) and b_1b_2 (q_1q_2).

Assuming that additive and dominant gene actions are present. Then, the genotypic value G of an F1 crossbred can be written (up to a common constant) as:

$$G_{B_1B_2} = a \qquad G_{B_1b_2} \text{ and } G_{b_1B_2} = d \qquad \text{and } G_{b_1b_2} = -a$$

Where a is the functional additive effect and is the same in the two parental populations and d is the functional value of both heterozygotes (B_1b_2 and b_1B_2). The genetic mean of the CB population is therefore

$$E(G) = (p_1p_2)a + (p_1q_2)d + (q_1p_2)d + (q_1q_2)(-a)$$
$$E(G) = (p_1p_2 - q_1q_2)a + (p_1q_2 + q_1p_2)d$$

If the difference in allele frequencies between the two parental populations is denoted by $y = p_1 - p_2 = q_2 - q_1$, the expression is, as in Falconer (1981, equation 14.6), equal to:

$$E(G) = (p_1 - q_1 - y)a + (2p_1q_1 + y(p_2 - q_2))d$$

The *average effect of alleles* (α_{B_1}, α_{b_1}, α_{B_2} and α_{b_2}) are derived from the genotypic values. Following the reasoning from Table A 1.1 (as in Table 7.2 in Falconer (1981)). If gametes carrying B_1 from population 1 are mated at random with gametes from population 2, the frequencies of the genotypes produced will be p_2 of B_1B_2 and q_2 of B_1b_2. The genotypic value of a hybrid B_1B_2 is $G_{B_1B_2} = a$, that of B_1b_2 is $G_{B_1b_2} = d$, and the average value of these two, taking into account of the proportions in which they occur is $E(G|B_1) = p_2(a) + q_2(d)$. The reasoning is the same for other alleles (Table A 1.1).

Table A 1.1 Average effect of a gene

	Values and frequencies of genotypes produced			
Type of gamete	B_1B_2 a	B_1b_2 d	b_1B_2 d	b_1b_2 $-a$
B_1	p_2	q_2		
b_1			p_2	q_2
B_2	p_1		q_1	
b_2		p_1		q_1

For instance, considering allele B_1, the difference between the mean value conditional on a particular genotype of the gamete (*e.g.* $E(G|B_1)$) and the population mean ($E(G)$) is the average effect of the allele B_1, α_{B_1}. The average effects of alleles (α_{B_1}, α_{b_1}, α_{B_2} and α_{b_2}) are derived from the genotypic values and following this table.

$$\alpha_{B_1} = p_2a + q_2d - E(G) = q_1[a + (q_2 - p_2)d]$$
$$\alpha_{b_1} = p_2d - q_2a - E(G) = -p_1[a + (q_2 - p_2)d]$$
$$\alpha_{B_2} = p_1a + q_1d - E(G) = q_2[a + (q_1 - p_1)d]$$
$$\alpha_{b_2} = p_1d - q_1a - E(G) = -p_2[a + (q_1 - p_1)d]$$

Then, the *average effect of the allele-substitution* can be derived. For instance, letting b_1 be substituted by B_1. From the b_1 alleles taken at random from the population for substitution, a proportion p_2 will be found in b_1B_2 genotypes and a proportion q_2 in b_1b_2. The substitution will respectively change the value from d to a and from $-a$ to d (see Figure A 1.6).

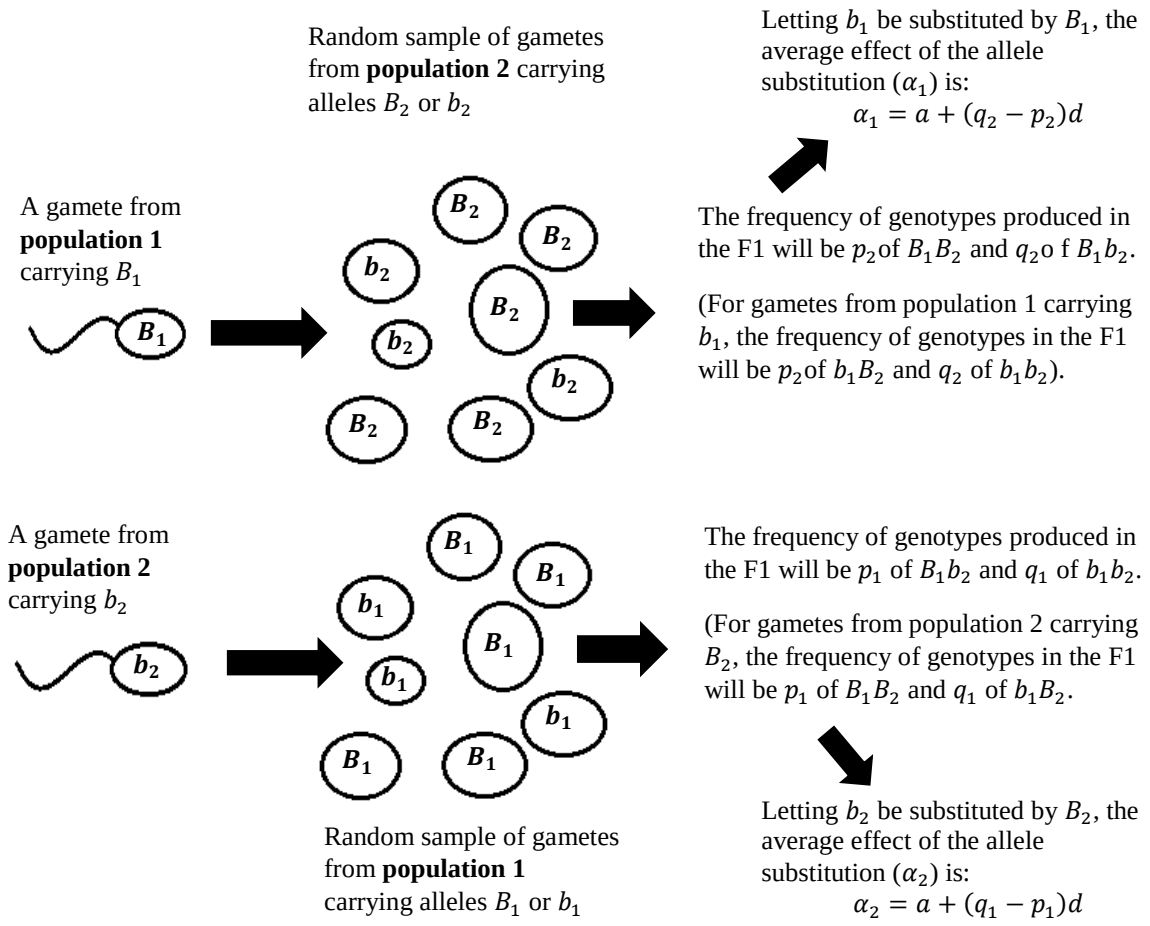

Figure A 1.6 Schematic representation of the average effect of allele-substitution for a single locus. B_1, b_1, (p_1, q_1) and B_2, b_2 (p_2, q_2) correspond to alleles (and their frequencies) from parental populations 1 and 2, respectively.

Thus, the average effect of the allele-substitution (α_1) of population 1 is

$$\alpha_1 = p_2(a - d) + q_2(d - (-a)) = a + (q_2 - p_2)d$$

The same result can be obtained as the difference between the average effects of alleles:

$$\alpha_1 = \alpha_{B_1} - \alpha_{b_1} = p_2 a + q_2 d - E(G) - [p_2 d - q_2 a - E(G)] = a + (q_2 - p_2)d$$

Note that the term $E(G)$ is canceled. For population 2 (α_2), it is

$$\alpha_2 = \alpha_{B_2} - \alpha_{b_2} = a + (q_1 - p_1)d$$

Note that allele-substitution effects involve both functional additive (a) and dominance (d) effects, and allele frequencies of the other parental population

Chapter 2

SNP-based mate allocation strategies to maximize total genetic value in pigs

David González-Diéguez[1*], Llibertat Tusell[1], Céline Carillier-Jacquin[1], Alban Bouquet[2,3], Zulma G Vitezica[1]

[1]GenPhySE, INRA, Université de Toulouse, Castanet-Tolosan 31326, France

[2]IFIP Institut du Porc, BP35104, 35651 Le Rheu, France

[3]France Génétique Porc, BP35104, 35651 Le Rheu, France

Genetic Selection Evolution (2019) 51:55

González-Diéguez *et al. Genet Sel Evol* (2019) 51:55
https://doi.org/10.1186/s12711-019-0498-y

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

SNP-based mate allocation strategies to maximize total genetic value in pigs

David González-Diéguez[1*], Llibertat Tusell[1], Céline Carillier-Jacquin[1], Alban Bouquet[2,3] and Zulma G. Vitezica[1]

Abstract

Background: Mate allocation strategies that account for non-additive genetic effects can be used to maximize the overall genetic merit of future offspring. Accounting for dominance effects in genetic evaluations is easier in a genomic context, than in a classical pedigree-based context because the combinations of alleles at loci are known. The objective of our study was two-fold. First, dominance variance components were estimated for age at 100 kg (AGE), backfat depth (BD) at 140 days, and for average piglet weight at birth within litter (APWL). Second, the efficiency of mate allocation strategies that account for dominance and inbreeding depression to maximize the overall genetic merit of future offspring was explored.

Results: Genetic variance components were estimated using genomic models that included inbreeding depression with and without non-additive genetic effects (dominance). Models that included dominance effects did not fit the data better than the genomic additive model. Estimates of dominance variances, expressed as a percentage of additive genetic variance, were 20, 11, and 12% for AGE, BD, and APWL, respectively. Estimates of additive and dominance single nucleotide polymorphism effects were retrieved from the genetic variance component estimates and used to predict the outcome of matings in terms of total genetic and breeding values. Maximizing total genetic values instead of breeding values in matings gave the progeny an average advantage of −0.79 days, −0.04 mm, and 11.3 g for AGE, BD and APWL, respectively, but slightly reduced the expected additive genetic gain, e.g. by 1.8% for AGE.

Conclusions: Genomic mate allocation accounting for non-additive genetic effects is a feasible and potential strategy to improve the performance of the offspring without dramatically compromising additive genetic gain.

Background

Mate allocation has been used in animal breeding schemes mainly to control inbreeding but also to increase connectedness among herds, to preserve genetic diversity, and to exploit dominance [1–4]. The total genetic merit of future progeny can be maximized by selecting pairs of mates to better exploit dominance, especially in crossbreeding programs (across breeds), but also within a population [5, 6].

In the classical pedigree context, obtaining accurate estimates of dominance effects is difficult because it requires large groups of individuals with dominance relationships (the probability of identical genotypes of two individuals at a locus), such as full-sibs, and also because estimations are computationally demanding [7, 8]. With the availability of high-density single nucleotide polymorphism (SNP) panels and large numbers of genotyped animals, genomic selection has improved the efficiency of animal breeding programs, essentially by reducing generation intervals and improving prediction accuracies of breeding values of difficult-to-measure and lowly heritable traits [9].

Accounting for dominance effects in a GBLUP (genomic best linear unbiased prediction) framework is easier than in a classical pedigree-based context because, instead of probabilities of identical genotypes, we have observations of heterozygous states at SNPs across the genome [10]. Thus, the availability of genomic information has renewed interest in non-additive genetic effects. Genomic models that include dominance have been implemented in several livestock populations, including dairy cattle, pigs, sheep, and layer chickens [11–17].

*Correspondence: david-omar.gonzalez-dieguez@inra.fr
[1] GenPhySE, INRA, Université de Toulouse, 31326 Castanet-Tolosan, France
Full list of author information is available at the end of the article

BMC

© The Author(s) 2019. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

The inclusion of dominance in genetic evaluation models has allowed the quantification of the magnitude of non-additive genetic effects and determination of the genetic structure of traits. In general, inclusion of dominance effects in genomic prediction models did not increase the accuracy of estimates of breeding values [18] but may be important for implementing mate allocation strategies to exploit non-additive genetic effects [2, 11, 19].

Mate allocation strategies are based on the idea that, although selection should be based on estimated breeding values (heritable effects), the animals used for commercial purposes can be the product of planned matings that maximize the total genetic merit (additive plus dominance effects) of the offspring [2]. Mate allocation accounting for dominance increased progeny performance in one simulated population [2] and when based on dairy cattle data [11, 19].

In pig breeding schemes, although the goal is to improve crossbreed performance (for commercial production), selection is traditionally performed within purebreds [20, 21]. Pig breeding schemes could benefit from mate allocation strategies and secure additional profit from across- and within-breed dominance variation [6]. In pigs, estimates of dominance variance based on pedigree-based estimation ranged from 11% of the additive genetic variance for backfat to 78% for litter weight at 21 days [22]. The large magnitude of these estimates suggests a potential gain in total genetic merit if matings are allocated to exploit dominance.

The aim of this study was twofold. First, the dominance variance for age at 100 kg, backfat depth at 140 days, and average piglet weight at birth within litter was estimated in Landrace pigs. Second, mate allocation strategies that focus on maximizing either the average breeding value or the average total genetic value for prospective progeny were compared.

Methods

Phenotypes and genotypes

Data were provided by a collective French Landrace pig program led by the breeding companies NUCLEUS (Le Rheu, France) and AXIOM (Azay-sur-Indre, France). Animal Care and Use Committee approval was not requested for this study because the data used were routinely collected data according to standard practices in breeding herds. Performances recorded in their selection herds and test stations are centralized to perform common genetic evaluations. The main selection objective of the collective program is the improvement of reproductive criteria. However, improving growth and feed efficiency traits without impairing carcass composition and meat quality is also considered in the breeding goal [23]. The Landrace breed is renowned for its high prolificacy, excellent maternal instinct, and milking abilities. It has also good rusticity, fattening performance, and meat quality and carcass composition characteristics [24]. Data on age at 100 kg (AGE), backfat depth (BD) at 140 days, and average piglet weight at birth within litter (APWL) were analyzed. Performances were recorded in breeding herds at around 100 kg live weight for AGE and BD. Backfat depth was measured ultrasonically at the shoulder, last rib and hip joint on each side of the animal (4 cm from the mid-dorsal spine line) and an average BD was calculated for each animal using these six measurements. AGE and BD were adjusted to a 100 kg live weight; AGE was adjusted to 100 kg based on individual age (age_{test}) and weight (w_{test}) at the test, and mean age (age_{batch}) and weight (w_{batch}) of the batch as: $AGE = age_{test} + [r - 0.0077 w_{batch} + 0.0047 * age_{batch}] * (100 - w_{test})$, with r equal to 1.05 and 1.125 for males and females, respectively; BD was adjusted to 100 kg as: $BD = averageBD + r * (100 - w_{test})$, with r equal to 0.1 and 0.12 for males and females, respectively. APWL was considered as a sow trait and was the average of individual birth weights of piglets. It was recorded for each litter, with an average of 2.7 parities per sow. Only Landrace animals with genotypes were included in the analysis. Average performance and number of animals and records used for analyses of the three traits are in Table 1.

Pedigrees were extracted from the collective French Landrace pig program and included all available ancestors of the genotyped animals in the datasets. Three generations were traced back from the pedigree for each trait, resulting in 5534, 6541, and 3331 animals for AGE, BD and APWL, respectively. Two medium-density panels were used for genotyping: the porcine SNP60 Illumina BeadChip (Illumina, San Diego, CA) and the GeneSeek Genomic Profiler HD 80 k (GeneSeek, Licoln, NE), on 98% and 2% of the genotyped animals, respectively. SNPs with a call rate lower than 0.98, a minor allele frequency lower than 0.05, and deviating from Hardy–Weinberg equilibrium (P value < 0.05) were removed. Six animals with a call rate lower than 0.98 were discarded. Three offspring that displayed Mendelian inconsistencies with

Table 1 Numbers of animals, records, and means (SD) for the three evaluated traits

Trait	Number of animals		Number of records	Mean
	Males	Females		
AGE (days)	789	2179	2968	149.0 (9.4)
BD (mm)	1007	2675	3682	11.2 (1.7)
APWL (g)	1446	1226	3297[a]	1322 (213)

AGE age at 100 kg, *BD* backfat depth, *APWL* average piglet weight at birth within litter

[a] Only for females as APWL is a maternal trait with repeated measurements

their parents were removed. After quality control, 39,353 SNPs that were in common between the two panels were retained for analyses and used to build the genomic relationship matrices.

Estimation of variance components

Three linear mixed models (one pedigree-based and two genomic-based models) were used to estimate variance components for each trait. The first model (model A) was a classical additive genetic model in which the genetic relationship matrix was computed based on pedigree information from genotyped animals (a sample of the French Landrace population). The genomic models included either additive genetic effects (model G) or additive and dominance genetic effects (model GD). The two genomic models included a genomic inbreeding depression parameter and were implemented using a GBLUP framework. Fixed effects for AGE were the combination of farm-year-sex (68 levels) and the covariate of birth weight. Fixed effects for BD included the combination of farm-year-sex (90 levels), parity of the dam (6 levels), and litter size as a covariate. The APWL trait included farrowing batch (283 levels) and parity of the dam (6 levels) as fixed effects and an animal permanent environmental random effect (1226 sows).

In matrix notation, model G can be represented as follows:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{f}b + \mathbf{Zu} + \mathbf{Zpe} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of phenotypes of a trait; $\boldsymbol{\beta}$ is the vector of fixed effects; $\mathbf{f}$ is the vector of genomic inbreeding coefficients, calculated as the proportion of homozygous SNP genotypes for each animal, and b is the inbreeding depression parameter; $\mathbf{u}$ is the vector of additive genetic effects (i.e. breeding values); $\mathbf{pe}$ is the vector of random permanent environmental effects specific to APWL $\left(\mathbf{pe} \sim N\left(0, \mathbf{I}\sigma^2_{pe}\right)\right)$; and $\mathbf{e}$ is the vector of residual effects $\left(\mathbf{e} \sim N\left(0, \mathbf{I}\sigma^2_e\right)\right)$. Breeding values were assumed distributed as $\mathbf{u} \sim N\left(0, \mathbf{G}\sigma^2_A\right)$, where $\mathbf{G}$ is the additive genomic relationship matrix (model G). The incidence matrix $\mathbf{X}$ relates observations to fixed effects, and $\mathbf{Z}$ is the incidence matrix for breeding values and permanent environmental random effects. Model A did not include the inbreeding depression parameter and breeding values were assumed distributed as $\mathbf{u} \sim N\left(0, \mathbf{A}\sigma^2_A\right)$, where $\mathbf{A}$ is the additive pedigree-based relationship matrix. Parameters σ^2_{pe}, σ^2_A, and σ^2_e refer to permanent environmental, additive genetic, and residual variances, respectively.

The additive genomic relationship matrix was calculated according to VanRaden [25] as:

$$\mathbf{G} = \frac{\mathbf{MM}'}{2\sum_{k=1}^{m} p_k q_k},$$

where $\mathbf{M}$ is a matrix with dimensions of number of animals (n) by number of SNPs (m), with elements equal to $(2 - 2p_k)$, $(1 - 2p_k)$ and $-2p_k$, for genotypes AA, Aa and aa respectively; p_k is the frequency of allele A of the kth SNP (with $k = 1, \ldots, m$), and $q_k = 1 - p_k$. Elements of $\mathbf{M}$ for missing SNP genotypes were set to the average of the population.

Model GD is an expansion of model G, which takes dominance genetic effects into account:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{f}b + \mathbf{Zu} + \mathbf{Zv} + \mathbf{Zpe} + \mathbf{e},$$

where $\mathbf{v}$ is the vector of random dominance deviation effects, assumed distributed as $\mathbf{v} \sim N\left(0, \mathbf{D}\sigma^2_D\right)$, where σ^2_D is the variance of dominance deviations. The dominance genomic relationship matrix $\mathbf{D}$ was built as in Vitezica et al. [10]:

$$\mathbf{D} = \frac{\mathbf{WW}'}{\sum_{k=1}^{m} (2p_k q_k)^2},$$

where $\mathbf{W}$ has the same dimension as $\mathbf{M}$, with elements equal to $-2q_k^2$, $2p_k q_k$ and $-2p_k^2$ for the AA, Aa and aa genotypes, respectively. Elements of $\mathbf{W}$ for missing SNP genotypes were set to the average of the population. Matrices $\mathbf{G}$ and $\mathbf{D}$ were scaled in order to improve numerical stability as $\mathbf{G}^* = 0.95\mathbf{G} + 0.05\mathbf{I}$, and $\mathbf{D}^* = 0.95\mathbf{D} + 0.05\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Genomic inbreeding was used to account for directional dominance [16, 26] by including it as a covariate in the model to obtain the correct estimate of dominance variance [16, 19].

Variance components for the three models described above were estimated by EM-REML with the *remlf90* software [27], which is available at http://nce.ads.uga.edu/wiki/doku.php. Asymptotic standard errors of variance components and variance proportion estimates were obtained as in Houle and Meyer [28].

The three models were compared using the Akaike Information Criterion (AIC) [29]. The model with the minimum AIC value was considered as the best model to fit the data.

Estimation of additive and dominance SNP effects

Marker effects were estimated by using a BLUP-SNP model that assumes that the variance components estimated above are known parameters (*GS3* software available at http://snp.toulouse.inra.fr/~alegarra/, Legarra et al. [30]). This marker-level model (in contrast to the previous animal-level models) is represented as follows:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{f}b + \mathbf{Ka} + \mathbf{Td}^* + \mathbf{Zpe} + \mathbf{e},$$

where **a** and **d*** are random vectors of additive and dominance marker effects, respectively, and all other terms are as described for the animal-level models. Xiang et al. [16] proved analytically that inclusion of genomic inbreeding

in genetic models accounts for directional dominance and inbreeding depression. The vector of dominance marker effects was defined as $\mathbf{d}^* = \mathbf{d} - E(\mathbf{d}) = \mathbf{d} - \mu_d$, where $\mu_d = -\frac{b}{m}$, which makes $E(\mathbf{d}^*) = 0$. Incidence matrices $\mathbf{K}$ and $\mathbf{T}$ relate data to additive and dominance marker effects, respectively, and they are coded as 1, 0 and -1 for additive effects and as 0, 1 and 0 for dominance effects of the three SNP genotypes *AA*, *Aa* and *aa*, respectively. Variances of the vectors of additive and dominance effects are $V(\mathbf{a}) = \mathbf{I}\sigma_a^2$ and $V(\mathbf{d}) = \mathbf{I}\sigma_d^2$, where $\mathbf{I}$ is the identity matrix, and σ_a^2 and σ_d^2 are the variances of additive and dominance SNP effects, respectively. Variance components at the marker level (σ_a^2 and σ_d^2) were obtained from genetic variance components [11] that were estimated by EM-REML with model GD (σ_A^2 and σ_D^2), as follows:

$$\sigma_d^2 = \frac{\sigma_D^2}{\sum\left(4p_k^2q_k^2\right)},$$

$$\sigma_a^2 = \frac{\sigma_A^2 - \sum\left[2p_kq_k(q_k - p_k)^2\right]\sigma_d^2}{\sum(2p_kq_k)}.$$

After estimating the SNP effects **a** and **d***, estimates of the marker dominance effects were recovered as $\mathbf{d} = \mathbf{d^*} - \frac{b}{m}$, where b is the inbreeding depression parameter estimated in this model and m is the number of SNPs [16].

Prediction of progeny total genetic and breeding values

From the estimates of additive and dominance SNP effects, the total genetic value (g_{ij}) of the progeny from a mating between the ith boar and jth sow, was predicted following Toro and Varona [2] as:

$$\widehat{g}_{ij} = \sum_k \left[P_{ijk}(AA)\widehat{a}_k + P_{ijk}(Aa)\widehat{d}_k + P_{ijk}(aa)(-\widehat{a}_k)\right],$$

where $P_{ijk}(AA)$, $P_{ijk}(Aa)$ and $P_{ijk}(aa)$ are the probabilities of genotypes *AA*, *Aa* and *aa* from the progeny of the mating between the ith and the jth individuals at the kth SNP, and $\widehat{a}_k$ and $\widehat{d}_k$ are the additive and dominance estimated effects of the kth SNP.

Analogously, the breeding value (u_{ij}) of the progeny of the same mating was predicted as:

$$\widehat{u}_{ij} = \sum_k \left[P_{ijk}(AA)(2-2p_k)\widehat{\alpha}_k + P_{ijk}(Aa)(1-2p_k)\widehat{\alpha}_k + P_{ijk}(aa)(-2p_k)\widehat{\alpha}_k\right],$$

where $\widehat{\alpha}_k$ is the allele substitution effect for the kth SNP, calculated as $\widehat{\alpha}_k = \widehat{a}_k + \widehat{d}_k(q_k - p_k)$ [31]. Missing genotypes were ignored and did not contribute to the prediction of progeny values.

Allocation of matings

The total genetic and the breeding values of the progeny were computed for all potential matings between 120 boars, representing the number of boars that are selected per year, and all available genotyped sows for each trait. In total, 789, 1007 and 1446 males were available for AGE, BD, and APWL (Table 1), respectively, of which the best 120 were selected based on their genomic estimated breeding value (GEBV) based on the estimated SNP effects. For example, for BD, the total number of possible matings was 321,000 (120 boars × 2675 sows). From all possible combinations, the best set of 600 matings was selected based either on the expected additive breeding value of the progeny $(\widehat{u})$ to maximize additive genetic gain of the expected progeny, equivalent to the traditional selection program for genetic improvement, or on the expected total genetic value of the progeny $(\widehat{g})$ to maximize total genetic gain of the expected progeny. The task of maximizing additive gain or total genetic gain of progeny was addressed via linear programming [32] using the R [33] *lpsolve* package [34]. Two constraints were set for the optimization: (1) each boar could be mated to up to 15 sows and (2) each sow could not be mated to more than one boar. In this setting, optimization was done for both boars and sows. The total number of sows was always equal to the number of selected matings (600), but the number of boars and their contribution to the progeny could differ depending on the mate allocation strategy. As a result, the objective function for the linear program, for instance for the total genetic value $(\widehat{g}_{ij})$, was defined as:

$$f_{optim}\left(\widehat{g}_{ij}\right) = \sum_{i=1}^{nb}\sum_{j=1}^{ns}\widehat{g}_{ij}x_{ij},$$

with constraint $x_{i1} + x_{i2} + x_{i3} + \cdots + x_{is} = 15 (i = 1, 2, \ldots, nb)$ for boar i and: $x_{1j} + x_{2j} + x_{3j} + \cdots + x_{bj} \leq 1\left(j = 1, 2, \ldots, ns\right)$ for sow j, where x_{ij} are binary decision variables, with 0 representing that the mating between boar i and sow j was not selected, and 1 that the mating was selected; *nb* and *ns* correspond to the number of boars (120) and sows (2179, 2675, and 1226 for AGE, BD and APWL (Table 1), respectively). In order to compare the results for these two mate allocation strategies (i.e. based on $\widehat{u}$ or on $\widehat{g}$), differences between the mean $\widehat{u}$ (or the mean $\widehat{g}$) of selected matings and the mean $\widehat{u}$ (or the mean $\widehat{g}$) of all possible matings were calculated. These differences were called expected additive genetic gain (ΔU) and expected total genetic superiority (ΔG).

Results

Variance components and heritabilities

Table 2 shows the estimates of variance components, narrow-sense heritabilities and variance ratios that were obtained with the three models for the three traits under

investigation. Our estimates of additive and dominance variances are presented in terms of breeding values and dominance deviations [10]. For AGE and BD, estimates of additive genetic variances were similar between the two genomic models (G and GD), regardless of whether non-additive genetic effects (dominance and inbreeding) were included. Estimates of additive genetic variance obtained with model A differed to some extent from those obtained with the genomic models for AGE, and especially for BD. In all cases, standard errors of the estimates of additive variance were higher in model A than in the genomic models. Estimates of heritabilities for AGE and BD were lower when based on the genomic models compared to the pedigree-based model, but these differences were not significant. For APWL, estimates of heritability were consistent across the three models.

The proportion of phenotypic variance explained by dominance genetic effects were 5, 2, and 4% for AGE, BD and APWL, respectively. The influence of dominance genetic effects was greater for AGE than for BD and APWL. In addition, AGE also had the highest ratio of dominance genetic variance to additive genetic variance (0.20), while this ratio was similar for BD and APWL (0.11 and 0.12, respectively). When dominance was included in model GD for analyses of AGE and BD, a small reduction in residual variance was observed for both traits compared to model G. This means that most of the dominance variance was absorbed in the residual variance for these traits in model G. For APWL with model G, where a repeatability model was used, most of the dominance variance was included in the permanent environmental variance.

Effect of inbreeding

The average proportion of homozygosity or genomic inbreeding per animal was 0.62 (SD = 0.015) in this Landrace population. Estimates of inbreeding depression (b), expressed as the change in phenotypic mean per 10% increase in genomic inbreeding (standard error), were equal to 4.29 (1.17) days, −0.01 (0.18) mm, and 42.38 (32.66) g for AGE, BD, and APWL, respectively. The impact of inbreeding was negative for AGE (i.e. inbred animals required more days to reach 100 kg), negligible for BD, and positive for APWL, but not different from zero for BD and APWL.

Goodness of fit

Table 3 shows the AIC value for each model. For all traits, genomic models fitted the data better than the pedigree-based model. The two genomic models had very similar AIC values for BD and APWL. For AGE, which was the trait with the highest ratio of dominance to additive variance, model GD fitted the data slightly better than model G but this improvement can be considered irrelevant.

Prediction of total genetic values of matings

Selection in pigs aims at reducing age at 100 kg (AGE) and backfat depth (BD), and increasing the average piglet weight at birth within litter (APWL). As expected, when matings were selected on the mean breeding value $(\widehat{u})$, 40 boars and 600 sows were selected, i.e. those with the best GEBV. Thus, mate selection on $\widehat{u}$ was equivalent to randomly mating males and females that were selected by standard truncation selection on GEBV. When the matings were selected on the mean total genetic value $(\widehat{g})$, most selected sows were the same as those selected when maximizing $\widehat{u}$, with the number of sows in common between the selection strategies being equal to 560, 583, and 590 out of 600 for AGE, BD, and APWL, respectively. The number of boars selected when maximizing $\widehat{g}$ was equal to 56, 44, and 55 for AGE, BD, and APWL, respectively. This indicates that different boars were contributing when mates were selected to optimize $\widehat{g}$. The

Table 2 Variance component estimates (standard error) obtained with models A, G and GD

Trait	Model	σ_A^2	σ_D^2	σ_{pe}^2	σ_e^2	h^2	σ_D^2/σ_A^2	σ_D^2/σ_P^2
AGE (d)	A	20.92 (3.22)			41.77 (2.42)	0.33 (0.05)		
	G	16.60 (2.18)			45.83 (1.67)	0.27 (0.03)		
	GD	16.26 (2.17)	3.31 (1.72)		42.80 (2.15)	0.26 (0.03)	0.20 (0.11)	0.05 (0.03)
BD (mm)	A	0.52 (0.07)			1.34 (0.06)	0.28 (0.04)		
	G	0.36 (0.05)			1.49 (0.04)	0.19 (0.02)		
	GD	0.36 (0.05)	0.04 (0.04)		1.45 (0.06)	0.19 (0.02)	0.11 (0.11)	0.02 (0.02)
APWL (g)	A	14,503.80 (2218.07)		3665.22 (1593.56)	21,499.64 (696.63)	0.37 (0.05)		
	G	14,476.03 (1860.53)		3793.68 (1128.60)	21,558.88 (694.24)	0.36 (0.04)		
	GD	14,312.09 (1845.51)	1700.24 (1597.04)	2234.86 (1792.05)	21,562.79 (684.33)	0.36 (0.04)	0.12 (0.11)	0.04 (0.04)

AGE age at 100 kg, *BD* backfat depth, *APWL* average piglet weight at birth within litter, *A* pedigree-based model including only additive genetic effects, *G* genomic model including additive genetic effects and inbreeding, *GD* genomic model including additive and dominance genetic effects and inbreeding, σ_A^2 additive genetic variance, σ_D^2 dominance genetic variance, σ_{pe}^2 permanent environmental variance, σ_e^2 residual variance, h^2 heritability, σ_P^2 phenotypic variance

Table 3 Value of the akaike information criterion (AIC) of the fitted models for each trait

Trait	Model		
	A	G	GD
AGE	20,205.4	20,158.1	20,155.6
BD	12,491.7	12,485.8	12,486.7
APWL	40,384.1	40,310.2	40,311.0

AGE age at 100 kg, *BD* backfat depth, *APWL* average piglet weight at birth within litter, *A* pedigree-based model including only additive genetic effects, *G* genomic model including additive genetic effects and inbreeding, *GD* genomic model including additive and dominance genetic effects and inbreeding

top 40 boars selected to maximize $\widehat{u}$ were also selected when maximizing $\widehat{g}$, but they contributed differently (i.e. they were mated to different numbers of sows) and they were mated with different sows. For example, for BD, 38 boars reached the restriction of 15 sows per boar and these boars had the highest GEBV. The remaining six boars were mated to 14, 6, 6, 2, 1, and 1 sows. For AGE and APWL, 32 and 34 boars, respectively, reached the restriction of 15 sows, and these boars had the highest GEBV as well.

The mean $\widehat{u}$ (or $\widehat{g}$) for all possible matings was −2.02 (−2.03) days, −0.35 (−0.35) mm, and 78.08 (78.08) g for AGE, BD, and APWL, respectively. The difference between the mean $\widehat{u}$ (or $\widehat{g}$) of selected matings and the mean of all matings was called expected additive genetic gain (ΔU) and expected total genetic superiority (ΔG), respectively (Table 4). The expected additive genetic gain and total genetic superiority that were obtained with the selected matings are presented for each mating strategy in Table 4. Note that negative values for AGE and BD are favorable, while a positive value is favorable for APWL. For all traits, the expected total genetic superiority (ΔG in Table 4) of the progeny was clearly higher when matings were selected based on $\widehat{g}$ compared to on $\widehat{u}$, giving the offspring an advantage of −0.79 days, −0.04 mm, and 11.34 g, for AGE, BD, and APWL, respectively. These advantages amounted to 0.20, 0.07, and 0.09 genetic standard deviations (SD) for AGE, BD, and APWL, respectively. The expected additive genetic gains (ΔU in Table 4) were very similar when matings were selected on $\widehat{u}$ or on $\widehat{g}$ for BD (−0.58 SD) and APWL (0.49 SD), but was slightly lower (1.8%) for AGE when matings were selected on total genetic value (−0.56 SD vs −0.55 SD). This indicates that selecting the candidates to maximize total genetic merit of progeny did not have a detrimental effect on the expected additive genetic gain for these traits.

Discussion

Variance components and heritabilities

The consistency of the estimates of additive genetic variance across genomic models (G and GD) is attributed to the orthogonality of the models by using the classical parameterization in terms of breeding values and allele substitution effects proposed by Vitezica et al. [35]. The genomic models resulted in lower estimates of additive genetic variance for AGE and BD than the pedigree-based estimates, although the differences were not relevant for practical purposes. This was also observed in other studies in pigs [13, 36, 37] and in dairy cattle [11]. For instance, estimates of genomic heritabilities were approximately half of those estimated with a pedigree-based model in a Duroc population [36] and in a Landrace population [13]. Using different marker densities and sequences, Zang et al. [37] attributed this difference to the so-called "missing heritability" (i.e. incomplete linkage disequilibrium between causal variants and markers), the genetic architecture of the traits and genotype-by-environment interactions. In our study, the main reason for this reduction is that genotyped animals represented a small proportion of the population.

For AGE, reported pedigree-based heritabilities for the French Landrace breed ranged from 0.23 to 0.36 [38–40], which are in agreement with our pedigree-based estimate

Table 4 Expected total (ΔG) and additive (ΔU) genetic gain obtained from matings selected on estimated breeding value ($\widehat{u}$) or on estimated total genetic value ($\widehat{g}$)

Trait	Expected gain	Selection on $\widehat{u}$		Selection on $\widehat{g}$	
		Trait units	Genetic SD units	Trait units	Genetic SD units
AGE (days)	ΔU	−2.24	−0.56	−2.21	−0.55
	ΔG	−2.19	−0.54	−2.98	−0.74
BD (mm)	ΔU	−0.35	−0.58	−0.35	−0.58
	ΔG	−0.35	−0.58	−0.39	−0.65
APWL (g)	ΔU	59.04	0.49	58.35	0.49
	ΔG	59.44	0.50	70.78	0.59

ΔG (or ΔU) is the difference between the mean $\widehat{u}$ (or the mean $\widehat{g}$) of selected matings and the mean $\widehat{u}$ (or the mean $\widehat{g}$) of all possible matings of the expected progeny
AGE age at 100 kg, *BD* backfat depth, *APWL* average piglet weight at birth within litter

of 0.33. To our knowledge, genomic estimates of heritability have not been published for Landrace or any other pig breed for this trait. The only comparable genomic estimate refers to lifetime daily gain, which was reported to be 0.27 by Lopes et al. [15] in a Landrace pig population. Pedigree-based estimates of the heritability for backfat reported for the French Landrace breed ranged from 0.46 to 0.55 [38–40] and the only reported genomic estimate in a Landrace population was 0.48 (see File S1 in additional information of Lopes et al. [41]). The heritabilities estimated for BD in our study are lower than those reported in the literature, which may be because the genotyped animals represented a small proportion of the animals in the population and/or because many causal variants for this trait are not in linkage disequilibrium (LD) with the SNPs that were used. Our heritability estimate for APWL was consistent with estimated in previous studies in the same French Landrace population using both genomic and pedigree-based models [42, 43].

Our estimates of the proportion of phenotypic variance explained by dominance effects (Table 2) show that dominance effects contributed only slightly to the phenotypic expression of the traits investigated, and their contributions were lower than the contributions of additive genetic effects. A study in a Yorkshire pig population, using a pedigree-based model that accounted for dominance genetic effects and inbreeding depression, reported proportions of dominance to additive variance of 31 and 11% for days to 104.5 kg and backfat at 104.5 kg, respectively [22]. Our genomic estimate of the proportion of dominance to additive variance for BD (0.11) agrees with this estimate, but our estimate for AGE (20%) was smaller. There are no published genomic estimates of dominance variance for AGE and APWL in any pig breed.

In the literature, estimates of dominance variance are reported based on models with or without inclusion of inbreeding as a covariate. When genomic inbreeding is not included in the model, the dominance variance estimate is inflated, as explained by Xiang et al. [7] and supported by other studies [7, 19, 26]. In our study, the ratio of dominance variance to additive variance was 0.11 for BD but estimates of 0.16 and 0.23 have been reported for backfat [41] using models that ignored inbreeding. This phenomenon has also been observed in cattle [19] for fertility and milk production traits.

In our analyses, estimates of dominance variance were less accurate than estimates of additive variance, which is the case in most studies that estimate dominance effects (e.g. [11, 17, 19]) possibly because of the low magnitude of this variance component, the reduced amount of data available, and the effect of the genotyping strategy. For instance, if only one or two piglets from each litter are genotyped, the proportion of full-sibs is small and, therefore, little dominance-specific information is available for the estimation [11, 44]. A larger number of genotyped individuals per litter allows the detection of identical genotypes among individuals and dominance relationships.

The observed reduction of the permanent environmental variance when dominance was included in the genomic model GD, suggests that a large part of the dominance variance is confounded with the permanent environmental effects when dominance genetic effects are not accounted for. Similar results have also been observed in other studies in cattle [12, 19]. The dominance variance can also be confounded with common-litter effect which was not included in the model as fixed effect.

Inbreeding effect

The average, across animals, of the proportion of homozygosity or genomic inbreeding in our Landrace population was similar to that (0.69, SD = 0.019) reported by Xiang et al. [16] for a Danish Landrace pig population. Genomic estimates of the effect of inbreeding for these traits are scarce in the literature. One study in a Yorkshire pig population based on a pedigree model reported estimates of inbreeding depression of around 2.10 (days) and 0.00 (mm) per 10% inbreeding, for days to 104.5 kg and backfat at 104.5 kg, respectively [22]. Our results agree with the values reported in the literature for AGE and BD. Our estimate of the effect of inbreeding depression was positive for APWL, but not significantly different from zero.

Goodness of fit

Genomic models fitted the data better than the pedigree-based model, probably because realized relationships among individuals are better captured by marker information. This is in agreement with previous findings (see e.g. de Los Campos et al. [45]). Differences in AIC between genomic models were small, which means that the inclusion of dominance effects did not significantly improve the goodness of fit of the data. This was in part because model G already included the inbreeding depression parameter, which improved the goodness of fit in a previous study [19]. Aliloo et al. [19] also did not find a difference in goodness of fit between an additive genomic model with average heterozygosity (or inbreeding) and a model that accounted for additive and dominance genetic effects with average heterozygosity.

Prediction of the total genetic value of matings

Our results confirm that within-breed selection on $\widehat{g}$ profits from dominance effects to maximize the total genetic value of the progeny (productive performance) without compromising the additive genetic gain (based

on breeding values). These results are important since the additive genetic value is the genetic component that can be accumulated and inherited by subsequent generations. The negligible reduction in additive genetic gain can be attributed to the fact that almost the same sows and the best boars (based on GEBV) were selected in the matings to optimize $\widehat{g}$ or $\widehat{u}$. In addition, the number of selected boars was larger when selecting on $\widehat{g}$ than when selecting on $\widehat{u}$, which could result in a reduction in the overall inbreeding of the population. Nevertheless, the part of the total genetic gain that is attributable to non-additive genetic effects obtained from mate allocation disappears in subsequent generations, as demonstrated in a simulation by Toro and Varona [2], because non-additive genetic effects depend on the specific allelic combinations that are present between the parents involved in a mating.

Other studies on mate allocation accounting for non-additive genetic effects have also reported increases in total genetic merit compared to selecting on estimated breeding values only. Using simulation, Toro and Varona [2] showed that mate allocation provides an additional selection response in expected progeny of up to 22% (assuming a narrow-sense heritability of 0.20 and a ratio of dominance to phenotypic variance of 0.10) over random mating. In dairy cattle, Sun et al. [46] reported an increase in total genetic gain for milk yield of 9.8 and 7.6% for Holstein and Jersey cows, respectively, from doing mating allocation. Ertl et al. [11] obtained an increase of 14.8 (milk yield) and 27.8% (protein yield) in expected total genetic superiority when matings were selected on $\widehat{g}$ instead of $\widehat{u}$, and, as in our results, only a slight reduction of 4.5% and 2.6% on additive genetic gain was observed. Aliloo et al. [19] reported increases in expected total genetic superiority of 27, 25, and 22% for milk, fat, and protein yield, respectively, in a Holstein population. Our results agree with those reported in the literature in the sense that a greater expected total genetic superiority can be obtained by including non-additive genetic effects to exploit dominance, with a slight or negligible reduction in expected additive genetic gain (1.8% for AGE and 0% for BD and APWL). The largest increase in total genetic gain was observed for AGE (0.20 genetic standard deviations), which had the highest ratio of dominance to additive genetic variance (20%). This greater increase in total genetic gain with a higher ratio of dominance variance was also observed in previous studies that used mate allocation [2, 11], except in [19].

Different mating design scenarios can be found in the literature. Ertl et al. [11] and Aliloo et al. [19] performed a pre-selection of bulls based on their estimated breeding values and only the selection and allocation of females were optimized based on the expected total genetic value of the offspring. This strategy minimizes loss of additive genetic gain since selection intensity is higher for males than for females. Nonetheless, pre-selection of bulls may preclude the opportunity to select other bulls that can potentially produce progeny with a higher total genetic merit when used in specific matings. In our study, optimization on $\widehat{g}$ was made for both males and females, which provides more opportunity to choose matings that have a high total genetic merit in the progeny.

To rule out the possibility that preselection of males based on the GEBV influenced our results, 120 boars were chosen at random and included in the optimization process to select the best set of 600 matings for each mating strategy. Optimization based on $\widehat{u}$ and $\widehat{g}$ yielded similar responses in ΔU (i.e. -2.93 vs. -2.91) but optimization based on $\widehat{g}$ resulted in a higher ΔG than optimization based on $\widehat{u}$ (i.e. -3.70 vs. -2.90). These results confirm that the optimization of the offspring based on $\widehat{g}$ is feasible without reducing ΔU, regardless of the preselection of males.

Our results show that mate allocation strategies for boosting total genetic gain could benefit production levels in the sow herds without placing a cost on additive genetic progress. However, its implementation on a large scale would require some organizational changes, for instance, additional genotyping when not all females are genotyped and high computational costs. An alternative to additional genotyping is to take the maternal grandsire SNP information into account, as was proposed by DeStefano and Hoeschele [5].

