Contributions to Genetic Diversity Management in Maize Breeding Programs using Genomic Selection

Antoine Allier

université PARIS-SACLAY

Contributions to Genetic Diversity Management in Maize Breeding Programs using Genomic Selection

Thèse de doctorat de l'université Paris-Saclay

Ecole doctorale n° 581 : Agriculture, Alimentation, Biologie, Environnement, Santé (ABIES)
Spécialité de doctorat : Sciences agronomiques
Unité de recherche : Université Paris-Saclay, INRAE, CNRS, AgroParisTech,
GQE