As stated above, genetic gain that is attributable to non-additive genetic effects is expressed in the immediate offspring and is not inherited in subsequent generations. Thus, selection and mating of parents that produce crossbred commercial animals are expected to benefit most from mating allocation strategies for improving total genetic gain in pig production schemes. We focused on single-trait selection but the use of mate allocation strategies can be extended to multiple traits by using an index as objective function to improve total genetic gain as in [19]. The feasibility and the success of mate allocation strategies in a multiple trait approach will depend highly on the dominance genetic correlation between the target traits. Hence, dominance genetic correlations different from 1 will indicate that matings will perform differently between traits. If this is the case, this will inevitably complicate the mating allocation strategy and a compromise between the expected total genetic gains of the traits will be needed.

Conclusions

In this study, we estimated non-additive genetic variance components and addressed their use to improve the performance of offspring of selected parents by means

of mate allocation strategies. Inclusion of non-additive genetic effects is straightforward in a genomic prediction context. For the traits analyzed here, estimates of dominance genetic variance were small, ranging from 11 to 20% of the additive genetic variance. Addition of dominance genetic effects to a model that included additive genetic effects and genomic inbreeding did not improve the goodness of fit of the model. Inbreeding depression was estimated to have an undesirable effect on AGE and no effect on other traits. Genomic mate allocation improved the performance of future offspring of −0.79 days, −0.04 mm, and 11.3 g for AGE, BD and APWL, respectively, but slightly reduced the expected additive genetic gain, e.g. by 1.8% for AGE. Our results show that genomic mate allocation, accounting for non-additive genetic effects, is a feasible and a potential strategy to improve the productive performance (total genetic value) of future offspring without compromising the additive genetic gain. Our conclusions are limited to a single population, to a few tested scenarios of mate allocation strategies and under a single-trait approach. Further research is very much needed to explore the benefits of mate allocation strategies in a broader context.

Acknowledgements
The authors thank France Génétique Porc for providing the data. We acknowledge two anonymous reviewers and the associate editor, whose comments improved our manuscript. We appreciate the contribution of Paulino Pérez-Rodríguez, who gave us advice on improving the efficiency of our R code.

Authors' contributions
DGD conducted the statistical analyses and wrote the first draft of the manuscript. ZV and LT proposed the models and methods of analysis. CCJ and AB provided and prepared the pig data. DGD, LT and ZV discussed the results, made suggestions and corrections. All authors read and approved the final manuscript.

Funding
DGD is funded by France Génétique Porc, RAGT and INRA (Contract 15000416, Project 00000837). Work was supported by INRA SELGEN metaprogram (Project OptiMaGics). Project was partly supported by Toulouse Midi-Pyrénées bioinformatics platform.

Availability of data and materials
The database used and analyzed in this study was provided by the national collective breeding scheme of French Landrace pig line and is not available. It comes from private companies and its privacy could be compromised.

Ethics approval and consent to participate
Animal Care and Use Committee approval was not necessary for this study because the data were obtained from an existing database.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] GenPhySE, INRA, Université de Toulouse, 31326 Castanet-Tolosan, France. [2] IFIP Institut du Porc, BP35104, 35651 Le Rheu, France. [3] France Génétique Porc, BP35104, 35651 Le Rheu, France.

Received: 5 October 2018 Accepted: 20 September 2019
Published online: 27 September 2019

References

1. Kinghorn BP. Mate selection by groups. J Dairy Sci. 1998;81:55–63.
2. Toro MA, Varona L. A note on mate allocation for dominance handling in genomic selection. Genet Sel Evol. 2010;42:33.
3. Pryce JE, Hayes BJ, Goddard ME. Novel strategies to minimize progeny inbreeding while maximizing genetic gain using genomic information. J Dairy Sci. 2012;95:377–88.
4. Liu H, Henryon M, Sørensen AC. Mating strategies with genomic information reduce rates of inbreeding in animal breeding schemes without compromising genetic gain. Animal. 2017;11:547–55.
5. DeStefano AL, Hoeschele I. Utilization of dominance variance through mate allocation strategies. J Dairy Sci. 1992;75:1680–90.
6. Hayes BJ, Miller SP. Mate selection strategies to exploit across- and within-breed dominance variation. J Anim Breed Genet. 2000;117:347–59.
7. De Boer IJM, Hoeschele I. Genetic evaluation methods for populations with dominance and inbreeding. Theor Appl Genet. 1993;86:245–58.
8. Fernández EN, Legarra A, Martínez R, Sánchez JP, Baselga M. Pedigree-based estimation of covariance between dominance deviations and additive genetic effects in closed rabbit lines considering inbreeding and using a computationally simpler equivalent model. J Anim Breed Genet. 2017;134:184–95.
9. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
10. Vitezica ZG, Varona L, Legarra A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics. 2013;195:1223–30.
11. Ertl J, Legarra A, Vitezica ZG, Varona L, Edel C, Emmerling R, et al. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genet Sel Evol. 2014;46:40.
12. Aliloo H, Pryce JE, González-Recio O, Cocks BG, Hayes BJ. Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. Genet Sel Evol. 2016;48:8.
13. Esfandyari H, Bijma P, Henryon M, Christensen OF, Sørensen AC. Genomic prediction of crossbred performance based on purebred Landrace and Yorkshire data using a dominance model. Genet Sel Evol. 2016;48:40.
14. Heidaritabar M, Wolc A, Arango J, Zeng J, Settar P, Fulton JE, et al. Impact of fitting dominance and additive effects on accuracy of genomic prediction of breeding values in layers. J Anim Breed Genet. 2016;133:334–46.
15. Lopes MS, Bastiaansen JWM, Janss L, Knol EF, Bovenhuis H. Genomic prediction of growth in pigs based on a model including additive and dominance effects. J Anim Breed Genet. 2016;133:180–6.
16. Xiang T, Christensen OF, Vitezica ZG, Legarra A. Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. Genet Sel Evol. 2016;48:92.
17. Moghaddar N, van der Werf JHJ. Genomic estimation of additive and dominance effects and impact of accounting for dominance on accuracy of genomic evaluation in sheep populations. J Anim Breed Genet. 2017;134:453–62.
18. Varona L, Legarra A, Toro MA, Vitezica ZG. Non-additive effects in genomic selection. Front Genet. 2018;9:78.
19. Aliloo H, Pryce JE, González-Recio O, Cocks BG, Goddard ME, Hayes BJ. Including nonadditive genetic effects in mating programs to maximize dairy farm profitability. J Dairy Sci. 2017;100:1203–22.
20. Clutter AC. Genetic of performance traits. In: Rothschild MF, Ruvinsky A, editors. The genetics of pig. 2nd ed. Wallingford: CABI International; 2011. p. 325–54.
21. Tusell L, Gilbert H, Riquet J, Mercat MJ, Legarra A, Larzul C. Pedigree and genomic evaluation of pigs using a terminal-cross model. Genet Sel Evol. 2016;48:32.
22. Culbertson MS, Mabry JW, Misztal I, Gengler N, Bertrand JK, Varona L. Estimation of dominance variance in purebred Yorkshire swine. J Anim Sci. 1998;76:448–51.
23. Institut Technique du Porc (ITP). La sélection porcine collective en France. 2003. https://www.ifip.asso.fr/sites/default/files/pdf-documentations/selection_collective.pdf. Accessed 3 May 2019.

24. IFIP. La génétique. Mémento de l'éleveur de porc. 7th ed. Paris: ITP; 2013. p. 86–102.
25. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
26. Varona L, Legarra A, Herring W, Vitezica ZG. Genomic selection models for directional dominance: an example for litter size in pigs. Genet Sel Evol. 2018;50:1.
27. Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee D. BLUPF90 and related programs (BGF90). In: Proceedings of the 7th world congress on applied livestock production: 19–23 August 2002. Montpellier; 2002.
28. Houle D, Meyer K. Estimating sampling error of evolutionary statistics based on genetic covariance matrices using maximum likelihood. J Evol Biol. 2015;28:1542–9.
29. Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. 1974;19:716–23.
30. Legarra A, Ricard A, Filangi O. GS3 Genomic Selection—Gibbs Sampling—Gauss Seidel (and Bayes Cπ). 2014. http://genoweb.toulouse.inra.fr/~alegarra/manualgs3_last.pdf. Accessed 2 Jan 2018.
31. Falconer DS, Mackay TFC. Introduction to quantitative genetics. 4th ed. Harlow: Pearson Education Limited; 1996.
32. Jansen GB, Wilton JW. Selecting mating pairs with linear programming techniques. J Dairy Sci. 1985;68:1302–5.
33. R Development Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2017.
34. Berkelaar M, Eikland K, Notebaert P. lpsolve 5.5, Open source (Mixed-Integer) Linear Programming system Software. 2004. http://lpsolve.sourceforge.net/5.5/. Accessed 16 Feb 2018.
35. Vitezica ZG, Legarra A, Toro MA, Varona L. Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. Genetics. 2017;206:1297–307.
36. Jiao S, Maltecca C, Gray KA, Cassady JP. Feed intake, average daily gain, feed efficiency, and real-time ultrasound traits in Duroc pigs: I. Genetic parameter estimation and accuracy of genomic prediction. J Anim Sci. 2014;92:2377–86.
37. Zhang C, Kemp RA, Stothard P, Wang Z, Boddicker N, Krivushin K, et al. Genomic evaluation of feed efficiency component traits in Duroc pigs using 80 K, 650 K and whole-genome sequence variants. Genet Sel Evol. 2018;50:14.
38. Bidanel JP, Ducos A, Guéblez R, Labroue F. Genetic parameters of backfat thickness, age at 100 kg and ultimate pH in on-farm tested French Landrace and Large White pigs. Livest Prod Sci. 1994;40:291–301.
39. Ducos A, Bidanel JP. Genetic correlations between production and reproductive traits measured on the farm, in the Large White and French Landrace pig breeds. J Anim Breed Genet. 1996;113:493–504.
40. Saintilan R, Merour I, Tribout T, Bidanel J, Brossard L, Van Milgen J, et al. Sélection pour l'efficacité alimentaire chez le porc en croissance: opportunités et conséquences de l'utilisation de la consommation moyenne journalière résiduelle dans les populations en sélection collective. In: Proceedings of the 44èmes Journées de la Recherche Porcine: 7–8 February 2012; Paris. 2012. p. 13–8.
41. Lopes MS, Bastiaansen JWM, Janss L, Knol EF, Bovenhuis H. Estimation of additive, dominance, and imprinting genetic variance using genomic data. G3 (Bethesda). 2015;5:2629–37.
42. Bouquet A, Delauney I, Merour I, Bellec T, Ligonesche B, Sourdioux M, et al. Conséquences d'une sélection sur l'homogénéité du poids des porcelets à la naissance sur la productivité numérique des truies Large White et Landrace Français. In: Proceedings of the 46èmes Journées de la Recherche Porcine: 4–5 February 2014; Paris. 2014. p. 19–24.
43. Bouquet A, Canaple M, Brenaut P, Bellec T, Flatres-Grall L, Ligonesche B, et al. Mise en place de la sélection génomique dans le schéma de sélection de la population Landrace Français. In: Proceedings of the 49èmes Journées de la Recherche Porcine: 31 January 2017; Paris. 2017. p. 31–6.
44. Varona L, Misztal I. Prediction of parental dominance combinations for planned matings, methodology, and simulation results. J Dairy Sci. 1999;82:2186–91.
45. de Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013;193:327–45.
46. Sun C, VanRaden PM, O'Connell JR, Weigel KA, Gianola D. Mating programs including genomic relationships and dominance effects. J Dairy Sci. 2013;96:8014–23.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

2.1 Appendices

Appendix 2.1

Accuracy of prediction of breeding values

The ability of prediction of the models A, G and GD (presented in the research article in Chapter 2, A = pedigree-based model, G = genomic additive model and GD = genomic additive + dominance) was assessed. The LR method for cross-validation proposed by Legarra and Reverter (2018) was used to compare predictions based on partial (old individuals) and whole (old + recent individuals) datasets. The prediction accuracy was evaluated for each trait and model as follows. First, BVs for all individuals were estimated with each model using the whole dataset available for each trait (2968 and 3682 records/animals for AGE and BD respectively; and 3297 records from 1226 females for APWL). The whole datasets involve animals from 2011-2017, 2009-2017 and 2009-2016 for AGE, BD and APWL, respectively. Then, for traits AGE and BD the cutoff was in 2016, whereas for APWL it was in 2015. Then, a partial dataset was created, in which phenotype values of individuals after the cutoff date (recent individuals) were set as missing values. Remaining individuals with records in the partial dataset were 1758, 2425 and 875 for AGE, BD and APWL. EBVs (or GEBVs) were then computed for all individuals based on the partial dataset. Then, candidates to selection were used as validation set. In the case of the trait APWL, as it had repeated measures, animals can have records in the whole and in the partial, which can bias the assessment of prediction accuracy. To avoid that, animals having records in both the whole and partial datasets were removed from the validation set. The correlation of EBV based on the partial and whole dataset ($\hat{\rho}_{p,w}$) were computed according to Legarra and Reverter (2018):

$$accuracy, \quad \hat{\rho}_{p,w} = \frac{cov(\widehat{\boldsymbol{u}}_p \widehat{\boldsymbol{u}}_w)}{\sqrt{var(\widehat{\boldsymbol{u}}_p) var(\widehat{\boldsymbol{u}}_w)}}$$

where, $\widehat{\boldsymbol{u}}_w$ and $\widehat{\boldsymbol{u}}_p$ are the vectors of EBVs (or GEBVs) of selection candidates from the whole and partial datasets, respectively.

Figure A 2.1 shows the comparison of the accuracy of prediction for each trait and model. The cross-validation results show that genomic-based models obtained higher accuracy than the pedigree-based model. The accuracy of prediction of BV did not improve when dominance was included in the genomic models. These results agree with those reported in the literature (Varona *et al.* 2018b).

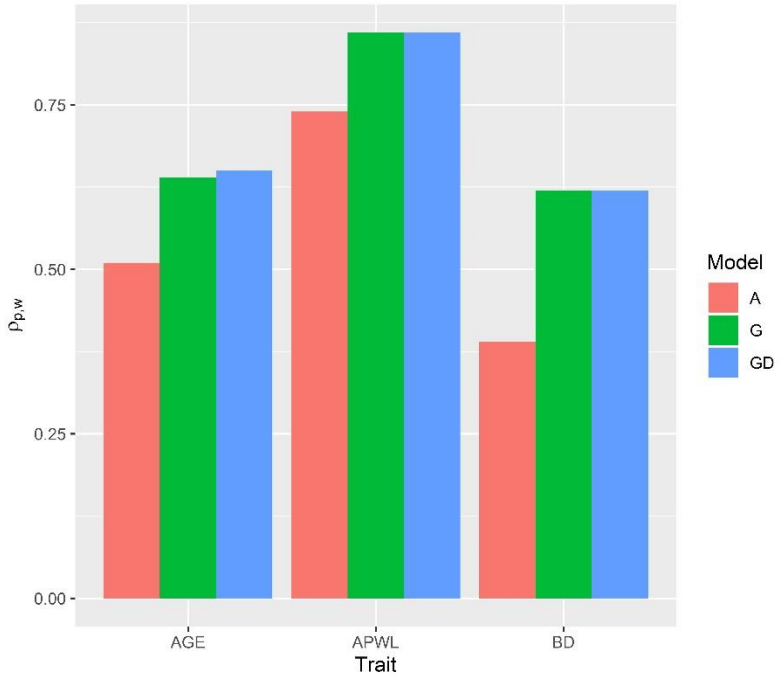

Figure A 2.1 Accuracy of prediction of models A, G and GD for traits AGE = age at 100 kg; BD = backfat depth; APWL = average piglet weight at birth within litter.

Appendix 2.2

Comparison of different strategies of pre-selection of males

Results presented in Table 4 in the previous article were based on pre-selection of 120 males based on its GEBV. Two additional strategies of pre-selection of males were tested to see how they influence the results of mate allocation on the additive and total genetic gain: (1) the top 40 males were selected based on its GEBV to minimize loss of additive genetic gain, and (2) 120 males were taken at random (so, no pre-selection and no selection pressure was applied). In the three cases, all the available genotyped females were included in the optimization. Table A 2.1 shows the results of the two additional strategies and also includes the results from of pre-selection of 120 males based on GEBV for easy comparison.

Table A 2.1 Expected total (ΔG) and additive (ΔU) genetic gain obtained from matings selected based on predicted breeding values ($\hat{u}$) or predicted total genetic value ($\hat{g}$) of expected progeny under different strategies of pre-selection of males.

Trait	Pre-selection of males	Mate allocation strategy	ΔU		ΔG		Increase in ΔG in Genetic SD units	Reduction in ΔU in Genetic SD units
			Trait units	Genetic SD units	Trait units	Genetic SD units		
AGE (days)	Top 40 males selected on GEBV	Selection on $\hat{u}$	-1.58	-0.39	-1.51	-0.37		
		Selection on $\hat{g}$	-1.56	-0.38	-2.3	-0.56	-0.20	0.01
	120 males pre-selected on GEBV	Selection on $\hat{u}$	-2.24	-0.56	-2.19	-0.54		
		Selection on $\hat{g}$	-2.21	-0.55	-2.98	-0.74	-0.20	0.01
	120 males taken at random.	Selection on $\hat{u}$	-2.93	-0.73	-2.90	-0.72		
		Selection on $\hat{g}$	-2.91	-0.72	-3.70	-0.92	-0.20	0.01
BD (mm)	Top 40 males selected on GEBV	Selection on $\hat{u}$	-0.25	-0.42	-0.25	-0.42		
		Selection on $\hat{g}$	-0.25	-0.42	-0.30	-0.49	-0.07	0.00
	120 males pre-selected on GEBV	Selection on $\hat{u}$	-0.35	-0.58	-0.35	-0.58		
		Selection on $\hat{g}$	-0.35	-0.58	-0.39	-0.65	-0.07	0.00
	120 males taken at random.	Selection on $\hat{u}$	-0.44	-0.74	-0.45	-0.74		
		Selection on $\hat{g}$	-0.44	-0.74	-0.49	-0.81	-0.07	0.00
APWL (g)	Top 40 males selected on GEBV	Selection on $\hat{u}$	39.77	0.33	40.29	0.34		
		Selection on $\hat{g}$	39.60	0.33	51.06	0.42	0.09	0.00
	120 males pre-selected on GEBV	Selection on $\hat{u}$	59.04	0.49	59.44	0.50		
		Selection on $\hat{g}$	58.35	0.49	70.78	0.59	0.09	0.00
	120 males taken at random.	Selection on $\hat{u}$	83.20	0.70	83.60	0.70		
		Selection on $\hat{g}$	82.89	0.69	94.68	0.79	0.09	0.00

AGE = age at 100 kg; BD = backfat depth; APWL = average piglet weight at birth within litter.
ΔG (or ΔU) is the difference between the mean $\hat{g}$ (or the mean $\hat{u}$) of selected matings and the mean $\hat{g}$ (or the mean $\hat{u}$) of all possible matings of the expected progeny.

Note that comparison of the pre-selection strategies should be made based on increase and reduction of the total genetic gain and additive genetic gain, respectively. This is because the genetic SD units are referred to the average value of all potential mating, which changes depending on the pre-selection strategy. Note that selection in pigs aims at reducing Age at 100 kg (AGE) and Backfat depth (BD), and increasing the Average piglet weight at birth within litter (APWL). Thus, negative values for AGE and BD are favorable, while a positive value is favorable for APWL. The results show that for all traits, optimization of matings based on selection on $\hat{g}$ resulted in higher total genetic gain than selection based on $\hat{u}$, with only a only a very slight reduction on the additive genetic gain on the trait AGE (0.01 genetic SD units), regardless of the pre-selection strategy. These results are discussed in Chapter 5.

Chapter 3

Purebred and crossbred genomic evaluation and mate allocation strategies to exploit dominance in pig crossbreeding schemes

David González-Diéguez[*1], Llibertat Tusell[*], Alban Bouquet[†,‡], Andres Legarra[*], Zulma G. Vitezica[*]

[*] GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France.

[†] IFIP Institut du Porc, BP35104, 35651 Le Rheu, France.

[‡] France Génétique Porc, BP35104, 35651 Le Rheu, France.

G3: Genes | Genomes | Genetics (2020) 10(8): 2829-2841

G3 Genes | Genomes | Genetics

GENOMIC PREDICTION

Purebred and Crossbred Genomic Evaluation and Mate Allocation Strategies To Exploit Dominance in Pig Crossbreeding Schemes

David González-Diéguez,*,[1] Llibertat Tusell,* Alban Bouquet,[†,‡] Andres Legarra,* and Zulma G. Vitezica*
*GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France, [†]IFIP Institut du Porc, BP35104, 35651 Le Rheu, France, and [‡]France Génétique Porc, BP35104, 35651 Le Rheu, France
ORCID ID: 0000-0002-8637-1835 (D.G.-D.)

ABSTRACT We investigated the effectiveness of mate allocation strategies accounting for non-additive genetic effects to improve crossbred performance in a two-way crossbreeding scheme. We did this by computer simulation of 10 generations of evaluation and selection. QTL effects were simulated as correlated across purebreds and crossbreds, and (positive) heterosis was simulated as directional dominance. The purebred-crossbred correlation was 0.30 or 0.68 depending on the genetic variance component used. Dominance and additive marker effects were estimated simultaneously for purebreds and crossbreds by multiple trait genomic BLUP. Four scenarios that differ in the sources of information (only purebred data, or purebred and crossbred data) and mate allocation strategies (mating at random, minimizing expected future inbreeding, or maximizing the expected total genetic value of crossbred animals) were evaluated under different cases of genetic variance components. Selecting purebred animals for purebred performance yielded a response of 0.2 genetic standard deviations of the trait "crossbred performance" per generation, whereas selecting purebred animals for crossbred performance doubled the genetic response. Mate allocation strategy to maximize the expected total genetic value of crossbred descendants resulted in a slight increase (0.8%, 4% and 0.5% depending on the genetic variance components) of the crossbred performance. Purebred populations increased homozygosity, but the heterozygosity of the crossbreds remained constant. When purebred-crossbred genetic correlation is low, selecting purebred animals for crossbred performance using crossbred information is a more efficient strategy to exploit heterosis and increase performance at the crossbred commercial level, whereas mate allocation did not improve crossbred performance.

KEYWORDS
Purebred and crossbred evaluation
non-additive genetic effects
heterosis
purebred-crossbred genetic correlation
Genomic Prediction
GenPred
Shared Data Resources

Crossbreeding schemes are widely used in almost all species of livestock production, especially in monogastric species (pigs and birds in particular). The main goal of crossbreeding is to improve the performance of crossbred (CB) animals by exploiting heterosis and breed complementarity (Falconer 1981). Dominance is one of the major genetic bases of heterosis and mate allocation can be used to maximize the total genetic merit of future progeny by exploiting dominance variation across-breeds (*e.g.*, crosses in a 2-way cross-breeding scheme) and within-breed (*e.g.*, in a purebred population) (DeStefano and Hoeschele 1992; Hayes and Miller 2000; Toro and Varona 2010). In pigs, although the selection is made within purebreds (PB) (Dekkers 2007), the commercial CB animals can be created by selecting specific pairs of mates between breeds that result in a superior CB descendants in terms of performance compared to random mating. With the advent of high-density single nucleotide polymorphism (SNP), genomic selection has become a standard practice in the genetic evaluation of livestock populations (Meuwissen *et al.* 2013). Moreover, nowadays, estimating dominance effects in genetic evaluations has become feasible in a genomic BLUP (best linear unbiased

Copyright © 2020 Gonzalez-Dieguez *et al.*
doi: https://doi.org/10.1534/g3.120.401376
Manuscript received February 12, 2020; accepted for publication June 16, 2020; published Early Online June 18, 2020.
This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Supplemental material available at figshare: https://doi.org/10.25387/g3.12504638.
[1]Corresponding author: INRAE Toulouse, UMR 1388 GenPhySE, 24 Chemin de Borde Rouge, 31326 CASTANET-TOLOSAN Cedex, France. E-mail: david-omar.gonzalez-dieguez@inrae.fr

prediction) context (Vitezica *et al.* 2013). Thus, SNP-based mate allocation strategies accounting for non-additive genetic effects have been developed (Toro and Varona 2010). Such strategies have provided encouraging results to maximize the expected total genetic merit of future progeny within breed, on computer simulation (Toro and Varona 2010) and on real data in dairy cattle and pigs (Ertl *et al.* 2014; Aliloo *et al.* 2017; González-Diéguez *et al.* 2019). However, the benefits of genomic mate allocation strategies to increase performance in a crossbreeding scheme have not been evaluated in the long term in a genomic scheme.

In crossbreeding schemes, the main limitation to improve the performance of CB animals by selection on PB is that the genetic correlation (r_{PC}) between PB and CB performances is lower than 1 (*e.g.*, 0.63 on average in pigs, with 50% of the estimates between 0.45 and 0.87 (Wientjes and Calus 2017)). This low genetic correlation between PB and CB may be due to genotype-by-environment interaction (GxE), and genotype-by-genotype interactions (GxG) (*i.e.*, dominance and/or epistasis). The effects of genetic causal variants depend on the environment where the animal is raised (GxE), and depend on the genetic backgrounds where the variants are expressed in (GxG). Both, GxE and GxG may result in a low r_{PC} between PB and CB (Wientjes and Calus 2017; Duenk *et al.* 2020). If r_{PC} is low, genetic merit of PB parents evaluated in a PB population are a poor predictor of the performance of their CB descendants (Dekkers 2007). Then, the integration of both PB and CB information is essential in genetic evaluation oriented to improve CB performance (Wei and van der Werf 1994). Although several genomic models have been proposed to address these issues (Dekkers 2007; Zeng *et al.* 2013; Esfandyari *et al.* 2015a), only Xiang *et al.* (2016) addressed most of the issues influencing the level of r_{PC}, in particular the inclusion of additive and dominance effects, and use of PB and CB data simultaneously. Their model fits purebred and crossbred data together fitting additive and dominance effects in a multivariate manner (Varona *et al.* 2010; Karoui *et al.* 2012). Xiang *et al.*'s model also includes a regression on "genomic inbreeding" (observed homozygosity) in PB and CB animals to measure individual "inbreeding depression". This can be seen as the opposite of heterosis and therefore it is possible to correct by, and to predict, individual heterosis (Iversen *et al.* 2019). In this manner, all essential aspects of the joint PB – CB prediction are considered in a single analysis.

Although the methodology is available for simultaneously accounting for all these aspects in genomic evaluations there is a need of addressing the benefits of its implementation in practice in a crossbreeding scheme. This is crucial because its implementation would require large organizational changes in the way pig breeding schemes are organized today.

The objective of this study was to investigate the effectiveness of mate allocation strategies and genomic evaluations that account for additive and dominance effects to improve CB performance. Breed-specific QTL and individual genomic heterosis effects were explicitly simulated in a two-way pig crossbreeding scheme. The effects of the sources of information used in the genetic evaluation (only PB data or PB and CB data), of several narrow and broad-sense heritability values, and of several options for mate allocation to produce the CB were examined.

MATERIALS AND METHODS

To evaluate the effectiveness of mate allocation strategies on crossbreeding, a two-way pig crossbreeding selection scheme for a maternal trait (*e.g.*, litter size) was simulated. The simulation was split into two steps. In the first step, the simulator QMSim (Sargolzaei and Schenkel 2009) was used to generate a historical population structure.

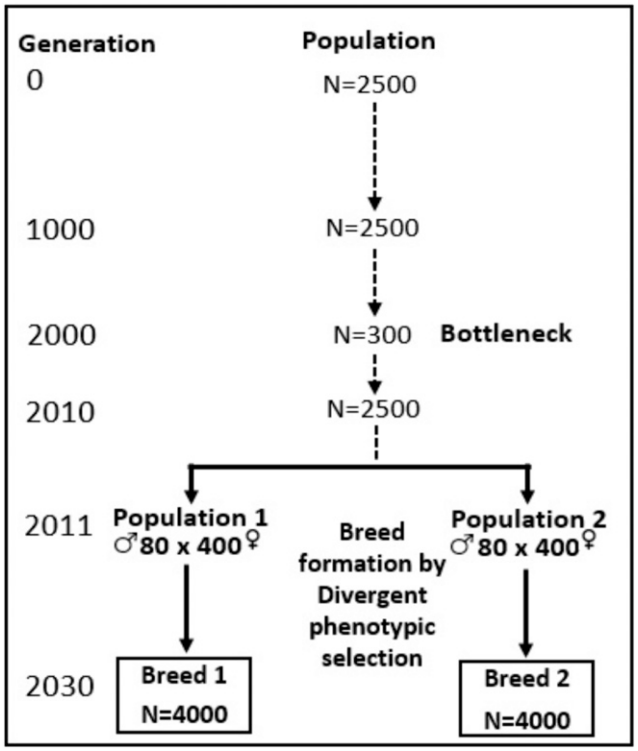

Figure 1 Schematic representation of the simulated historical population (First step). N stands for number of animals.

In the second step, a self-made Fortran program was developed to simulate a two-way maternal pig crossbreeding scheme across ten generations, and evaluate four different scenarios. The scenarios differed in the sources of information used to evaluate the selection candidates (PB or PB and CB) and in the use or not of mate allocation strategies to produce the CB descendants. Different values of genetic (co)variance components were also tested.

Historical and recent populations

To create the historical population (HP) (undergoing drift and mutation), the simulator QMSim (Sargolzaei and Schenkel 2009) was used. Figure 1 shows a schematic representation of the simulated historical population. First, a constant population size of 2500 individuals was generated over 1000 generations of random mating. Second, from generation 1001 to 2000, the population size was gradually reduced to 300 individuals in order to simulate a bottleneck and generate initial linkage disequilibrium (LD). Then, 10 additional generations were simulated to gradually expand the size of the population from 300 to 2500 individuals. At the historical generation (2010), there were equal numbers of males and females (1250 each sex). Two samples of 480 animals (80 males and 400 females) were drawn from the generation 2010 to create two breeds, P1 and P2. Then, the two breeds were divergently selected based on phenotype for 20 generations. Within each breed, selected animals were mated with a restriction to reduce inbreeding during breed formation. A litter size of 10 was used. In the last generation of breed formation (generation 2030), 4000 animals (with an equal number of males and females) composed each breed. Note that a dummy phenotype with a narrow-sense heritability of 0.5 was used in the first step, only in order to create the linkage disequilibrium structure in parental breeds.

In the second step, a two-way pig crossbreeding scheme with 10 generations of selection was simulated (Figure 2). The initial

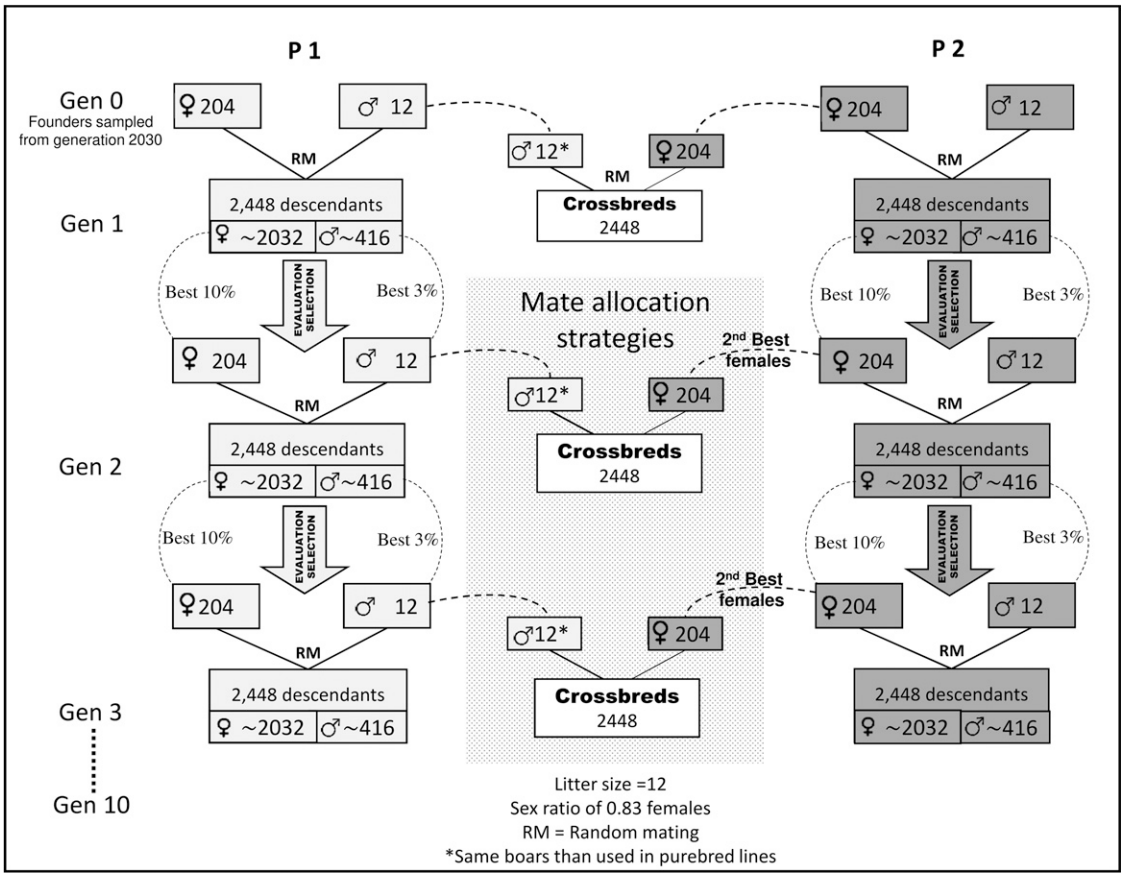

Figure 2 Schematic representation of the simulated two-way crossbreeding scheme (second step). Genetic evaluations and selection were carried out in each generation within breed. Best purebred animals were selected to be mated at random to produce the next generations within breed. The same best males selected to create the next generations within P1 were also crossed with the second-best females from P2 to create the crossbred descendants. Mate allocation strategies were implemented only to create the crossbred animals.

generations of breeds P1 and P2 were formed by randomly sampling 12 males and 204 females (founders) from each respective breeds in generation 2030. First generation of P1 (and P2) was mated at random to produce the first progeny of PB animals. From generation 1 to 10, PB animals were evaluated and selected based on different models and criteria depending on each scenario (see description below). Within each breed, selected animals were mated at random to produce the next generation. Real pigs breeding schemes are complex with several steps of selection, and its simulation is not straightforward. To simplify programming, we used a litter size of 12 with an artificial sex ratio of 0.83 females, resulting in 2448 descendants in each generation (~2032 females and ~416 males, Figure 2). Selection intensity was 3% and 10% with respect to the simulated number of males and females, respectively, which could correspond to those used in a pig breeding scheme after performing pre-selection of individuals on other traits (*e.g.*, morphological defects, disease resistance, etc). Hence, the best 12 males and 204 females were chosen within each breed at each generation to be parents of the next generation of PB animals. These animals were selected at birth (before having any own record) based on selection criteria that will be detailed later. Figure 3 shows the closest sources of information available to evaluate the candidates to selection of the two parental breeds.

To create the CB population, P1 was used as sire breed and P2 was used as dam breed. The first CB progeny was created by crossing at random the 12 founder males from P1 with the 204 founder females from P2. These CB animals are half-sibs of PB animals in generation 1. In subsequent generations, CB progenies were created mating the 12 best males selected within P1 with the 204 second best candidate females from P2. Note that since the best 204 best females were already used within P2, they were not available for generating CB animals. Then, different mate allocation strategies to generate the CB progeny were implemented depending on each scenario (see below). All PB and CB females were eventually phenotyped and genotyped, but at the time of selection, candidates to selection were not phenotyped. The information available for the estimation of breeding values at the time of selection differed between breeds. For a selection candidate in P1, the most related CB animal with a performance record was an offspring of the paternal grandsire (*i.e.*, with an additive genetic relationship of 0.125). For a selection candidate in P2, the most related CB animal with a performance record was an offspring of a paternal grand-grand-sire (*i.e.*, with an additive genetic relationship of 0.03125). This difference occurs because in P2, different sets of dams were selected to produce PB and CB offspring, while in P1, the same set of sires was used to create both PB and CB offspring (Figure 3). Hence, the breeding scheme was not symmetric. All the simulation steps (historical and recent populations) were replicated 10 times.

Genome

The genome consisted of 18 chromosomes (autosomes) of 120 cM each. In the first generation of the historical population, biallelic

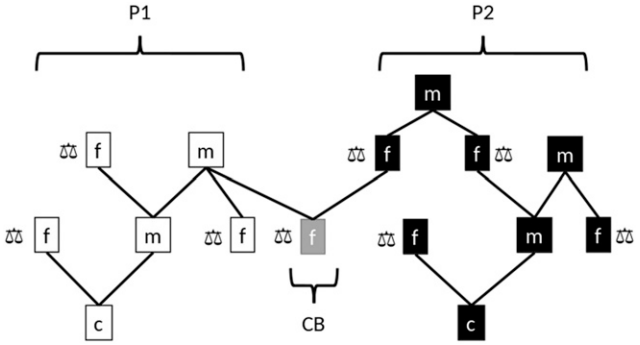

Figure 3 Close sources of information available through pedigree to evaluate the candidates to selection (c) via males (m) and females (f) in the two breeds (P1 and P2). A "balance" symbol means recorded phenotype. CB stands for crossbreds.

markers (72,000) and QTL (7,200) were distributed at random along the chromosomes with 0.5 allele frequencies. The mutation rate (recurrent mutation process) for markers and QTL was assumed to be 2.5×10^{-4} per locus per generation.

Marker quality control was done in each breed (generation 2030) independently. Markers with minor allele frequency (MAF) lower than 0.05 and out of Hardy-Weinberg equilibrium (t > 0.15, based on Wiggans *et al.* (2009)) were unqualified. Only markers that passed quality control in both breeds were kept. In that way, the number and order of markers were similar in both breeds. In founder generation (generation 0) of P1 and P2 breeds, around 50K SNPs (in each replicate) were segregating in both breeds.

Linkage disequilibrium patterns (LD) were also evaluated in the two breeds (P1 and P2) in generation 2030. LD was measured by calculating the squared correlation coefficient (r^2) between all pairs of markers (Hill 1974). The LD decay was measured for increasing distances between markers by calculating the mean r^2 within each distance interval. The resulting average over replicates of LD of SNP with an interval distance, assuming a 1Mbp:1cM ratio, of 0.9 to 1 Mbp, was the same for both parental breeds, $r^2 = 0.15$, close to values observed in real data in Landrace and Yorkshire breeds at the same distance (Boré *et al.* 2018).

Furthermore, the relationship between the two breeds (Cockerham 1969; Robertson 1975), defined as the covariance of allele frequencies was calculated as $8cov(p^{P1}, p^{P2})$ (Garcia-Baccino *et al.* 2017), where p^{P1} and p^{P2} are the allele frequencies across all loci of P1 and P2, respectively. The average of the relationship over replicates was equal to 0.24, close to that estimated between Landrace and Yorkshire base populations (0.26) (Xiang *et al.* 2017). Hence, the simulated breeds (recent populations) mimicked the structure of the real ones both within and between breeds.

Simulation of heterosis and QTL effects

QTL positions were the same in P1, P2 and CB, but with different allele frequencies. The positions and effects of 2500 QTL were sampled anew in each replicate. A maternal trait (*e.g.*, litter size) controlled by additive and dominance QTL action was simulated. The assumed genetic model is the same as Xiang *et al.* (2016), which involves additive and dominance effects correlated among PB and CB populations, and genomic heterosis.

To simulate heterosis (superiority of heterozygotes over homozygotes), a possible model is directional dominance, which states that dominance effects tend (on average across all QTL) to be favorable for the trait, *i.e.*, $E(d) = \mu_d$. For ease of simulation, we split the dominance effect at each locus as $d = \mu_d + d^*$ such that μ_d is constant across all QTL loci and d^* is a random deviation normally distributed with zero mean. From Xiang *et al.* (2016) it is known that the value of μ_d is obtained from estimates of heterosis (or of inbreeding depression) as $\mu_d = -\frac{b}{m}$, where b is the inbreeding depression parameter (or the value of heterosis if the sign is switched), and in our simulation m is the number of QTL. Note that there is still individual heterosis as for each individual the heterosis is equal to $(1-f_g)m\mu_d$, where $(1-f_g)$ is individual heterozygosity, and f_g is the vector of genomic inbreeding coefficients, calculated as the proportion of homozygous QTL per individual. Here, we assumed that inbreeding depression parameters were equal to -10 (equivalent to -1 (*e.g.*, piglet) per 10% increase in genomic inbreeding) in P1, P2 and CB (b^{P1}, b^{P2}, b^{c}), and were distributed across all QTL. As reference, estimates of inbreeding depression b reported in the literature for litter size ranged from -2 to -9 piglets per 100% of observed homozygosity (Xiang *et al.* 2016; Iversen *et al.* 2019).

Then we simulated breed-specific QTL effects, but correlated across the three populations (P1, P2 and CB) to account for differences due to GxE and gen-by-gen (GxG) (epistasis) interactions. To do that, additive (σ^2_u) and dominance (σ^2_v) estimated genetic variances from Xiang *et al.* (2016) for total number of piglets born were used. The values were equal to 0.86, 0.54 and 0.28 for σ^2_u; and 0.04, 0.06 and 0.02 for σ^2_v in Landrace, Yorkshire and their crossbreds, respectively. Then, for each of the QTL locus, two 3 by 3 QTL covariance matrices (one for additive a and one for dominance d^*) were built using the estimated variances of Xiang *et al.* (2016) from above and assuming a correlation between QTL effects (r_{QTL}) of 0.5 across populations as described in the Supplemental Material, S1. The QTL additive and dominance effects ($a^{P1}_{QTL}, a^{P2}_{QTL}, a^{c}_{QTL}$ and $d^{*P1}_{QTL}, d^{*P2}_{QTL}, d^{*c}_{QTL}$) were sampled from two multivariate normal distributions with zero mean and covariance matrices described in the Supplemental

Table 1 Description of simulated scenarios

Scenario	Evaluation model	Source information	Selection criterion within breed	Creation of crossbred animals
S1	GBLUP	PB	EBV_P	RM
S2	GBLUP	PB	EBV_P	MA $_{min}EFI_{ij}$
S3	Trivariate SNP-BLUP	PB and CB	EBV_C	RM
S4	Trivariate SNP-BLUP	PB and CB	EBV_C	MA $_{max}ETGV_{ij}$

EBV_P genomic estimated breeding value on the purebred scale.
EBV_C genomic estimated breeding value on the crossbred scale.
PB purebred.
CB crossbred.
RM random mating.
MA $_{min}EFI_{ij}$ mate allocation strategy that minimizes the average expected genomic inbreeding.
MA $_{max}ETGV_{ij}$ mate allocation strategy that maximizes the average expected total genetic value.

Material, S1. The "overall" value of the QTL dominance effect was obtained summing back the random deviate to its mean as $d = \mu_d + d^*$.

Residual variances ($\sigma^2_{e^{P1}}, \sigma^2_{e^{P2}}, \sigma^2_{e^c}$) were assumed uncorrelated across populations and were adjusted to reach a heritability in narrow-sense (h^2) and the resulting proportions of dominance variance to phenotypic variance (h^2_d). Three cases with different values for h^2 and h^2_d were simulated: (1) $h^2 = 0.1$ and $h^2_d = 0.01$ (as in Xiang *et al.* (2016)), (2) $h^2 = 0.1$ and $h^2_d = 0.1$ (large variation due to dominance), and (3) $h^2 = 0.3$ and $h^2_d = 0.1$ (more heritable trait).

Simulation of true genetic values and phenotype

Each individual in the two parental breeds (P1 and P2) has two true breeding values, one on the PB scale (TBV_P) and one on the CB scale (TBV_C). From the additive and dominance QTL effects, TBV_P were computed according to Falconer (1981) for each individual i within each parental breed. For example for P1 ($TBV^{P1}_{P_i}$) was: $TBV^{P1}_{P_i} = \sum_{j=1}^{nQTL}(z_{ij} - 2p^{P1_f}_{QTL_j})\alpha^{P1}_j$ where α^{P1}_j is the allele substitution effect for the jth QTL, in P1, calculated as $\alpha^{P1}_j = a^{P1}_{QTL_j} + d^{P1}_{QTL_j}(q^{P1}_{QTL_j} - p^{P1}_{QTL_j})$ where $d^{P1}_{QTL_j} = d^{*P1}_{QTL_j} + \mu^{P1}_d$ is the dominance QTL effect from P1 breed including directional dominance μ^{P1}_d in addition to the random deviate $d^{*P1}_{QTL_j}$; and z_{ij} is equal to 2, 1 or 0 when the QTL genotype for animal i is AA, Aa or aa, respectively. The allele frequency $p^{P1_f}_{QTL_j}$ for A was obtained from the founder population in P1, and the allele frequencies $p^{P1}_{QTL_j}$ and $q^{P1}_{QTL_j}$ for A and a, respectively, were computed at each generation.

The TBV_C for one breed (*e.g.*, P1 $TBV^{P1}_{C_i}$) depends on the allele frequencies of the other breed (P2 with allele frequencies $p^{P2}_{QTL_j}$, $q^{P2}_{QTL_j}$). For the pure breed P1, the substitution effect is $\alpha^{P1}_{C_j} = a^c_{QTL_j} + (q^{P2}_{QTL_j} - p^{P2}_{QTL_j})d^c_{QTL_j}$, whereas for P2 this is $\alpha^{P2}_{C_j} = a^c_{QTL_j} + (q^{P1}_{QTL_j} - p^{P1}_{QTL_j})d^c_{QTL_j}$ where $\alpha^{P1}_{C_j}$ is the additive effect of the gametes from P1 when crossed with P2, and $\alpha^{P2}_{C_j}$ is the additive effect of gametes from P2 when crossed with P1 (Vitezica *et al.* 2016). Note that here, additive ($a^c_{QTL_j}$) and dominance ($d^c_{QTL_j}$) QTL effects are from the CB population. Then, the breeding value on the CB scale for each individual i in P1 was computed as $TBV^{P1}_{C_i} = \sum_{j=1}^{nQTL}(z_{ij} - 2p^{P1}_{QTL_j})\alpha^{P1}_{C_j}$, with a similar expression for P2 but using $p^{P2}_{QTL_j}$ and $\alpha^{P2}_{C_j}$ instead.

The true total genetic value ($TTGV$) was computed for CB animals. For instance, for an individual i in CB its $TTGV$ was computed as:

$$TTGV_i = \sum_{j=1}^{nQTL} z_{ij}a^c_{QTL_j} + \sum_{j=1}^{nQTL} z_{ij}(2 - z_{ij})d^c_{QTL_j}$$

Note that $z_{ij}(2 - z_{ij})$ in the second term of the equation is equal to 0 or 1 for homozygous and heterozygous genotypes, respectively. The phenotype of the trait was calculated by adding a general mean equal to 10 and a random error to the total genetic value of each CB animal. Only females had records (like for litter size).

Scenarios and prediction models

Four scenarios (S1, S2, S3 and S4) of selection were simulated. They differed in the sources of information used for genomic evaluation, in selection criteria within-breed and in mate allocation strategies used to create the CB descendants. Table 1 shows a short description of these scenarios. Scenarios S1 and S2 can be considered "classical" schemes since only purebred information was used to evaluate the selection candidates within each breed (P1 and P2). The evaluation model was an univariate GBLUP including additive genetic effects and genomic inbreeding. Only PB information (genotypes and phenotypes) was considered, and evaluations were performed within each parental breed. The selection criterion of PB animals was the genomic estimated breeding value on the PB scale (EBV_P). The CB populations were generated using either random mating (RM) (S1) or a mate allocation strategy (S2) that minimizes the average expected genomic inbreeding (EFI_{ij}). Scenarios S3 and S4 were "combined" schemes because they use PB and CB information (genotypes and phenotypes) and additive and dominance effects to perform the genetic evaluation. In these two scenarios, we used a multivariate model based on "biological" (genotypic) additive and dominance effects of SNPs and including genomic inbreeding. Both PB (P1, P2) and CB performances were treated as different but genetically correlated traits. The selection criterion of PB animals was the estimated genomic breeding value on the CB scale (EBV_C). To create the CB population, S3 used RM, whereas S4 used a mate allocation strategy that maximized the average expected total genetic value ($ETGV_{ij}$) of the CB descendants. Genomic evaluation models for all scenarios are detailed in Supplemental Material, S2.

As explained before, the 4 scenarios were tested across three cases of genetic parameters: (1) $h^2 = 0.1$ and $h^2_d = 0.01$, (2) $h^2 = 0.1$ and $h^2_d = 0.1$, and (3) $h^2 = 0.3$ and $h^2_d = 0.1$. In all cases, the r_{QTL} across loci for all pairs of populations (P1, P2 and CB) of functional additive and dominance effects was 0.5 which resulted in r_{PC} (*i.e.*, $cor(TBV_P, TBV_C)$) of 0.46, 0.30 and 0.42 (cases 1 to 3) in the founders generation. These r_{PC} are the average of the two parental breeds, but values of r_{PC} were very similar for P1 and P2.

Three extra scenarios were considered. Scenarios S1 (first) and S3 (second) were evaluated under case 3 considering a r_{QTL} across loci of 0.8 (leading to $r_{PC} = 0.68$). This was also explored to have a situation where there is little GxG or GxE interaction. The third one (S4*) considered S4, but using true QTL effects and genotypes instead of estimated SNP effects and genotypes, only to perform the mate allocation strategy. This gives the upper bound of the possibilities of mate allocation strategy.

Mate allocation strategies

After selecting PB animals at each generation, two mate allocation strategies were tested in order to define the matings between males (from P1) and females (from P2) to create the CB animals. The first strategy (S2), aimed to minimize the average genomic inbreeding of the CB population. This strategy is commonly known as minimum coancestry mating. The second strategy (S4), was focused on maximizing the average total genetic value of the CB population by exploiting non-additive genetic effects. These two strategies were compared with the random mating used in S1 and S3.

For each of 2448 (12 males × 204 females) possible matings, the expected future inbreeding and the expected total genetic value of CB progeny were calculated. The expected future inbreeding (EFI_{ij}) of the progeny from a mating between the ith male (from P1) and the jth female (from P2), was calculated as the expected proportion of homozygous SNP loci across the genome, as follows:

$$EFI_{ij} = \sum_k \left[P_{ijk}(AA) + P_{ijk}(aa)\right] \Big/ N$$

where $P_{ijk}(AA)$ and $P_{ijk}(aa)$ are the probabilities of homozygous SNP genotypes AA, and aa, at the kth SNP; and N is the total number of SNP. Note that expected future heterozygosity is equal to $1 - EFI_{ij}$,

Table 2 Genetic correlation (r_{PC}) between purebred and crossbred performances in the founder and last generation of selection for each scenario, under three cases of genetic variance component and r_{QTL}= 0.5 and 0.8

Case	h^2	h_d^2	r_{QTL}	r_{PC} in founder generation*	r_{PC} in the last generation			
					S1	S2	S3	S4
Case 1	0.1	0.01	0.5	0.46	0.36	0.36	0.32	0.34
Case 2	0.1	0.1	0.5	0.30	0.15	0.15	0.04	0.03
Case 3	0.3	0.1	0.5	0.42	0.28	0.28	0.18	0.22
Case 3	0.3	0.1	0.8	0.68	0.44		0.36	

* It is the same in all scenarios.
h^2 heritability in narrow-sense.
h_d^2 is the proportion of dominance variance to phenotypic variance.
r_{QTL} is the correlation at the QTL level between purebred (P1, P2) and crossbred populations.
Scenarios: purebred selection is based on estimated breeding values on the purebred scale and the crossbreds are generated by either random mating (**S1**) or with mate allocation to minimizing the expected genomic inbreeding (**S2**); purebred selection is based on estimated breeding value on the crossbred scale accounting for dominance and the crossbreds are generated either by random mating (**S3**) or with mate allocation to maximize the average expected total genetic value of the progeny (**S4**).

so minimizing genomic inbreeding is equivalent to maximizing heterozygosity.

The expected total genetic value ($ETGV_{ij}$) of the CB progeny from the same mating was predicted according to Toro & Varona (2010), as follows:

$$ETGV_{ij} = \sum_k \left[P_{ijk}(AA)\hat{a}_k + P_{ijk}(Aa)\hat{d}_k + P_{ijk}(aa)(-\hat{a}_k)\right]$$

where $P_{ijk}(Aa)$ is the probability of SNP genotype Aa from the progeny of mating ij at the kth SNP; $\hat{a}_k$ and $\hat{d}_k$ are the additive and dominance estimated effects of the kth SNP estimated in the CB. Again, $\hat{d}_k = \hat{d}_k^* + \hat{\mu}_d$ includes inbreeding depression (or heterosis) in the form of μ_d.

From all possible combinations of matings, we selected the best 204 matings that either minimized *EFI* (S2) or maximized *ETGV* (S4) of the CB descendants, depending on which mate allocation strategy was used. Optimization of matings was addressed via linear programming (Jansen and Wilton 1985) using the R (R-Core Team 2017) *lpsolve* package (Berkelaar *et al.* 2004). Two constraints were used in the optimization: (1) each male was mated to 17 females, and (2) each female could not be mated to more than one male. For instance, the linear programming function for EFI_{ij} was defined as:

$$f_{min}\left(EFI_{ij}\right) = \sum_{i=1}^{nm}\sum_{j=1}^{nf} EFI_{ij}\times\theta_{ij},$$

where θ_{ij} are binary variables of decision, where 1 and 0 means that the mating between male i and female j is selected or not selected, respectively. Constrains for male i can be written as: $\theta_{i1} + \theta_{i2} + \theta_{i3} + \ldots + \theta_{i,nf} = 17$ $(i = 1, 2, \ldots, nm)$, and for female j: $\theta_{1j} + \theta_{2j} + \theta_{3j} + \ldots + \theta_{nm,j} = 1$ $(j = 1, 2, \ldots, nf)$, where nm and nf are the number of males (12) and females (204), respectively.

Scenarios were compared in terms of response to selection. The mean *TTGV* (or *TBV*) of CB (PB) animals was computed for each generation expressed relative to the mean *TTGV* (or *TBV*) at the generation 0 to evaluate the realized cumulative response to selection. Increased TTGV is primarily of interest, whereas TBV of PB animals is of secondary interest, meaning that although the most important objective in the crossbreeding scheme is boosting CB performance, PB lines need to be good enough to ensure its viability (*e.g.*, PB animals need a large litter size to produce enough dams for the next generations). Heterosis (H) was measured in the first and last generation for S1 and S3 in all cases of genetic variance component, as: $H = \sum_{j=1}^{nQTL} d^c_{QTL_j}(p^{P1}_{QTL_j} - p^{P2}_{QTL_j})^2$ (Falconer 1981). The realized genomic inbreeding was also calculated for the three populations. Results were the average of the 10 replicates of each scenario.

Data availability

Programs and simulated data are available at http://genoweb.toulouse.inra.fr/~zvitezic/simuPB-CB_G3. A README file contains a description of the files, codes and programs; and general instructions to run the simulation. Supplemental material available at figshare: https://doi.org/10.25387/g3.12504638.

RESULTS

Genetic correlation between PB and CB

The r_{PC} values (*i.e.*, $cor(TBV_P, TBV_C)$) in the founder and last generation of selection are presented in Table 2, for the four scenarios and all cases of genetic variance components. The r_{PC} was the same in the founder generation for all scenarios within each case. In all cases and scenarios, r_{PC} values in the last generation were lower than those in the founder generation.

As selection proceeds, the r_{PC} value decreases over generations because the difference in allele frequencies between breeds increases (Wientjes and Calus 2017; Duenk *et al.* 2020). For instance, in case 2, the average absolute difference in allele frequencies between breeds in the founder generation was 0.21 in both S1 and S3, whereas in the last generation of selection it was 0.36 and 0.38 in S1 and S3, respectively. The most important reduction in r_{PC} was observed in case 2 when PB animals were selected based on EBV_C (from 0.30 to 0.04). This result agrees with Duenk *et al.* (2020) that showed that an increase in magnitude of dominance (as in case 2) results in a reduction in r_{PC}.

Response to selection in crossbred performance

Figure 4 shows the mean *TTGV* of the CB animals accumulated across generations for the four scenarios (S1, S2, S3, S4), for case 1 $h^2 =$ (0.1 and h_d^2 =0.01) and $r_{QTL} = 0.5$. Similar results were observed for the other cases of genetic variance components ($h^2 = 0.1$ and $h_d^2 = 0.1$, $h^2 = 0.3$ and $h_d^2 = 0.1$), they were not plotted here. Results (in Figure 4) show that scenarios S3 and S4, where PB and CB information was used and PB animals were selected on EBV_C, clearly outperform scenarios S1 and S2, where only PB information was used and the selection criteria was based on EBV_P, regardless of whether a mate allocation strategy was used or

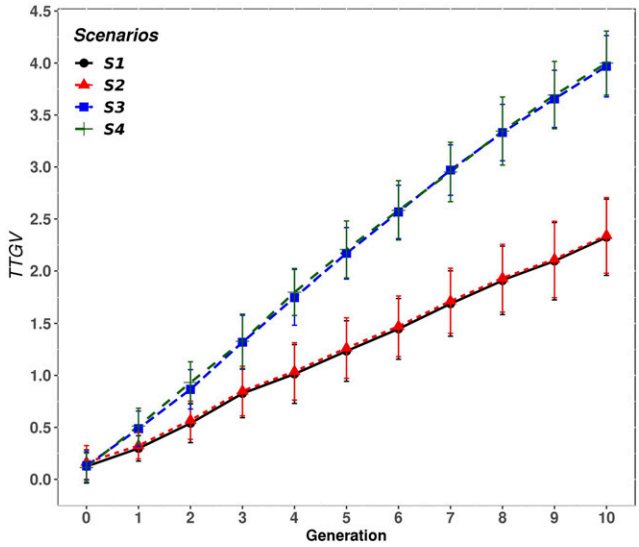

Figure 4 Accumulated response to selection (*TTGV*) of the crossbred animals for each generation and scenario expressed relative to generation 0, for case 1 ($h^2 = 0.1$ and $h_d^2 = 0.01$) and $r_{QTL} = 0.5$. Scenarios: purebred selection is based on estimated breeding values on the purebred scale and the crossbreds are generated by either random mating (S1) or with mate allocation to minimizing the expected genomic inbreeding (S2); purebred selection is based on estimated breeding value on the crossbred scale accounting for dominance and the crossbreds are generated either by random mating (S3) or with mate allocation to maximize the average expected total genetic value of the progeny (S4).

not. This advantage in S3 and S4 over S1 and S2 was observed from the first generations and the gap increased through generations.

The mean *TTGV* of CB animals accumulated in the last generation for each scenario in all cases of genetic variance components are presented in Table 3. The improvement in CB performance due to the selection criteria used (EBV_C instead of EBV_P) was observed by contrasting S3 *vs.* S1. For $r_{QTL} = 0.5$ the improvement was equal to 71%, 271% and 123%, for cases 1, 2 and 3, respectively (Table 3). An increase in heritability (*i.e.*, case 2 *vs.* case 3) resulted in a higher response to selection in CB performance in all scenarios, whereas the increase of dominance variance (*i.e.*, case 1 *vs.* case 2) was more advantageous for S3 and S4 than for S1 and S2. In fact, in S3 and S4, the proportion of improvement in CB performance was roughly tripled in case 2 (271%), where the ratio between dominance and additive variances was equal to 1 ($h^2 = 0.1$ and $h_d^2 = 0.1$) and r_{PC} (0.30) was lower than the other cases. Thus, the benefit of using both PB and CB information and select PB animals based on EBV_C (*i.e.*, S3 and S4) improved with the ratio dominance/additive variance. In scenarios where only PB information was used and the selection criteria was based on EBV_P (*i.e.*, S1 and S2), some proportion of the additive genetic gain obtained on the PB breeds is transmitted to the CB progeny depending on the purebred-crossbred genetic correlation r_{PC}. For instance, in S1 and S2, where only PB information was used, a reduction in the mean *TTGV* of CB animals was observed when h_d^2 was increased from 0.01 to 0.1 and h^2 was held at 0.1 (*i.e.*, case 1 *vs.* case 2). This is because, under case 2, the increase in the proportion of dominance variance lowers the r_{PC} from 0.46 to 0.30 in the founder generation (Table 2). This reduction in r_{PC}, in turn, reduces the proportion of genetic gain that is transmitted from PB parents to their crossbred descendants. However, when h^2 was increased from 0.1 to 0.3 and h_d^2 was held at 0.1 (*i.e.*, case 2 *vs.* case 3), and the r_{PC} was 0.42, the mean *TTGV* of CB animals was higher compared to the case 1. These results show the importance of including CB information in the model to evaluate PB animals for CB performance, especially in cases where additive effects are low and dominance effects are relevant (like in case 2).

A correlation (at the QTL level) between PB and CB populations (r_{QTL}) equal to 0.5 was assumed in the results mentioned earlier, which resulted in relatively low r_{PC} correlations of 0.3 – 0.46. For case 3, a $r_{QTL} = 0.8$, resulting in $r_{PC} = 0.68$ was also examined. Figure 5 compares the effect of increasing the r_{QTL} from 0.5 to 0.8 on the mean *TTGV* of CB animals for S1 and S3 evaluated under case 3. For $r_{QTL} = 0.8$ ($r_{PC} = 0.68$) and case 3, the CB performance accumulated in the last generation for S1 and S3 were higher compared when $r_{QTL} = 0.5$. The CB performances, accumulated in the last generations, were 2.48 and 3.97 in S1, and 5.54 and 5.72 in S3, for r_{QTL} equal to 0.5 and 0.8, respectively. Thus, the increase in r_{QTL} was reflected in an improvement in the CB performance in both scenarios, S1 and S3, but the improvement was greater in S1, where the selection was based on EBV_P. The advantage in CB performance of S3 over S1 was reduced from 123% at $r_{QTL} = 0.5$ to 44% at $r_{QTL} = 0.8$.

These results show, on one hand, that the proportion of the genetic progress that is transmitted from the parental breeds to their CB descendants, by selecting PB on EBV_P, depends on the value of r_{PC}, and on the second hand, the use of crossbred information to evaluate the PB for CB performance greatly helps the genetic improvement of the CB, especially if r_{PC} is low.

Change in genomic inbreeding and heterosis in purebreds and crossbreds

Figure 6 shows the average, across animals, of genomic inbreeding (proportion of observed homozygosity) in P1 and in the CB population, for the four scenarios under case 1. Similar results were obtained for P2 and for the other cases (not shown). The genomic inbreeding of the two parental populations was around 0.62 in generation 0. The increase in genomic inbreeding per generation in the two parental breeds was almost the same in all scenarios, regardless of whether the selection criteria was in EBV_P (S1 and S2) or EBV_C (S3 and S4). For instance, the genomic inbreeding in P1 in the last generation was around 0.80 in all scenarios (Figure 6 A).

In the CB population, the initial average genomic inbreeding was around 0.56, which reflects a reduction of 0.06 in genomic inbreeding compared to the parental breeds (0.62 in PB – 0.56 in CB). Regardless of the scenarios and cases under investigation, genomic inbreeding remained almost the same across generations in the CB population. In scenarios where mate allocation strategies were implemented, a very small reduction in the genomic inbreeding was observed (Figure 6 B).

Estimates of inbreeding depression (or heterosis if the sign is changed) were close to the simulated value (-10 per 100% homozygosity) in the two parental breeds and lower in CB population. For instance, inbreeding depression was estimated in case 3 (standard deviation), as -13.3 (1.05) and -10.6 (1.0) for S1 and S3, respectively, in P1, and -5.94 (2.1) for CB in S3.

Heterosis was also computed at the first and last generations from QTL frequencies and the dominance effect. Both scenarios (S1 and S3) had the same initial amount of heterosis, 0.71, 0.75 and 0.73 for cases 1 to 3, respectively. The heterosis in the last generation for S1 was 2.24, 1.88 and 1.99, and for S3 was 2.49, 5.77, 3.66, for cases 1 to 3, respectively. The absolute value of the QTL frequency differences $\left|(p_{QTL_j}^{P1} - p_{QTL_j}^{P2})\right|$ in the last generation were 0.367, 0.359 and 0.357 for

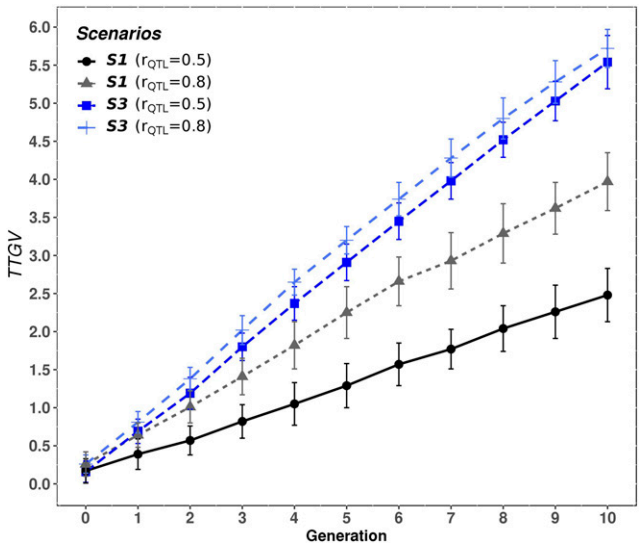

Figure 5 Accumulated response to selection in crossbred performance (*TTGV*) for scenarios S1 and S3 at different r_{QTL} (0.5 and 0.8) evaluated under case 3 ($h^2 = 0.3$ and $h_d^2 = 0.1$). In S1, purebred selection is based on estimated breeding values on the purebred scale and the crossbreds are generated by random mating. In S3, purebred selection is based on estimated breeding value on the crossbred scale accounting for dominance and the crossbreds are generated by random mating.

S1 and 0.372, 0.379 and 0.373 for S3 in the last generation. Differences in heterosis are due to differences in QTL allele frequencies (between scenarios) and QTL dominance effects (among cases). Results showed that selecting PB on EBV_C produced higher heterosis than selection on EBV_P, and it was higher when the ratio between dominance and additive variances was equal to 1 (*i.e.*, case 2).

Effect of mate allocation strategies

Scenarios that differed in the use or not of mate allocation, but shared similar genomic evaluation model and selection criterion (*e.g.*, S2 *vs.* S1) were contrasted in order to measure the effect of using a mate allocation strategy. The advantage of using mate allocation strategies (S2 and S4) to improve CB performance was negligible compared to random mating (S1 and S3). On one hand, S2 based on minimizing the average expected genomic inbreeding of the CB progeny produced a small improvement on the CB performance (0.9, 2.2 and 0.8% for cases 1 to 3, respectively) compared to random mating (S1) (Table 3). This is probably because a very small reduction (0.5% in all cases) in the average genomic inbreeding (or increase in individual heterosis) of the CB population was observed in S2 compared to S1 (Figure 6). On the other hand, S4 based on maximizing the average expected total genetic value of the CB progeny, produced a slight increase in the CB performance (0.8, 4 and 0.5% for cases 1 to 3, respectively) compared to random mating (S3) (Table 3). The most important advantage of mate allocation was observed in S4 (4%) and in case 2, where the ratio dominance/additive variance was higher.

In order to get more insight on the mate allocation strategies, we created an extra scenario by replicating the S4 but instead of using the SNP genotypes and estimated SNP effects, we used the genotypes and effects of QTL to perform the mate allocation strategy (S4*). This situation represents the maximum achievable gain in CB performance by implementing this mate allocation strategy. Figure 7 shows the mean *TTGV* of CB animals accumulated across generations for S3, S4, and the extra scenario S4* where, to simplify, we only presented cases 1 and 3. The results show that assuming the QTL information as known (S4*), an increase of 0.13 and 0.24 in the CB performance was observed in the first generation in both cases, however, that improvement was gradually decreasing across generations of selection. Thus, in the last generation, the CB performance obtained in S4* (4.06, 5.29 and 5.56, for case 1, 2 and 3, respectively) was very similar to those obtained using the SNP effects and genotypes to perform mate allocation (S4, Table 3). Even when h^2 increased from 0.1 to 0.3 (case 3 *vs.* case 1) (Figure 7), the CB performance obtained in S4 was closer to that obtained in S4* (by using QTL effects), in the last generations. These results suggest that, even if the QTL genotypes and effects are known and used to perform mate allocation, the improvement in CB performance would be only in the first few generations of selection when the selection criteria is based on EBV_c.

Response to selection in purebred animals

Figure 8 illustrates the accumulated response to selection (TBV_P) of PB animals from P1 and P2 across generations, for the four scenarios, case 3 (h^2=0.3 and h_d^2=0.1) and $r_{QTL} = 0.5$. Since results were similar for the other cases of genetic variance components, they were not plotted. When PB animals were selected based on EBV_C, *i.e.*, S3 and S4, the accumulated response to selection was clearly lower compared to when they were selected based on EBV_P (*i.e.*, S1 and S2). Compared to P1, the accumulated response in P2 was lower in all scenarios. In S1 and S2 this difference between P1 and P2 was due to differences in genetic variances (P2 has smaller genetic variances than P1), whereas in S3 and S4, where CB information was included, this difference was also because the scheme was not symmetric (P2 males had no crossbred daughters, see Figure 3).

The genetic response (TBV_P) for the two parental breeds accumulated in the last generation for the four scenarios under the three cases of genetic variance components are presented in Table 4. Response to selection in PB was the same between S1 and S2, and also changed slightly between S3 and S4 (in case 1) due to the different mate allocation strategies used to produce crossbreds. No genetic response in S3 and S4 was observed in PB when the proportion of dominance variance h_d^2 increased (*i.e.*, case 1 *vs.* case 2). This absence of genetic response can be explained by a r_{PC} equal to 0.04 and 0.03 in the last generation for S3 and S4 respectively (see Table 2). An increase in h^2 (*i.e.*, case 2 *vs.* case 3) resulted in a higher genetic response in P1 and P2 breeds in all scenarios.

Additionally, Figure 9 shows the accumulated genetic response in P1 for two values of r_{QTL} (0.5, 0.8) in S1 and S3 under the case 3 (h^2=0.3 and h_d^2=0.1). The accumulated genetic response in both parental breeds in S1 did not change with r_{QTL}, whereas an important increase was observed in S3 (from 1.13 to 2.84 in P1 and from 0.84 to 2.41 in P2) when r_{QTL} was increased. Thus, the loss in TBV_P by selecting PB animals on EBV_C was reduced by almost half (*e.g.*, from 80 to 42% in P2) when r_{QTL} increased from 0.5 to 0.8. This is as expected – selection for CB performance in S3 and S4 results in a correlated response in PB performance, and the magnitude of the response in PB performance depends on the value of r_{PC}.

DISCUSSION

The purpose of this work was to investigate, in a crossbreeding scheme, genomic evaluation models and the potential benefit of applying mate allocation. Novelties of our work from previous works include: (1) we simulate QTL effects that are correlated, but not identical, across populations, (2) we optimize matings in the crossbreds and (3) we consider in the simulation genetic evaluation using a

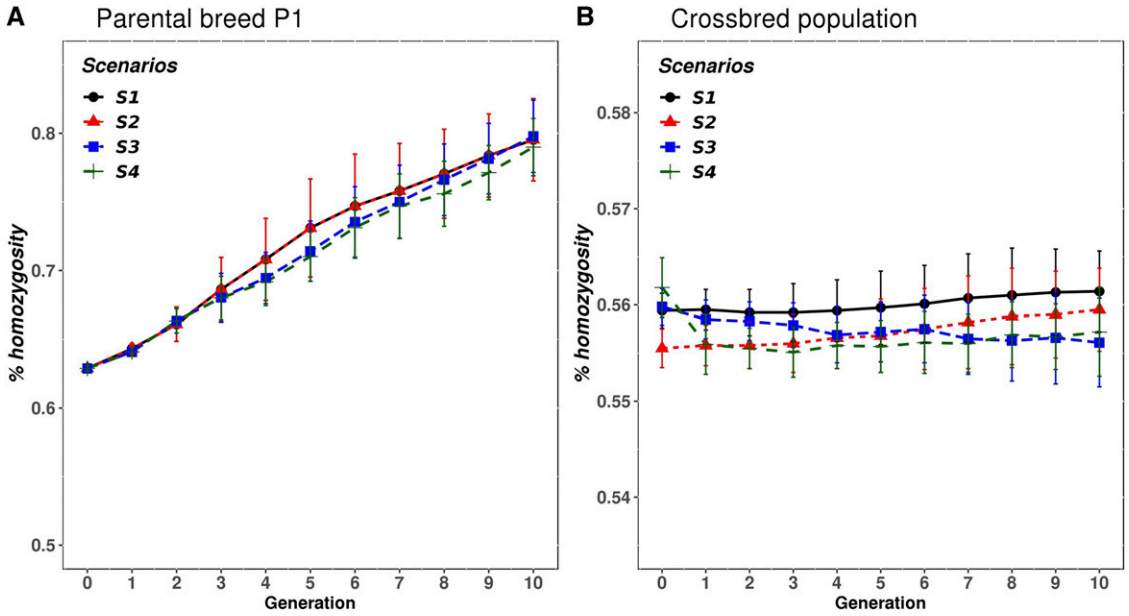

Figure 6 Mean genomic inbreeding for A) purebred animals from parental breed P1, and B) CB animals, for the four scenarios across each generation, for case 1 and $r_{QTL} = 0.5$. Scenarios: purebred selection is based on estimated breeding values on the purebred scale and the crossbreds are generated by either random mating (S1) or with mate allocation to minimizing the expected genomic inbreeding (S2); purebred selection is based on estimated breeding value on the crossbred scale accounting for dominance and the crossbreds are generated either by random mating (S3) or with mate allocation to maximize the average expected total genetic value of the progeny (S4).

multiple-trait SNP-BLUP across the three populations (purebreds and crossbreds) including additive and dominance effects.

To date, breed-specific QTL effects correlated across PB and CB have not been accounted for in simulations. Our results show, according to the literature, that a genomic evaluation model that explicitly includes dominance and phenotype and genotype data of PB and CB animals improves CB performance faster than a model considering PB data only, in particular for low values of r_{pc}. Finally, using mate allocation provides a negligible extra response in CB performance, even for cases where dominance variance was high.

All selection steps were within parental PB and mate allocation was only implemented to create CB population. Four scenarios that combined different sources of information (PB or PB and CB), selection criteria and mate allocation strategies were evaluated

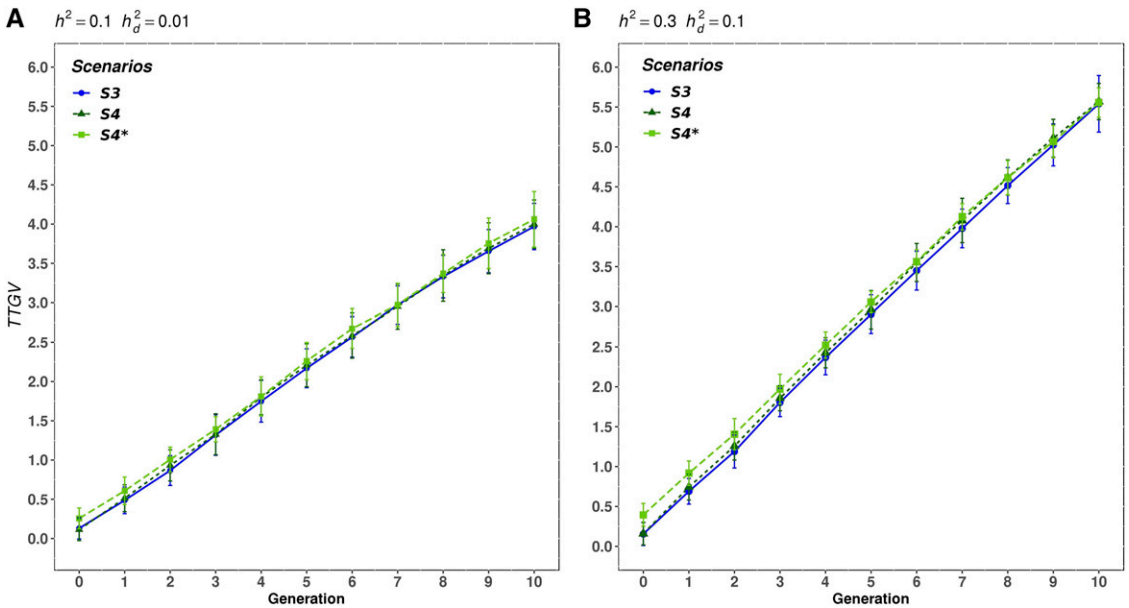

Figure 7 Mean true total genetic value of the CB animals for each generation for scenarios 3, 4, and the extra scenario (S4*), for A) case 1, B) case 3. Scenarios: purebred selection is based on estimated breeding value on the crossbred scale accounting for dominance and the crossbreds are generated by random mating (S3) or by mate allocation to maximize the average expected total genetic value of the progeny (S4). S4* uses the genotypes and effects of QTL, instead of the estimated SNP effects (as in S4), to implement the mate allocation strategy.

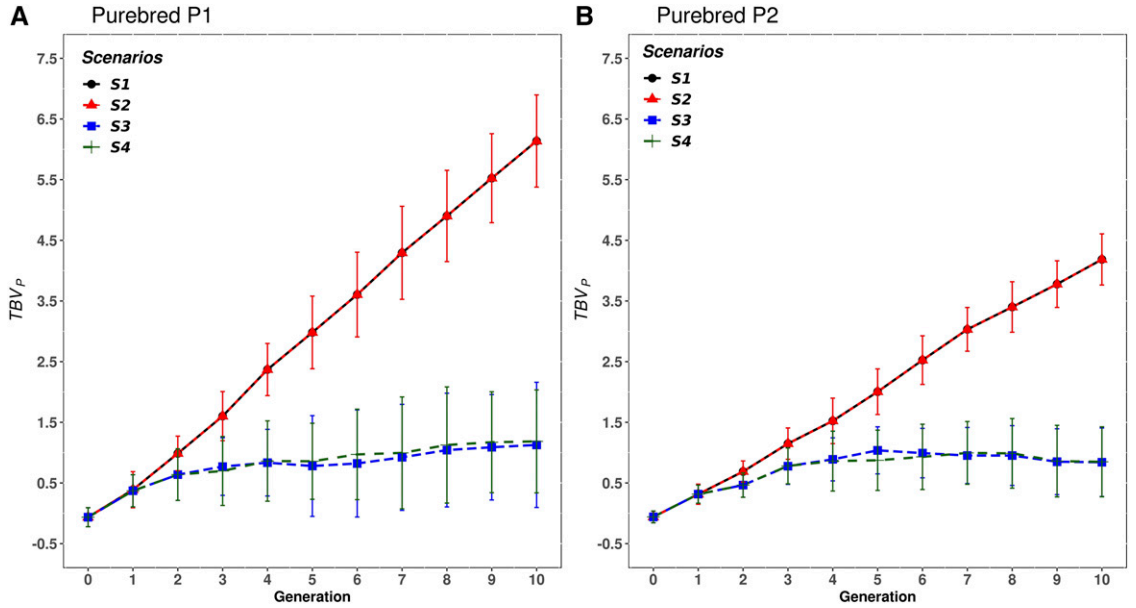

Figure 8 Mean true breeding value (TBV_P) of the purebred animals from population A) P1 and B) P2, for each scenario and generation, for case 3 ($h^2 = 0.3$ and $h_d^2 = 0.1$) and a correlation at the QTL level of $r_{QTL} = 0.5$. Scenarios: purebred selection is based on estimated breeding values on the purebred scale and the crossbreds are generated by either random mating (S1) or with mate allocation to minimizing the expected genomic inbreeding (S2); purebred selection is based on estimated breeding value on the crossbred scale accounting for dominance and the crossbreds are generated either by random mating (S3) or with mate allocation to maximize the average expected total genetic value of the progeny (S4).

(over 10 generations) under different values of genetic variance components.

Comparison of selection criteria

Our results show that selecting PB animals based on the EBV_C produced higher CB performance compared to selection on the EBV_P, but at the cost of reducing the genetic response in the parental breeds. Our results agree with previous studies reported in the literature (Dekkers 2007; Esfandyari *et al.* 2015a; Wientjes and Calus 2017). Thus, choosing the selection criteria depends on the breeding objectives and the r_{PC} correlation. If the PB trait is of interest, it is possible to weight both traits (PB and CB performance) in a selection index (*e.g.*, Esfandyari *et al.* (2018)).

The selection criteria (selecting on EBV_P or on EBV_C) differ in how the allele substitution effects are calculated. For calculating the EBV_C of PB, the allele substitution effects were computed using the SNP effects estimated for the CB population from the trivariate model. Thus, the effects of PB alleles are estimated on the genetic background and environment where they will be expressed in (Dekkers 2007; Ibáñez-Escriche *et al.* 2009). Moreover, as the trivariate SNP-BLUP model provides estimates for both additive and dominance effects, it allows calibrating the substitution effect of one parent breed based on the allele frequencies of the opposite breed, with which it will be crossed, and conversely (Falconer 1981; Zeng *et al.* 2013; Esfandyari *et al.* 2018). The resulting allele substitution effects are breed-specific and can be interpreted as additive genetic effects of the gamete from one breed when crossed with the opposite breed (Vitezica *et al.* 2016). Previous studies in the literature reported an extra response in CB performance when allele substitution effects were estimated using additive and dominance effects and calibrated with the allele frequencies of the opposite breed, regardless of whether the training was based on PB or CB data (Zeng *et al.* 2013; Esfandyari *et al.* 2015b 2018).

Furthermore, the conventional EBV_P of PB were estimated from performances and genotypes of each parental breed, which are supposed to be under the influence of the nucleus environment where PB animals are raised. Furthermore, allele frequencies observed within the breed were used in the genetic evaluation. Thus, the resulting estimates of EBV_P are relevant to improve the PB performances in the nucleus environment (Dekkers 2007). However, the genetic response obtained within PB populations is only partially

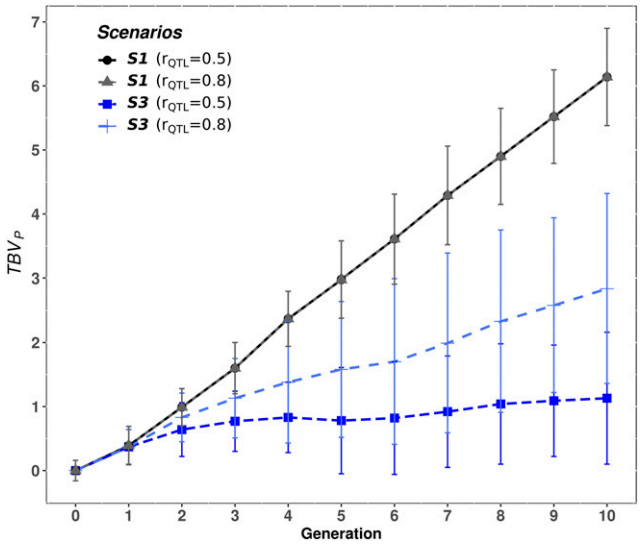

Figure 9 Accumulated response to selection (TBV_P) in purebred P1 for scenarios S1 and S3, at different values of correlation at QTL level (r_{QTL} equal to 0.5 and 0.8) evaluated under case 3 ($h^2 = 0.3$ and $h_d^2 = 0.1$). In S1, purebred selection is based on estimated breeding values on the purebred scale and the crossbreds are generated by random mating. In S3, purebred selection is based on estimated breeding value on the crossbred scale accounting for dominance and the crossbreds are generated by random mating.

Table 3 Mean true total genetic value (standard deviation) of crossbred animals at the last generation for each scenario, under three cases of genetic variance component and r_{QTL}= 0.5

Case	h^2	h_d^2	r_{PC}	S1	S2	S3	S4
Case 1	0.1	0.01	0.46	2.32 (0.37)	2.34 (0.36)	3.97 (0.30)	4.00 (0.31)
Case 2	0.1	0.1	0.30	1.35 (0.40)	1.38 (0.41)	5.01 (0.40)	5.21 (0.44)
Case 3	0.3	0.1	0.42	2.48 (0.35)	2.50 (0.35)	5.54 (0.35)	5.57 (0.23)

h^2 heritability in narrow-sense.
h_d^2 is the proportion of dominance variance to phenotypic variance.
r_{PC} is the average purebred-crossbred genetic correlation of the two parental breeds calculated in the founder generation.
Scenarios: purebred selection is based on estimated breeding values on the purebred scale and the crossbreds are generated by either random mating (**S1**) or with mate allocation to minimizing the expected genomic inbreeding (**S2**); purebred selection is based on estimated breeding value on the crossbred scale accounting for dominance and the crossbreds are generated either by random mating (**S3**) or with mate allocation to maximize the average expected total genetic value of the progeny (**S4**).

transmitted to the CB progeny according to the PB-CB genetic correlation r_{PC}. Therefore, the r_{PC} is a relevant parameter for breeders to decide if data at the CB commercial level should be collected.

Allele substitution effects are a function of allele frequencies and additive and dominance effects (Falconer 1981). If there are QTL that expresses overdominance in CB, the performance of CB animals will be maximized when opposite alleles are fixed in both parental breeds (Esfandyari *et al.* 2018). In this study, the proportions of QTL that exhibited overdominance in the CB population were 15, 38 and 25% in case 1 to 3, respectively. For these QTL, alternate alleles will tend to be fixed in parental populations when selecting PB on EBV_C, but at the same time, alleles that are unfavorable for PB performance will tend to be fixed alternatively in the PB lines. That could explains why selection on EBV_C produced a loss in response to selection in PB. This phenomenon of allele fixation has been well examined in previous simulation studies in animal crossbreeding context (Zeng *et al.* 2013; Esfandyari *et al.* 2018), but also in hybrid breeding in crops, like maize (Technow *et al.* 2014).

Comparison of genetic evaluations models

In our simulation, genetic evaluation integrated PB and CB phenotypes and genotypes by using a trivariate SNP-BLUP model with additive and dominance effects (*e.g.*, S3). This model explicitly distinguishes between PB and CB data by modeling additive and dominance marker effects correlated among populations, therefore, the resulting estimated SNP effects are breed-specific. As dominance is involved in heterosis, its inclusion in the model was expected to be more efficient than pure additive models when PB animals are selected for CB performance, as it was demonstrated by Zeng *et al.* (2013) and Xiang *et al.* (2016). Previous studies in the literature and in a genomic context, based on simulation or real data, have never investigated simultaneously the effect of accounting for PB and CB data and additive and dominance effects on genetic gains in the long term (Dekkers 2007; Esfandyari *et al.* 2015b, 2015a 2018; Lopes *et al.* 2017). All these authors concluded that training on CB increases CB performance compared to training on PB separately or in a combined way. Nevertheless, none of them used all sources of information in the genomic evaluation. To our knowledge, only Xiang *et al.* (2016) used both PB and CB information and estimated additive and non-additive genetic effects on real data. One reason for the lack of studies even with simulation is its complexity.

Furthermore, inbreeding was taken into account (as a covariate) in genetic evaluation models, to correctly estimate dominance variance and heterosis / inbreeding depression, but also because it produces an improvement in the prediction of breeding values (Xiang *et al.* 2016). Another advantage of the SNP-based model is that solutions are SNP effect estimates which can be used directly into the mate allocation algorithms. Furthermore, we assumed as known the parameters of correlation of SNP effects (identical to the simulated correlation at QTL), however, these can be estimated from data.

If the objective is to improve the performance of CB, the trivariate genetic evaluation model is the optimal approach to select PB based on EBV_c. One of the limits of this approach is that its implementation requires phenotypic and genotypic data collected at the commercial CB level, which is not commonly available due to logistics and costs. Genotyping cost is still high to have CB females genotyped in routine but the decreasing trend may open some opportunities. Alternatively, to avoid recording phenotypes and genotypes of CB in routine, marker effects can be estimated using phenotypes and genotypes from a random sample of CB. These estimated SNP effects can be used for a few generations of selection (Dekkers 2007), but at a cost of reducing the selection response in CB performance due to loss of LD between SNP and QTL (Toro and Varona 2010; Esfandyari *et al.* 2015a).

Genomic inbreeding in purebreds and heterosis in crossbreds

A substantial increase in genomic inbreeding in PB was observed in all scenarios. The main causes of this can be attributed to the small number of selected males per generation (12) in our study, and because in absence of information (*e.g.*, maternal scheme context), *EBVs* tend to be shrunk toward family means, so that related individuals are selected. Note that we used random mating to produce next generations within each PB. However, minimum coancestry mating can be used to manage PB inbreeding.

The reduction in genomic inbreeding (or increase in heterozygosity) (0.06) in CB animals respect the parental populations (0.62 in PB – 0.56 in CB), shows the benefit of crossbreeding schemes for exploiting the phenomenon of heterosis, even if PB inbreeding increases over generations. However, the amount of heterosis across generations of selection on PB, can vary depending on the selection criterion. In the last generation, the amount of heterosis was higher when PB animals were selected on EBV_C than on EBV_P. These results have also been observed in previous simulation studies (Zeng *et al.* 2013; Esfandyari *et al.* 2015a).

Mate allocation strategies

The effectiveness of mate allocation strategies accounting for non-additive genetic effects to improve CB performance was evaluated. On the one hand, minimizing the expected future inbreeding of the progeny (*i.e.*, S2) does not seem a promising mate allocation strategy to boost CB performance. A slight reduction (0.5% for the three cases) in the CB genomic inbreeding was observed compared to random mating (S1) which resulted in an almost negligible improvement of

Table 4 Mean true breeding value (standard deviation) of purebred animals in the purebred scale (TBV_P) for the two parental breeds, at the last generation and for each scenario and all cases of genetic variance components and r_{QTL} = 0.5

Purebred	Case	h^2	h_d^2	r_{PC}	S1	S2	S3	S4
P1	Case 1	0.1	0.01	0.46	4.30 (0.82)	4.30 (0.82)	1.06 (0.48)	1.35 (0.58)
	Case 2	0.1	0.1	0.30	3.25 (0.82)	3.25 (0.82)	−2.17 (1.37)	−2.30 (1.09)
	Case 3	0.3	0.1	0.42	6.14 (0.76)	6.14 (0.76)	1.13 (1.03)	1.19 (0.85)
P2	Case 1	0.1	0.01	0.46	3.10 (0.45)	3.10 (0.45)	1.32 (0.60)	0.93 (0.62)
	Case 2	0.1	0.1	0.30	1.42 (0.74)	1.42 (0.74)	−1.99 (0.57)	−2.04 (0.87)
	Case 3	0.3	0.1	0.42	4.19 (0.42)	4.19 (0.42)	0.84 (0.56)	0.85 (0.58)

h^2 heritability in narrow-sense.
h_d^2 proportion of dominance variance to phenotypic variance.
r_{PC} is the average purebred-crossbred genetic correlation of the two parental breeds calculated in the founder generation.
Scenarios: purebred selection is based on estimated breeding values on the purebred scale and the crossbreds are generated by either random mating (**S1**) or with mate allocation to minimizing the expected genomic inbreeding (**S2**); purebred selection is based on estimated breeding value on the crossbred scale accounting for dominance and the crossbreds are generated either by random mating (**S3**) or with mate allocation to maximize the average expected total genetic value of the progeny (**S4**).

CB performance across generations. During the optimization process of the mate allocation, it was observed that the distribution of the expected future inbreeding from all the potential matings between the selected parents, had a very small standard deviation (*e.g.*, 0.558 ± 0.005 for case 3 in generation 1). Due to this small variation in the inbreeding, the reduction in realized genomic inbreeding achieved by this mate allocation strategy was very small. On the other hand, maximizing the average expected total genetic value of the CB (S4) through mate allocation might promote SNP heterozygosity, especially in those regions where there are QTL with favorable dominance effects and, hence, increase CB performance. The CB performance in S4 was slightly better compared to random mating in S3, except in case 2 where the improvement was 4% in the last generation. These results show that the benefit of implement mate allocation is better when the ratio of dominance to additive variances is higher. That was also observed by Toro and Varona (2010).

Furthermore, when QTL information was assumed as known, the increase in CB performance was notable in the first generations, but then, the improvement was gradually decreasing until reaching a CB performance close to that of S4 in the last generation. Thus, the potential improvement of mate allocation decrease in the long term when PB animals are selected on EBV_C. One reason that could explain such gradual decrease in CB performance is because opposite alleles that increase the proportion of favorable heterozygotes in the CB are already fixed in both parental breeds when PB animals are selected on EBV_C. Thus, the number of loci with favorable heterozygotes that are potentially optimizable is reduced over the generations of selection on EBV_C. Toro and Varona (2010) also found that the advantage of mate allocation over random mating disappears in subsequent generations of selection, but in a PB population.

In our study, a larger genome with 18 chromosomes of 120 cM each and 2500 QTL was simulated compared to other studies in the literature (*e.g.*, Toro and Varona 2010; Esfandyari *et al.* 2015a). Our more realistic genetic architecture led to very small effects of QTL, especially when genetic variances were small (as in case 1), in such a way that a large training set is required to estimate genetic effects (Zeng *et al.* 2013; Esfandyari *et al.* 2015a). In addition, our simulation included a maternal trait (only females had phenotype) and CB training population came from previous generations. Hence, in our simulation, it was more difficult to capture QTL effects through SNP than in previous studies, but it supposed to be closer to real situations. Our results agree with those obtained by Toro & Varona (2010) and Fernández *et al.* (2018), in the sense that the success of mate allocation strategies depends on the proportion of additive and dominance variance of the trait and the ability of estimating additive and dominance SNP effects.

The average expected increase in total genetic value of CB using mate allocation in S4, expressed in genetic (additive) standard deviations (SD), were 0.04, 0.21 and 0.13, for case 1, 2 and 3, respectively. Recently, we reported an increase in expected progeny performance of 0.09 SD, for the average piglet weight at birth within litter ($h^2 = 0.36$ and $h_d^2 = 0.04$) (González-Diéguez *et al.* 2019). Other studies in dairy cattle show increases from 0.8 to 0.13 SD and 0.1 to 0.22 SD for milk production and protein yield, respectively, with mate allocation (Ertl *et al.* 2014; Aliloo *et al.* 2017). None of these articles studied the benefits of implementing mating optimizations in the long-term in a crossbreeding scheme like ours. In our simulation, expected increases in CB performance were similar to those reported in the literature in single populations (PB).

In this study, first, PB animals were selected based on breeding values, and second, they were mated following different strategies to produce the CB animals. However, in this way, some matings with high progeny merit can be excluded (Hayes and Miller 2000). Another interesting strategy is mate selection (*i.e.*, deciding at the same time the selected individuals and their mates).

To conclude, our results show that a genomic evaluation model that simultaneously accounts for both PB and CB phenotype and genotype data and additive and dominance effects improves CB performance faster than a model considering PB data only, in particular for low values of r_{pc}. When r_{PC} is low, selecting PB animals for CB performance using CB information is a more efficient strategy to exploit heterosis and increase performance at the CB commercial level. Furthermore, the benefit of mate allocation strategies for response in CB performance was negligible, even for cases where dominance variance was high. In addition, mate allocation implementation is not straightforward and will require some organizational changes (*e.g.*, semen logistic). The cost-benefit of implementing mate allocation is not clear.

ACKNOWLEDGMENTS

DGD is funded by France Génétique Porc, RAGT and INRA (contract 15000416, project 00000837). Work was supported by INRA SELGEN metaprogram (project OptiMaGics). Project was partly supported by Toulouse Midi-Pyrénées bioinformatics platform.

LITERATURE CITED

Aliloo, H., J. E. Pryce, O. González-Recio, B. G. Cocks, M. E. Goddard *et al.*, 2017 Including nonadditive genetic effects in mating programs to

maximize dairy farm profitability. J. Dairy Sci. 100: 1203–1222. https://doi.org/10.3168/jds.2016-11261

Berkelaar, M., K. Eikland, and P. Notebaert, 2004 lpsolve 5.5, Open source (Mixed-Integer) Linear Programming system Software.

Boré, R., L. F. Brito, M. Jafarikia, A. Bouquet, L. Maignel *et al.*, 2018 Genomic data reveals large similarities among Canadian and French maternal pig lines. Can. J. Anim. Sci. 98: 809–817. https://doi.org/10.1139/cjas-2017-0103

Cockerham, C. C., 1969 Variance of Gene Frequencies. Evolution 23: 72–84. https://doi.org/10.1111/j.1558-5646.1969.tb03496.x

Dekkers, J. C. M., 2007 Marker-assisted selection for commercial crossbred performance. J. Anim. Sci. 85: 2104–2114. https://doi.org/10.2527/jas.2006-683

DeStefano, A. L., and I. Hoeschele, 1992 Utilization of Dominance Variance Through Mate Allocation Strategies. J. Dairy Sci. 75: 1680–1690. https://doi.org/10.3168/jds.S0022-0302(92)77925-9

Duenk, P., P. Bijma, M. P. L. Calus, Y. C. J. Wientjes, and J. H. J. van der Werf, 2020 The impact of non-additive effects on the genetic correlation between populations. G3 (Bethesda) 10: 783–795. https://doi.org/10.1534/g3.119.400663

Ertl, J., A. Legarra, Z. G. Vitezica, L. Varona, C. Edel *et al.*, 2014 Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genet. Sel. Evol. 46: 40. https://doi.org/10.1186/1297-9686-46-40

Esfandyari, H., P. Berg, and A. C. Sørensen, 2018 Balanced selection on purebred and crossbred performance increases gain in crossbreds. Genet. Sel. Evol. 50: 8. https://doi.org/10.1186/s12711-018-0379-9

Esfandyari, H., A. C. Sørensen, and P. Bijma, 2015a A crossbred reference population can improve the response to genomic selection for crossbred performance. Genet. Sel. Evol. 47: 76. https://doi.org/10.1186/s12711-015-0155-z

Esfandyari, H., A. C. Sørensen, and P. Bijma, 2015b Maximizing crossbred performance through purebred genomic selection. Genet. Sel. Evol. GSE 47: 16. https://doi.org/10.1186/s12711-015-0099-3https://doi.org/10.1186/s12711-015-0099-3

Falconer, D. S., 1981 *Introduction to quantitative genetics*, Longman Goup Ltd, New York.

Fernández, J., B. Villanueva, and M. A. Toro, 2018 Diseño de apareamientos óptimos para aprovechar la dominancia en esquemas de selección genómica en especies de acuicultura, in *Proccedings*, XIX Reunión Nacional de Mejora Genética Animal, León, Spain. https://acteon.webs.upv.es

Garcia-Baccino, C. A., A. Legarra, O. F. Christensen, I. Misztal, I. Pocrnic *et al.*, 2017 Metafounders are related to Fst fixation indices and reduce bias in single-step genomic evaluations. Genet. Sel. Evol. 49: 34. https://doi.org/10.1186/s12711-017-0309-2

González-Diéguez, D., L. Tusell, C. Carillier-Jacquin, A. Bouquet, and Z. G. Vitezica, 2019 SNP-based mate allocation strategies to maximize total genetic value in pigs. Genet. Sel. Evol. 51: 55. https://doi.org/10.1186/s12711-019-0498-y

Hayes, B. J., and S. P. Miller, 2000 Mate selection strategies to exploit across- and within-breed dominance variation. J. Anim. Breed. Genet. 117: 347–359. https://doi.org/10.1046/j.1439-0388.2000.00252.x

Hill, W. G., 1974 Linkage disequilibrium between neutral genes in finite populations. Adv. Appl. Probab. 6: 13–15. https://doi.org/10.2307/1426201

Ibánẽz-Escriche, N., R. L. Fernando, A. Toosi, and J. C. Dekkers, 2009 Genomic selection of purebreds for crossbred performance. Genet. Sel. Evol. 41: 12. https://doi.org/10.1186/1297-9686-41-12

Iversen, M. W., Ø. Nordbø, E. Gjerlaug-Enger, E. Grindflek, M. S. Lopes *et al.*, 2019 Effects of heterozygosity on performance of purebred and crossbred pigs. Genet. Sel. Evol. 51: 8. https://doi.org/10.1186/s12711-019-0450-1

Jansen, G. B., and J. W. Wilton, 1985 Selecting Mating Pairs with Linear Programming Techniques. J. Dairy Sci. 68: 1302–1305. https://doi.org/10.3168/jds.S0022-0302(85)80961-9

Karoui, S., M. J. Carabaño, C. Díaz, and A. Legarra, 2012 Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genet. Sel. Evol. 44: 39. https://doi.org/10.1186/1297-9686-44-39

Lopes, M. S., H. Bovenhuis, A. M. Hidalgo, J. A. M. van Arendonk, E. F. Knol *et al.*, 2017 Genomic selection for crossbred performance accounting for breed-specific effects. Genet. Sel. Evol. 49: 51. https://doi.org/10.1186/s12711-017-0328-z

Meuwissen, T., B. Hayes, and M. Goddard, 2013 Accelerating Improvement of Livestock with Genomic Selection. Annu. Rev. Anim. Biosci. 1: 221–237. https://doi.org/10.1146/annurev-animal-031412-103705

R-Core Team, 2017 *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

Robertson, A., 1975 Gene Frequency Distributions as a Test of Selective Neutrality. Genetics 81: 775–785.

Sargolzaei, M., and F. S. Schenkel, 2009 QMSim: a large-scale genome simulator for livestock. Bioinformatics 25: 680–681. https://doi.org/10.1093/bioinformatics/btp045

Technow, F., T. A. Schrag, W. Schipprack, E. Bauer, H. Simianer *et al.*, 2014 Genome Properties and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program of Maize. Genetics 197: 1343–1355. https://doi.org/10.1534/genetics.114.165860

Toro, M. A., and L. Varona, 2010 A note on mate allocation for dominance handling in genomic selection. Genet. Sel. Evol. 42: 33. https://doi.org/10.1186/1297-9686-42-33

Varona, L., C. Moreno, N. Ibáñez-Escriche, and J. Altarriba, 2010 Whole genome evaluation for related populations, in *Proceedings, 9th World Congress on Genetics Applied to Livestock Production*, Leipzig, Germany. http://wcgalp.org/proceedings/2010/whole-genome-evaluation-related-populations

Vitezica, Z. G., L. Varona, J.-M. Elsen, I. Misztal, W. Herring *et al.*, 2016 Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. Genet. Sel. Evol. 48: 6. https://doi.org/10.1186/s12711-016-0185-1

Vitezica, Z. G., L. Varona, and A. Legarra, 2013 On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope. Genetics 195: 1223–1230. https://doi.org/10.1534/genetics.113.155176

Wei, M., and J. H. J. van der Werf, 1994 Maximizing genetic response in crossbreds using both purebred and crossbred information. Anim. Sci. 59: 401–413. https://doi.org/10.1017/S0003356100007923

Wientjes, Y. C. J., and M. P. L. Calus, 2017 BOARD INVITED REVIEW: The purebred-crossbred correlation in pigs: A review of theory, estimates, and implications. J. Anim. Sci. 95: 3467–3478.

Wiggans, G. R., T. S. Sonstegard, P. M. VanRaden, L. K. Matukumalli, R. D. Schnabel *et al.*, 2009 Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J. Dairy Sci. 92: 3431–3436. https://doi.org/10.3168/jds.2008-1758

Xiang, T., O. F. Christensen, and A. Legarra, 2017 Technical note: Genomic evaluation for crossbred performance in a single-step approach with metafounders. J. Anim. Sci. 95: 1472–1480.

Xiang, T., O. F. Christensen, Z. G. Vitezica, and A. Legarra, 2016 Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. Genet. Sel. Evol. 48: 92. https://doi.org/10.1186/s12711-016-0271-4

Zeng, J., A. Toosi, R. L. Fernando, J. C. M. Dekkers, and D. J. Garrick, 2013 Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet. Sel. Evol. GSE 45: 11. https://doi.org/10.1186/1297-9686-45-11

Communicating editor: D.-J. de Koning

3.1 Appendices

Supplemental material

S1. Simulation of QTL effects for two parental breeds and their crossbred

At each locus, three (P1, P2 for the two parental breeds and C for the CB progeny) QTL additive and three (for P1, P2 and C) QTL dominance effects were sampled from two normal distributions with 0 mean and covariance matrices

$$Var\begin{pmatrix} a^{P1} \\ a^{P2} \\ a^{c} \end{pmatrix} = \begin{pmatrix} \sigma^2_{a^{P1}_{QTL}} & \sigma_{a^{P1,P2}_{QTL}} & \sigma_{a^{P1,c}_{QTL}} \\ \sigma_{a^{P2,P1}_{QTL}} & \sigma^2_{a^{P2}_{QTL}} & \sigma_{a^{P2,c}_{QTL}} \\ \sigma_{a^{c,P1}_{QTL}} & \sigma_{a^{c,P2}_{QTL}} & \sigma^2_{a^{c}_{QTL}} \end{pmatrix}$$

$$Var\begin{pmatrix} d^{*P1} \\ d^{*P2} \\ d^{*c} \end{pmatrix} = \begin{pmatrix} \sigma^2_{d^{*P1}_{QTL}} & \sigma_{d^{*P1,P2}_{QTL}} & \sigma_{d^{*P1,c}_{QTL}} \\ \sigma_{d^{*P2,P1}_{QTL}} & \sigma^2_{d^{*P2}_{QTL}} & \sigma_{d^{*P2,c}_{QTL}} \\ \sigma_{d^{*c,P1}_{QTL}} & \sigma_{d^{*c,P2}_{QTL}} & \sigma^2_{d^{*c}_{QTL}} \end{pmatrix}$$

Variances in the diagonal of these matrices were obtained backsolving from estimates in the animal scale in Xiang *et al.* (2016). For instance, additive ($\sigma^2_{a^{P1}_{QTL}}$) and dominance ($\sigma^2_{d^{*P1}_{QTL}}$) QTL variances in the P1 breed with allele frequencies $p^{P1}_{QTL_j}$ and $q^{P1}_{QTL_j}$ at the jth QTL, are equal to:

$$\sigma^2_{d^{*P1}_{QTL}} = \sigma^2_{v^{P1}} / \sum_{j=1}^{nQTL} \left(2p^{P1}_{QTL_j} q^{P1}_{QTL_j}\right)^2$$

$$\sigma^2_{a^{P1}_{QTL}} = (\sigma^2_{u^{P1}} - \sum_{j=1}^{nQTL} (2p^{P1}_{QTL_j} q^{P1}_{QTL_j} \left(q^{P1}_{QTL_j} - p^{P1}_{QTL_j}\right)^2) \sigma^2_{d^{*P1}_{QTL}}) / \sum_{j=1}^{nQTL} (2p^{P1}_{QTL_j} q^{P1}_{QTL_j})$$

where $\sigma^2_{u^{P1}}$ and $\sigma^2_{v^{P1}}$ are the additive and dominance genetic variances in P1. QTL variances were computed in the same manner for P2. In the CB population, additive ($\sigma^2_{a^{c}_{QTL}}$) and dominance ($\sigma^2_{d^{*c}_{QTL}}$) QTL variances were obtained from Eq. (5), (6) and (7) in Vitezica *et al.* (2016), as:

$$\sigma^2_{d^{*c}_{QTL}} = \sigma^2_{v^{c}} / \sum_{j=1}^{nQTL} (4p^{P1}_{QTL_j} q^{P1}_{QTL_j} p^{P2}_{QTL_j} q^{P2}_{QTL_j})$$

And

$$\sigma^2_{a^c_{QTL}} = \frac{\left[\sigma^2_{u^c} - \left[\sum\left(p^{P1}_{QTL_j} q^{P1}_{QTL_j}\left(q^{P2}_{QTL_j} - p^{P2}_{QTL_j}\right)^2\right) + \sum\left(p^{P2}_{QTL_j} q^{P2}_{QTL_j}\left(q^{P1}_{QTL_j} - p^{P1}_{QTL_j}\right)^2\right)\right]\sigma^2_{d^{*c}_{QTL}}\right]}{\left[\sum\left(p^{P1}_{QTL_j} q^{P1}_{QTL_j}\right) + \sum\left(p^{P2}_{QTL_j} q^{P2}_{QTL_j}\right)\right]}$$

where $p^{P2}_{QTL_j}$ and $q^{P2}_{QTL_j}$ refers to the frequencies in the second parental breed (P2). Note that QTL allelic frequencies differ between parental breeds.

The covariance matrix for residual effects was as follows:

$$Var\begin{pmatrix}\boldsymbol{e_1}\\ \boldsymbol{e_2}\\ \boldsymbol{e_c}\end{pmatrix} = \begin{pmatrix}\sigma^2_{e1} & 0 & 0\\ 0 & \sigma^2_{e2} & 0\\ 0 & 0 & \sigma^2_{ec}\end{pmatrix} \otimes \boldsymbol{I}$$

S2. Genomic evaluation models

Genomic evaluations were performed based on two models, depending on each scenario. The first model was a univariate GBLUP model included only additive genetic effects plus genomic inbreeding, and was written as follows:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{f}b + \boldsymbol{Zu} + \boldsymbol{e},$$

where $\boldsymbol{y}$ is a vector of phenotypes of PB females; μ is the general mean; $\boldsymbol{f}$ is the vector of genomic inbreeding coefficients, calculated as the average homozygosity per individual; and b is the inbreeding depression parameter. Further, $\boldsymbol{u}$ is the vector of additive genetic effects and $\boldsymbol{e}$ is a vector of random residual effects ($\boldsymbol{e}\sim N(0, \boldsymbol{I}\sigma^2_e)$). Breeding values were distributed as $\boldsymbol{u}\sim N(0, \boldsymbol{G}\sigma^2_u)$, where $\boldsymbol{G}$ is the additive genomic relationship matrix, and $\boldsymbol{Z}$ is the incidence matrix that related phenotypes to breeding values. Terms σ^2_u and σ^2_e refer to the additive genetic and residual variances, respectively, which were assumed known.

The additive genomic relationship matrix was calculated according to VanRaden (2008), as follows:

$$\boldsymbol{G} = \frac{\boldsymbol{MM'}}{2\sum_{k=1}^{m} p_k q_k},$$

where $\boldsymbol{M}$ is a matrix with dimensions equal to the number of animals (n) by number of SNPs (m), with elements equal to $(2-2p_k)$, $(1-2p_k)$ and $-2p_k$, for genotypes $AA, Aa,$ and aa respectively; p_k is the frequency for allele A of the kth SNP (for $k = 1, \dots, m$), and $q_k = 1 - p_k$. This model was implemented with *Blupf90* software (Misztal *et al.* 2002). The EBV_P for candidates were obtained directly from the solutions of the mixed model equations.

The second model was a trivariate SNP-BLUP, with correlated SNP effects across the three populations. This model can be written according to Xiang *et al.* (2016a) as:

$$\boldsymbol{y}^{P1} = \mathbf{1}\mu^{P1} + \boldsymbol{f}^{P1}b^{P1} + \boldsymbol{Z}^{P1}\boldsymbol{a}^{P1} + \boldsymbol{W}^{P1}\boldsymbol{d}^{*P1} + \boldsymbol{e}^{P1}$$

$$\boldsymbol{y}^{P2} = \mathbf{1}\mu^{P2} + \boldsymbol{f}^{P2}b^{P2} + \boldsymbol{Z}^{P2}\boldsymbol{a}^{P2} + \boldsymbol{W}^{P2}\boldsymbol{d}^{*P2} + \boldsymbol{e}^{P2}$$

$$\boldsymbol{y}^{c} = \mathbf{1}\mu^{c} + \boldsymbol{f}^{c}b^{c} + \boldsymbol{Z}^{c}\boldsymbol{a}^{c} + \boldsymbol{W}^{c}\boldsymbol{d}^{*c} + \boldsymbol{e}^{c}$$

where $\boldsymbol{y}^{P1}$, $\boldsymbol{y}^{P2}$, and $\boldsymbol{y}^{c}$ are the vector containing the phenotype of females for P1 and P2 breeds, and CB population, respectively; μ^{P1}, μ^{P2}, and μ^{c} are the respective general mean; $\boldsymbol{a}^{P1}$, $\boldsymbol{a}^{P2}$, $\boldsymbol{a}^{c}$ and $\boldsymbol{d}^{*P1}$, $\boldsymbol{d}^{*P2}$, $\boldsymbol{d}^{*c}$ are the functional additive and dominance SNP effects for each population, respectively. Note that estimated dominance SNP effects d^* are centered ($\boldsymbol{d}^* = \boldsymbol{d} - \mu_d$). Then $\boldsymbol{f}^{P1}$, $\boldsymbol{f}^{P2}$ and $\boldsymbol{f}^{c}$ are the genomic inbreeding coefficient as explained above; and b^{P1}, b^{P2} and b^{c} are the inbreeding depression parameters for P1, P2, and CB, respectively, which are transformed into directional depression estimates $\mu_d = -\frac{b}{m}$ so that the final estimate of the dominance effects is $\widehat{\boldsymbol{d}} = \widehat{\boldsymbol{d}^*} + \mathbf{1}\widehat{\mu_d}$. $\boldsymbol{Z}$ and $\boldsymbol{W}$ are incidence matrices that related the data to additive and dominance SNP effects, respectively. SNP genotypes were coded as 1, 0, and -1 for additive effects and 0, 1 and 0 for dominance effects, for genotypes $AA, Aa,$ and aa, respectively. The covariance structures for the additive and dominance SNP effects across populations are:

$$Var\begin{pmatrix}\boldsymbol{a}^{P1}\\ \boldsymbol{a}^{P2}\\ \boldsymbol{a}^{c}\end{pmatrix} = \begin{pmatrix}\sigma^2_{a^{P1}_{SNP}} & \sigma_{a^{P1,P2}_{SNP}} & \sigma_{a^{P1,c}_{SNP}}\\ \sigma_{a^{P2,P1}_{SNP}} & \sigma^2_{a^{P2}_{SNP}} & \sigma_{a^{P2,c}_{SNP}}\\ \sigma_{a^{c,P1}_{SNP}} & \sigma_{a^{c,P2}_{SNP}} & \sigma^2_{a^{c}_{SNP}}\end{pmatrix} \otimes \boldsymbol{I}$$

$$Var\begin{pmatrix}\boldsymbol{d}^{*P1}\\ \boldsymbol{d}^{*P2}\\ \boldsymbol{d}^{*c}\end{pmatrix}=\begin{pmatrix}\sigma^2_{d^{*P1}_{SNP}} & \sigma_{d^{*P1,P2}_{SNP}} & \sigma_{d^{*P1,c}_{SNP}}\\ \sigma_{d^{*P2,P1}_{SNP}} & \sigma^2_{d^{*P2}_{SNP}} & \sigma_{d^{*P2,c}_{SNP}}\\ \sigma_{d^{*c,P1}_{SNP}} & \sigma_{d^{*c,P2}_{SNP}} & \sigma^2_{d^{*c}_{SNP}}\end{pmatrix}\otimes\boldsymbol{I}$$

The additive and dominance SNP (co)variances were assumed known (backsolving from estimates in the animal scale in Xiang *et al.* (2016a)). This SNP-BLUP was implemented using own Fortran program. The EBV_C was obtained from the equations used to simulate the TBV_C, but instead of using QTL genotypes and effects, the SNP frequencies from P1 and P2 and the estimated SNP effects of the CB progeny were used. The SNP effects were estimated each generation by using the current allele frequencies.

Genomic evaluations were performed each generation, and animals in previous generations were included in the analyses. For instance, for S1 and S2, the number of PB animals evaluated by GBLUP in the last generation was equal to 27,144. However, when the trivariate SNP-BLUP model was implemented, a maximum of 5 generations was included in the evaluations, due to the high computing demanding resources and time. For example, in S3 and S4, the total number of animals evaluated in the last generation was 36,720 (12,240 animals from each of the three populations). Note that, as CB animals were created after selecting PB animals, CB information used in the genetic evaluation came from the previous generations. For instance, for the first genomic evaluation in generation 1, we included PB information (genotypes) from generation 1 and founders (genotypes and records of females), but the CB information (genotypes) comes only from the CB reference population derived from founder animals (see Figure 2 in the manuscript Chapter 3).

Chapter 4

Genomic prediction of hybrid crops allows disentangling dominance and epistasis

D. GONZÁLEZ-DIÉGUEZ [*, 1], A. LEGARRA [*], A. CHARCOSSET [†], L. MOREAU [†], C. LEHERMEIER [‡], S. TEYSSÈDRE [‡], Z.G. VITEZICA [*]

* INRAE, INP, UMR 1388 GenPhySE, F-31326 Castanet-Tolosan, France

† GQE-Le Moulon, INRAE, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, Gif-sur-Yvette, France

‡ Genetics and Analytics Unit, RAGT2n, Druelle, France

Submitted to Genetics

14/08/2020

[1] Corresponding author: INRAE, INP, UMR 1388 GenPhySE, 24 Chemin de Borde Rouge, 31326 Castanet-Tolosan Cedex, France. E-mail. david-omar.gonzalez-dieguez@inrae.fr

Running title: Genomic models for Single-Cross Hybrids

Key words: dominance, epistasis, genetic variance, heterosis, genomic models

Corresponding author: D. González-Diéguez

INRA, UMR 1388 GenPhySE

24 Chemin de Borde Rouge

31326 Castanet Tolosan Cedex, France

Phone: + 33 5 61 28 53 59

E-mail: david-omar.gonzalez-dieguez@inrae.fr

ABSTRACT

We revisited, in a genomic context, the theory of hybrid genetic evaluation models of hybrid crosses of pure lines issued from different groups (as in maize), as the current practice is largely based on infinitesimal model assumptions. Expressions for covariances between hybrids due to additive substitution effects and dominance and epistatic deviations were analytically derived. Using dense markers in a GBLUP analysis, it is possible to split specific combining ability (SCA) into dominance and across-groups epistatic deviations, and to split general combining ability (GCA) into within-line additive effects and within-line additive by additive (and higher order) epistatic deviations. We analyzed a publicly available maize data set of Dent × Flint hybrids using our new model up to additive by additive epistasis. To model higher order interactions, we also fitted "remaining" "unrelated" Dent and Flint effects. We compared to the currently used model and to another genomic model which assumes a common effect of genes across origins. Most variation in hybrids is accounted for by GCA. Variances due to dominance and di-genic epistatic deviations have similar magnitudes. Models based on defining effects either differently or identically across Dent and Flint groups resulted in similar predictive abilities for hybrids. The currently used model inflates the estimated additive genetic variance and the additive part of the estimated GCAs. This is not important for hybrid predictions but has consequences for the breeding scheme – e.g. overestimation of the genetic gain. Therefore we recommend fitting within-group epistasis or, if not possible, unrelated "remaining" line effects.

INTRODUCTION

Many plant species are presently cultivated in the form of single-cross hybrid varieties, especially when a strong heterosis effect is observed for yield related traits, (e.g. maize, sunflower, sugarbeet, etc.). These hybrids are generally obtained by crossing inbred lines originated from two complementary populations, called heterotic groups. Breeders' objective is therefore to identify (i) the best single-cross hybrids among all possible crosses between existing inbred lines from the two groups and (ii) create new lines within heterotic group, from crosses of existing lines, that will improve the performance of candidate hybrids at a next generation. Models for genetic improvement of hybrid crops (e.g. maize) across two heterotic groups are typically based on the notions of General Combining Ability (GCA) and Specific Combining Ability (SCA) (Griffing 1962; Stuber and Cockerham 1966; Bernardo 2010). The genotypic value G_{ij} of the cross of lines *i* and *j*, as a function of uniting gametes from *i* and *j*, can be written as:

$$G_{ij} = \mu + GCA_i + GCA_j + SCA_{ij} \quad (1)$$

where GCA of line *i* is the average effect of a gamete when ideally crossed to all gametes from the reciprocal heterotic group. SCA of the combination of line *i* and *j* is the remainder.

It is important to notice, for readers not familiar with hybrid crops, that in many hybrid crops such as maize, parents are pure homozygous individuals (inbred lines). Thus, all gametes produced by *i* (and *j*) are identical, and all F1 descendants of *i* and *j* are identical. This is different from crosses of other species such as animals (pigs for instance) where full-sibs show genetic variation. As a result, GCA contains single locus (additive, in the statistical sense) and multiple loci (additive by additive and higher additive interactions) effects. This is because the whole genotype (gamete) of the pure line is transmitted to the F_1 descendants, including any possible epistatic combination, and regardless of whether loci in interaction are in the same or in different chromosomes.

Informally, the GCAs within group 1 (group 2) are the sum of additive, additive x additive, additive x additive x additive… deviations *within* group 1 (group 2), whereas SCA are the sum of dominance, all epistatic interaction involving dominance, and any epistatic additive interaction across both groups.

Stuber and Cockerham (1966) presented the covariance across GCAs of different lines and SCAs of different pairs of lines as a function of probabilities of alleles at loci being identical by descent. These probabilities are, in their work, implicitly based on pedigree, and are the same across all loci or pairs of loci. For this reason, and based on pedigree of lines alone, some components of the variance cannot be distinguished; for instance, dominance and across-groups additive epistasis cannot be separated (Bernardo 2010).

With the advent of molecular markers (SNPs), it is possible to obtain better predictions based, ultimately, on similarity across lines (or pairs of lines) measured using markers (using for instance kinship matrices based on VanRaden (2008). Bernardo (1994) transposed the concepts of Stuber and Cockerham (1966) to the prediction of hybrid lines of corn using molecular markers. These ideas have been used extensively (e.g. Massman *et al.* 2013; Technow *et al.* 2014; Bouvet *et al.* 2016; Kadam *et al.* 2016; Acosta-Pech *et al.* 2017; Westhues *et al.* 2017; Schrag *et al.* 2018). However, the infinitesimal assumptions of Stuber and Cockerham (1966) are not needed or completely pertinent any more. SNP markers allow finer distinction of patterns of relationships across and within regions. For instance, it is possible to distinguish relationship within locus and across loci – thus making it possible to split dominance and across-groups epistasis. Also, relationships across dominance deviations are not simple expressions built from additive relationships (Vitezica *et al.* 2013, 2017). Therefore, it is important to properly re-define statistical models for genomic prediction, in order to have models that are more adequate to the physical nature of the genome (finite and not infinitesimal), better understood, and potentially more accurate.

The model including a GCA component for each group is convenient because individuals are selected within groups. However, there is a different model proposed by Stuber and Cockerham (1966) where gene effects are defined "uniquely" *within* hybrids regardless of the origin. In this model the effect of allele, say, base "G" at one locus is considered the same either from group 1 or from group 2. The model can be decomposed in an additive value for the hybrid (A), a dominant deviation (D), an additive by additive epistatic interaction (AxA), and so on. Informally:

$$G_{ij} = A_{ij} + D_{ij} + (AxA)_{ij} \ldots$$

This model can be used for prediction of hybrid performance. However, the model does not explicitely produce estimates for pure lines, and these are needed for selection within

heterotic groups. This model has sometimes been borrowed from animal breeding and then used (Fritsche-Neto *et al.* 2018) but the particularities of hybrid crops have not been considered: for instance, the population of hybrids is *not* in HWE, which leads to lack of orthogonality (Vitezica *et al.* 2017) and to misleading estimates of variance components and of marker effects. A proper decomposition of effects can be achieved using NOIA (natural and orthogonal interactions) approach (Vitezica *et al.* 2017) but it has not been presented in hybrid crops.

In this work we rederive models for genomic prediction in hybrid crops using the notions of effects defined "according to parental origin" (GCAs and SCAs) and "uniquely" (at the hybrid genotype level ignoring group origin of alleles) (Stuber and Cockerham, 1966), using quantitative genetics theory and considering the substitution effects of markers. We present expressions for additive, dominant and epistatic relationships across pure lines and hybrids. Then we test the resulting models using a publicly available data set (Technow *et al.* 2014), both for estimation of variance components and of predictive abilities.

THEORY

Here we derive the theory for genomic relationship matrices in the analysis of hybrid crosses of inbred lines from two populations. We draw on the tradition of separately modelling effects of gametes coming from each heterotic group (Sprague and Tatum 1942; Griffing 1962; Stuber and Cockerham 1966; Bernardo 2010). The two parental populations or heterotic groups (*e.g.* Dent and Flint in the case of North European maize), are named 1 and 2. Extension to more than two heterotic groups is immediate.

Our aim is to split the total genotypic value of a single-cross hybrid in two statistical additive effects (one from each group), a single dominance deviation (particular to each cross) and epistatic interactions (either intra-group or across-groups). Ideally this partition is orthogonal and all components should be estimable. To our knowledge, this partition and its use in hybrid crops using markers has not been presented elsewhere.

Additive substitution effects and dominance deviations in hybrid crops

We start from a genotypic model to derive statistical effects (Falconer 1981; Bernardo 2010). Then we extend our reasoning to epistatic interactions in the manner of NOIA (Álvarez-Castro and Carlborg 2007; Vitezica *et al.* 2017). Consider a biallelic single locus/gene with alleles *B/b* and the allele origin denoted as *i*=1 and *j*=2, thus population 1 has B_1 and b_1 with respective frequencies p_1 and $q_1 = 1 - p_1$ and population 2 has B_2 and b_2 with frequencies p_2 and q_2. The hybrid population has genotypes (frequencies) B_1B_2 (p_1p_2), B_1b_2 (p_1q_2), b_1B_2 (q_1p_2) and b_1b_2 (q_1q_2).

We assume additive and dominant gene action, and separate effects of gametes coming from 1 and 2. Then the genotypic value G of a hybrid can be written (up to a common constant) as

$$G_{B_1B_2} = a_1 + a_2 \qquad G_{B_1b_2} = a_1 + d \qquad G_{b_1B_2} = a_2 + d \qquad G_{b_1b_2} = 0.$$

where a_1 is the functional additive effect for B_1 from P1, a_2 is the functional additive effect for B_2 from P2, and d is the functional value of both heterozygotes (B_1b_2 and b_1B_2). The genetic mean of the hybrid population is therefore

$$E(G) = p_1p_2(a_1 + a_2) + p_1q_2(a_1 + d) + q_1p_2(a_2 + d) = p_1a_1 + p_2a_2 + (p_1q_2 + q_1p_2)d$$

Classically, the genotypic values of a hybrid are split into *statistical* additive values, one per parent, and a dominant deviation for the hybrid, as in (Bernardo 2010):

$$G = E(G) + g_{A^{(1)}} + g_{A^{(2)}} + g_D$$

where $g_{A^{(1)}}$ ($g_{A^{(2)}}$) is the additive effect of a gamete from population 1 (from population 2) combined with a gamete from population 2 (population 1), whereas g_D is the dominant deviation.

Additive values $g_{A^{(1)}}$ and $g_{A^{(2)}}$ of the gametes include average effects of each gene/allele. The *average effect of alleles* (α_{B_1}, α_{b_1}, α_{B_2} and α_{b_2}) are derived from the genotypic values. The reasoning is set out in Table 1 (following Table 7.2 in Falconer (1981)). If gametes carrying B_1 from population 1 are mated at random with gametes from population 2, the frequencies of the genotypes produced will be p_2 of B_1B_2 and q_2 of B_1b_2. The genotypic value of a hybrid B_1B_2

is $G_{B_1B_2} = (a_1 + a_2)$, that of B_1b_2 is $G_{B_1b_2} = (a_1 + d)$, and the mean of these two, taking into account of the proportions in which they occur is $E(G|B_1) = p_2(a_1 + a_2) + q_2(a_1 + d)$.

Table 1 Representation of the average effect of a gene

Type of gamete	Values and frequencies of genotypes produced			
	B_1B_2 $(a_1 + a_2)$	B_1b_2 $(a_1 + d)$	b_1B_2 $(a_2 + d)$	b_1b_2 0
B_1	p_2	q_2		
b_1			p_2	q_2
B_2	p_1		q_1	
b_2		p_1		q_1

For instance, considering allele B_1, the difference between the mean value conditional on a particular genotype of the gamete (*e.g.* $E(G|B_1)$) and the population mean ($E(G)$) is the average effect of the allele, α_{B_1}. The average effects of alleles (α_{B_1}, α_{b_1}, α_{B_2} and α_{b_2}) are therefore

$$\alpha_{B_1} = p_2(a_1 + a_2) + q_2(a_1 + d) - E(G) = q_1[a_1 + (q_2 - p_2)d]$$
$$\alpha_{b_1} = p_2(a_2 + d) - E(G) = -p_1[a_1 + (q_2 - p_2)d]$$
$$\alpha_{B_2} = p_1(a_1 + a_2) + q_1(a_2 + d) - E(G) = q_2[a_2 + (q_1 - p_1)d]$$
$$\alpha_{b_2} = p_1(a_1 + d) - E(G) = -p_2[a_2 + (q_1 - p_1)d]$$

Now, we can derive the *average effect of the allele-substitution*, for instance, letting b_1 be substituted by B_1. From the b_1 alleles taken at random from the population for substitution, a proportion p_2 will be found in b_1B_2 genotypes and a proportion q_2 in b_1b_2. The substitution will respectively change the value from $(a_2 + d)$ to $(a_1 + a_2)$ and from 0 to $(a_1 + d)$ (see Table 1). Thus, the average effect of the allele-substitution (α_1) of population 1 is

$$\alpha_1 = p_2[(a_1 + a_2) - (a_2 + d)] + q_2[(a_1 + d) - 0] = a_1 + (q_2 - p_2)d$$

The same result can be obtained as the difference between average effects: $\alpha_1 = \alpha_{B_1} - \alpha_{b_1}$. For population 2 (α_2), it is $\alpha_2 = \alpha_{B_2} - \alpha_{b_2} = a_2 + (q_1 - p_1)d$. The average effects of alleles can also be rewritten as function of the allele substitution effect as follows: $\alpha_{B_1} = q_1\alpha_1$, $\alpha_{b_1} = -p_1\alpha_1$, $\alpha_{B_2} = q_2\alpha_2$ and $\alpha_{b_2} = -p_2\alpha_2$.

Note that allele-substitution effects involve both functional additive $(a_1 + a_2)$ and dominant (d) effects, and allele frequencies of the other parental population. Similar expressions were presented by Vitezica *et al.* (2016) but an identical functional additive effect $(a = a_1 = a_2)$ was assumed in both parental populations, ignoring the origin of the allele.

The *statistical additive effects* of a gamete are equal to the sum of the average effects of the alleles it carries. Thus, for a single locus, the statistical additive effects are

$$g_{A^{(1)}_{B_1}} = q_1\alpha_1$$

$$g_{A^{(1)}_{b_1}} = -p_1\alpha_1$$

$$g_{A^{(2)}_{B_2}} = q_2\alpha_2$$

$$g_{A^{(2)}_{b_2}} = -p_2\alpha_2$$

This can also be written as $g_{A^{(1)}} = z_1\alpha_1$ and $g_{A^{(2)}} = z_2\alpha_2$ with

$z_1 = \begin{cases} (1-p_1) \\ -p_1 \end{cases}$ for gametes $\begin{cases} B_1 \\ b_1 \end{cases}$, and $\qquad z_2 = \begin{cases} (1-p_2) \\ -p_2 \end{cases}$ for gametes $\begin{cases} B_2 \\ b_2 \end{cases}$

Subtracting statistical additive effects from genotypic values ($G_{B_1B_2}$, $G_{B_1b_2}$, $G_{b_1B_2}$ and $G_{b_1b_2}$) gives *dominance deviations* which are interactions between the alleles received from parental populations. This is detailed in the Appendix, and the dominance deviation of the hybrid according to its genotype is

$$g_{D_{B_1B_2}} = -2q_1q_2d$$

$$g_{D_{B_1b_2}} = 2q_1p_2d$$

$$g_{D_{b_1B_2}} = 2p_1q_2d$$

$$g_{D_{b_1b_2}} = -2p_1p_2d$$

So, the dominance deviation of a hybrid individual can be written as $g_D = wd$ with

$$w = \begin{cases} -2q_1q_2 \\ 2q_1p_2 \\ 2p_1q_2 \\ -2p_1p_2 \end{cases} \quad \text{for genotypes} \quad \begin{cases} B_1B_2 \\ B_1b_2 \\ b_1B_2 \\ b_1b_2 \end{cases}$$

Or, equivalently, $w = -2z_1z_2$ for z_1, z_2 defined as above.

Finally, the model for analysis of hybrid crosses considering additive and dominance effects can be written in matrix form for a set of crosses as

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{g}_{A^{(1)}} + \boldsymbol{g}_{A^{(2)}} + \boldsymbol{g}_D$$

Where, for a single locus, $\boldsymbol{g}_{A^{(1)}} = \boldsymbol{z}_1\alpha_1$, $\boldsymbol{g}_{A^{(2)}} = \boldsymbol{z}_2\alpha_2$, $\boldsymbol{g}_D = \boldsymbol{w}d$

Derivation of additive and dominance genomic relationships

Now we extend the analysis to multiple markers, using $\boldsymbol{g}_{A^{(1)}} = \boldsymbol{Z}_1\boldsymbol{\alpha}_1$, $\boldsymbol{g}_{A^{(2)}} = \boldsymbol{Z}_2\boldsymbol{\alpha}_2$, $\boldsymbol{g}_D = \boldsymbol{W}\boldsymbol{d}$, where $\boldsymbol{Z}_1 = \left(\boldsymbol{z}_{1_1} \ldots \boldsymbol{z}_{1_{nsnp}}\right)$ and $\boldsymbol{Z}_2 = \left(\boldsymbol{z}_{2_1} \ldots \boldsymbol{z}_{2_{nsnp}}\right)$ are matrices with lines of each heterotic group in rows and number of markers, $nsnp$, in columns. The matrix $\boldsymbol{W} = \left(\boldsymbol{w} \ldots \boldsymbol{w}_{nsnp}\right)$ has as many rows as hybrid individuals and as many columns as markers.

Genotypes in pure lines are in matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ which contain zero for genotypes b_1b_1 and b_2b_2, respectively; and 1 for genotypes B_1B_1 and B_2B_2, respectively. The observed *B* allele frequencies for marker *j* in the heterotic groups composed by *n* lines can be computed as $p_j = \frac{\sum_{i=1}^{n} M_{ij}}{n}$. Matrices $\boldsymbol{Z}$ are obtained subtracting $\boldsymbol{p}$ (which is equal to centering if $\boldsymbol{p}$ is computed from observed genotypes), as $\boldsymbol{Z}_1 = \boldsymbol{M}_1 - \mathbf{1}\boldsymbol{p}'$ for population 1. It is analogous for population 2.

Now, we can set up the covariance matrices. *Additive covariance matrix* for population 1 (for $\boldsymbol{g}_{A^{(1)}}$) is

$$Var\left(\boldsymbol{g}_{A^{(1)}}\right) = \boldsymbol{Z}_1\boldsymbol{Z}_1'Var(\boldsymbol{\alpha}_1) = \boldsymbol{Z}_1\boldsymbol{Z}_1'\sigma^2_{\alpha_1}$$

where $\sigma^2_{\alpha_1}$ is the variance of the allele-substitution effect (α_1) of population 1. For one locus, if the population 1 is a population of pure lines (individuals are homozygotes), the genetic variance *of its gametes* $g_{A^{(1)}}$ is

$$\sigma^2_{A^{(1)}} = Var\left(g_{A^{(1)}}\right) = E\left(g^2_{A^{(1)}}\right) - E\left(g_{A^{(1)}}\right)^2$$

By construction of the matrices, $E\left(g_{A^{(1)}}\right) = 0$ and then we have for one locus the following table of gametes and their effects

Genotype	Frequency	$g^2_{A^{(1)}}$	$g_{A^{(1)}}$
B_1	p_1	$(q_1\alpha_1)^2$	$q_1\alpha_1$
b_1	q_1	$(-p_1\alpha_1)^2$	$-p_1\alpha_1$

So, $E\left(g^2_{A^{(1)}}\right) = p_1(q_1\alpha_1)^2 + q_1(-p_1\alpha_1)^2 = p_1q_1(p_1+q_1)\alpha_1^2 = p_1q_1\alpha_1^2$ and $Var\left(g_{A^{(1)}}\right) = p_1q_1\alpha_1^2$.

Assuming linkage equilibrium, we generalize this result to all $nsnp$ markers

$$\sigma^2_{A^{(1)}} = Var\left(g_{A^{(1)}}\right) = \sum_i^{nsnp} p_{1_i}q_{1_i}\ \sigma^2_{\alpha_1}$$

and

$$\sigma^2_{\alpha_1} = \frac{\sigma^2_{A^{(1)}}}{\sum_i^{nsnp} p_{1_i}q_{1_i}}$$

Therefore, we can now divide $Var\left(\boldsymbol{g}_{A^{(1)}}\right)$ above by this variance and we have

$$Var\left(\boldsymbol{g}_{A^{(1)}}\right) = \frac{\boldsymbol{Z}_1\boldsymbol{Z}_1'}{\sum_i^{nsnp} p_{1_i}q_{1_i}}\sigma^2_{A^{(1)}} = \boldsymbol{G}_{A^{(1)}}\sigma^2_{A^{(1)}}$$

where $\boldsymbol{G}_{A^{(1)}} = \frac{\boldsymbol{Z}_1\boldsymbol{Z}_1'}{\sum_i^{nsnp} p_{1_i}q_{1_i}}$ is the additive genomic relationship matrix across lines in population 1 of size $n_1 \times n_1$. The reasoning is identical for population 2 and only the allele frequencies change.

These results are similar, but not identical, to VanRaden (2008). In particular, using VanRaden's method 1 directly, while coding genotypes in pure lines as 0/2, results in $\boldsymbol{G}_{VR^{(1)}} = 2\boldsymbol{G}_{A^{(1)}}$, a matrix that does not refer to the correct genetic variance component. Commonly, plant

breeders use VanRaden's additive relationship matrix divided by two to obtain kinship matrix (or coancestry matrix).

For the *dominance deviations*, the covariance matrix for hybrids is

$$Var(\boldsymbol{g}_D) = \boldsymbol{WW'}Var(\boldsymbol{d}) = \boldsymbol{WW'}\sigma_d^2$$

where σ_d^2 is the variance of the dominant effect at the locus level, defined at the hybrid population.

We know from Reif *et al.* (2007) and Vitezica *et al*. (2016) that the variance of dominance deviations for F1 hybrids is, for one locus, $\sigma_D^2 = 4p_1q_1p_2q_2d^2$. Alternatively, because gametes are uncorrelated, the proof follows as $Var(g_D) = Var(wd) = Var(-2z_1z_2d) = 4d^2Var(z_1z_2) = 4d^2Var(z_1)Var(z_2) = 4d^2p_1q_1p_2q_2 = 4p_1q_1p_2q_2d^2$. From here and assuming LE across loci we have

$$\sigma_D^2 = Var(g_D) = \sum_i^{nsnp} \left(4p_{1_i}q_{1_i}p_{2_i}q_{2_i}\right) \sigma_d^2$$

which results in the covariance matrix

$$Var(\boldsymbol{g}_D) = \frac{\boldsymbol{WW'}}{\sum_i^{nsnp}(4p_{1_i}q_{1_i}p_{2_i}q_{2_i})}\sigma_D^2 = \boldsymbol{D}\sigma_D^2$$

Where $\boldsymbol{D} = \frac{\boldsymbol{WW'}}{\sum_i^{nsnp}(4p_{1_i}q_{1_i}p_{2_i}q_{2_i})}$ is the dominance relationship matrix across hybrids of size $n \times n$. Note that we assume that allele substitution effects and dominance effects are random with respective variances σ_a^2 and σ_d^2.

Technow *et al.* (2014) modelled specific combining abilities using element-by-element products of matrices $\boldsymbol{G}_{A^{(1)}}$ and $\boldsymbol{G}_{A^{(2)}}$, following Stuber and Cockerham (1966). Clearly this is not the same as our matrix $\boldsymbol{D}$ above, that results directly from modelling dominance deviations. We will show later that Technow *et al.* (2014) approach only models additive by additive across-heterotic groups epistasis, that this is an approximation to our $\boldsymbol{D}$, and that the method of Stuber and Cockerham (1966) to obtain relationships across specific combining abilities assumes an infinitesimal model, and should not be directly transposed to marker-based models.

Some properties of the additive and dominance relationship matrices

Matrices $\boldsymbol{G}_{A^{(1)}}$, $\boldsymbol{G}_{A^{(2)}}$ and $\boldsymbol{D}$ have an average diagonal equal to 1 and an average value equal to 0 across the whole matrix. This implies that estimates of variance components can be interpreted as genetic variances (Legarra 2016). Also, z and w, the underlying incidence matrices to $\boldsymbol{G}_{A^{(1)}}$, $\boldsymbol{G}_{A^{(2)}}$ and $\boldsymbol{D}$ are orthogonal (see Appendix for the proof), which implies that statistical estimates are independent from each other – lack of orthogonality plagues models that are not carefully constructed and hampers interpretation of results. Also, because the basic bricks *z* and w are orthogonal, extension to higher order of interaction (epistasis) is immediate and also orthogonal (as briefly mentioned in the Appendix). In the next section we present epistatic relationship matrices.

Epistasis in hybrid populations

So far, we have written down the model for analysis of hybrid crosses including additive and dominance relationships. Now we extend this model to consider interactions across loci, using previous results and the ideas of the NOIA system (Álvarez-Castro and Carlborg 2007; Vitezica *et al.* 2017).

The classical model (1) including general combining ability (GCA) and specific combining ability (SCA) effects for a hybrid individual can be written as $y_{ij} = \mu + gca_i + gca_j + sca_{ij} + e_{ij}$, where y_{ij} is the phenotypic value of the hybrid, μ is the population mean, gca_i is the GCA of line i, gca_j is the GCA of line j and sca_{ij} is the SCA which depends of the combination of alleles received from *i* and *j*.

Epistasis intrapopulation

Stuber and Cockerham (1966) showed that the GCA-term (gca_i) includes, in addition to the additive gametic effects, the additive-by-additive epistasis across loci for alleles present in the line (equation 1, page 1279), and all higher order additive interactions in the line. This is because the lines are inbred – so exactly the same gamete is always transmitted to the hybrid, contrary to animal breeding where recombination breaks down epistatic combinations. So, for

instance the GCA of a gamete from population 1, considering two loci, *k* and *m* and second-order epistasis,

$$gca_i = z_{1(k)}\alpha_1^k + z_{1(m)}\alpha_1^m + z_{1(k)} \otimes z_{1(m)}(\alpha\alpha)_1^{k,m}$$

where $(\alpha\alpha)_1^{k,m}$ is the deviation due to epistatic interaction across loci in population 1. The coefficients of the incidence matrix, $z_{1(k)} \otimes z_{1(m)}$, for second-order epistatic effects between two loci can be computed as the Kronecker products ($\otimes$) of the respective incidence matrices for single locus effects (Álvarez-Castro and Carlborg 2007). For multiple loci, the matrix $\boldsymbol{Z}_{11}$ of additive-by-additive interaction effects can be written using Kronecker products of each row (corresponding to each line) of the preceding matrices as

$$\boldsymbol{Z}_{11} = \begin{pmatrix} \boldsymbol{z}_{1_1} \otimes \boldsymbol{z}_{1_1} \\ \cdots \\ \boldsymbol{z}_{1_i} \otimes \boldsymbol{z}_{1_i} \\ \cdots \\ \boldsymbol{z}_{1_{n_1}} \otimes \boldsymbol{z}_{1_{n_1}} \end{pmatrix}$$

Matrix $\boldsymbol{Z}_{11}$ is of large size (the number of rows is $nsnp^2$) but it is not explicitely used. Following Vitezica *et al.* (2017), we know that $\boldsymbol{Z}_{11}\boldsymbol{Z}_{11}' = \boldsymbol{Z}_1\boldsymbol{Z}_1' \odot \boldsymbol{Z}_1\boldsymbol{Z}_1'$ where $\odot$ is the Hadamard product; following developments in Vitezica *et al.* (2017) the genomic additive-by-additive epistatic relationship matrices of lines of population 1 with themselves is thus

$$\boldsymbol{G}_{AA^{(1,1)}} = \frac{(\boldsymbol{Z}_1\boldsymbol{Z}_1' \odot \boldsymbol{Z}_1\boldsymbol{Z}_1')}{tr(\boldsymbol{Z}_1\boldsymbol{Z}_1' \odot \boldsymbol{Z}_1\boldsymbol{Z}_1')/n_1} = \frac{\boldsymbol{G}_{A^{(1)}} \odot \boldsymbol{G}_{A^{(1)}}}{tr(\boldsymbol{G}_{A^{(1)}} \odot \boldsymbol{G}_{A^{(1)}})/n_1}$$

Where tr is the trace and n_1 is the number of lines in population 1. Therefore, the covariance matrix for the additive-by-additive interaction within population 1 ($\boldsymbol{g}_{AA^{(11)}}$) is:

$$Var(\boldsymbol{g}_{AA^{(1,1)}}) = \boldsymbol{G}_{AA^{(1,1)}}\sigma^2_{AA^{(1,1)}}$$

The reasoning for population 2 is the same resulting in

$$\boldsymbol{G}_{AA^{(2,2)}} = \frac{(\boldsymbol{Z}_2\boldsymbol{Z}_2' \odot \boldsymbol{Z}_2\boldsymbol{Z}_2')}{tr(\boldsymbol{Z}_2\boldsymbol{Z}_2' \odot \boldsymbol{Z}_2\boldsymbol{Z}_2')/n_2} = \frac{\boldsymbol{G}_{A^{(2)}} \odot \boldsymbol{G}_{A^{(2)}}}{tr(\boldsymbol{G}_{A^{(2)}} \odot \boldsymbol{G}_{A^{(2)}})/n_2}$$

and $Var(\boldsymbol{g}_{AA^{(2,2)}}) = \boldsymbol{G}_{AA^{(2,2)}}\sigma^2_{AA^{(2,2)}}$

The dimensions of $\boldsymbol{G}_{AA^{(2,2)}}$ and $\boldsymbol{G}_{AA^{(2,2)}}$ are $n_1 \times n_1$ and $n_2 \times n_2$, respectively.

Epistasis across populations

According to Stuber and Cockerham (1966), the SCA-term (sca_{ij}) includes the additive-by-additive epistasis across loci in alleles coming from different populations (equation 2, page 1279), the additive-by-dominant and dominant-by-dominant interactions, in addition to the dominant deviation effects. So, SCA for a hybrid from populations 1 and 2, considering two loci, *k* and *m*,

$$sca_{ij} = w_{(k)}d^k + w_{(m)}d^m + z_{1(k)} \otimes z_{2(m)}(\alpha\alpha)_{1,2}^{k,m} + z_{1(k)} \otimes w_{(m)}(\alpha d)_{1,2}^{k,m} + w_{(k)} \otimes z_{2(m)}(\alpha d)_{1,2}^{m,k} + w_{(k)} \otimes w_{(m)}(dd)_{1,2}^{k,m}$$

The different $z_1$and z_2 come from two parental lines *i* and *j*. Let $\boldsymbol{T}_1$ be a matrix relating hybrids to lines in population 1 with 1 in the *k,l* position if the *k*-th hybrid comes from the *l*-th line in population 1 and $\boldsymbol{T}_2$ a similar matrix linking hybrids to lines in population 2. The covariance matrix for the additive-by-additive interaction between populations 1 and 2 ($\boldsymbol{g}_{AA^{(12)}}$) can be calculated as:

$$\boldsymbol{G}_{AA^{(1,2)}} = \frac{\boldsymbol{T}_1\boldsymbol{G}_{A^{(1)}}\boldsymbol{T}_1' \odot \boldsymbol{T}_2\boldsymbol{G}_{A^{(2)}}\boldsymbol{T}_2'}{tr(\boldsymbol{T}_1\boldsymbol{G}_{A^{(1)}}\boldsymbol{T}_1' \odot \boldsymbol{T}_2\boldsymbol{G}_{A^{(2)}}\boldsymbol{T}_2')/n}$$

$$Var(\boldsymbol{g}_{AA^{(12)}}) = \boldsymbol{G}_{AA^{(1,2)}}\sigma^2_{AA^{(1,2)}}$$

where n is the number of hybrids and the matrix $\boldsymbol{G}_{AA^{(1,2)}}$ has size $n \times n$. In other words, the matrix $\boldsymbol{G}_{AA^{(1,2)}}$ is formed as follows:

1. For each pair of hybrids i, j with respective parents $parent1(i)$ (from population 1), $parent2(i)$ (from population 2) and $parent1(j)$ (from population 1) and $parent2(j)$ (from population 2) do:
   $\boldsymbol{G}_{AA^{(1,2)}}[i,j] = \boldsymbol{G}_{A^{(1)}}[parent1(i), parent1(j)] \times \boldsymbol{G}_{A^{(2)}}[parent2(i), parent2(j)]$
2. Scale the resulting matrix to an average diagonal of 1.

Technow *et al.* (2014) used $\boldsymbol{G}_{AA^{(1,2)}}$ to model the relationship matrix of SCA's. We have shown that this is incorrect because $\boldsymbol{G}_{AA^{(1,2)}}$ models across-population epistasis (interactions across loci) but it does not model dominance deviations (interactions within loci). We will see in the discussion section that $\boldsymbol{G}_{AA^{(1,2)}}$ is in fact an *approximation* of $\boldsymbol{D}$.

Relationships for the other pairwise epistatic interactions (all of them present in the SCA and of size $n \times n$) are:

- Additive in population 1 by dominant: $\boldsymbol{G}_{A^{(1)}D} = \frac{\boldsymbol{T}_1\boldsymbol{G}_{A^{(1)}}\boldsymbol{T}_1' \odot \boldsymbol{D}}{tr\left(\boldsymbol{T}_1\boldsymbol{G}_{A^{(1)}}\boldsymbol{T}_1' \odot \boldsymbol{D}\right)/n}$
- Additive in population 2 by dominant: $\boldsymbol{G}_{A^{(2)}D} = \frac{\boldsymbol{T}_2\boldsymbol{G}_{A^{(2)}}\boldsymbol{T}_2' \odot \boldsymbol{D}}{tr\left(\boldsymbol{T}_2\boldsymbol{G}_{A^{(2)}}\boldsymbol{T}_2' \odot \boldsymbol{D}\right)/n}$
- Dominant by dominant: $\boldsymbol{G}_{DD} = \frac{\boldsymbol{D} \odot \boldsymbol{D}}{tr(\boldsymbol{D} \odot \boldsymbol{D})/n}$

In the same manner, it is possible to derive relationships for third and higher order interactions, using Hadamard products of $\boldsymbol{G}_{A^{(1)}}$, $\boldsymbol{G}_{A^{(2)}}$ and $\boldsymbol{D}$ including the incidence matrices $\boldsymbol{T}$ for across-population interactions.

Since the two gametes in each hybrid are uncorrelated, the genotypic variance (Stuber and Cockerham, 1966) (ignoring third order epistatic terms) is

$$\sigma_G^2 = \sigma_{A^{(1)}}^2 + \sigma_{A^{(2)}}^2 + \sigma_D^2 + \sigma_{AA^{(1,1)}}^2 + \sigma_{AA^{(2,2)}}^2 + \sigma_{AA^{(1,2)}}^2 + \sigma_{A^{(1)}D}^2 + \sigma_{A^{(2)}D}^2 + \sigma_{DD}^2$$

where

$$\sigma_{GCA^{(1)}}^2 = \sigma_{A^{(1)}}^2 + \sigma_{AA^{(1,1)}}^2, \qquad \sigma_{GCA^{(2)}}^2 = \sigma_{A^{(2)}}^2 + \sigma_{AA^{(2,2)}}^2$$

and

$$\sigma_{SCA}^2 = \sigma_D^2 + \sigma_{AA^{(1,2)}}^2 + \sigma_{A^{(1)}D}^2 + \sigma_{A^{(2)}D}^2 + \sigma_{DD}^2.$$

In our analysis on real data, we will ignore the epistatic interaction terms $\sigma_{A^{(1)}D}^2$, $\sigma_{A^{(2)}D}^2$ and σ_{DD}^2 as their estimate is very inaccurate.

We remark that a breeder is interested in σ^2_{GCA} (because it indicates how much variation is expected in hybrids) but also in σ^2_A, which determines the genetic progress of hybrids that is achievable by selecting *within* the inbred populations (Stuber and Cockerham 1966). This is because epistasis combinations are broken down by recombination when creating new source populations for line development within each heterotic group.

MATERIALS AND METHODS

As illustration of the genomic relationship matrices developed here, variance components were estimated using the publicly available data set from the breeding program of the University of Hohenheim (https://doi.org/10.1534/genetics.114.165860) (Technow *et al.* 2014).

Phenotypes and genotypes

Here, a brief description of the phenotypic and genotypic data set is given (more details in Technow *et al.* (2014)). We analyzed the adjusted entry means of $n = 1254$ single-cross hybrids for grain yield expressed in quintals per hectare (q ha^{-1}), representing an incomplete factorial design between $n_1 = 123$ Dent and $n_2 = 86$ Flint inbred lines (two genetically divergent heterotic groups) with high linkage disequilibrium (LD) within heterotic groups. The hybrid data were collected in 14 years (1999-2012) and on average, 95 hybrids produced from 15 Dent and 11 Flint lines, were tested each year.

All parental inbred lines were genotyped with the Illumina Maize SNP50 BeadChip (Ganal *et al.* 2011). Here, we used the 35,478 SNP available after quality control (see details in Technow *et al.* 2014). Markers that were monomorphic in one group but segregating in the other group were kept. Genotypes of tested single-cross hybrids were derived from parental genotypes.

Genomic evaluation models

We used two different genomic models for analysis of hybrid records (either estimation of genetic parameters or cross-validation): the GCA-model (allele effects defined "according to origin") and the G-model (allele effects "uniquely defined"). The difference from ours to previous studies is that we use genomic relationship matrices that we have completely developed from theory (and not by transposition of pedigree-based concepts) in both the GCA-model and the G-model. For instance, Technow *et al.* (2014), in a GCA-model, modeled SCA's with a relationship matrix that ignores dominant deviations; Fritsche-Neto *et al.* (2018), working with the G-model, modeled SCA's with a relationship matrix that does not model dominance deviations (rather, they model genotypic effects).

GCA model. Here, effects were defined "according to origin" (Sprague and Tatum 1942; Griffing 1962; Stuber and Cockerham 1966), but with relationship matrices developed in the Theory section. So, GCA-model for the (i, j) hybrid resulting from the combination of parental lines i and j can be written as:

$$y_{ij} = \mu + f_{ij}b + gca_i^{(1)} + gca_j^{(2)} + sca_{ij} + e_{ij}$$

where y_{ij} is the hybrid phenotype (entry mean), μ is the overall mean, f_{ij} is the genomic inbreeding coefficient of hybrid i, j included as a covariate; and b is the inbreeding depression parameter (or the value of heterosis if the sign is switched – see explanation below). Then we make several choices of effects to include in GCA and SCA, leading to several models that will be described later. The most complete model includes $gca^{(1)} = g_{A^{(1)}} + g_{AA^{(1,1)}}$, $gca^{(2)} = g_{A^{(2)}} + g_{AA^{(2,2)}}$, $sca = g_D + g_{AA^{(1,2)}}$. We ignore second-order epistasis including dominance and higher order epistatic interactions. Thus, the most complete model is:

$$y_{ij} = \mu + f_{ij}b + g_{A^{(1)}\,i} + g_{A^{(2)}\,j} + g_{D\,ij} + g_{AA^{(1,1)}\,i} + g_{AA^{(2,2)}\,j} + g_{AA^{(1,2)}\,ij} + e_{ij}$$

which in vectorial form (all phenotypes in vector $\boldsymbol{y}$) is:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{f}b + \boldsymbol{T}_1\boldsymbol{g}_{A^{(1)}} + \boldsymbol{T}_2\boldsymbol{g}_{A^{(2)}} + \boldsymbol{g}_D + \boldsymbol{T}_1\boldsymbol{g}_{AA^{(1,1)}} + \boldsymbol{T}_2\boldsymbol{g}_{AA^{(2,2)}} + \boldsymbol{g}_{AA^{(1,2)}} + \boldsymbol{e}$$

Where the $\boldsymbol{T}$ incidence matrices assign hybrids to parents in each heterotic group.

The additive effects of gametes from each inbred line are assumed distributed as $\boldsymbol{g}_{A^{(1)}} \sim MVN\left(\mathbf{0}, \boldsymbol{G}_{A^{(1)}}\sigma^2_{A^{(1)}}\right)$ and $\boldsymbol{g}_{A^{(2)}} \sim MVN\left(\mathbf{0}, \boldsymbol{G}_{A^{(2)}}\sigma^2_{A^{(2)}}\right)$, the dominance deviation effects for each hybrid combination, as $\boldsymbol{g}_D \sim MVN(\mathbf{0}, \boldsymbol{D}\sigma^2_D)$; the epistatic interaction effects within each heterotic group are $\boldsymbol{g}_{AA^{(1,1)}} \sim MVN\left(\mathbf{0}, \boldsymbol{G}_{AA^{(1,1)}}\sigma^2_{AA^{(1,1)}}\right)$, $\boldsymbol{g}_{AA^{(2,2)}} \sim MVN\left(\mathbf{0}, \boldsymbol{G}_{AA^{(2,2)}}\sigma^2_{AA^{(2,2)}}\right)$ and between heterotic groups is $\boldsymbol{g}_{AA^{(1,2)}} \sim MVN\left(\mathbf{0}, \boldsymbol{G}_{AA^{(1,2)}}\sigma^2_{AA^{(1,2)}}\right)$. Finally, $\boldsymbol{e}$ is the vector of random residual effects $\boldsymbol{e} \sim MVN(\mathbf{0}, \boldsymbol{I}\sigma^2_e)$. All the relationship matrices have been defined in the Theory section.

G model. This model ignores the origin of the gametes and uses a "uniquely defined" effect per hybrid (Stuber and Cockerham 1966), as developed in a genomic context by Vitezica *et al.* (2017) using the NOIA approach, to correctly model dominance deviations under the constraint that hybrids are not in HWE. The G-model for single-cross hybrid individuals can be written as:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{f}b + \boldsymbol{g}_{A^{(H)}} + \boldsymbol{g}_{D^{(H)}} + \boldsymbol{g}_{AA^{(H)}} + \boldsymbol{e}$$

where $\boldsymbol{g}_{A^{(H)}}$ are the additive genetic effects of hybrids distributed as $\boldsymbol{g}_{A^{(H)}} \sim MVN(\mathbf{0}, \boldsymbol{G}_{A^{(H)}}\sigma^2_{A^{(H)}})$, the dominant genetic effects are $\boldsymbol{g}_{D^{(H)}} \sim MVN\left(\boldsymbol{O}, \boldsymbol{D}_H\sigma^2_{D^{(H)}}\right)$, and the additive-by-additive epistatic interaction effects are $\boldsymbol{g}_{AA^{(H)}} \sim MVN(\boldsymbol{O}, \boldsymbol{G}_{AA^{(H)}}\sigma^2_{AA^{(H)}})$.

The matrices $\boldsymbol{G}_{A^{(H)}}$, $\boldsymbol{D}_H$ and $\boldsymbol{G}_{AA^{(H)}}$ are the additive, dominant and additive-by-additive genomic relationship matrices, defined as (Vitezica *et al.* 2017)

$$\boldsymbol{G}_{A^{(H)}} = \frac{\boldsymbol{H}_a\boldsymbol{H}'_a}{tr(\boldsymbol{H}_a\boldsymbol{H}'_a)/n}$$

where the matrix $\boldsymbol{H}_a$ has elements equal to $(2-2p_k), (1-2p_k), -2p_k$ for genotypes BB, Bb and bb, and p_k is the frequency of B at the kth marker of the hybrid population. It is the same as VanRaden's **G** but with a different denominator to account for lack of HWE. The dominance matrix $\boldsymbol{D}_H$ is

$$\boldsymbol{D}_H = \frac{\boldsymbol{H}_d\boldsymbol{H}'_d}{tr(\boldsymbol{H}_d\boldsymbol{H}'_d)/n}$$

where $\boldsymbol{H}_d$ contains elements h_d for each individual and locus equal to

$$h_d = \begin{cases} -2[p_{BB} + p_{bb} - (p_{BB} - p_{bb})^2]^{-1} p_{Bb} p_{bb} \\ 4[p_{BB} + p_{bb} - (p_{BB} - p_{bb})^2]^{-1} p_{BB} p_{bb} \\ -2[p_{BB} + p_{bb} - (p_{BB} - p_{bb})^2]^{-1} p_{BB} p_{Bb} \end{cases} \quad \text{for genotypes} \begin{cases} BB \\ Bb \\ bb \end{cases}$$

according to Vitezica *et al.* (2017). This is different from the $\boldsymbol{D}$ matrix proposed by Su *et al.* (2012) which does not correctly model dominance deviations (and captures part of the GCA), and it is also different from the $\boldsymbol{D}$ matrix in Vitezica *et al.* (2013) (which assumes HWE). The additive-by-additive epistatic relationship matrix can be written as

$$\boldsymbol{G}_{AA^{(H)}} = \frac{\boldsymbol{G}_{A^{(H)}} \odot \boldsymbol{G}_{A^{(H)}}}{tr(\boldsymbol{G}_{A^{(H)}} \odot \boldsymbol{G}_{A^{(H)}})/n}$$

Modelling of heterosis. Typically, genetic models assume that additive ($\boldsymbol{a}$) and dominance ($\boldsymbol{d}$) effects are normally distributed with zero means, which is not true for $\boldsymbol{d}$ when directional dominance exists (i.e. a higher percentage of positive than negative dominance effects, $E(\boldsymbol{d}) \neq 0$). So, inbreeding depression (b, or heterosis) should be considered in the genetic evaluation model. This models correctly the fact that more diverse individual show positive heterosis at the trait, and avoids spurious overestimation of dominance variation (de Boer and Hoeschele 1993; Xiang *et al*. 2016, Aliloo *et al.* 2017). Therefore, the heterosis (or the inbreeding depression) of hybrids was also investigated to account for potential positive directional dominance, by including a regression on genomic inbreeding of hybrids (Xiang *et al.* 2016; Varona *et al.* 2018b). Genomic inbreeding was calculated as the proportion of homozygous SNP genotypes per hybrid individual (Silió *et al.* 2013).

Fit of "remaining" unrelated GCA effects. As discussed before, the GCA effect conceptually contains within-population additive effects and additive epistatic interactions. If these interactions are not explicitly modelled, and because pure lines are repeated in hybrids, model fitting tends to assign to the additive effects in the model all the remaining effect in the GCA attributable to the lines; namely, within-group epistasis not explicitly modelled. To verify this hypothesis, in addition to all relationship matrices and effects, we fit "remaining" unrelated Dent and Flint effects (similar to a permanent environmental effect in animal breeding) in GCA-models. For instance, if only additive effects are fit, this "remaining" effect should account for all additive epistasis within-group. Indeed, historically GCAs have been estimated as unrelated effects, for instance in diallel designs (Sprague and Tatum 1942; Hallauer *et al.* 1988). Fitting remaining unrelated GCA effects is similar to fitting individual permanent environmental

effects in animal breeding (e.g. for a cow that gives repeated performances of milk yield). In animal breeding, this effect captures non-genetic effects (e.g. poor nurturing at birth) but also genetic effects not explicitly modelled such as dominance or epistasis (Vitezica et al., 2018). This second source of effects is also plausible in a hybrid plant context. Thus, in a complementary series of models, we ran all the models detailed above (shown in Table 2) with an extra "remaining" unrelated GCA effect term.

Sub models and model comparison

Variance components were estimated for nested models (GCA-model and G-model) that added, in succession, additive effects (A), dominance effects (A+D), additive-by-additive genetic effects (A+D+(AA)). Also in another set of models "remaining effects" of lines were added to all the nested models. The additive-by-additive epistatic effects can be interactions of loci within line 1 $(AA)^{(11)}$, within line 2 $(AA)^{(22)}$ and interactions between loci across lines 1 and 2 $(AA)^{(12)}$, or within hybrids $(AA)^{(H)}$. For details see Table 2. Also, the variance attributable to GCA's and SCA's is $\sigma^2_{GCA^{(1)}} = \sigma^2_{A^{(1)}} + \sigma^2_{AA^{(1,1)}}$, $\sigma^2_{GCA^{(2)}} = \sigma^2_{A^{(2)}} + \sigma^2_{AA^{(2,2)}}$ and $\sigma^2_{SCA} = \sigma^2_D + \sigma^2_{AA^{(1,2)}}$ (plus the variance of "remaining" unrelated GCA effect, if fit).

Goodness-of-fit of models was compared based on the Deviance Information Criterion (DIC), which balances model fit and model complexity to avoid overfitting (Spiegelhalter *et al.* 2002). The lower the DIC value, the better fit of the model to the data.

Predicted ability of phenotypes of "untested hybrids" for the different models was tested performing a fivefold cross-validation. The data set was randomly divided into 5 folds of ~250 hybrids each (20% of all hybrids). Hybrid phenotypes from each fold in turn (testing set) were masked to exclude them from the training population. The remaining hybrids (80%) from the four-folds (training set) were used as training population to estimate genetic effects and then, predict the hybrid phenotypes from the fold in turn. The same training-testing sets were used to fit each of the models. Predictive performance of hybrids in the testing set (e.g. untested hybrids) was assessed by using Pearson's correlation coefficient ($r_{y,\hat{y}} = cor(y, \hat{y})$) between entry means ($y$) and predicted phenotypes ($\hat{y}$) ($\hat{y} = \hat{\mu} + f\hat{b} + \hat{g}$) which is equal to the sum of the estimated mean and covariate, and genetic values ($\hat{g}$). The mean squared error ($MSE = \sum_{i=1}^{nc}(y_i - \hat{y}_i)^2/n$) was also computed. We calculated correlations and MSE in the five cross-validation runs.

Table 2 Definition of genomic models for maize single-cross hybrids

Models	Effects	Model Code	Variances Additive	Dominance	Epistasis
GCA	$g_{A^{(1)}} + g_{A^{(2)}}$	$GCA\colon A$	$\sigma^2_{A^{(1)}}, \sigma^2_{A^{(2)}}$		
	$g_{A^{(1)}} + g_{A^{(2)}} + g_D$	$GCA\colon AD$	$\sigma^2_{A^{(1)}}, \sigma^2_{A^{(2)}}$	σ^2_D	
	$g_{A^{(1)}} + g_{A^{(2)}} + g_{AA^{(1,2)}}$	$GCA\colon A(AA)^{(1,2)}$	$\sigma^2_{A^{(1)}}, \sigma^2_{A^{(2)}}$		$\sigma^2_{AA^{(1,2)}}$
	$g_{A^{(1)}} + g_{A^{(2)}} + g_D + g_{AA^{(1,2)}}$	$GCA\colon AD(AA)^{(1,2)}$	$\sigma^2_{A^{(1)}}, \sigma^2_{A^{(2)}}$	σ^2_D	$\sigma^2_{AA^{(1,2)}}$
	$g_{A^{(1)}} + g_{A^{(2)}} + g_D +$ $+g_{AA^{(1,1)}} + g_{AA^{(2,2)}} + g_{AA^{(1,2)}}$	$GCA\colon AD(AA)^{(1,1)}(AA)^{(2,2)}(AA)^{(1,2)}$	$\sigma^2_{A^{(1)}}, \sigma^2_{A^{(2)}}$	σ^2_D	$\sigma^2_{AA^{(1,1)}}, \sigma^2_{AA^{(1,2)}}, \sigma^2_{AA^{(2,2)}}$
G	$g_{A^{(H)}}$	$G\colon A$	$\sigma^2_{A^{(H)}}$		
	$g_{A^{(H)}} + g_{D^{(H)}}$	$G\colon AD^{(H)}$	$\sigma^2_{A^{(H)}}$	$\sigma^2_{D^{(H)}}$	
	$g_{A^{(H)}} + g_{AA^{(H)}}$	$G\colon A(AA)^{(H)}$	$\sigma^2_{A^{(H)}}$		$\sigma^2_{AA^{(H)}}$
	$g_{A^{(H)}} + g_{D^{(H)}} + g_{AA^{(H)}}$	$G\colon AD^{(H)}(AA)^{(H)}$	$\sigma^2_{A^{(H)}}$	$\sigma^2_{D^{(H)}}$	$\sigma^2_{AA^{(H)}}$

GCA-model (effects (g) and variances (σ^2) defined within heterotic group): additive ($A^{(1)}$ and $A^{(2)}$), dominance (D), and additive-by-additive epistasis (AA) within heterotic groups ((1,1) and (2,2)) and between heterotic groups (1,2).

All the models detailed above were run without and with an extra "remaining" unrelated GCA effect term.

Estimation of variance components and cross-validation were performed in a Bayesian approach using the BGLR R-package (Pérez-Rodríguez and de los Campos 2014). The software can be downloaded from https://cran.r-project.org/web/packages/BGLR/index.html. To speed up computation, the eigenvalue decomposition of the variance-covariance matrices was done according to Acosta-Pech *et al.* (2017) and modeled as Bayesian Ridge Regression (BRR). For each model, inferences were based on 30,000 samples collected from 60,000 iterations after discarding 30,000 for burn-in and thinning of 10. Convergence of variance parameters was inspected by trace plots and convergence diagnostic was assessed using the BOA R-package (Smith 2007). A program to build the genomic matrices is available at http://genoweb.toulouse.inra.fr/~zvitezic/maize.

RESULTS

Variance components estimates and heritabilities

Variance components and broad-sense heritabilities (H^2) for GCA- and G-models are shown in Tables 3 and 4, and for the model including remaining GCA effects, in Table A3 and A4. Remember that the variance attributable to GCA's and SCA's is $\sigma^2_{GCA^{(1)}} = \sigma^2_{A^{(1)}} + \sigma^2_{AA^{(1,1)}}$, $\sigma^2_{GCA^{(2)}} = \sigma^2_{A^{(2)}} + \sigma^2_{AA^{(2,2)}}$ and $\sigma^2_{SCA} = \sigma^2_D + \sigma^2_{AA^{(1,2)}}$.

For GCA-models, estimates of GCA variance ($\sigma^2_{GCA^{(1)}} + \sigma^2_{GCA^{(2)}}$) were from 51.03 (in $GCA{:}\,AD(AA)^{(1,1)}(AA)^{(2,2)}(AA)^{(1,2)}$ model) to 55.91 (in $GCA{:}\,A$ model). In the full model, the inclusion of within-group epistatic variances ($\sigma^2_{AA^{(1,1)}}$ and $\sigma^2_{AA^{(2,2)}}$) reduced additive variances from 30.89 to 22.25 for $\sigma^2_{A^{(1)}}$, and from 21.84 to 14.27 for $\sigma^2_{A^{(2)}}$ (Table 3). However, the reduction was much less dramatic when remaining GCA effects were fit, from 22.58 to 19.38 and 12.76 to 10.65 for the same cases (Table A3).

The correlation between $\boldsymbol{G}_{A^{(1)}}$ and $\boldsymbol{G}_{AA^{(1,1)}}$ (and regression coefficient of $\boldsymbol{G}_{A^{(1)}} \sim \boldsymbol{G}_{AA^{(1,1)}}$) was 0.39 (0.92); and the correlation between $\boldsymbol{G}_{A^{(2)}}$ and $\boldsymbol{G}_{AA^{(2,2)}}$ (and regression coefficient of $\boldsymbol{G}_{A^{(2)}} \sim \boldsymbol{G}_{AA^{(2,2)}}$) was 0.55 (1.63). These results suggest a redundancy between these relationship matrices and explain that the additive effects tend to capture additive by additive effects if the

latter are not explicitly fit, something that is empirically partially corrected fitting remaining GCA effects.

In Table 3, the SCA variance estimates increased from 3.87 (in $GCA{:}\,AD$ model) to 6.57 (in $GCA{:}\,AD(AA)^{(1,2)}$ model). The estimates of σ^2_D (in the $GCA{:}\,AD$ model) and of $\sigma^2_{AA^{(1,2)}}$ (in the $GCA{:}\,A(AA)^{(1,2)}$) were equal to 3.87 and 5.04, respectively. However, the variance attributable to SCA (in $GCA{:}\,AD(AA)^{(1,2)}$) was lower than the addition of both variances (6.57 *vs.* 8.91). This reduction can be explained because $\boldsymbol{D}$ and $\boldsymbol{G}_{AA^{(1,2)}}$ matrices were highly correlated (0.88), suggesting that it is difficult to separate them completely. To avoid redundancy between the $\boldsymbol{D}$ and $\boldsymbol{G}_{AA^{(1,2)}}$ matrices, we corrected the $\boldsymbol{G}_{AA^{(1,2)}}$ matrix by subtracting the contribution of dominance as in Alves *et al.* (2019). However, similar correlation was obtained. Also the regression coefficient of $\boldsymbol{D}\sim\boldsymbol{G}_{AA^{(1,2)}}$ was 1.02, indicating that elements of $\boldsymbol{G}_{AA^{(1,2)}}$ are unbiased (but shrunken) estimators of the elements of $\boldsymbol{D}$.

For G-models, Table 3 shows that the additive variance estimate in the $G{:}\,A$ model (51.29) was lower than the addition of additive variance estimates for Dent and Flint groups (55.91) in the $GCA{:}\,A$ model. Although, estimates of dominance ($\sigma^2_{D^{(H)}}$) and epistasis within hybrids ($\sigma^2_{AA^{(H)}}$) variances were higher than in GCA-models. The inclusion of the additive-by-additive epistasis effects in the $G{:}\,AD(AA)^{(H)}$ model reduced the estimate of additive variance ($\sigma^2_{A^{(H)}}$) from 51.29 to 42.7. Furthermore, similar to GCA-models, the sum of estimates of $\sigma^2_{D^{(H)}}$ and $\sigma^2_{AA^{(1,2)}}$ (10.46) obtained with the $G{:}\,AD^{(H)}(AA)^{(H)}$ model was lower than the sum (14.37) of the estimates of $\sigma^2_{D^{(H)}}$ (in the $G{:}\,AD^{(H)}$ model) and of $\sigma^2_{AA^{(H)}}$ (in the $G{:}\,A(AA)^{(H)}$). These results can be explained because there was a correlation of 0.55 between $\boldsymbol{G}_{A^{(H)}}$ and $\boldsymbol{G}_{AA^{(H)}}$ and of 0.67 between $\boldsymbol{D}_H$ and $\boldsymbol{G}_{AA^{(H)}}$.

Estimates of genomic broad-sense heritability H^2 for grain yield ranged from 0.75 to 0.81, and from 0.74 to 0.79 in the full GCA- and G-models, respectively (Table 4). Estimates of residual variances were similar between GCA-models and G-models (Table 3). The estimates of residual variances decreased as the non-additive genetic effects were added in both GCA- and G-models.

Table 3 Estimated posterior means and standard deviation (in parenthesis) of genetic variance components obtained with the two types of models for maize grain yield

	Additive		Dominance	Epistasis			
Model Code	$\sigma^2_{A^{(1)}}, \sigma^2_{A^{(2)}}$ or $\sigma^2_{A^{(H)}}$		σ^2_D or $\sigma^2_{D^{(H)}}$	$\sigma^2_{AA^{(1,1)}}$	$\sigma^2_{AA^{(2,2)}}$	$\sigma^2_{AA^{(1,2)}}$ or $\sigma^2_{AA^{(H)}}$	σ^2_e
$GCA: A$	32.98 (5.50)	22.93 (4.54)					17.90 (0.78)
$GCA: AD$	31.53 (5.16)	23.05 (4.60)	3.87 (0.72)				15.19 (0.80)
$GCA: A(AA)^{(1,2)}$	30.99 (5.12)	22.37 (4.40)				5.04 (0.93)	13.91 (0.84)
$GCA: AD(AA)^{(1,2)}$	30.89 (5.10)	21.84 (4.36)	2.72 (0.59)			3.85 (0.84)	13.59 (0.82)
$GCA: AD(AA)^{(1,1)}(AA)^{(2,2)}(AA)^{(1,2)}$	22.25 (5.09)	14.27 (4.26)	2.32 (0.55)	6.70 (2.46)	7.81 (2.98)	3.43 (0.80)	13.70 (0.82)
$G: A$	51.29 (6.78)						17.84 (0.78)
$G: AD^{(H)}$	48.65 (6.42)		5.24 (0.97)				15.16 (0.79)
$G: A(AA)^{(H)}$	42.63 (6.18)					9.13 (1.78)	14.12 (0.86)
$G: AD^{(H)}(AA)^{(H)}$	42.7 (6.12)		3.65 (0.76)			6.81 (1.66)	13.75 (0.81)

Estimates of additive ($\sigma^2_{A^{(1)}}, \sigma^2_{A^{(2)}}$ or $\sigma^2_{A^{(H)}}$), dominance (σ^2_D or $\sigma^2_{D^{(H)}}$), additive-by-additive ($\sigma^2_{AA^{(1,1)}}$, $\sigma^2_{AA^{(2,2)}}$, $\sigma^2_{AA^{(1,2)}}$ or $\sigma^2_{AA^{(H)}}$) and residual (σ^2_e) variances for GCA- and G-models that included genomic inbreeding and successively added additive effects (A), dominance effects (AD), additive-by-additive effects (AD(AA)). Superscripts 1 and 2 in parenthesis refer to dent and flint heterotic groups respectively. The additive-by-additive epistatic effects can be interactions between loci within group ($(AA)^{(11)}$ and $(AA)^{(22)}$), across groups $(AA)^{(12)}$ or within hybrids $(AA)^{(H)}$.

Table 4 Estimated posterior means and standard deviation (in parenthesis) of broad-sense heritability and inbreeding depression obtained with the two types of models for maize grain yield

Model Code	H^2	b
$GCA: A$	0.75 (0.03)	-69.91 (23.56)
$GCA: AD$	0.79 (0.02)	-96.29 (26.79)
$GCA: A(AA)^{(1,2)}$	0.81 (0.02)	-52.62 (6.66)
$GCA: AD(AA)^{(1,2)}$	0.81 (0.02)	-14.82 (5.99)
$GCA: AD(AA)^{(1,1)}(AA)^{(2,2)}(AA)^{(1,2)}$	0.81 (0.02)	-26.78 (11.12)
$G: A$	0.74 (0.03)	-94.42 (19.84)
$G: AD^{(H)}$	0.78 (0.02)	-96.07 (10.33)
$G: A(AA)^{(H)}$	0.78 (0.02)	-66.16 (18.91)
$G: AD^{(H)}(AA)^{(H)}$	0.79 (0.02)	-35.85 (25.06)

GCA- and G-models are models that successively added additive effects (A), dominance effects (AD), and additive-by-additive genetic effects ($AD(AA)$). The additive-by-additive epistatic effects can be interactions between loci within group ($(AA)^{(11)}$ and $(AA)^{(22)}$), across groups $(AA)^{(12)}$ or within hybrids $(AA)^{(H)}$. Superscripts 1 and 2 in parenthesis refer to dent and flint heterotic groups respectively. H^2 is the genomic broad-sense heritability or narrow-sense heritability in pure additive models (GCA: A and G: A). b is the inbreeding depression parameter expressed as the change in phenotypic mean per 100% increase in genomic inbreeding.

Estimates of variance components in GCA- and G-models without the inclusion of genomic inbreeding are shown in Table A2. Ignoring the genomic inbreeding in the models resulted in higher estimates of dominance (σ^2_D or $\sigma^2_{D(H)}$), across-group epistasis ($\sigma^2_{AA^{(1,2)}}$), and within-hybrid epistasis ($\sigma^2_{AA^{(H)}}$) variances compared to models that included genomic inbreeding to account for directional dominance.

Effect of inbreeding

The average proportion of homozygosity or genomic inbreeding in this set of hybrids was $\bar{f}$=0.64 (SD=0.009). Estimates of inbreeding depression parameter b (or the value of heterosis if the sign is switched) are shown in Table 4. Results show that estimates of inbreeding depression parameter were negative, but with a wide range of variation among models. Inbreeding depression was significantly different from zero in most of the cases, except in model $G{:}\,AD^{(H)}(AA)^{(H)}$. Expressed as a change in phenotypic mean per 1% increase in genomic inbreeding coefficient, it varies from -0.14 to -0.96 (q ha^{-1}). Hence, heterosis has a positive impact on grain yield, as expected. However, variation of heterosis across hybrids is small as genomic homozygosity varies very little, because most heterozygosity is due to crossing heterotic groups and little to the specific lines that are being crossed.

Goodness of fit

Table 5 shows the Deviance Information Criteria (DIC) values for each model with and without genomic inbreeding. Inclusion of non-additive effects in GCA- and G- models, regardless of whether genomic inbreeding was included or not, improved the goodness of fit of both models. DIC values between GCA- and G-models were very similar. Among all models, the best model (with lower DIC value) was the $GCA{:}\,A(AA)^{(1,2)}$ without genomic inbreeding. Among the GCA-models, the best ones were those that accounted only for additive and additive-by-additive epistatic effects. Models including both dominance and epistatic effects had slightly worse DIC values than the best ones. This can be explained because increasing the number of parameters in the model penalizes DIC values.

Based on the inspection of the trace plot and convergence diagnostic with BOA R-package (Smith 2007), each of the variance parameter in all models converged.

Table 5 Deviance Information Criteria (DIC) values models with or without including genomic inbreeding

	Deviance Information Criteria (DIC)	
Model Code	With Inbreeding	Without inbreeding
$GCA: A$	7329.89	7337.56
$GCA: AD$	7264.52	7264.79
$GCA: A(AA)^{(1,2)}$	7209.10	7205.19
$GCA: AD(AA)^{(1,2)}$	7214.80	7210.35
$GCA: AD(AA)^{(1,1)}(AA)^{(2,2)}(AA)^{(1,2)}$	7211.20	7207.53
$G: A$	7325.36	7335.75
$G: AD^{(H)}$	7261.58	7260.88
$G: A(AA)^{(H)}$	7208.79	7206.89
$G: AD^{(H)}(AA)^{(H)}$	7214.48	7212.46

GCA- and G-models are models that successively added additive effects (A), dominance effects (AD), and additive-by-additive genetic effects ($AD(AA)$). The additive-by-additive epistatic effects can be interactions between loci within group (($AA)^{(11)}$ and $(AA)^{(22)}$), across groups $(AA)^{(12)}$ or within hybrids $(AA)^{(H)}$. Superscripts 1 and 2 in parenthesis refer to dent and flint heterotic groups respectively.

Cross-validation

The results for cross-validation are shown in Table 6. In general, prediction accuracy was considerably high for maize grain yield (~0.87) and similar values were obtained with the G and GCA models in this data set. The inclusion of non-additive genetic effects, as well as, the inclusion of genomic inbreeding in the models did not improve the prediction accuracy of

hybrid values in the testing set compared to models including only additive effects and without genomic inbreeding.

Table 6 Five-fold cross-validation results obtained for models with or without including of genomic inbreeding

	With Inbreeding		Without Inbreeding	
Model Code	MSE	$r_{y,\hat{y}}$	MSE	$r_{y,\hat{y}}$
$GCA: A$	20.741	0.869	20.854	0.868
$GCA: AD$	20.544	0.870	20.575	0.870
$GCA: A(AA)^{(1,2)}$	20.007	0.874	20.030	0.873
$GCA: AD(AA)^{(1,2)}$	20.362	0.871	20.406	0.871
$GCA: AD(AA)^{(1,1)}(AA)^{(2,2)}(AA)^{(1,2)}$	20.178	0.872	20.251	0.872
$G: A$	20.746	0.869	20.797	0.868
$G: AD^{(H)}$	20.430	0.871	20.544	0.870
$G: A(AA)^{(H)}$	19.846	0.875	19.899	0.874
$G: AD^{(H)}(AA)^{(H)}$	20.191	0.872	20.269	0.872

GCA- and G-models are models that successively added additive effects (A), dominance effects (AD), and additive-by-additive genetic effects ($AD(AA)$). The additive-by-additive epistatic effects can be interactions between loci within group ($(AA)^{(11)}$ and $(AA)^{(22)}$), across groups $(AA)^{(12)}$ or within hybrids $(AA)^{(H)}$. Superscripts 1 and 2 in parenthesis refers to dent and flint heterotic groups respectively. $r_{y,\hat{y}} = cor(y, \hat{y})$ is the Pearson's correlation coefficient between entry means (y) and predicted phenotypes ($\hat{y}$). $MSE = \sum_{i=1}^{nc}(y - \hat{y})^2/nc$ is the mean of squared errors, where nc is the number of hybrids.

DISCUSSION

In this study, the theory in the analysis of hybrid crosses of inbred lines from two populations using relationship matrices was revisited in a genomic context. Models for genomic

prediction in hybrid crops using the notions of effects defined "according to origin" (GCAs and SCAs) were rederived and expressions for additive, dominant and epistatic relationships for hybrids were presented. These models were applied to a public data set to exemplify the theory and its consequences in real life.

Insights into relationships for Dominance and across-population pairwise Epistasis

A surprising fact (to us) is that, in the classical pedigree-based methods, it is not possible to disentangle dominance deviations from across-population epistasis, whereas using markers it is possible. This seems not to have been recognized by previous researchers, leading to the wrong conclusion that in a genomic setting the relationship of dominance deviations is a cross-product of additive relationships of parental lines. In this section we try to explain why such a difference.

Stuber and Cockerham (1966) used the notion of identity by descent (IBD) coefficients to model relationships, where the starting block is the use of coancestries Φ - the probability that two alleles drawn at random from each of two pure lines are identical by descent. Although in our work we use genomic relationships (that are not probabilities), the concept of IBD is useful in the following.

For two hybrids, the IBD dominance relationship coefficient at locus *k* (say $\delta^{(k)}$) is the probability that two complete genotypes at locus *k* in hybrids (*i* and *j*) are identical, and because the lines are fully inbred, this is the joint probability that both "parents1" (ancestors from population 1) are IBD at locus *k* (with probability $\Phi^{(k)}_{parent1(i)parent1(j)}$) and both "parents2" (ancestors from population 2) are IBD at locus *k* (with probability $\Phi^{(k)}_{parent2(i)parent2(j)}$). See Figure 1. This results in $\delta^{(k)}_{ij} = \Phi^{(k)}_{parent1(i)parent1(j)}\Phi^{(k)}_{parent2(i)parent2(j)}$. Across all *m* loci, $\delta_{ij} = \frac{1}{m}\sum_{k=1,m}\delta^{(k)}_{ij} = \frac{1}{m}\sum_{k=1,m}\left(\Phi^{(k)}_{parent1(i)parent1(j)}\Phi^{(k)}_{parent2(i)parent2(j)}\right)$. However, in practice, pedigree-based coancestries at specific loci are not observable and they are replaced by infinitesimal coancestries:

$$\delta_{ij} = \frac{1}{m}\sum_{k=1,m}\left(\Phi^{(k)}_{parent1(i)parent1(j)}\Phi^{(k)}_{parent2(i)parent2(j)}\right)$$

$$\approx \frac{1}{m}\sum_{k=1,m}\left(\Phi_{parent1(i)parent1(j)}\,\Phi_{parent2(i)parent2(j)}\right)$$

$$= \Phi_{parent1(i)parent1(j)}\,\Phi_{parent2(i)parent2(j)}$$

resulting in $\delta_{ij} \approx \Phi_{parent1(i)parent1(j)}\,\Phi_{parent2(i)parent2(j)}$ which is the expression presented by Stuber and Cockerham (1966). The approximation results from the fact that the genome is finite. For instance, if there were m=3 loci, there would be 3 local IBD, one at each locus, whose average will not, in general, be the same as the pedigree-based IBD which does assume infinite loci (Hill and Weir 2011). In an infinitesimal model, the approximation is exact.

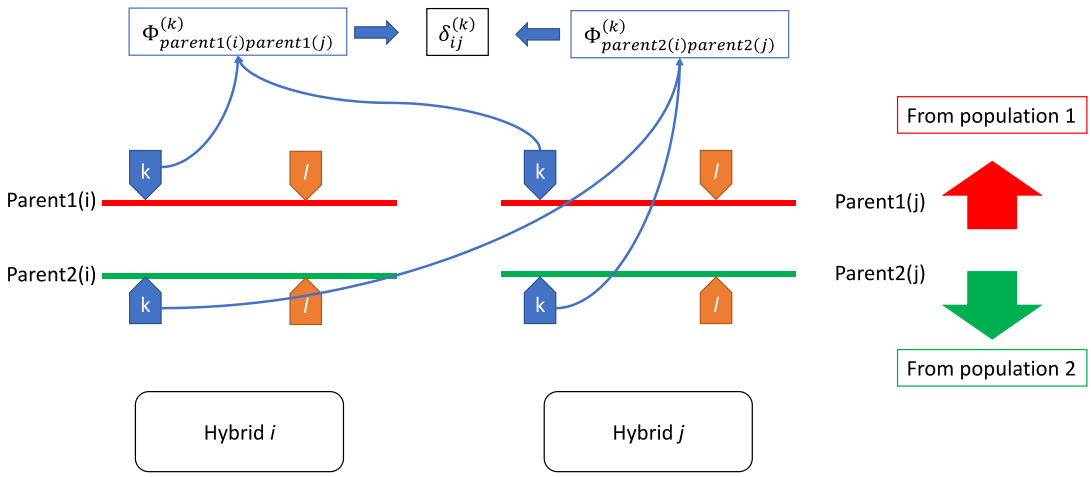

Figure 1. Dominance relationship across two hybrids for locus k.

Now we address the across-population epistatic additive by additive relationships. Consider two loci *k* and *l*. In an IBD framework, across-population epistatic additive by additive relationship for hybrids *i* and *j* at two loci *k* and *l* (say $\psi_{ij}^{(k,l)}$) is the joint probability that both "parents1" (ancestors from population 1) have the same genotype at locus *k* and that both "parents2" (ancestors from population 2) have the same genotype at locus *l*. See Figure 2. Thus

$\psi_{ij}^{(k,l)} = \Phi_{parent1(i)parent1(j)}^{(k)} \Phi_{parent2(i)parent2(j)}^{(l)}$. Whole-genome epistatic relationship would therefore be

$$\psi_{ij} = \frac{1}{m(m-1)} \sum_{k,l,k\neq l} \Phi_{parent1(i)parent1(j)}^{(k)} \Phi_{parent2(i)parent2(j)}^{(l)}$$

However, based on pedigree, the different $\Phi_{parent1(i)parent1(j)}^{(l)}$ and $\Phi_{parent2(i)parent2(j)}^{(k)}$ are not observable and they are replaced by infinitesimal coancestries resulting in the approximation $\psi_{ij} \approx \Phi_{parent1(i)parent1(j)} \Phi_{parent2(i)parent2(j)}$. Again, in an infinitesimal model the approximation is exact.

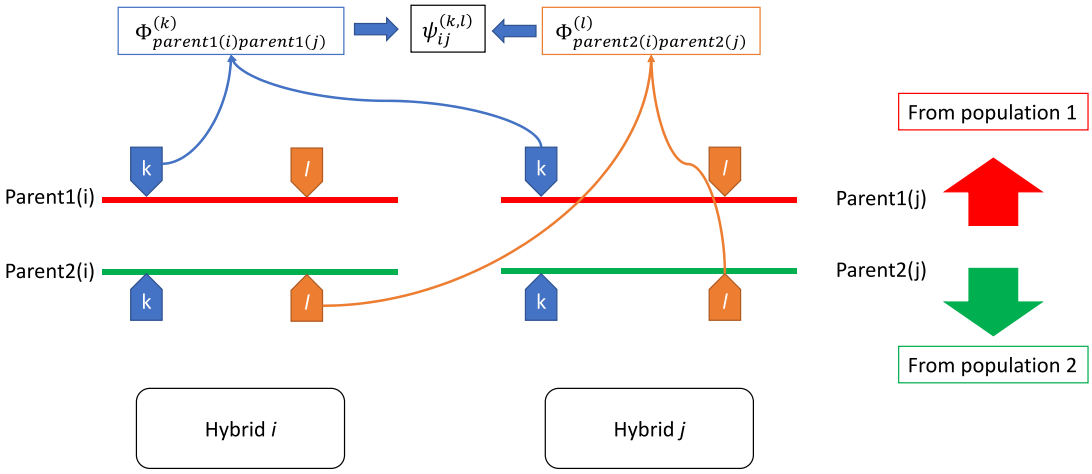

Figure 2. Additive by additive epistatic across-population relationship across two hybrids for locus k and l.

It is worth noting that $\psi_{ij} = \frac{1}{m(m-1)} \sum_{k,l,k\neq l} \Phi_{parent1(i)parent1(j)}^{(l)} \Phi_{parent2(i)parent2(j)}^{(k)}$ involves relationships *across pairs* of loci whereas $\delta_{ij} = \frac{1}{m} \sum_{k=1,m} \left(\Phi_{parent1(i)parent1(j)}^{(k)} \Phi_{parent2(i)parent2(j)}^{(k)} \right)$ involves relationships *within single* loci. On average, pairs of loci are transmitted in a manner similar to transmission of single locus (for instance two neighboring markers are often transmitted together), which explains why ψ_{ij} is an estimator (albeit not necessarily a good one) of δ_{ij}, and it explains why the elements of $\boldsymbol{G}_{AA^{(1,2)}}$

are unbiased (but not necessarily accurate) estimators of the elements of $\boldsymbol{D}$, as shown by the results.

Thus, we have shown that the Stuber and Cockerham (1966) relationships assuming pedigrees are only exact under infinitesimal models. In previous sections we have shown that observing the genome (i.e. with markers), different relationships can be formed for each, additive substitution and dominant and epistatic deviation. Thus, contrary to pedigree-based formulations, a marker-based formulation allows a better disentangling of the different variance components.

Thus, in pedigree-based models the dominance and across-population epistatic relationships are conceptually different, but the lack of other information forces to use the same estimator for both. This is not the case in marker-based models, where we can actually observe different relationships within locus or across loci from two populations.

Partition of genetic variance components and heritability

The partition of the genetic variance in terms of statistical additive effects, and dominance and epistatic deviations effects, was possible using the relationship matrices developed here. In our model, estimates of additive genetic variance based on allele substitution effects are useful for selection or in the prediction of potential selection response in pool improvement. Vitezica *et al.* (2013) compared a classical model (in terms of statistical values for breeding purposes) with a genotypic model (biological values at the gene level) proposed by Su *et al.* (2012). When the genotypic model is used, additive and dominant genotypic variances are obtained. Both models are able to explain the data but their results and interpretation are different (Vitezica *et al.* 2013; Varona *et al.* 2018a). The genotypic model has been used for hybrid genomic prediction (Werner *et al.* 2018; Fritsche-Neto *et al.* 2018; Alves *et al.* 2019; Ramstein *et al.* 2020), but estimates of genotypic additive variance should not be interpreted for breeding purposes. Furthermore, the G-model based on the NOIA approach and accounting for deviations from HWE is able to obtain estimates of additive and dominance deviations genetic variances (Vitezica *et al.* 2017, 2018). The GCA- (proposed here) and G-models are equivalent models to explain the data only if all relevant gene actions (i.e., high order interactions) are included (Stuber and Cockerham, 1966), but it is impossible to ascertain if all relevant

interactions are included. In our results, both definitional systems perform similarly for prediction. However, as the G-model assumes gene effects uniquely within hybrids and does not provide additive values within pool, it can not be directly used for the selection of inbred lines within pools for recurrent pool improvement.

When within-group epistatic effects are not fit in the GCA-model, additive variance is overestimated. Similar results were observed by Bernardo (1995). He attributed this to multicollinearity between the additive and within-group epistatic relationships, as we observe. After fitting the remaining unrelated GCA effect, in all the models, additive variances were similar (~22 and ~12 for group 1 and 2, respectively: see Table A3) showing empirical orthogonality (Hill and Mäki-Tanila 2015; Vitezica *et al.* 2017). For planning the breeding scheme (to estimate genetic gain and selection of within pools crosses), it is important to obtain good estimates of the genetic variance, and therefore we recommend fitting within-group epistasis or, if not possible, unrelated "remaining" line effects, in order to avoid overestimation of the genetic variance. The latter option is only possible if each line contributed to several phenotyped hybrids. However, with the availability of markers, the trend is to increase the number of lines tested by reducing their contribution to hybrids.

In the G-model, we also observed a drop (but much smaller) in estimated genetic variance when fitting epistasis, that we attribute to essentially the same phenomena. However, it was not possible to fit "unrelated" hybrid effects in this model because in our dataset each hybrid has a single record (adjusted entry means).

Similarly, a reduction in dominance variance and epistasis, either across groups (in the GCA-model) or within hybrids (in the G-model), was observed when both were simultaneously estimated. We explain this because the $\boldsymbol{G}_{AA^{(1,2)}}$ matrix is very similar to $\boldsymbol{D}$, and not fitting one of the effects leads to the other one capturing the variance.

Genomic relationship matrices for within and across groups epistasis (in the full GCA-model) allows to partition the genetic variance in terms of GCA and SCA effects, as was originally defined by Stuber and Cockerham (1966) in an infinitesimal context. With our model, it is possible to split the GCA effect into the additive gametic effect and the additive-by-additive epistasis interaction within the line; and split the SCA effect into dominance deviation effect and additive-by-additive epistasis across groups. This has practical implications in hybrid breeding programs that will be discussed later.

Compared to the estimates of genetic variance component from Technow *et al.* (2014), we obtained similar estimates with the GCA: $A(AA)^{(1,2)}$ model without genomic inbreeding (see Table A2). This makes sense because, as indicated in the theory, the estimate of SCA variation from Technow *et al.* (2014), is in fact the estimate of epistasis variation across populations $\sigma^2_{AA^{(1,2)}}$. Their entry-mean heritability was 0.87, whereas our genomic estimate of broad-sense heritability was slightly lower (0.81). Differences between our and their estimates are mainly because we used entry means (publicly available) instead of the whole data set, which can be seen through estimated residual variance which was much lower (e.g. ~17) than their estimated values (179).

Effect of heterosis on estimates of non-additive genetic variance

The individual heterosis (or inbreeding depression) was also investigated to account for positive directional dominance. In maize, previous studies have reported some evidence of directional dominance, particularly for grain yield (Reif *et al.* 2003; Ramstein *et al.* 2020). In our results, estimates of dominance variance and epistasis across groups or within hybrids were inflated when genomic inbreeding was ignored in the model (Table A2), compared to the model that included it (Table 3). Our results agree with previous studies in the literature. If there is directional dominance and it is not accounted for in genetic evaluations, estimates of dominance variance may be overestimated, as was proved analytically by Boer and Hoeschele (1993) in a pedigree context and Xiang *et al.* (2016) in a genomic context, and empirically by other studies (Aliloo *et al.* 2017; Varona *et al.* 2018a; Vitezica *et al.* 2018). Here, we included the genomic inbreeding coefficient, measured as the proportion of observed homozygous marker genotypes in hybrids, to estimate the inbreeding depression (Varona *et al.* 2018b). Alternatively, the heterozygosity can be measured as the genetic distance between parental lines of a hybrid, which lead to expected heterozygosity of the cross (Bernardo 1992; Charcosset and Essioux 1994). A previous study in maize hybrids including the genetic distance between parental lines (expected heterozygosity) also reported a reduction in non-additive variances (Larièpe *et al.* 2017).

Inbreeding depression effect

The average proportion of homozygosity or genomic inbreeding in this set of hybrids showed a very small standard deviation ($\bar{f}$=0.64, SD=0.009). Our genomic estimates of inbreeding depression (ranging from -0.96 to -0.14 (q ha^{-1}) expressed as a change in phenotypic mean per 1% increase in inbreeding) are in range with those reported in the literature in different maize populations based on pedigree (from -0.78 to -0.2 (q ha^{-1})) (Hallauer *et al.* 2010). If we change the sign, this can be seen as the positive effect of heterosis. Estimates of inbreeding depression in hybrid populations and in genomic context are scarce. Recently, Ramstein *et al.* (2020) reported a genomic estimate of inbreeding depression for grain yield in a maize hybrid population (-0.69 (q ha^{-1}). Our results agree with the literature in the sense that inbreeding depression in hybrids has a negative impact on grain yield. Note that the genomic inbreeding coefficient is interpreted as fixation of alleles that results in increased homozygosity or loci identical by state. Therefore, it is possible to estimate inbreeding depression even when hybrid individuals are derived from two populations unrelated by pedigree. In the literature, estimates of inbreeding depression (or of its opposite, heterosis) have also been reported in crossbred animals (e.g. pigs and tilapia) derived from two unrelated breeds (Xiang *et al.* 2016; Iversen *et al.* 2019; Joshi *et al.* 2020).

Goodness of fit

Models with lower DIC values better fit the data, and a difference of less than 7 units is often considered irrelevant (Plummer *et al.* 2006). In general, the inclusion of non-additive genetic effects improved the goodness of fit to the data in both GCA- and G- models in this set of hybrids. This result agrees with previous studies in maize hybrids (Ferrão *et al.*; Alves *et al.* 2019; Hunt *et al.* 2020). DIC values obtained with the G-model were similar to those obtained in GCA-models, indicating that they are equivalent models in terms of fitting the data. The model with the best balance between goodness of fit and model complexity was the $GCA{:}\,A(AA)^{(1,2)}$ without genomic inbreeding, which corresponds to the frequently used model in genomic prediction of hybrids (Technow *et al.* 2014). That means this model is efficient to fit the data. However, fit to the data is not the only aspect that should be considered – interpretation of the model in a genetic context is important.

Inclusion of genomic inbreeding or heterozygosity improved goodness of fit only in additive model. This has also been reported in previous studies in animal breeding (Aliloo *et al.* 2017). The inclusion of genomic inbreeding in models that accounted for non-additive genetic effects resulted in similar DIC values in this set of hybrids – neither improvement nor deterioration.

Cross-validation

Overall, cross-validation analyses yielded a high prediction accuracy of hybrid performance (~0.87). This is because a high heritability generally results in high prediction accuracy, as was demonstrated theoretically and empirically (Daetwyler *et al.* 2010; Combs and Bernardo 2013). Inclusion of non-additive genetic effects did not show improvement in prediction accuracy. This result agrees with other studies using real data where virtually no benefit was observed by including SCA effects in genomic prediction models of inter-heterotic-group hybrids (Bernardo 1994; Schrag *et al.* 2006, 2018; Maenhout *et al.* 2010; Kadam *et al.* 2016). This is because in inter-heterotic-group hybrids the proportion of SCA variance is often low and GCA high (Reif *et al.* 2007).

Assigning the fivefold at random is comparable to the T2 method presented by Technow *et al.* (2014), which corresponds to groups of hybrids with two parents known in the training set. Dividing our correlations by $\sqrt{H^2}$, we obtain correlations comparables to those reported by Technow *et al.* (2014). For instance, the correlation obtained with the $GCA{:}\,AD(AA)^{(1,2)}$ results in a value of 0.97 by dividing by $\sqrt{H^2}$, which is close to the correlation of 0.92 (for 450 hybrids in the training set) reported by Technow *et al.* (2014) for grain yield.

Assuming equal marker effects between groups (G-models) gave similar prediction accuracy than assuming gene effects according to origin (GCA-models). This result was also reported by Technow *et al.* (2014) with the same data set, but also by Alves *et al.* (2019) who analyzed a population of hybrids derived from a convergent population. Thus, GCA- and G-models are equivalent in terms of predictive ability of hybrid performance.

Practical implications in hybrid breeding

The way of partitioning the genetic variance is a matter of convenience. Partitioning in terms of GCA within group is more convenient because inbred lines are selected within group. The magnitude of the GCA variance, gives to the breeder an idea of how much variation is expected in the hybrids. Further, splitting the GCA variance into additive and epistasis within group may be relevant at the moment of planning the genetic progress in maize breeding programs. The genetic improvement in hybrid performance is through the selection of inbred lines. So that, breeders create new segregating (e.g. F2) populations by crossing elite lines within groups followed by subsequent generation of inbreeding to develop new inbred lines. Thereby, the additive-by-additive epistasis across loci in alleles present in the elite lines is not transmitted to their F2 (and further selfing) progeny because recombination breaks down the original epistatic combinations present in the elite inbred lines and creates new epistatic combinations. Thus, the use of the additive variance, instead of the GCA variance, is more appropriate for the prediction of genetic progress that is achievable by selecting *within* heterotic pools (Stuber and Cockerham 1966). Besides, variance of epistasis within groups is expected to be converted in new additive genetic variance in the long term by random drift, thus, it affects the long-term selection response indirectly (Hill 2017). Also, for pool improvement, it is better to use estimates of additive effects instead of estimates of GCA, because the first reflects better expected genetic progress.

Splitting the SCA variance into dominance deviations and epistasis across groups could also have practical implications. Estimates of additive and dominance effects might be important for hybrid pool development. For instance, Zhao *et al.* (2015) suggested to use additive and dominance effects from an incomplete factorial in order to develop heterotic pools in wheat. Further, estimates of dominance deviation variance are relevant in the definition of mate allocation procedures (Varona *et al.* 2018a); for instance, it could be used to maximize hybrid performance or maintain diversity for long-term genetic gain in hybrid breeding programs (e.g. Allier *et al.* 2019).

CONCLUSIONS

Models using the genomic relationships developed here allow for a proper partition of statistical additive effects (one per parent) and dominance and epistasis deviations in hybrids derived from inbred lines from two populations. Contrary to common belief, using SNP genotypes, it is possible to split specific combining ability (SCA) into dominance deviations and across-groups epistasis, and to split general combining ability (GCA) into within-line additive effects and within-line additive by additive epistasis effects. Our model is more appropriate for using genomic data and estimates of genetic variance components can be interpreted for breeding purposes.

Furthermore, the G-model following the NOIA approach is equivalent to the GCA model in terms of goodness of fit and predictive ability of hybrid performance, but they split the genetic variance differently and their interpretations must be different. Inbreeding depression has a negative impact on grain yield or, equivalently, molecular heterozygosity translates into heterosis in grain yield. Genomic models accounting for non-additive genetic effects must include heterozygosity (or genomic inbreeding) to account for directional dominance and obtain unbiased estimates of variance components. Inclusion of non-additive genetic effects improved the goodness of fit of models, whereas genomic inbreeding only improved the pure additive models.

ACKNOWLEDGMENTS

DGD is funded by France Génétique Porc, RAGT and INRA (contract 15000416, project 00000837). Work was partly supported by INRA SELGEN metaprogram (project OptiMaGics) and by the "Investissement d'Avenir" 112 project "Amaizing" (Amaizing, ANR-10-BTBR-0001). Project was partly supported by Toulouse Midi-Pyrénées bioinformatics platform.

LITERATURE CITED

Acosta-Pech R., J. Crossa, G. de los Campos, S. Teyssèdre, B. Claustres, *et al.*, 2017 Genomic models with genotype × environment interaction for predicting hybrid performance: an application in maize hybrids. Theor. Appl. Genet. 130: 1431–1440. https://doi.org/10.1007/s00122-017-2898-0

Aliloo H., J. E. Pryce, O. González-Recio, B. G. Cocks, M. E. Goddard, *et al.*, 2017 Including nonadditive genetic effects in mating programs to maximize dairy farm profitability. J. Dairy Sci. 100: 1203–1222. https://doi.org/10.3168/jds.2016-11261

Allier A., C. Lehermeier, A. Charcosset, L. Moreau, and S. Teyssèdre, 2019 Improving Short- and Long-Term Genetic Gain by Accounting for Within-Family Variance in Optimal Cross-Selection. Front. Genet. 10. https://doi.org/10.3389/fgene.2019.01006

Álvarez-Castro J. M., and Ö. Carlborg, 2007 A Unified Model for Functional and Statistical Epistasis and Its Application in Quantitative Trait Loci Analysis. Genetics 176: 1151–1167. https://doi.org/10.1534/genetics.106.067348

Alves F. C., Í. S. C. Granato, G. Galli, D. H. Lyra, R. Fritsche-Neto, *et al.*, 2019 Bayesian analysis and prediction of hybrid performance. Plant Methods 15: 14. https://doi.org/10.1186/s13007-019-0388-x

Bernardo R., 1992 Relationship between single-cross performance and molecular marker heterozygosity. Theor. Appl. Genet. 83: 628–634. https://doi.org/10.1007/BF00226908

Bernardo R., 1994 Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. Crop Sci. 34: 20–25. https://doi.org/10.2135/cropsci1994.0011183X003400010003x

Bernardo R., 1995 Genetic Models for Predicting Maize Single-Cross Performance in Unbalanced Yield Trial Data. Crop Sci. 35: cropsci1995.0011183X003500010026x. https://doi.org/10.2135/cropsci1995.0011183X003500010026x

Bernardo R., 2010 *Breeding for quantitative traits in plants*. Stemma Press, Woodbury.

Boer I. J. M. de, and I. Hoeschele, 1993 Genetic evaluation methods for populations with dominance and inbreeding. Theor. Appl. Genet. 86: 245–258. https://doi.org/10.1007/BF00222086

Bouvet J.-M., G. Makouanzi, D. Cros, and P. Vigneron, 2016 Modeling additive and non-additive effects in a hybrid population using genome-wide genotyping: prediction accuracy implications. Heredity 116: 146–157. https://doi.org/10.1038/hdy.2015.78

Charcosset A., and L. Essioux, 1994 The effect of population structure on the relationship between heterosis and heterozygosity at marker loci. Theor. Appl. Genet. 89: 336–343. https://doi.org/10.1007/BF00225164

Cockerham C. C., 1954 An Extension of the Concept of Partitioning Hereditary Variance for Analysis of Covariances among Relatives When Epistasis Is Present. Genetics 39: 859–882.

Combs E., and R. Bernardo, 2013 Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers. Plant Genome 6: plantgenome2012.11.0030. https://doi.org/10.3835/plantgenome2012.11.0030

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

Falconer D. S., 1981 *Introduction to quantitative genetics*. Longman Goup Ltd, New York.

Ferrão L. F. V., C. D. Marinho, P. R. Munoz, and M. F. R. R. Jr, Improvement of predictive ability in maize hybrids by including dominance effects and marker × environment models. Crop Sci. n/a. https://doi.org/10.1002/csc2.20096

Fritsche-Neto R., D. Akdemir, and J.-L. Jannink, 2018 Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. Theor. Appl. Genet. 131: 1153–1162. https://doi.org/10.1007/s00122-018-3068-8

Ganal M. W., G. Durstewitz, A. Polley, A. Bérard, E. S. Buckler, *et al.*, 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: e28334. https://doi.org/10.1371/journal.pone.0028334

Griffing B., 1962 Prediction Formulae for General Combining Ability Selection Methods Utilizing One or Two Random-Mating Populations. Aust. J. Biol. Sci. 15: 650–665. https://doi.org/10.1071/bi9620650

Hallauer A. R., W. A. Russell, and K. R. Lamkey, 1988 Corn Breeding, pp. 463–564 in *Corn and Corn Improvement*, edited by Sprague G. F., Dudley J. W. ASA-CSSA-SSSA, Madison, WI.

Hallauer A. R., M. J. Carena, and J. B. M. Filho, 2010 *Quantitative Genetics in Maize Breeding*. Springer-Verlag, New York.

Hill W. G., and B. S. Weir, 2011 Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. 93: 47–64. https://doi.org/10.1017/S0016672310000480

Hill W. G., and A. Mäki-Tanila, 2015 Expected influence of linkage disequilibrium on genetic variance caused by dominance and epistasis on quantitative traits. J. Anim. Breed. Genet. 132: 176–186. https://doi.org/10.1111/jbg.12140

Hill W. G., 2017 “Conversion” of epistatic into additive genetic variance in finite populations and possible impact on long-term selection response. J. Anim. Breed. Genet. 134: 196–201. https://doi.org/10.1111/jbg.12270

Hunt C. H., B. J. Hayes, F. A. van Eeuwijk, E. S. Mace, and D. R. Jordan, 2020 Multi-environment analysis of sorghum breeding trials using additive and dominance genomic relationships. Theor. Appl. Genet. https://doi.org/10.1007/s00122-019-03526-7

Iversen M. W., Ø. Nordbø, E. Gjerlaug-Enger, E. Grindflek, M. S. Lopes, *et al.*, 2019 Effects of heterozygosity on performance of purebred and crossbred pigs. Genet. Sel. Evol. 51: 8. https://doi.org/10.1186/s12711-019-0450-1

Joshi R., T. H. E. Meuwissen, J. A. Woolliams, and H. M. Gjøen, 2020 Genomic dissection of maternal, additive and non-additive genetic effects for growth and carcass traits in Nile tilapia. Genet. Sel. Evol. 52: 1. https://doi.org/10.1186/s12711-019-0522-2

Kadam D. C., S. M. Potts, M. O. Bohn, A. E. Lipka, and A. J. Lorenz, 2016 Genomic Prediction of Single Crosses in the Early Stages of a Maize Hybrid Breeding Pipeline. G3 GenesGenomesGenetics 6: 3443–3453. https://doi.org/10.1534/g3.116.031286

Larièpe A., L. Moreau, J. Laborde, C. Bauland, S. Mezmouk, *et al.*, 2017 General and specific combining abilities in a maize (Zea mays L.) test-cross hybrid panel: relative importance of population structure and genetic divergence between parents. Theor. Appl. Genet. 130: 403–417. https://doi.org/10.1007/s00122-016-2822-z

Legarra A., 2016 Comparing estimates of genetic variance across different relationship models. Theor. Popul. Biol. 107: 26–30. https://doi.org/10.1016/j.tpb.2015.08.005

Maenhout S., B. De Baets, and G. Haesaert, 2010 Prediction of maize single-cross hybrid performance: support vector machine regression versus best linear prediction. Theor. Appl. Genet. 120: 415–427. https://doi.org/10.1007/s00122-009-1200-5

Massman J. M., A. Gordillo, R. E. Lorenzana, and R. Bernardo, 2013 Genomewide predictions from maize single-cross data. Theor. Appl. Genet. 126: 13–22. https://doi.org/10.1007/s00122-012-1955-y

Pérez-Rodríguez P., and G. de los Campos, 2014 Genome-Wide Regression and Prediction with the BGLR Statistical Package. Genetics 198: 483–495. https://doi.org/10.1534/genetics.114.164442

Plummer M., N. Best, K. Cowles, and K. Vines, 2006 CODA: convergence diagnosis and output analysis for MCMC. R News 6: 7–11.

Ramstein G. P., S. J. Larsson, J. P. Cook, J. W. Edwards, E. S. Ersoz, *et al.*, 2020 Dominance Effects and Functional Enrichments Improve Prediction of Agronomic Traits in Hybrid Maize. Genetics. https://doi.org/10.1534/genetics.120.303025

Reif J. C., A. E. Melchinger, X. C. Xia, M. L. Warburton, D. A. Hoisington, *et al.*, 2003 Genetic Distance Based on Simple Sequence Repeats and Heterosis in Tropical Maize Populations. Crop Sci. 43: 1275–1282. https://doi.org/10.2135/cropsci2003.1275

Reif J. C., F.-M. Gumpert, S. Fischer, and A. E. Melchinger, 2007 Impact of Interpopulation Divergence on Additive and Dominance Variance in Hybrid Populations. Genetics 176: 1931–1934. https://doi.org/10.1534/genetics.107.074146

Schrag T. A., A. E. Melchinger, A. P. Sørensen, and M. Frisch, 2006 Prediction of single-cross hybrid performance for grain yield and grain dry matter content in maize using AFLP markers associated with QTL. Theor. Appl. Genet. 113: 1037–1047. https://doi.org/10.1007/s00122-006-0363-6

Schrag T. A., M. Westhues, W. Schipprack, F. Seifert, A. Thiemann, *et al.*, 2018 Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize. Genetics 208: 1373–1385. https://doi.org/10.1534/genetics.117.300374

Silió L., M. c. Rodríguez, A. Fernández, C. Barragán, R. Benítez, *et al.*, 2013 Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics. J. Anim. Breed. Genet. 130: 349–360. https://doi.org/10.1111/jbg.12031

Smith B. J., 2007 boa: An R Package for MCMC Output Convergence Assessment and Posterior Inference. J. Stat. Softw. 21: 1–37. https://doi.org/10.18637/jss.v021.i11

Spiegelhalter D. J., N. G. Best, B. P. Carlin, and A. V. D. Linde, 2002 Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B Stat. Methodol. 64: 583–639. https://doi.org/10.1111/1467-9868.00353

Sprague G. F., and L. A. Tatum, 1942 General vs. Specific Combining Ability in Single Crosses of Corn1. Agron. J. 34: 923–932. https://doi.org/10.2134/agronj1942.00021962003400100008x

Stuber C. W., and C. C. Cockerham, 1966 Gene Effects and Variances in Hybrid Populations. Genetics 54: 1279–1286.

Su G., O. F. Christensen, T. Ostersen, M. Henryon, and M. S. Lund, 2012 Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. PLOS ONE 7: e45293. https://doi.org/10.1371/journal.pone.0045293

Technow F., T. A. Schrag, W. Schipprack, E. Bauer, H. Simianer, *et al.*, 2014 Genome Properties and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program of Maize. Genetics 197: 1343–1355. https://doi.org/10.1534/genetics.114.165860

VanRaden P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423. https://doi.org/10.3168/jds.2007-0980

Varona L., A. Legarra, M. A. Toro, and Z. G. Vitezica, 2018a Non-additive Effects in Genomic Selection. Front. Genet. 9. https://doi.org/10.3389/fgene.2018.00078

Varona L., A. Legarra, W. Herring, and Z. G. Vitezica, 2018b Genomic selection models for directional dominance: an example for litter size in pigs. Genet. Sel. Evol. GSE 50. https://doi.org/10.1186/s12711-018-0374-1

Vitezica Z. G., L. Varona, and A. Legarra, 2013 On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope. Genetics 195: 1223–1230. https://doi.org/10.1534/genetics.113.155176

Vitezica Z. G., L. Varona, J.-M. Elsen, I. Misztal, W. Herring, *et al.*, 2016 Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. Genet. Sel. Evol. 48: 6. https://doi.org/10.1186/s12711-016-0185-1

Vitezica Z. G., A. Legarra, M. A. Toro, and L. Varona, 2017 Orthogonal Estimates of Variances for Additive, Dominance, and Epistatic Effects in Populations. Genetics 206: 1297–1307. https://doi.org/10.1534/genetics.116.199406

Vitezica Z. G., A. Reverter, W. Herring, and A. Legarra, 2018 Dominance and epistatic genetic variances for litter size in pigs using genomic models. Genet. Sel. Evol. 50: 71. https://doi.org/10.1186/s12711-018-0437-3

Werner C. R., L. Qian, K. P. Voss-Fels, A. Abbadi, G. Leckband, *et al.*, 2018 Genome-wide regression models considering general and specific combining ability predict hybrid performance in oilseed rape with similar accuracy regardless of trait architecture. Theor. Appl. Genet. 131: 299–317. https://doi.org/10.1007/s00122-017-3002-5

Westhues M., T. A. Schrag, C. Heuer, G. Thaller, H. F. Utz, *et al.*, 2017 Omics-based hybrid prediction in maize. Theor. Appl. Genet. 130: 1927–1939. https://doi.org/10.1007/s00122-017-2934-0

Xiang T., O. F. Christensen, Z. G. Vitezica, and A. Legarra, 2016 Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. Genet. Sel. Evol. 48: 92. https://doi.org/10.1186/s12711-016-0271-4

Zhao Y., Z. Li, G. Liu, Y. Jiang, H. P. Maurer, *et al.*, 2015 Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. Proc. Natl. Acad. Sci. 112: 15624–15629. https://doi.org/10.1073/pnas.1514547112

4.1 Appendices

Appendix A

Dominance deviations

The table of genotypic values is

	Genotype at P2		
Genotype at P1		B_2B_2	b_2b_2
	B_1B_1	$a_1 + a_2$	$a_1 + d$
	b_1b_1	$a_2 + d$	0

The mean of the genotypic value G of the crossbred population is

$$E(G) = p_1p_2(a_1 + a_2) + p_1q_2(a_1 + d) + q_1p_2(a_2 + d) = p_1a_1 + p_2a_2 + (p_1q_2 + q_1p_2)d$$

After centering, the table of centered genotypic values is

	Genotype at P2		
Genotype at P1		B_2B_2	b_2b_2
	B_1B_1	$q_1a_1 + q_2a_2 - (p_1q_2 + q_1p_2)d$	$q_1a_1 - p_2a_2 + (1 - p_1q_2 - q_1p_2)d$
	b_1b_1	$q_2a_2 - p_1a_1 + (1 - p_1q_2 - q_1p_2)d$	$-p_1a_1 - p_2a_2 - (p_1q_2 + q_1p_2)d$

The sum of the breeding values of the different gametes at the hybrid is

	Genotype at P2		
Genotype at P1		B_2B_2	b_2b_2
	B_1B_1	$q_1\alpha_1 + q_2\alpha_2$	$q_1\alpha_1 - p_2\alpha_2$
	b_1b_1	$-p_1\alpha_1 + q_2\alpha_2$	$-p_1\alpha_1 - p_2\alpha_2$

Subtracting this table from the centered genotypic values gives dominance deviations. If we go genotype by genotype:

$$\delta_{B_1B_2} = q_1a_1 + q_2a_2 - (p_1q_2 + q_1p_2)d - q_1\alpha_1 - q_2\alpha_2 =$$

$$q_1a_1 + q_2a_2 - (p_1q_2 + q_1p_2)d - q_1(a_1 + (q_2 - p_2)d) - q_2(a_2 + (q_1 - p_1)d) =$$

$$-(p_1q_2 + q_1p_2)d - q_1(q_2 - p_2)d - q_2(q_1 - p_1)d =$$

$$(-p_1q_2 - q_1p_2 - q_1q_2 + q_1p_2 - q_1q_2 + q_2p_1)d = -2q_1q_2d$$

$$\delta_{B_1b_2} = q_1a_1 - p_2a_2 + (1 - p_1q_2 - q_1p_2)d - (q_1\alpha_1 - p_2\alpha_2) =$$

$$q_1a_1 - p_2a_2 + (1 - p_1q_2 - q_1p_2)d - q_1(a_1 + (q_2 - p_2)d) + p_2(a_2 + (q_1 - p_1)d) =$$

$$(1 - p_1q_2 - q_1p_2)d - q_1(q_2 - p_2)d + p_2(q_1 - p_1)d =$$

$$(1 - p_1q_2 - q_1p_2 - q_1q_2 + q_1p_2 + p_2q_1 - p_1p_2)d =$$

$$(1 - p_1(q_2 + p_2) - q_1q_2 + p_2q_1)d =$$

$$(q_1 - q_1q_2 + p_2q_1)d = q_1(1 - q_2 + p_2)d = 2q_1p_2d$$

$$\delta_{b_1B_2} = q_2a_2 - p_1a_1 + (1 - p_1q_2 - q_1p_2)d - (-p_1\alpha_1 + q_2\alpha_2) =$$

$$q_2a_2 - p_1a_1 + (1 - p_1q_2 - q_1p_2)d - \big(-p_1(a_1 + (q_2 - p_2)d) + q_2(a_2 + (q_1 - p_1)d)\big) =$$

$$(1 - p_1q_2 - q_1p_2)d + p_1(q_2 - p_2)d - q_2(q_1 - p_1)d =$$

$$(1 - p_1q_2 - q_1p_2 + p_1q_2 - p_1p_2 - q_2q_1 + q_2p_1)d =$$

$$(1 - p_2(q_1 + p_1) + p_1q_2 - q_2q_1)d =$$

$$(q_2 + p_1q_2 - q_2q_1)d = q_2(1 + p_1 - q_1)d = q_2(1 - q_1 + p_1)d = 2p_1q_2d$$

$$\delta_{b_1b_2} = -p_1a_1 - p_2a_2 - (p_1q_2 + q_1p_2)d - (-p_1\alpha_1 - p_2\alpha_2) =$$

$$-p_1a_1 - p_2a_2 - (p_1q_2 + q_1p_2)d + p_1(a_1 + (q_2 - p_2)d) + p_2(a_2 + (q_1 - p_1)d) =$$

$$-(p_1q_2 + q_1p_2)d + p_1(q_2 - p_2)d + p_2(q_1 - p_1)d =$$

$$-(p_1q_2 + q_1p_2)d + (p_1q_2 - p_1p_2 + p_2q_1 - p_1p_2)d = -2p_1p_2d$$

See Table A1 for more details.

Table A1 Values of genotypes in a two-allele system, measured as deviation from the population mean

		Assigned	Deviations from population mean $E(G)$		
Hybrid Genotypes	Frequency	genotypic values	G^*	$g_{A^{(1)}} + g_{A^{(2)}}$	g_D
B_1B_2	p_1p_2	$a_1 + a_2$	$q_1a_1 + q_2a_2 - (p_1q_2 + q_1p_2)d$	$q_1\alpha_1 + q_2\alpha_2$	$-2q_1q_2d$
B_1b_2	p_1q_2	$a_1 + d$	$q_1a_1 - p_2a_2 + (1 - p_1q_2 - q_1p_2)d$	$q_1\alpha_1 - p_2\alpha_2$	$2q_1p_2d$
b_1B_2	q_1p_2	$a_2 + d$	$q_2a_2 - p_1a_1 + (1 - p_1q_2 - q_1p_2)d$	$-p_1\alpha_1 + q_2\alpha_2$	$2p_1q_2d$
b_1b_2	q_1q_2	0	$-p_1a_1 - p_2a_2 - (p_1q_2 + q_1p_2)d$	$-p_1\alpha_1 - p_2\alpha_2$	$-2p_1p_2d$

G^* is the total genotypic value of a hybrid deviated from the population mean.

$(g_{A^{(1)}} + g_{A^{(2)}})$ is the additive-effect portion of a hybrid's genotypic value.

g_D is the dominance deviation of the hybrid.

Properties of relationship matrices

Note that $\boldsymbol{G}_{A^{(1)}}$ and $\boldsymbol{G}_{A^{(2)}}$ have the following properties: the average value of the diagonal is 1, and the average value of the entire matrix is 0. For instance, the diagonal of $\boldsymbol{G}_{A^{(1)}}$ sums to

$$\frac{\sum_i p_{1_i}(1 - p_{1_i})^2 + q_{1_i}(-p_{1_i})^2}{\sum_i p_{1_i}q_{1_i}}$$

which is equal to 1. In addition, the sum of the elements of $\boldsymbol{G}_{A^{(1)}}$ is 0. Indeed, this sum can be written as

$$\frac{\sum_i (p_{1_i} \quad q_{1_i}) \begin{pmatrix} 1 - p_{1_i} \\ -p_{1_i} \end{pmatrix} \begin{pmatrix} 1 - p_{1_i} \\ -p_{1_i} \end{pmatrix}' (p_{1_i} \quad q_{1_i})'}{\sum_i p_{1_i}q_{1_i}}$$

which sums to 0. The same proof holds for $\boldsymbol{G}_{A^{(2)}}$.

The diagonal of $\boldsymbol{D}$ sums to

$$\frac{\sum_i p_{1_i}p_{2_i}\left(-2q_{1_i}q_{2_i}\right)^2 + p_{1_i}q_{2_i}\left(2q_{1_i}p_{2_i}\right)^2 + q_{1_i}p_{2_i}\left(2p_{1_i}q_{2_i}\right)^2 + q_{1_i}q_{2_i}\left(-2p_{1_i}p_{2_i}\right)^2}{\sum_i 4p_{1_i}q_{1_i}p_{2_i}q_{2_i}}$$

which is equal to 1. In addition, the average value of the entire matrix $\boldsymbol{D}$ is 0. In effect, this sum can be written as

$$\frac{\sum_i (p_{1_i}p_{2_i} \quad p_{1_i}q_{2_i} \quad q_{1_i}p_{2_i} \quad q_{1_i}q_{2_i}) \begin{pmatrix} -2q_{1_i}q_{2_i} \\ 2q_{1_i}p_{2_i} \\ 2p_{1_i}q_{2_i} \\ -2p_{1_i}p_{2_i} \end{pmatrix} \begin{pmatrix} -2q_{1_i}q_{2_i} \\ 2q_{1_i}p_{2_i} \\ 2p_{1_i}q_{2_i} \\ -2p_{1_i}p_{2_i} \end{pmatrix}' (p_{1_i}p_{2_i} \quad p_{1_i}q_{2_i} \quad q_{1_i}p_{2_i} \quad q_{1_i}q_{2_i})}{\sum_i 4p_{1_i}q_{1_i}p_{2_i}q_{2_i}}$$

which sums to 0.

Orthogonality

Next we prove orthogonality. In this model, z_1, z_2 and w are shifted to have mean zero for a population with these frequencies ($\boldsymbol{f} = [p_1p_2, p_1q_2, q_1p_2, q_1q_2]$). Thus, the mean of additive value is zero because

$$\sum_j z_{1_j} f_j = (1-p_1)p_1p_2 + (1-p_1)p_1q_2 + (-p_1)q_1p_2 + (-p_1)q_1q_2 = 0$$

$$\sum_j z_{2_j} f_j = (1-p_2)p_1p_2 + (-p_2)p_1q_2 + (1-p_2)q_1p_2 + (-p_2)q_1q_2 = 0$$

and the mean of dominant deviations is also zero because

$$\sum_j w_j f_j = (-2q_1q_2)p_1p_2 + (2q_1p_2)p_1q_2 + (2p_1q_2)q_1p_2 + (-2p_1p_2)q_1q_2 = 0$$

These equations correspond to the first requirement of orthogonality in Cockerham's (1954) model.

The second requirement can be expressed as $\sum_{i,j} f_{i,j} z_{1_i} z_{1_j} = 0$, $\sum_{i,j} f_{i,j} z_{2_i} z_{2_j} = 0$, $\sum_{i,j} f_{i,j} w_i w_j = 0$, e.g. the contrasts have 0 mean across all possible pairs of genotypes (Cockerham 1954). For the first two, this can be written as $\sum_{i,j} f_{i,j} z_{1_i} z_{1_j} = \sum_{i,j} f_i z_i z_j f_j =$ $(p_1 \quad q_1)\begin{pmatrix} 1-p_1 \\ -p_1 \end{pmatrix}\begin{pmatrix} 1-p_1 \\ -p_1 \end{pmatrix}' (p_1 \quad q_1)' = 0$ where $(p_1 \quad q_1)$ are frequencies of each genotype at the pure line and $\begin{pmatrix} 1-p_1 \\ -p_1 \end{pmatrix}$ are the values of z for each genotype. Similarly, $\sum_{i,j} f_{i,j} w_i w_j = \sum_{i,j} f_i w_i w_j f_j =$

$$(p_1p_2 \quad p_1q_2 \quad q_1p_2 \quad q_1q_2)\begin{pmatrix} -2q_1q_2 \\ 2q_1p_2 \\ 2p_1q_2 \\ -2p_1p_2 \end{pmatrix}\begin{pmatrix} -2q_1q_2 \\ 2q_1p_2 \\ 2p_1q_2 \\ -2p_1p_2 \end{pmatrix}' (p_1p_2 \quad p_1q_2 \quad q_1p_2 \quad q_1q_2)' = 0$$

Once the orthogonality of the one-locus formulation is proved, the orthogonal scales for the interactions in the multi-locus case can be generated by the Kronecker product. The extension of model using the Kronecker product guarantees the orthogonality of the multi-locus formulation (Álvarez-Castro and Carlborg 2007).

Appendix B

Variance components estimates and heritabilities for GCA- and G- models without the inclusion of genomic inbreeding are shown in Table A2.

Variance components estimates for GCA-models including unrelated Dent and Flint effects (similar to a permanent environmental effect in animal breeding) are shown in Table A3. Heritability estimates, inbreeding depression parameter estimates and the Deviance Information Criteria (DIC) values are presented in Table A4.

TABLE A2 Estimated posterior means and standard deviation (in parenthesis) of genetic variance component and heritabilities obtained with the two types of models without including genomic inbreeding for maize grain yield

	Additive		Dominance	Epistasis				
Model Code	$\sigma^2_{A^{(1)}}, \sigma^2_{A^{(2)}}$ or $\sigma^2_{A^{(H)}}$		σ^2_D or $\sigma^2_{D^{(H)}}$	$\sigma^2_{AA^{(1,1)}}$	$\sigma^2_{AA^{(2,2)}}$	$\sigma^2_{AA^{(1,2)}}$ or $\sigma^2_{AA^{(H)}}$	σ^2_e	H^2
$GCA: A$	33.89 (5.52)	23.35 (4.56)					18.01 (0.79)	0.76 (0.02)
$GCA: AD$	31.89 (5.20)	22.56 (4.46)	4.38 (0.77)				15.03 (0.80)	0.80 (0.02)
$GCA: A(AA)^{(1,2)}$	31.38 (5.13)	22.53 (4.42)				5.58 (0.96)	13.68 (0.85)	0.81 (0.02)
$GCA: AD(AA)^{(1,2)}$	31.08 (5.06)	22.11 (4.42)	2.97 (0.58)			4.20 (0.87)	13.36 (0.81)	0.82 (0.02)
$GCA: AD(AA)^{(1,1)}(AA)^{(2,2)}(AA)^{(1,2)}$	22.30 (5.20)	14.42 (4.11)	2.55 (0.56)	7.19 (2.50)	8.06 (2.89)	3.63 (0.82)	13.53 (0.81)	0.81 (0.02)
$G: A$	51.77 (6.75)						18.02 (0.79)	0.74 (0.03)
$G: AD^{(H)}$	47.81 (6.35)		6.18 (1.06)				14.97 (0.78)	0.78 (0.02)
$G: A(AA)^{(H)}$	42.22 (6.30)					10.2 (1.76)	13.89 (0.82)	0.79 (0.02)
$G: AD^{(H)}(AA)^{(H)}$	42.26 (6.08)		4.14 (0.81)			7.19 (1.55)	13.59 (0.80)	0.80 (0.02)

GCA- and G-models are models that successively added additive effects (A), dominance effects (AD), and additive-by-additive genetic effects ($AD(AA)$). The additive-by-additive epistatic effects can be interactions between loci within group (($AA)^{(11)}$ and $(AA)^{(22)}$), across groups $(AA)^{(12)}$ or within hybrids $(AA)^{(H)}$. Superscripts 1 and 2 in parenthesis refer to dent and flint heterotic groups respectively.

In GCA-model, the variances are: additive ($\sigma^2_{A^{(1)}}$ and $\sigma^2_{A^{(2)}}$), dominance (σ^2_D), and additive-by-additive epistasis within groups ($\sigma^2_{AA^{(1,1)}}$ and $\sigma^2_{AA^{(2,2)}}$) and additive-by-additive epistasis between groups ($\sigma^2_{AA^{(1,2)}}$). In G-model, the additive, dominance and additive-by-additive epistasis variances within hybrids are: $\sigma^2_{A^{(H)}}$, $\sigma^2_{D^{(H)}}$, and $\sigma^2_{AA^{(H)}}$, respectively. σ^2_e is the residual variance and H^2 is the genomic broad-sense heritability (or narrow-sense heritability in pure additive models).

TABLE A3 Estimates of variance components including unrelated effects from Dent and Flint groups using GCA-model with inbreeding

	Additive		Dominance	Epistasis			Remaining unrelated effect		Residual
Model Code	$\sigma^2_{A^{(1)}}$	$\sigma^2_{A^{(2)}}$	σ^2_D	$\sigma^2_{AA^{(1,1)}}$	$\sigma^2_{AA^{(2,2)}}$	$\sigma^2_{AA^{(1,2)}}$	$\sigma^2_{pe^{(1,1)}}$	$\sigma^2_{pe^{(2,2)}}$	σ^2_e
$GCA: A$	22.92 (4.72)	12.35 (3.51)					5.47 (1.53)	6.11 (1.81)	17.52 (0.76)
$GCA: AD$	22.81 (4.75)	12.87 (3.59)	3.31 (0.7)				4.92 (1.49)	5.66 (1.69)	15.09 (0.77)
$GCA: A(AA)^{(1,2)}$	22.6 (4.78)	12.92 (3.77)				4.53 (0.87)	4.84 (1.41)	5.62 (1.71)	13.83 (0.81)
$GCA: AD(AA)^{(1,2)}$	22.58 (4.81)	12.76 (3.67)	2.33 (0.53)			3.49 (0.8)	4.5 (1.32)	5.4 (1.66)	13.55 (0.81)
$GCA: AD(AA)^{(1,1)}(AA)^{(2,2)}(AA)^{(1,2)}$	19.38 (4.76)	10.65 (3.64)	2.03 (0.49)	4.98 (1.84)	5.42 (2.08)	3.28 (0.79)	3.54 (1.19)	4.36 (1.54)	13.72 (0.81)

$\sigma^2_{pe^{(1,1)}}$: remaining unrelated effect from Dent lines, $\sigma^2_{pe^{(2,2)}}$: remaining unrelated effect from Flint lines. Superscripts 1 and 2 in parenthesis refers to dent and flint heterotic groups respectively.

TABLE A4 Estimates of inbreeding depression and heritability, and goodness-of-fit criterion of the models with "remaining" unrelated effects

Model Code	b	H^2	DIC
$GCA: A$	-77.53 (14.57)	0.55 (0.05)	7311.65
$GCA: AD$	-49.04 (18.06)	0.6 (0.05)	7254.92
$GCA: A(AA)^{(1,2)}$	-40.68 (15.06)	0.62 (0.05)	7201.73
$GCA: AD(AA)^{(1,2)}$	-68.57 (18.87)	0.63 (0.05)	7200.78
$GCA: AD(AA)^{(1,1)}(AA)^{(2,2)}(AA)^{(1,2)}$	-48.03 (10.54)	0.68 (0.04)	7206.7

DIC: Deviance Information Criteria

Superscripts 1 and 2 in parenthesis refer to dent and flint heterotic groups respectively.

Chapter 5

General discussion and perspectives

5.1 Introduction

Pig and maize crossbreeding schemes rely on the exploitation of heterosis and breed complementarity by crossing breeds/lines. The main goal of these crossbreeding schemes is to improve the performance of CB populations through genetic selection of PB animals or inbred lines in parental populations. Pig and maize crossbreeding schemes have different limitations to achieve the ultimate goal of improving the CB performance. Selection of PB animals in current pig crossbreeding schemes is based on PB performance, with the expectation that part of the genetic progress obtained will be transmitted to the CB progeny through a correlated response (Wei *et al.* 1991). However, this strategy depends on the r_{PC} value: the attained correlated response in the CB population can be substaintially diminished if the r_{PC} differs from unity. Alternatively, pig breeders may improve CB performance directly by including CB data in the genetic evaluations of the PB candidates to selection (Wei and van der Werf 1994). However, pig breeding schemes have practical limitations for routine collection of phenotypes and reliable pedigree information from CB animals in field conditions required for such genetic evaluation. In maize crossbreeding, inbred lines are selected explicitly to improve hybrid performance by using hybrid records as selection criterion. However, the number of potential hybrids to evaluate far exceeds the capacity of any phenotypic hybrid breeding program. Thus, the identification of promising pairs of lines that produce the best performing hybrids is still challenging (Kadam and Lorenz 2018).

Low cost genotyping in recent years has allowed the development of genomic-based approaches, which, in turn, has brought the opportunity to overcome some of the limitations encountered in pig and maize crossbreeding schemes. GS has led to revisit quantitative genetics under a genomic context and the efficiency of animal and plant breeding programs has been increased. The decreasing trend in genotyping cost alleviates the need of recording pedigree data from CB animals and has renewed the interest of breeders for GS for CB performance in pig breeding. Maize breeding has profited from the use of high-density markers for genomic prediction of hybrid performance as well. Moreover, with the availability of SNP marker genotypes, estimation of non-additive genetic effects is straightforward compared with pedigree-based methods. Therefore, there is a renewed interest for including non-additive genetic effects in genomic models used in plants and animals. Also, the implementation of other genomic strategies (e.g. mate allocation) has become interesting and straighforward to boost

the performance of future progeny, especially in crossbreeding where the objective is to improve the overall genetic merit of CB individuals.

The overall objective of this thesis was to explore and develop methodology to exploit non-additive genetic effects on genomic selection models in pig and maize crossbreeding schemes. The Chapter 2 was focused on estimating and exploiting within-breed dominance variance through mate allocation strategies to boost the overall genetic merit in a Landrace pig population. Estimates of dominance variance were smaller compared to estimates of additive variance. Moreover, mate allocation accounting for non-additive genetic effects resulted in an improvement of the performance of expected offspring without compromising the additive genetic gain compared to random mating. In Chapter 3, the effectiveness of mate allocation strategies and genomic evaluations accounting for non-additive genetic effects to improve the performance in a two-way pig crossbreeding scheme in the long term was evaluated. Further, the effects of the sources of information used in the genetic evaluation (only PB data or both PB and CB data) were investigated. It was found that when r_{PC} is low, selecting PB animals for CB performance using CB information (phenotypes and genotypes) is a more efficient strategy to exploit heterosis and increase CB performance, whereas mate allocation did not improve CB performance. Finally, in Chapter 4, the theory of statistical models for hybrid prediction accounting for additive and non-additive (dominance and epistasis) effects was revisited under a genomic context. It was demonstrated theoretically and confirmed afterward in real data that, with SNP genotypes, it is possible to split SCA into dominance deviations and across-groups epistasis and to split GCA into within-line additive effects and within-line additive by additive epistasis effects. This partition of the genetic variance has practical implications when planning the breeding scheme, especially for estimating genetic gain and selection of within pools crosses.

The objective of the present chapter is to discuss the main findings of this thesis in a broader and joint perspective, highlighting the strengths and weaknesses of the models and strategies used and guiding future research.

5.2 Genomic selection accounting for non-additive genetic effects

It is well stablished that the incorporation of genomic information in genetic evaluations has led to a greater selection response for complex traits in animal and plant breeding compared to pedigree-based information (de los Campos *et al.* 2013; Crossa *et al.* 2017; VanRaden 2020). Quantitative genetics has played an important role in developing methodologies for an appropriate inclusion of marker information. In this thesis, several genomic models and methods were explored and developed to exploit non-additive genetic effects in pig and maize crossbreeding schemes. Through chapters 2 to 4, dominance and epistatic deviations were modeled using SNP markers and included in genomic evaluation models for pig and maize crossbreeding schemes. Those models are discussed here in a broad context.

5.2.1 Genomic selection in pig crossbreeding schemes

Chapter 2 consisted of estimating and exploiting dominance variance through mate allocation strategies to maximize the performance of PB progeny in a Landrace pig population. The within-breed selection can be considered the traditional approach for evaluating and selecting PB animals based only on PB performance. Pedigree-based and genomic-based models were compared to estimate variance components and predict BV of PB animals. The genomic models outperformed the pedigree-based models in terms of goodness of fit and accuracy of prediction of BV for all the traits evaluated, as shown in Figure A 2.1. These results agree with the literature (e.g. de los Campos *et al.* 2013). So, it is expected that a higher accuracy of BV will lead to a greater response to selection, especially for sex-limited traits (e.g. APWL in pigs). This genetic gain encourages pig breeding companies to increase the proportion of genotyped PB animals. Furthermore, estimates of dominance variance were obtained and accounted for 2 – 5% of the phenotypic variance depending on the trait. This is in agreement with what it has been estimated previously in livestock populations (Varona *et al.* 2018) (less than 10% of the phenotypic variance, Varona et al. 2018). The inclusion of dominance in the genomic model did not improve the accuracy of BV, as it was shown in Figure A 2.1. With the exception of Aliloo et al. (2016) (for fat yield in Holstein), in most studies, the inclusion of dominance effects in a GBLUP model did not improve the predictive ability of the model (Su *et al.* 2012; Ertl *et al.* 2014; Esfandyari *et al.* 2016; Xiang *et al.* 2016; Moghaddar and van der Werf 2017). Nevertheless, although dominance does not improve the accuracy of prediction, it

can be exploited through mate allocation to boost overall gen etic merit of the progeny as it will be discussed later in this chapter.

In this study, only genotyped animals were involved in genetic evaluations. Modeling of dominance is straightforward in a GBLUP context. In practice, genetic evaluations in pigs are based on single-step GBLUP approach (Bouquet *et al.* 2017) to include genotyped and non-genotyped animals. Even if the cost of genotyping decreases and the number of genotyped animals increases, it is expected that there will always be non-genotyped animals. Thus, single-step approach is more suitable to estimate BV from pedigree and genomic information jointly. Recently, a reduced single-step GBLUP model accounting for dominance was proposed in a dairy cattle context (Ertl *et al.* 2018). This much more computationally expensive model did not provide better BV accuracies than the one accounting only for additive effects.

The traditional approach of selecting PB animals based on PB performance maybe suboptimal to improve CB performance because r_{PC} is generally lower than one for many traits of interest in pig production (Wientjes and Calus 2017). In Chapter 3, PB and CB information (genotypes and phenotypes) from a two-way crossbreeding scheme were used simultaneously to evaluate PB lines across 10 generations of selection by simulation. Genetic evaluations were performed at each generation including animals from previous generations in the analyses. When r_{PC} was low (less than 0.5), selecting PB animals for CB performance using PB and CB information was a more efficient strategy to improve performance at the CB level in the short and long term compared to selecting PB candidates using only PB data, but at the cost of reducing the genetic response in the parental breeds. In addition, an increase in inbreeding was observed in PB populations. In this study, PB progeny was generated by random mating between selected candidates, but a mating strategy called "minimum coancestry mating" can be used to manage and alleviate inbreeding.

Given that the ultimate goal in a crossbreeding scheme is to improve CB performance at the commercial level, and with a r_{PC} lower than 1, the challenge is how to improve performance at the CB and PB level simultaneously. This will depend, among other factors, on the economic relevance of CB and PB animals and the value of r_{PC}. Considering the r_{PC} value and based on the main findings of chapters 2 and 3, two alternatives can be proposed to manage simultaneous improvement of PB and CB performance. If the r_{PC} is not too low (e.g. ≥ 0.8), selecting animals for PB performance will improve CB performance. If r_{PC} differs markedly from unity (e.g. $<$

0.8), a better strategy could be weighting both traits (PB and CB performances) in a selection index to balance the improvement across PB and CB performance (Esfandyari *et al.* 2018). In Chapter 3, a trivariate SNP-BLUP model including additive and dominance marker (genotypic) effects was used. Such SNP-BLUP model provides the estimates of SNP effects, which can be used directly into mate allocation. The trivariate SNP-BLUP model distinguishes explicitly between PB and CB data. Note that the differences between PB and CB performances are due to genotype-by-environment and genotype-by-genotype interactions. A limitation of this model is that only genotyped animals can be included in the analysis. In practice, it is not often the case for all animals, especially for CB animals. Alternatively, to avoid recording phenotypes and genotypes of CB in routine, a random sample of CB animals can be used as a reference population. Then, the estimated SNP effects can be exploited for a few generations of selection (Dekkers 2007). Further, the implementation of this SNP-BLUP model with additive and dominance inheritance requires previous knowledge of (co)variances of SNP effects between PB and CB populations. These (co)variances can be estimated from data using a trivariate GBLUP model including dominance as in Xiang *et al.* (2016a) and then switching from a breeding parameterization (in terms of BV and dominance deviations) to a genotypic parameterization (in terms of additive and dominant SNP effects), as we did in Chapter 2 and 3. In this thesis, a maternal two-way crossbreeding scheme was analyzed, but selection for CB performance can be implemented in three- or four-way crossbreeding schemes. Recently, Christensen *et al.* (2019) proposed a genomic model with additive and dominance effects for the case of a PB sire line and three-way crossbred pigs. It would be interesting to evaluate the selection for CB performance in a three- or four-way crossbreeding schemes in the long term.

5.2.2 Genomic selection in maize crossbreeding scheme

In Chapter 4, the theory of hybrid genetic evaluation was revisited. Based on SNP genotypes, it is possible to split the SCA into dominance and across-group epistasis deviations, and to split GCA into within-line additive effect and within-line additive-by-additive epistasis deviations. This has been demonstrated analytically and on real data in this thesis. Thus, the partition of the hybrid's genotypic value, as was originally defined by Stuber and Cockerham (1966), is now feasible using the proposed model (GCA-model) under a genomic context. Moreover, another genomic model assuming a common effect of genes across origins (Vitezica

et al. 2017) was used for the analysis of hybrids crops (G-model). To account for directional dominance, heterosis was fitted in the models as a regression on observed homozygosity in the hybrids. In addition, the concept of permanent environmental effect was introduced into hybrid genomic prediction as "remaining" "unrelated" Dent and Flint effects. This effect is commonly used in animal breeding to capture non-genetic effects but also genetic effects not explicitly modeled such as dominance or epistasis. If within-group epistasis is not explicitly fitted in the model, this permanent effect captures the remaining effect in the GCA attributable to the lines i.e. within-group epistasis. Note that inbred lines transmit within-line epistasis combination to their hybrid progeny contrary to animals where recombination breaks down epistatic combinations. Thus, GCA contains, in addition to the additive effects, the additive-by-additive within-line epistasis.

Models based on defining effects either differently (e.g. in hybrid crops) or identically across origins (e.g. in breeds in livestock) were equivalent in terms of goodness of fit and predictive ability of hybrid performance. However, these models split the genetic variance differently leading to different interpretations. The partition of the genetic variance has practical relevance especially for planning the hybrid breeding scheme and selection of within pools lines for hybrid improvement. On one hand, the sum of the GCA and SCA variance (including all epistatic interactions) gives to the breeder and idea of how much variation can be expected in hybrids and it should be considered for genomic prediction of hybrid performance. On the other hand, for planning the breeding scheme, i.e. prediction of genetic progress and selection of lines to be parents of the next generation within heterotic groups, only the additive effects from the GCA should be considered. This is because the additive-by-additive within-line epistatic combinations are lost by recombination during the development of new inbred lines. This idea was also pointed out by Stuber and Cockerham (1966). Thereby, using a model that does not include all sources of variation (as the traditional model used in hybrid crops e.g. Technow *et al.* 2014) tends to inflate the additive genetic variance and the additive part of the estimated GCAs. Therefore, the recommended model needs to fit within-group epistasis or, if this is not possible, unrelated "remaining" line effects in order to avoid overestimation of the genetic gain.

Estimates of genetic variance obtained in this study confirmed that most of the variation in hybrids is accounted for by GCA variance. Dominance and across-group epistatic variances have similar magnitudes. These results provided evidence that epistasis is contributing to heterosis, at least at the same proportion than dominance in this set of hybrids. Estimates of

non-additive effects might be important for hybrid pool development. For instance, Zhao *et al.* (2015) suggested to use additive and dominance effects from an incomplete factorial in order to develop heterotic pools in wheat. Moreover, estimates of non-additive effects may be used in mate allocation strategies in maize hybrids. Some applications of mate allocation including dominance are discussed later. Finally, the new model proposed here is more appropriate for using genomic data since it allows a better understanding of the architecture of heterotic traits and the obtained estimates of genetic variance components can be interpreted for breeding purposes.

5.3 Considering directional dominance in genomic evaluations

The quantitative genetics theory argues that the manifestation of inbreeding depression or heterosis depends on the existence of directional dominance (Falconer 1981). Animal breeders know that in the presence of directional dominance genetic evaluations must consider inbreeding depression to obtain proper estimates of dominance variance (de Boer and Hoeschele 1993; Xiang *et al.* 2016a) (see Appendix 1.1). In plant breeding, the notion of accounting for directional dominance is, to some extent, implicit in some studies addressing the relationship between genetic distance between parental inbred lines and hybrid performance (Charcosset *et al.* 1991; Bernardo 1991; Charcosset and Essioux 1994; Reif *et al.* 2003; Larièpe *et al.* 2012, 2017). However, its impact on the estimation of non-additive genetic variance has been overlooked in most of them. In these studies, the modified Roger's distance (Rogers 1972) was used to compute the marker-based genetic distance between two parental lines, which is an estimator of the expected average heterozygosity of the hybrid progeny (Charcosset and Essioux 1994; Larièpe *et al.* 2012). With the availability of genomic information, the realized heterozygosity of hybrids, computed from SNP markers, could be used instead of the expected one (Roger's distance) to account for directional dominance in genomic models for hybrids crops.

In this thesis, directional dominance was accounted for in all genomic models in chapters 2, 3 and 4, by including a regression on genomic inbreeding. In addition, a model without genomic inbreeding was also evaluated (Chapter 4) to see the impact on estimates of non-additive genetic variances in maize hybrids. Results showed that when genomic inbreeding was not included in the model, estimates of non-additive genetic variances, both dominance and

epistasis, were overestimated compared to the model that included it. These results agree with the theory and show that genomic inbreeding must be included in genomic evaluations of hybrids crops to avoid biased estimates of non-additive variances.

Furthermore, here directional epistasis was ignored. However, in reality, directional epistasis effects may exist. If it is the case, the relationship between the genomic inbreeding and fitness is expected to be non-linear (Crow and Kimura 1970). To capture such non-linearity (e.g. dominance-by-dominance epistasis), the quadratic effect of the inbreeding coefficient (f^2), together with its linear effect (f), could be fitted in genomic models (e.g. Ramstein *et al.* 2020). Here, the genomic inbreeding coefficients were calculated based on the proportion of homozygous SNPs (Silió *et al.* 2013). Alternatively, genomic inbreeding coefficients based on ROH (McQuillan *et al.* 2008) could be used.

5.4 Inbreeding depression in crossbreds

In animal breeding, estimates of inbreeding depression, based on either pedigree-based or genomic-based information, have been obtained in dairy cattle, pigs and sheep (e.g. Pryce *et al.* 2014; Saura *et al.* 2015; Rodríguez-Ramilo *et al.* 2019). Recently, estimates of (genomic) inbreeding depression were reported in CB pigs (Xiang *et al.* 2016; Iversen *et al.* 2019; Tusell *et al.* 2019). Under the conventional view of maize hybrid breeders, it is not expected to observe inbreeding depression in hybrids derived from two divergent heterotic groups. However, even when the parental populations are unrelated by pedigree, they might be related by a more distant common ancestor. With genomic information, it is possible to measure the relationship between two unrelated (by pedigree) individuals due to a recent or very old common ancestor (VanRaden 2008). Therefore, estimates of inbreeding depression can also be obtained for CB populations. However, in maize hybrid populations, estimates of inbreeding depression are scarce.

Here, inbreeding depression was investigated in pigs and maize crossbreeding schemes. In Chapter 2, inbreeding depression was estimated in a Landrace pig population, which had a negative impact on the age at 100 kg trait (AGE) and no effect on the other analyzed traits (backfat depth (BD), average piglet weight at birth within litter (APWL)), which was in agreement with the literature (Culbertson *et al.* 1998). In Chapter 3, estimates of inbreeding depression were close to those simulated in both PB and CB populations. Finally, in Chapter 4,

inbreeding depression was estimated in a population of inter-heterotic group maize hybrids. Contrary to expectations, results indicated that inbreeding depression exists in hybrids and it had a negative impact on grain yield in this hybrid population. The estimates of inbreeding depression were within the range of those reported in the literature based on either genomic-based estimated in a population of hybrids (Ramstein *et al.* 2020) or pedigree-based estimates in different maize populations (Hallauer *et al.* 2010).

The existence of inbreeding depression in inter-heterotic group hybrids might indicate that there have been some recent or old migrations between the two heterotic groups, which may happen in practice in maize hybrid breeding, as it was pointed out by Bernardo (1995).

5.5 Mate allocation strategies

Crossbreeding is the main way to exploit heterosis in animal and plant breeding. More sophisticated breeding programs are interested in optimizing the use of SCA or non-additive genetic effects through planned matings or mate allocation. In this section, the relevance and limitations of the implementation of mate allocation in pig breeding schemes are discussed. Also, potential applications of mate allocation accounting for dominance in maize crossbreeding are presented.

5.5.1 Mate allocation in pig breeding schemes

In this thesis, the effectiveness of mate allocation strategies accounting for dominance effects and inbreeding depression to improve progeny performance in a pig breeding scheme was evaluated in a context where matings were optimized either within a French Landrace population or in two-way crossbred pigs by simulation.

5.5.1.1 Within-breed mate allocation

Dominance also exists at the breed level although it is generally ignored. In Chapter 2, a mating strategy that maximized the total genetic value of the expected progeny ($\hat{g}$) was

compared to another mating strategy based on selection on predicted breeding values of the expected progeny ($\hat{u}$) in a purebred population. The last one is equivalent to perform standard truncation selection followed by random mating. Results show that optimization on $\hat{g}$ results in a higher expected progeny performance than optimization on $\hat{u}$. Further, the improvement on the expected progeny performance was higher when the ratio of dominance to additive variance was higher. Therefore, the benefit of implementing mate allocation will depend on the ratio of dominance to additive variance of the trait: the greater the ratio, the greater total genetic gain at the progeny can be attained. These results agree with other studies reported in the literature that implemented mate allocation accounting for dominance effects in dairy cattle (Sun *et al.* 2013; Ertl *et al.* 2014; Aliloo *et al.* 2017) and in simulated data (Toro and Varona 2010).

Furthermore, when mate allocation is implemented within a PB population, where selection takes place (i.e. at the nucleus), the main breeder's concern is whether the optimization of progeny performance may result in a detrimental effect on the additive genetic gain. This is a valid concern because the additive component is the inherited part that will be transmitted to subsequent generations. In the study of Chapter 2, a higher number of boars (120) than the required under truncation selection (40) were pre-selected based on their GEBV. In this case, optimization was performed in both males and females (i.e. simultaneous selection and mate allocation). Results showed that optimization on $\hat{g}$ resulted only in a very slight reduction in the additive genetic gain on the trait AGE (1.8%) (Figure 5.1). Because the 120 boars were pre-selected based on its GEBV, the low reduction in additive genetic gain by optimization on $\hat{g}$ could be due to the pre-selection of those males. In order to discard the effect of pre-selection of males, two additional mating allocation strategies were tested. Given that males have the greatest intensity of selection compared to females, one strategy consisted of selecting the top 40 males based on their GEBV to minimize the loss of additive genetic gain. This mate allocation strategy has been used in dairy cattle, which resulted in a very low reduction in the additive genetic gain (Ertl *et al.* 2014; Aliloo *et al.* 2017). In this case, only females are optimized and the pre-selection of males ignores other non-selected males that could have a good specific combining ability. A third case consisted of selecting 120 males at random among all candidates to selection (i.e. no pre-selection and no selection intensity was applied on both, male and females sides). In all the cases, all the available genotyped females were included in the optimization. Results for the trait AGE are presented in Table 5.1. For the other trait, see Table A 2.1 in Appendix 2.2. Note that comparison of the pre-selection strategies should be

made based on the increase of total genetic gain (ΔG) and reduction of additive genetic grain (ΔU). This is because the genetic SD units are referred to the average value of all potential matings, which changes depending on the choice of pre-selection strategy. Results show that the pre-selection of males does not affect neither the additive genetic gain nor the total genetic gain. This can be explained because almost the same sows and best boars were selected in matings to optimize the additive or total genetic gain. That means that optimization on $\hat{g}$ maximizes both the total genetic and the additive genetic gain in contrast with optimization on $\hat{u}$ which only maximized the additive genetic gain. These results confirm that, for the analyzed traits, optimization of progeny performance by selecting on $\hat{g}$ is feasible without compromising the expected additive genetic gain within a purebred population, regardless of the pre-selection of males.

Table 5.1 Expected total (ΔG) and additive (ΔU) genetic gain obtained based on optimization on $\hat{u}$ or $\hat{g}$ under different strategies of pre-selection of males for the trait Age at 100 kg (AGE).

Trait	Pre-selection of males	Mate allocation strategy	ΔU	ΔG	Increase in ΔG in Genetic SD units	Reduction in ΔU in Genetic SD units
			Genetic SD units	Genetic SD units		
AGE (days)	Top 40 males selected on GEBV	Selection on $\hat{u}$	-0.39	-0.37		
		Selection on $\hat{g}$	-0.38	-0.56	-0.20	0.01
	120 males pre-selected on GEBV	Selection on $\hat{u}$	-0.56	-0.54		
		Selection on $\hat{g}$	-0.55	-0.74	-0.20	0.01
	120 males taken at random.	Selection on $\hat{u}$	-0.73	-0.72		
		Selection on $\hat{g}$	-0.72	-0.92	-0.20	0.01

Specific combining ability variance can be partitioned into the average effect of inbreeding depression or heterosis (caused by directional dominance), the dominance effects beyond inbreeding, and the epistasis effects (DeStefano and Hoeschele 1992). In the analysis of Chapter 2, estimates of dominant effects ($\boldsymbol{d}$) with $E(\boldsymbol{d}) = \mu_d$ were used to compute the total genetic value of the expected progeny. That means that inbreeding depression was implicitly included in the mating optimizations. Ignoring epistasis, the improvement obtained by the exploitation of non-additive genetic effects with mate allocation may also be due to the effect of inbreeding. To evaluate the contribution of the inbreeding effect and dominance to the improvement of progeny performance, the total genetic value of progeny ($\hat{g}_{ij}$) from sire i and dam j was decomposed in its additive breeding value ($\hat{u}_{ij}$), dominance deviation ($\hat{v}_{ij}$) and average effect of inbreeding ($f\hat{b}$) i.e. $\hat{g}_{ij} = \hat{u}_{ij} + \hat{v}_{ij} + f\hat{b}$. Figure 5.1 shows the contribution

of each component contributing to the total genetic gain (ΔG) for each trait when the mating optimization was either on $\hat{g}$ or $\hat{u}$.

The results confirm that part of the improvement in progeny performance by optimization on $\hat{g}$ is due to the average effect of inbreeding. The contribution of preventing inbreeding is related to the magnitude of inbreeding depression effect. For instance, the benefit of preventing inbreeding by optimization on $\hat{g}$ was higher than the benefit of dominance for the trait AGE. The AGE had the highest estimate of inbreeding depression among the analyzed traits (1.05, -0.02 and 0.35 expressed in genetic SD units per 10% increase of inbreeding for AGE, BD, APWL, respectively). For BD, the contribution of dominance to ΔG was higher than the benefit of preventing inbreeding, as the effect of inbreeding depression was negligible for BD. Further, results show that the benefit of dominance depends on the size of the ratio of dominance variance to additive variance, as it was discussed before. These results show the importance of accounting for directional dominance in genetic evaluations which results in an improvement of progeny performance. Preventing inbreeding, after standard truncation selection, is a simple strategy to improve progeny performance within PB population.

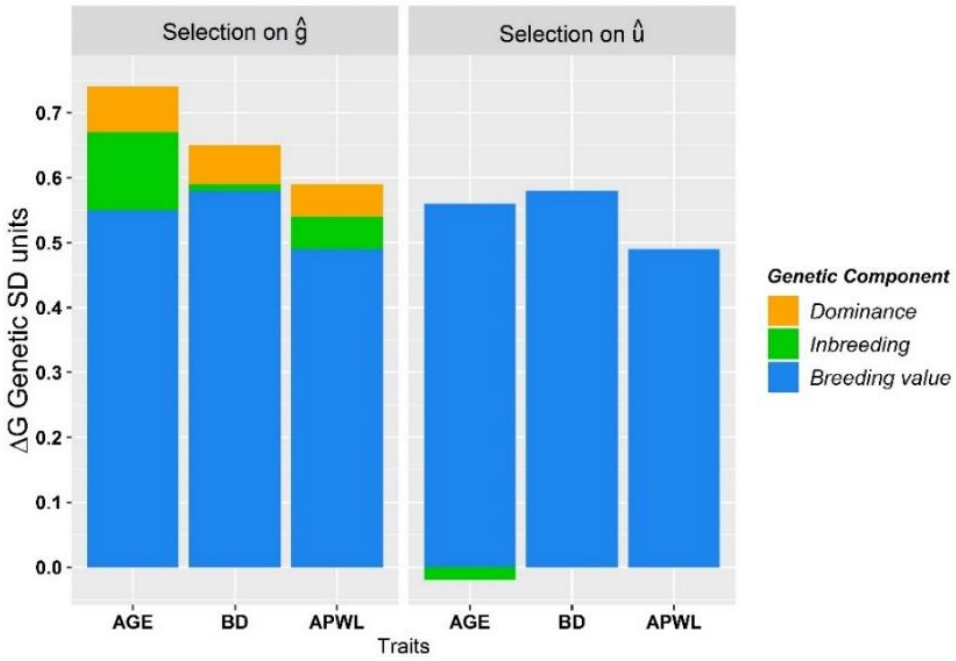

Figure 5.1 Contribution of breeding values, dominance deviations and average effect of inbreeding to the total genetic gain (ΔG) expressed in genetic SD units for traits AGE, BD and APWL (positive values for AGE and BD were plotted for easy comparison) when selecting matings on $\hat{g}$ or $\hat{u}$.

The conclusions from this study were limited to a few scenarios of mate allocation in a single purebred population and for one generation. Results from a simulation study in a single population provide evidence that the benefit from mate allocation profiting from dominance disappears in subsequent generations (Toro and Varona 2010). Therefore, mate allocation has to be implemented recurrently to profit of non-additive effects (Howard *et al.* 2017). In Chapter 3, the effectiveness of mate allocation strategies to optimize the performance of two-way CB pigs was investigated in the long term. The results are discussed in a broader context in the next section.

5.5.1.2 Mate allocation in a two-way crossbreeding scheme

In crossbreeding schemes, the commercial CB animals can be the product of planned crosses instead of random mating between individuals from different breeds to exploit heterosis. Results in Chapter 3 show that the advantage of using mate allocation strategies (minimizing the average expected genomic inbreeding or maximizing the average expected total genetic value of CB progeny) to boost CB performance was negligible compared to random mating; even for cases where the ratio dominance to additive variance was high. Thus, mate allocation does not improve CB performance under the tested scenarios. However, the selection of PB animals for CB performance using PB and CB information simultaneously is a more efficient strategy to improve CB performance in the short and long term (especially if r_{PC} is low).

The conclusions of this study are based on a simulation inspired in a two-way maternal crossbreeding scheme, and simulated genetic parameters were similar to those observed in maternal traits, which have low heritability and low dominance effects. Further investigation is needed to explore mate allocation in different scenarios; for instance, using different criteria to select PB animals (e.g. PB and CB combined indexes), traits with different genetic determinism and in a three-way crossbreeding scheme. In this thesis, random mating was used to produce the next generations within PB populations, in order not to exceed the complexity of the simulation and computation time. Other point is that PB animals were selected prior to mate allocation. Another interesting strategy could be mate selection where selection and mate allocation are performed simultaneously.

5.5.1.3 Practical implications of mate allocation in a pig crossbreeding scheme

From a practical point of view, implementing SNP-mate allocation in pigs is not straightforward. Mate allocation has to be implemented recurrently to profit from non-additive effects. That implies a continuous collection of phenotypic and genotypic data (of PB or of PB and CB) and the reevaluation of marker effects. For males is not a big deal at the nucleus level, phenotypes and genotypes are often available. For females, the situation is different. An alternative to have genotyped females is to take the maternal grandsire SNP information into account, as was proposed by (DeStefano and Hoeschele 1992). The cost-benefit of implementing mate allocation at CB level is not clear, our results are not encouraging compared to its implementation within-breed a least at the long term.

5.5.2 Potential applications of mate allocation in maize crossbreeding

Mate allocation aimed at profiting from dominance have different objectives in pig and maize crossbreeding, which are determined in part by the specific features of each species. In animal breeding, in general, as the number of descents by mate is low compared to plants, a set of matings is required to produce enough individuals either for breeding or commercial purposes. In plant breeding, it is possible to generate a large number of genetically identical individuals (e.g. hybrids, inbred lines or clones), a productive farm of e.g. 1000 hectares can be planted with the same hybrid genotype. The main task of maize hybrid breeders is to identify the mating between parental inbred lines that produces the best performing hybrid (Bernardo 1994). The latter can be addressed by testing as many hybrid combinations as possible and using predictive approaches. Note that all hybrid combinations cannot be tested in different environments and a training set (a sample of hybrids) is used for prediction instead. Mate allocation could play an important role in the definition of the training set to increase the reliability of genomic predictions.

Based on the theory and results from Chapter 4, now it is possible to split all sources of genetic variation in hybrids (GCA-model). One application of mate allocation in maize breeding is the optimization of the mating designs to train the prediction models for hybrid performances. In maize crossbreeding the number of potential hybrids still far exceeds the capacity of any phenotypic breeding program. Genomic prediction has been used to predict the

performance of untested hybrids. However, one of the critical factors affecting the accuracy of genomic predictions is the population used to calibrate the model (Rincent *et al.* 2012; Kadam and Lorenz 2018). The question is how to define which hybrids should be included in the calibration. Mate allocation can be used to better define the hybrid combinations to be phenotyped and used as training set to maximize the reliability of prediction of non-phenotyped hybrids (Figure 5.2). Rincent *et al.* (2012) used the generalized coefficient of determination (CD) (Laloë 1993) to optimize the training set, which is a function of the prediction error variance and of the genetic variance. Thus, it takes into account the genetic covariances between individuals to avoid the selection of closely related hybrids within the training set. Rincent *et al.* (2012) used this approach within heterotic group, for hybrids that shared the same tester line. They used only the additive genomic relationships among potential hybrids. However, this approach could be extended to hybrids derived from a factorial design considering both additive and dominance genomic relationships among hybrids, which may result in higher reliability of hybrid prediction. This extension could be made by deriving the CD criteria from the G- or GCA-model accounting for additive and dominance effects. The G-model seems to be more appropriate for this purpose than the GCA-model because it has only one additive relationship among hybrids instead of two additive relationships for each group of parental lines. More investigations are needed to derive the CD criteria based on the G- and GCA-models and evaluate the reliability of prediction of untested hybrids.

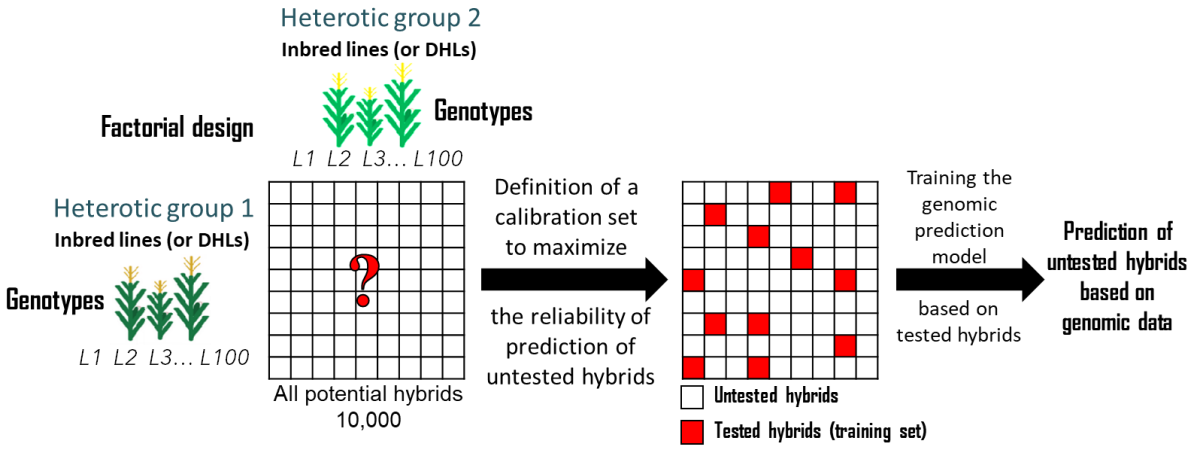

Figure 5.2 Optimization of the mating designs to train the prediction models for hybrid performances (modified from Rincent *et al.* 2012).

Furthermore, recently, some studies have implemented mate allocation strategies to improve genetic gain in plant breeding programs. For instance, Lehermeier *et al.* (2017) and Allier *et al.* (2019) used mate allocation strategies to define parental-crosses to maximum selection progress within a population. Encouraging results to maximize progeny performance and maintain diversity for long-term genetic gain were reported in an inbred line breeding program but considering only additive effects. This approach could be extended to a hybrid breeding program, by including additive and non-additive effects (dominance deviations and genomic inbreeding), to jointly improve the two heterotic groups for hybrid performance.

The ideas proposed above apply for hybrid crops in general. Therefore, subsequent investigations should be focused on exploiting dominance effects in mate allocation strategies to improve hybrid performance and maximizing the reliability of genomic prediction by optimizing the training set.

5.6 General conclusion

Theory of quantitative genetics, and most of the methods and genomic tools used in this thesis can be shared by animal and plant breeding, but particularities of each species should be considerd. For instance, crop inbred lines transmit within-line epistasis to the hybrid progeny contrary to animals where recombination breaks down epistatic combinations. Probably, biased by my family history of plant breeders, one of the main contributions of this thesis is the development of methodology useful for genetic evaluation of hybrid crops (GCA-model), which allows disentangling dominance and epistasis effects. Also, the introduction of the concept of "remaining" permanent effect into hybrid genomic prediction, which is commonly used in animal breeding, is a new idea for plant breeding. Further, directional dominance must be accounted for in genomic models in both pigs and maize hybrids.

Finally, animal and plant breeding share the common goal to improve traits of economic importance. Genomic selection brings the opportunity to reduce the gap between animal and plant breeding. This thesis is an example. Working in collaboration, sharing experience and learning from each other is a good strategy to overcome the challenges and maximize the genetic gain in both animals and plants.

References

Acosta-Pech, R., J. Crossa, G. de los Campos, S. Teyssèdre, B. Claustres *et al.*, 2017 Genomic models with genotype × environment interaction for predicting hybrid performance: an application in maize hybrids. Theor. Appl. Genet. 130: 1431–1440.

Ahlborn-Breier, G., and W. D. Hohenboken, 1991 Additive and Nonadditive Genetic Effects on Milk Production in Dairy Cattle: Evidence for Major Individual Heterosis. J. Dairy Sci. 74: 592–602.

Albrecht, T., V. Wimmer, H.-J. Auinger, M. Erbe, C. Knaak *et al.*, 2011 Genome-based prediction of testcross values in maize. Theor. Appl. Genet. 123: 339.

Aliloo, H., J. E. Pryce, O. González-Recio, B. G. Cocks, M. E. Goddard *et al.*, 2017 Including nonadditive genetic effects in mating programs to maximize dairy farm profitability. J. Dairy Sci. 100: 1203–1222.

Allier, A., C. Lehermeier, A. Charcosset, L. Moreau, and S. Teyssèdre, 2019 Improving Short- and Long-Term Genetic Gain by Accounting for Within-Family Variance in Optimal Cross-Selection. Front. Genet. 10:.

Álvarez-Castro, J. M., and Ö. Carlborg, 2007 A Unified Model for Functional and Statistical Epistasis and Its Application in Quantitative Trait Loci Analysis. Genetics 176: 1151–1167.

Alves, F. C., Í. S. C. Granato, G. Galli, D. H. Lyra, R. Fritsche-Neto *et al.*, 2019 Bayesian analysis and prediction of hybrid performance. Plant Methods 15: 14.

Baselga, M., 2004 Genetic improvement of meat rabbits. Programmes and diffusion. Programmes and diffusion. Proc. 8th World Rabbit Congr., Puebla, Mexico 1, 1-13.

Bell, A. E., C. H. Moore, and D. C. Warren, 1950 Systems of breeding designed to give maximum heterosis in chickens. Poult. Sci. 29: 749.

Bernardo, R., 1996 Best Linear Unbiased Prediction of Maize Single-Cross Performance. Crop Sci. 36: 50–56.

Bernardo, R., 2010 *Breeding for quantitative traits in plants*. Stemma Press, Woodbury.

Bernardo, R., 1991 Correlation between testcross performance of lines at early and late selfing generations. Theor. Appl. Genet. 82: 17–21.

Bernardo, R., 1995 Genetic Models for Predicting Maize Single-Cross Performance in Unbalanced Yield Trial Data. Crop Sci. 35: cropsci1995.0011183X003500010026x.

Bernardo, R., 2003 Parental selection, number of breeding populations, and size of each population in inbred development. Theor. Appl. Genet. 107: 1252–1256.

Bernardo, R., 1994 Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. Crop Sci. 34: 20–25.

Bernardo, R., and J. Yu, 2007 Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Sci. 47: 1082.

Bidanel, J., 1993 Estimation of crossbreeding parameters between Large White and Meishan porcine breeds. III. Dominance and epistatic components of heterosis on reproductive traits. Genet. Sel. Evol. 25: 263.

Bijma, P., and J. W. Bastiaansen, 2014 Standard error of the genetic correlation: how much data do we need to estimate a purebred-crossbred genetic correlation? Genet. Sel. Evol. 46: 79.

Bijma, P., J. A. Woolliams, and J. A. M. van Arendonk, 2001 Genetic gain of pure line selection and combined crossbred purebred selection with constrained inbreeding. Anim. Sci. 72: 225–232.

de Boer, I. J. M., and I. Hoeschele, 1993 Genetic evaluation methods for populations with dominance and inbreeding. Theor. Appl. Genet. 86: 245–258.

Bouquet, A., M. Canaple, P. Brenaut, T. Bellec, L. Flatres-Grall *et al.*, 2017 Mise en place de la sélection génomique dans le schéma de sélection de la population Landrace Français.

Bouvet, J.-M., G. Makouanzi, D. Cros, and P. Vigneron, 2016 Modeling additive and non-additive effects in a hybrid population using genome-wide genotyping: prediction accuracy implications. Heredity 116: 146–157.

Bruce, A. B., 1910 The Mendelian Theory of Heredity and the Augmentation of Vigor. Science 32: 627–628.

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

Chaikam, V., W. Molenaar, A. E. Melchinger, and P. M. Boddupalli, 2019 Doubled haploid technology for line development in maize: technical advances and prospects. TAG Theor. Appl. Genet. Theor. Angew. Genet. 132: 3227–3243.

Charcosset, A., and L. Essioux, 1994 The effect of population structure on the relationship between heterosis and heterozygosity at marker loci. Theor. Appl. Genet. 89: 336–343.

Charcosset, A., M. Lefort-Buson, and A. Gallais, 1991 Relationship between heterosis and heterozygosity at marker loci: a theoretical computation. Theor. Appl. Genet. 81: 571–575.

Christensen, O. F., A. Legarra, M. S. Lund, and G. Su, 2015 Genetic evaluation for three-way crossbreeding. Genet. Sel. Evol. 47: 98.

Christensen, O. F., and M. S. Lund, 2010 Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42: 2.

Christensen, O. F., P. Madsen, B. Nielsen, and G. Su, 2014 Genomic evaluation of both purebred and crossbred performances. Genet. Sel. Evol. 46: 23.

Christensen, O. F., B. Nielsen, G. Su, T. Xiang, P. Madsen *et al.*, 2019 A bivariate genomic model with additive, dominance and inbreeding depression effects for sire line and three-way crossbred pigs. Genet. Sel. Evol. 51: 45.

Cockerham, C. C., 1954 An Extension of the Concept of Partitioning Hereditary Variance for Analysis of Covariances among Relatives When Epistasis Is Present. Genetics 39: 859–882.

Comstock, R. E., H. F. Robinson, and P. H. Harvey, 1949 A Breeding Procedure Designed To Make Maximum Use of Both General and Specific Combining Ability1. Agron. J. 41: 360–367.

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño *et al.*, 2010 Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics 186: 713–724.

Crossa, J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín *et al.*, 2017 Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends Plant Sci. 22: 961–975.

Crow, J. F., 1998 90 Years Ago: The Beginning of Hybrid Maize. Genetics 148: 923–928.

Crow, J. F., 1999 Dominance and Overdominance, pp. 49–58 in *Genetics and Exploitation of Heterosis in Crops*, edited by J. G. Coors and S. Pandey. American Society of Agronomy, Crop Science Society of America, Madison, WI.

Crow, J. F., and M. Kimura, 1970 *An introduction to population genetics theory*. Harper and Row, New York.

Cui, Y., R. Li, G. Li, F. Zhang, T. Zhu *et al.*, 2020 Hybrid breeding of rice via genomic selection. Plant Biotechnol. J. 18: 57–67.

Culbertson, M. S., J. W. Mabry, I. Misztal, N. Gengler, J. K. Bertrand *et al.*, 1998 Estimation of dominance variance in purebred Yorkshire swine. J. Anim. Sci. 76: 448–451.

Darwin, C., 1876 *The Effects of Cross and Self Fertilization in the Vegetable Kingdom*. Murry, London.

Davenport, C. B., 1908 Degeneration, Albinism and Inbreeding. Science 28: 454–455.

Dekkers, J. C. M., 2007 Marker-assisted selection for commercial crossbred performance. J. Anim. Sci. 85: 2104–2114.

Dekkers, J. C. M., P. K. Mathur, and E. F. Knol, 2011 Genetic Improvement of the Pig, pp. 390–425 in *The genetics of the pig*, edited by M. F. Rothschild and A. Ruvinsky. CABI International, Wallingford.

DeStefano, A. L., and I. Hoeschele, 1992 Utilization of Dominance Variance Through Mate Allocation Strategies. J. Dairy Sci. 75: 1680–1690.

Dias, K. O. D. G., S. A. Gezan, C. T. Guimarães, A. Nazarian, L. da C. e Silva *et al.*, 2018 Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. Heredity 121: 24–37.

Doekes, H. P., R. F. Veerkamp, P. Bijma, G. de Jong, S. J. Hiemstra *et al.*, 2019 Inbreeding depression due to recent and ancient inbreeding in Dutch Holstein–Friesian dairy cattle. Genet. Sel. Evol. 51: 54.

Duenk, P., P. Bijma, M. P. L. Calus, Y. C. J. Wientjes, and J. H. J. van der Werf, 2020 The Impact of Non-additive Effects on the Genetic Correlation Between Populations. G3 Genes Genomes Genet. 10: 783–795.

Duvick, D. N., 1999 Heterosis: Feeding People and Protecting Natural Resources, pp. 19–29 in *Genetics and Exploitation of Heterosis in Crops*, edited by J. G. Coors and S. Pandey. American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc., Madison, WI.

Duvick, D. N., J. S. C. Smith, and M. Cooper, 2010 Long-term selection on a commercial hybrid maize breeding program. Plant Breed. Rev.

East, E. M., 1908 Inbreeding in corn., pp. 419–428 in *Connecticut Agricultural Experiments Station Report 1907*,.

Ertl, J., C. Edel, E. C. G. Pimentel, R. Emmerling, and K.-U. Götz, 2018 Considering dominance in reduced single-step genomic evaluations. J. Anim. Breed. Genet. 135: 151–158.

Ertl, J., A. Legarra, Z. G. Vitezica, L. Varona, C. Edel *et al.*, 2014 Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genet. Sel. Evol. 46: 40.

Esfandyari, H., P. Berg, and A. C. Sørensen, 2018 Balanced selection on purebred and crossbred performance increases gain in crossbreds. Genet. Sel. Evol. 50: 8.

Esfandyari, H., P. Bijma, M. Henryon, O. F. Christensen, and A. C. Sørensen, 2016 Genomic prediction of crossbred performance based on purebred Landrace and Yorkshire data using a dominance model. Genet. Sel. Evol. 48: 40.

Esfandyari, H., A. C. Sørensen, and P. Bijma, 2015 A crossbred reference population can improve the response to genomic selection for crossbred performance. Genet. Sel. Evol. 47: 76.

Falconer, D. S., 1981 *Introduction to quantitative genetics*. Longman Goup Ltd, New York.

Fernández, E. N., A. Legarra, R. Martínez, J. P. Sánchez, and M. Baselga, 2017 Pedigree-based estimation of covariance between dominance deviations and additive genetic effects in closed rabbit lines considering inbreeding and using a computationally simpler

equivalent model. J. Anim. Breed. Genet. Z. Tierzuchtung Zuchtungsbiologie 134: 184–195.

Fritsche-Neto, R., D. Akdemir, and J.-L. Jannink, 2018 Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. Theor. Appl. Genet. 131: 1153–1162.

González-Recio, O., E. López de Maturana, and J. P. Gutiérrez, 2007 Inbreeding Depression on Female Fertility and Calving Ease in Spanish Dairy Cattle. J. Dairy Sci. 90: 5744–5752.

Gregory, K. E., L. V. Cundiff, and R. M. Koch, 1992 Breed effects and heterosis in advanced generations of composite populations for reproduction and maternal traits of beef cattle. J. Anim. Sci. 70: 656–672.

Griffing, B., 1956 Concept of General and Specific Combining Ability in Relation to Diallel Crossing Systems. Aust. J. Biol. Sci. 9: 463–493.

Hallauer, A. R., M. J. Carena, and J. B. M. Filho, 2010 *Quantitative Genetics in Maize Breeding*. Springer-Verlag, New York.

Hayes, B. J., and S. P. Miller, 2000 Mate selection strategies to exploit across- and within-breed dominance variation. J. Anim. Breed. Genet. 117: 347–359.

Hayes, B. J., P. M. Visscher, and M. E. Goddard, 2009 Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. 91: 47–60.

Henderson, C. R., 1975 Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics 31: 423–447.

Hickey, J. M., T. Chiurugwi, I. Mackay, and W. Powell, 2017 Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nat. Genet. 49: 1297–1303.

Hidalgo, A. M., J. W. M. Bastiaansen, M. S. Lopes, M. P. L. Calus, and D. J. de Koning, 2016 Accuracy of genomic prediction of purebreds for cross bred performance in pigs. J. Anim. Breed. Genet. 133: 443–451.

Hill, W. G., and B. S. Weir, 2011 Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. 93: 47–64.

Howard, J. T., J. E. Pryce, C. Baes, and C. Maltecca, 2017 Invited review: Inbreeding in the genomics era: Inbreeding, inbreeding depression, and management of genomic variability. J. Dairy Sci. 100: 6009–6024.

Hull, F. H., 1945 Recurrent Selection for Specific Combining Ability in Corn. Agron. J. 37: 134–145.

Hunt, C. H., B. J. Hayes, F. A. van Eeuwijk, E. S. Mace, and D. R. Jordan, 2020 Multi-environment analysis of sorghum breeding trials using additive and dominance genomic relationships. Theor. Appl. Genet.

Ibánẽz-Escriche, N., R. L. Fernando, A. Toosi, and J. C. Dekkers, 2009 Genomic selection of purebreds for crossbred performance. Genet. Sel. Evol. 41: 12.

Ibáñez-Escriche, N., S. Forni, J. L. Noguera, and L. Varona, 2014 Genomic information in pig breeding: Science meets industry needs. Livest. Sci. 166: 94–100.

IFIP, 2013 Memento de l'éleveur de porc. La génétique.

Iversen, M. W., Ø. Nordbø, E. Gjerlaug-Enger, E. Grindflek, M. S. Lopes *et al.*, 2019 Effects of heterozygosity on performance of purebred and crossbred pigs. Genet. Sel. Evol. 51: 8.

Jenkins, M. T., 1940 The Segregation of Genes Affecting Yield of Grain in Maize. Agron. J. 32: 55–63.

Jenkins, M. T., and A. M. Brunson, 1932 Methods of Testing Inbred Lines of Maize in Crossbred Combinations. J Am Soc Agron 24: 523–530.

Jones, D. F., 1917 Dominance of Linked Factors as a Means of Accounting for Heterosis. Genetics 2: 466–479.

Kadam, D. C., and A. J. Lorenz, 2018 Toward Redesigning Hybrid Maize Breeding Through Genomics-Assisted Breeding, pp. 367–388 in *The Maize Genome*, edited by J. Bennetzen, S. Flint-Garcia, C. Hirsch, and R. Tuberosa. Compendium of Plant Genomes, Springer International Publishing, Cham.

Kadam, D. C., S. M. Potts, M. O. Bohn, A. E. Lipka, and A. J. Lorenz, 2016 Genomic Prediction of Single Crosses in the Early Stages of a Maize Hybrid Breeding Pipeline. G3 GenesGenomesGenetics 6: 3443–3453.

Kinghorn, B. P., 2011 An algorithm for efficient constrained mate selection. Genet. Sel. Evol. 43: 4.

Kinghorn, B. P., 1998 Mate Selection by Groups. J. Dairy Sci. 81: 55–63.

Knol, E. F., B. Nielsen, and P. W. Knap, 2016 Genomic selection in commercial pig breeding. Anim. Front. 6: 15–22.

Kosba, M. A., 1978 Heterosis and phenotypic correlations for shank length, body weight and egg production traits in the Alexandria strains and their crosses with Fayoumi chickens. Beitr. Trop. Landwirtsch. Veterinarmed. 16: 187–198.

Laloë, D., 1993 Precision and information in linear models of genetic evaluation. Genet. Sel. Evol. 25: 557.

Larièpe, A., B. Mangin, S. Jasson, V. Combes, F. Dumas *et al.*, 2012 The Genetic Basis of Heterosis: Multiparental Quantitative Trait Loci Mapping Reveals Contrasted Levels of Apparent Overdominance Among Traits of Agronomical Interest in Maize (Zea mays L.). Genetics 190: 795–811.

Larièpe, A., L. Moreau, J. Laborde, C. Bauland, S. Mezmouk *et al.*, 2017 General and specific combining abilities in a maize (Zea mays L.) test-cross hybrid panel: relative importance of population structure and genetic divergence between parents. Theor. Appl. Genet. 130: 403–417.

Legarra, A., I. Aguilar, and I. Misztal, 2009 A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92: 4656–4663.

Legarra, A., and A. Reverter, 2018 Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. Genet. Sel. Evol. 50: 53.

Lehermeier, C., S. Teyssèdre, and C.-C. Schön, 2017 Genetic Gain Increases by Applying the Usefulness Criterion with Improved Variance Prediction in Selection of Crosses. Genetics 207: 1651–1661.

Lo, L. L., R. L. Fernando, R. J. C. Cantet, and M. Grossman, 1995 Theory for modelling means and covariances in a two-breed population with dominance inheritance. Theor. Appl. Genet. 90: 49–62.

Lo, L. L., R. L. Fernando, and M. Grossman, 1993 Covariance between relatives in multibreed populations: additive model. Theor. Appl. Genet. 87: 423–430.

Lo, L. L., R. L. Fernando, and M. Grossman, 1997 Genetic evaluation by BLUP in two-breed terminal crossbreeding systems under dominance. J. Anim. Sci. 75: 2877–2884.

Lopes, M. S., H. Bovenhuis, A. M. Hidalgo, J. A. M. van Arendonk, E. F. Knol *et al.*, 2017 Genomic selection for crossbred performance accounting for breed-specific effects. Genet. Sel. Evol. 49: 51.

Lutaaya, E., I. Misztal, J. W. Mabry, T. Short, H. H. Timm *et al.*, 2001 Genetic parameter estimates from joint evaluation of purebreds and crossbreds in swine using the crossbred model. J. Anim. Sci. 79: 3002–3007.

Lv, A., H. Zhang, Z. Zhang, Y. Tao, B. Yue *et al.*, 2012 Conversion of the Statistical Combining Ability into a Genetic Concept. J. Integr. Agric. 11: 43–52.

Maenhout, S., B. De Baets, and G. Haesaert, 2010 Prediction of maize single-cross hybrid performance: support vector machine regression versus best linear prediction. Theor. Appl. Genet. 120: 415–427.

Massman, J. M., A. Gordillo, R. E. Lorenzana, and R. Bernardo, 2013 Genomewide predictions from maize single-cross data. Theor. Appl. Genet. 126: 13–22.

McQuillan, R., A.-L. Leutenegger, R. Abdel-Rahman, C. S. Franklin, M. Pericic *et al.*, 2008 Runs of Homozygosity in European Populations. Am. J. Hum. Genet. 83: 359–372.

Melchinger, A. E., and R. K. Gumber, 1998 Overview of Heterosis and Heterotic Groups in Agronomic Crops, pp. 29–44 in *Concepts and Breeding of Heterosis in Crop Plants*, edited by K. R. Larnkey and J. E. Staub. CSSA, Madison, WI.

Melchinger, A. E., H.-P. Piepho, H. F. Utz, J. Muminović, T. Wegenast *et al.*, 2007 Genetic Basis of Heterosis for Growth-Related Traits in Arabidopsis Investigated by Testcross Progenies of Near-Isogenic Lines Reveals a Significant Role of Epistasis. Genetics 177: 1827–1837.

Meuwissen, T., B. Hayes, and M. Goddard, 2013 Accelerating Improvement of Livestock with Genomic Selection. Annu. Rev. Anim. Biosci. 1: 221–237.

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

Misztal, I., 1997 Estimation of Variance Components with Large-Scale Dominance Models. J. Dairy Sci. 80: 965–974.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet *et al.*, 2002 BLUPF90 and related programs (BGF90), pp. 28 in *Proceedings of the 7th World Congress Applied to Livestock Production.*, INRA, Montpellier, France.

Moghaddar, N., and J. H. J. van der Werf, 2017 Genomic estimation of additive and dominance effects and impact of accounting for dominance on accuracy of genomic evaluation in sheep populations. J. Anim. Breed. Genet. 134: 453–462.

Powers, L., 1944 An Expansion of Jones's Theory for the Explanation of Heterosis. Am. Nat. 78: 275–280.

Pryce, J. E., M. Haile-Mariam, M. E. Goddard, and B. J. Hayes, 2014 Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. Genet. Sel. Evol. 46: 71.

Ramstein, G. P., S. J. Larsson, J. P. Cook, J. W. Edwards, E. S. Ersoz *et al.*, 2020 Dominance Effects and Functional Enrichments Improve Prediction of Agronomic Traits in Hybrid Maize. Genetics.

Reif, J. C., F.-M. Gumpert, S. Fischer, and A. E. Melchinger, 2007 Impact of Interpopulation Divergence on Additive and Dominance Variance in Hybrid Populations. Genetics 176: 1931–1934.

Reif, J. C., A. R. Hallauer, and A. E. Melchinger, 2005 Heterosis and heterotic patterns in maize. Maydica 50: 215–223.

Reif, J. C., A. E. Melchinger, X. C. Xia, M. L. Warburton, D. A. Hoisington *et al.*, 2003 Genetic Distance Based on Simple Sequence Repeats and Heterosis in Tropical Maize Populations. Crop Sci. 43: 1275–1282.

Richey, F. D., 1924 Effects of selection on the yield of a cross between varieties of corn. USDA Bull 1209.

Richey, F. D., 1942 Mock-Dominance and Hybrid Vigor. Science 96: 280–281.

Rincent, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel *et al.*, 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: 715–728.

Rodríguez-Ramilo, S. T., J. M. Elsen, and A. Legarra, 2019 Inbreeding and effective population size in French dairy sheep: Comparison between genomic and pedigree estimates. J. Dairy Sci. 102: 4227–4237.

Rogers, J., 1972 Measures of genetic similarity and genetic distance, pp. 145–153 in The University of Texas publications, Austin.

Roos, A. P. W. de, B. J. Hayes, and M. E. Goddard, 2009 Reliability of Genomic Predictions Across Multiple Populations. Genetics 183: 1545–1553.

Samorè, A. B., and L. Fontanesi, 2016 Genomic selection in pigs: state of the art and perspectives. Ital. J. Anim. Sci. 15: 211–232.

Saura, M., A. Fernández, L. Varona, A. I. Fernández, M. Á. R. de Cara *et al.*, 2015 Detecting inbreeding depression for reproductive traits in Iberian pigs using genome-wide data. Genet. Sel. Evol. 47: 1.

Schön, C. C., and H. Simianer, 2015 Resemblance between two relatives – animal and plant breeding. J. Anim. Breed. Genet. 132: 1–2.

Schulthess, A. W., Y. Zhao, and J. C. Reif, 2017 Genomic Selection in Hybrid Breeding, pp. 149–183 in *Genomic Selection for Crop Improvement: New Molecular Breeding Strategies for Crop Improvement*, edited by R. K. Varshney, M. Roorkiwal, and M. E. Sorrells. Springer International Publishing, Cham.

Sellier, P., 1976 The basis of crossbreeding in pigs; A review. Livest. Prod. Sci. 3: 203–226.

Sevillano, C. A., H. Bovenhuis, and M. P. L. Calus, 2019 Genomic Evaluation for a Crossbreeding System Implementing Breed-of-Origin for Targeted Markers. Front. Genet. 10:.

Sevillano, C. A., J. Vandenplas, J. W. M. Bastiaansen, and M. P. L. Calus, 2016 Empirical determination of breed-of-origin of alleles in three-breed cross pigs. Genet. Sel. Evol. 48: 55.

Shen, G., W. Zhan, H. Chen, and Y. Xing, 2014 Dominance and epistasis are the main contributors to heterosis for plant height in rice. Plant Sci. 215–216: 11–18.

Shull, G. H., 1909 A Pure-Line Method in Corn Breeding. J. Hered. os-5: 51–58.

Shull, G. H., 1914 Duplicate genes for capsule-form inBursa bursa-pastoris. Z. Für Indukt. Abstamm.- Vererbungslehre 12: 97–149.

Shull, G. H., 1910 Hybridization Methods in Corn Breeding. J. Hered. 1: 98–107.

Shull, G. H., 1908 The Composition of a Field of Maize. J. Hered. os-4: 296–301.

Silió, L., M. c. Rodríguez, A. Fernández, C. Barragán, R. Benítez *et al.*, 2013 Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics. J. Anim. Breed. Genet. 130: 349–360.

Smith, O. S., 1986 Covariance between Line per se and Testcross Performance1. Crop Sci. 26: cropsci1986.0011183X002600030023x.

Smith, C., 1964 The use of specialised sire and dam lines in selection for meat production. Anim. Sci. 6: 337–344.

Smith, J. S. C., D. N. Duvick, O. S. Smith, A. Grunst, and S. J. Wall, 1999 Effect of Hybrid Breeding on Genetic Diversity in Maize, pp. 119–126 in *The Genetics and Exploitation of Heterosis in Crops*, edited by J. G. Coors and S. Pandey. American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc., Madison, WI.

Smith, J. S. C., T. Hussain, E. S. Jones, G. Graham, D. Podlich *et al.*, 2008 Use of doubled haploids in maize breeding: implications for intellectual property protection and genetic diversity in hybrid crops. Mol. Breed. 22: 51–59.

Smith, C., and J. W. B. King, 1964 Crossbreeding and litter production in British pigs. Anim. Sci. 6: 265–271.

Sprague, G. F., 1946 Early Testing of Inbred Lines of Corn1. J Am Soc Agron 38: 108–117.

Sprague, G. F., and L. A. Tatum, 1942 General vs. Specific Combining Ability in Single Crosses of Corn1. Agron. J. 34: 923–932.

Stich, B., A. E. Melchinger, H.-P. Piepho, S. Hamrit, W. Schipprack *et al.*, 2007 Potential causes of linkage disequilibrium in a European maize breeding program investigated with computer simulations. Theor. Appl. Genet. 115: 529–536.

Stuber, C. W., and C. C. Cockerham, 1966 Gene Effects and Variances in Hybrid Populations. Genetics 54: 1279–1286.

Su, G., O. F. Christensen, T. Ostersen, M. Henryon, and M. S. Lund, 2012 Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. PLOS ONE 7: e45293.

Sun, C., P. M. VanRaden, J. R. O'Connell, K. A. Weigel, and D. Gianola, 2013 Mating programs including genomic relationships and dominance effects. J. Dairy Sci. 96: 8014–8023.

Technow, F., C. Riedelsheimer, T. A. Schrag, and A. E. Melchinger, 2012 Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor. Appl. Genet. 125: 1181–1194.

Technow, F., T. A. Schrag, W. Schipprack, E. Bauer, H. Simianer *et al.*, 2014 Genome Properties and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program of Maize. Genetics 197: 1343–1355.

Toro, M., 1993 A new method aimed at using the dominance variance in closed breeding populations. Genet. Sel. Evol. 25: 63.

Toro, M. A., and L. Varona, 2010 A note on mate allocation for dominance handling in genomic selection. Genet. Sel. Evol. 42: 33.

Tusell, L., H. Gilbert, J. Riquet, M.-J. Mercat, A. Legarra *et al.*, 2016 Pedigree and genomic evaluation of pigs using a terminal-cross model. Genet. Sel. Evol. 48: 32.

Tusell, L., H. Gilbert, Z. G. Vitezica, M. J. Mercat, A. Legarra *et al.*, 2019 Dissecting total genetic variance into additive and dominance components of purebred and crossbred pig traits. animal 13: 2429–2439.

Tusell, L., P. Pérez-Rodríguez, S. Forni, X.-L. Wu, and D. Gianola, 2013 Genome-enabled methods for predicting litter size in pigs: a comparison. Anim. Int. J. Anim. Biosci. 7: 1739–1749.

Vandenplas, J., M. P. L. Calus, C. A. Sevillano, J. J. Windig, and J. W. M. Bastiaansen, 2016 Assigning breed origin to alleles in crossbred animals. Genet. Sel. Evol. 48: 61.

Vandenplas, J., J. J. Windig, and M. P. L. Calus, 2017 Prediction of the reliability of genomic breeding values for crossbred performance. Genet. Sel. Evol. 49: 43.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

VanRaden, P. M., 2020 Symposium review: How to implement genomic selection. J. Dairy Sci. 103: 5291–5301.

VanRaden, P. M., and L. A. Smith, 1999 Selection and Mating Considering Expected Inbreeding of Future Progeny. J. Dairy Sci. 82: 2771–2778.

Varona, L., A. Legarra, W. Herring, and Z. G. Vitezica, 2018a Genomic selection models for directional dominance: an example for litter size in pigs. Genet. Sel. Evol. GSE 50:.

Varona, L., A. Legarra, M. A. Toro, and Z. G. Vitezica, 2018b Non-additive Effects in Genomic Selection. Front. Genet. 9:.

Visscher, P., R. Pong-Wong, C. Whittemore, and C. Haley, 2000 Impact of biotechnology on (cross)breeding programmes in pigs. Livest. Prod. Sci. 65: 57–70.

Vitezica, Z. G., A. Legarra, M. A. Toro, and L. Varona, 2017 Orthogonal Estimates of Variances for Additive, Dominance, and Epistatic Effects in Populations. Genetics 206: 1297–1307.

Vitezica, Z. G., L. Varona, J.-M. Elsen, I. Misztal, W. Herring *et al.*, 2016 Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. Genet. Sel. Evol. 48: 6.

Vitezica, Z. G., L. Varona, and A. Legarra, 2013 On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope. Genetics 195: 1223–1230.

Wang, H., X. Zhang, H. Yang, X. Liu, H. Li *et al.*, 2016 Identification of heterotic loci associated with grain yield and its components using two CSSL test populations in maize. Sci. Rep. 6: 38205.

Wegenast, T., C. F. H. Longin, H. F. Utz, A. E. Melchinger, H. P. Maurer *et al.*, 2008 Hybrid maize breeding with doubled haploids. IV. Number versus size of crosses and importance of parental selection in two-stage selection for testcross performance. Theor. Appl. Genet. 117: 251–260.

Wei, M., and H. A. M. van der Steen, 1991 Comparison of reciprocal recurrent selection with pure-line selection systems in animal breeding (A review). Anim. Breed. Abstr. 59: 281–298.

Wei, M., J. H. J. V. der Werf, and E. W. Brascamp, 1991 Relationship between purebred and crossbred parameters. J. Anim. Breed. Genet. 108: 262–269.

Wei, M., and J. H. J. van der Werf, 1994 Maximizing genetic response in crossbreds using both purebred and crossbred information. Anim. Sci. 59: 401–413.

Wientjes, Y. C. J., and M. P. L. Calus, 2017 BOARD INVITED REVIEW: The purebred-crossbred correlation in pigs: A review of theory, estimates, and implications. J. Anim. Sci. 95: 3467–3478.

Williams, W., 1959 Heterosis and the Genetics of Complex Characters. Nature 184: 527–530.

Xiang, T., O. F. Christensen, and A. Legarra, 2017 Technical note: Genomic evaluation for crossbred performance in a single-step approach with metafounders. J. Anim. Sci. 95: 1472–1480.

Xiang, T., O. F. Christensen, Z. G. Vitezica, and A. Legarra, 2016a Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. Genet. Sel. Evol. 48: 92.

Xiang, T., B. Nielsen, G. Su, A. Legarra, and O. F. Christensen, 2016b Application of single-step genomic evaluation for crossbred performance in pig. J. Anim. Sci. 94: 936–948.

Zeng, J., A. Toosi, R. L. Fernando, J. C. M. Dekkers, and D. J. Garrick, 2013 Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet. Sel. Evol. GSE 45: 11.

Zhao, Y., Z. Li, G. Liu, Y. Jiang, H. P. Maurer *et al.*, 2015 Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. Proc. Natl. Acad. Sci. 112: 15624–15629.

Training and education during the PhD thesis

Year	Training course	Hours
2020	▪ Genotype by environment interaction, uniformity and resilience. Wageningen University and Research, NL, 3rd – 7th Feb 2020. Dr. Han Mulder, Dr. Piter Bijma and Dr. John Bastiaansen.	40
2019	▪ Modify and extract information from large text files using AWK and Sed programming Languages. Genotoul bioinformatics platform, INRA, Toulouse France, 16th Oct 2019.	8
	▪ Ethics, Science and Society UNESCO. Université Fédérale Toulouse Midi-Pyrénées, Toulouse. 17 th Jun 2019.	6
	▪ Master courses of Animal Genetics and Breeding, and Reproductive Biotechnology: Quantitative Genetics I, Quantitative Genetics II, Quantitative Genetics III and Conservation and management of animal genetic resources. Universitat Politècnica de València, Animal Science Depto. Valencia, Spain. 28th Jan – 29th Mar 2019. Dr. María Santacreu, Dr. Noelia Ibáñez-Escriche, Dr. Agustín Blasco and Dr. Miguel Toro.	412
	▪ TOEIC score 770. 18th July 2019.	-
2018	▪ Systems Biology and Gene Networks Inference: Application to Livestock Breeding and Genetics. INRA Toulouse, 23th Nov. 2018. Dr. Antonio Reverter (CSIRO).	20
	▪ Programming and Computer Algorithms in animal Breeding with focus on genomic Selection and Single-Step GBLUP. UGA, US, 7th – 25th May 2018. Dr. Ignacy Misztal and Collaborators.	75
	▪ French as a foreign language course, level A2/B1. Alliance Française, Toulouse France, 8th Oct – 26th Nov. 2018.	18
	▪ French as a foreign language course, level A2. Alliance Française, Toulouse France, 26th Feb – 2th Mar. 2018.	25
2017	▪ Design of Breeding Programs with Genomic Selection. Wageningen University and Research, NL, 16th – 20th Oct. 2017. Dr. Julius van der Werf, Dr. Piter Bijma, Dr. Mario Calus and Dr. Han Mulder.	40
	▪ Animal Breeding and Genetics. Wageningen University and Research (MOOC – EdX), 26th Oct. 2017.	28
	▪ Intensive English course, level A2/B1. l'Université Fédérale Toulouse Midi-Pyrénées, Toulouse France. 30-31th Oct. and 2nd-3rd Nov. 2017.	25
	Total hours of training courses	697

Titre: Sélection génomique tenant compte des effets génétiques non additifs dans les programmes de croisement de porcs et de maïs

Mots clés: Sélection génomique, croisement, porc, maïs

Résumé

Cette thèse explore et développe des méthodes pour exploiter les effets génétiques de dominance ou/et d'épistasie sur des modèles de sélection génomique dans les schémas de sélection utilisant le croisement chez le porc et le maïs. Le Chapitre 2 a consisté à estimer et à exploiter la variance de dominance intra-race à travers des stratégies de planification des accouplements pour maximiser le progrès génétique global de l'âge à 100 kg (AGE), de l'épaisseur de lard dorsal (BD) et du poids moyen des porcelets par portée (APWL), dans une population porcine de race Landrace française. La maximisation de la valeur génétique totale au lieu de la seule partie additive dans la planification des accouplements a donné à la descendance un avantage moyen de -0,79 jour, -0,04 mm et 11,3 g pour AGE, BD et APWL, respectivement. En contrepartie, le gain génétique additif attendu a légèrement été réduit (1,8% pour AGE par exemple). Ces résultats indiquent que la planification des accouplements peut améliorer les performances des descendants sans compromettre considérablement le progrès génétique. Dans le Chapitre 3, l'efficacité de la planification des accouplements et de l'évaluation génomique, en tenant compte des effets additifs et de dominance, pour améliorer les performances des individus croisés (CB) a été étudiée par simulation dans un croisement à deux voies chez le porc. Les effets de l'utilisation de différentes sources d'informations dans l'évaluation génétique (uniquement des données de race pure (PB) ou des données PB et CB), de plusieurs valeurs d'héritabilité au sens étroit et large, et de plusieurs stratégies d'accouplement pour produire les animaux CB (accouplements au hasard, minimisant la consanguinité future ou maximisant la valeur génétique totale attendue des animaux CB) ont été évaluées. La sélection des animaux PB sur leurs performances en PB a donné un gain génétique de 0,2 écart-type génétique par génération pour le caractère « performance en croisement ». Ce gain a été doublé lorsque les animaux PB étaient sélectionnés sur leur performance en croisés. Les stratégies d'accouplement ont entraîné une légère augmentation des performances des animaux CB. Lorsque la corrélation génétique entre les performances exprimées chez les animaux PB et CB est faible, la sélection des animaux PB pour leur performance en croisés en utilisant les informations CB est une stratégie plus efficace pour exploiter l'hétérosis et augmenter les performances des animaux CB au niveau commercial. Dans le Chapitre 4, la théorie des modèles d'évaluation génétique chez des hybrides à partir du croisement de lignées pures (comme pour le maïs) a été revue dans un contexte génomique. La covariance entre les hybrides due aux effets de substitution additifs, à la dominance et à l'épistasie a été dérivée analytiquement. En utilisant les marqueurs SNP, il est possible de décomposer l'aptitude spécifique à la combinaison (SCA) en dominance et épistasie inter-groupes, et de décomposer les aptitudes générales à la combinaison (GCA) en effets additifs intra-lignée et une épistasie additive par additive intra-ligne. Un jeu de données publiques sur des hybrides Dent × Flint a été analysé. Le modèle proposé a été comparé à d'autres modèles génomiques en termes d'estimations des composantes de variance et de capacité prédictive, y compris un modèle supposant un effet commun des gènes des lignées pures. L'étude confirme que la majeure partie de la variabilité observée chez les hybrides est expliquée par la GCA et que les variances dues à la dominance et à l'épistasie sont moins importantes et du même ordre de grandeur. Les modèles basés sur la définition d'effets différents (comme traditionnellement considérés chez le maïs), ou commun aux origines (comme considérés intra-race en sélection animale) ont abouti à des capacités prédictives similaires pour les hybrides.

Title: Genomic selection accounting for non-additive genetic effects in pig and corn crossbreeding schemes

Keywords: Genomic selection, crossbreeding, pigs, maize

Summary

This thesis explores and develops methodology to exploit dominance or/and epistasis genetic effects on genomic selection models in pig and maize crossbreeding schemes. The Chapter 2 consisted of estimating and exploiting within-breed dominance variance through mate allocation strategies to maximize the overall genetic merit of the traits age at 100 Kg (AGE), backfat depth (BD) and average piglet weight per litter (APWL), in a French Landrace pig population. Maximizing total genetic values instead of breeding values in matings gave to the progeny an average advantage of − 0.79 days, − 0.04 mm, and 11.3 g for AGE, BD and APWL, respectively, but slightly reduced the expected additive genetic gain (e.g. 1.8 % for AGE). These results indicate that genomic mate allocation can improve the performance of the offspring without dramatically compromising the additive genetic gain. In Chapter 3, the effectiveness of mate allocation strategies and genomic evaluations, accounting for additive and dominance effects, to improve crossbred (CB) performance were investigated by simulation in a two-way pig crossbreeding scheme. Effects of the sources of information used in the genetic evaluation (only purebred (PB) data or PB and CB data), of several narrow and broad-sense heritability values, and of several options for mate allocation to produce the CB (mating at random, minimizing expected future inbreeding, or maximizing the expected total genetic value of crossbred animals) were evaluated. Selecting PB animals for PB performance yielded a genetic gain of 0.2 genetic standard deviations of the trait "CB performance" per generation, whereas selecting PB animals for CB performance doubled the genetic response. Mate allocation strategy resulted in a slight increase of the CB performance. When the genetic correlation between PB and CB is low, selecting PB animals for CB performance using CB information is a more efficient strategy to exploit heterosis and increase performance at the CB commercial level. In Chapter 4, the theory of hybrid genetic evaluation models from single-cross of pure lines (as in maize) was revisited in a genomic context. Covariance between hybrids due to additive substitution effects and dominance and epistatic deviations were analytically derived. Using SNP genotypes, it is possible to split specific combining ability (SCA) into dominance and across-groups epistasis, and to split general combining ability (GCA) into within-line additive effects and within-line additive by additive epistasis. A publicly available maize data set of Dent × Flint hybrids was analyzed. The proposed model was compared to other genomic models in terms of variance components estimation and predictive ability, including a model assuming a common effect of genes across origins. The study confirms that most variation in hybrids is accounted for by GCA, and that variances due to dominance and epistasis are small and have similar magnitudes. Models based on defining effects either differently (as it is traditionally done in maize) or identically across origins (as it is done in single breeds in livestock) resulted in similar predictive abilities for hybrids.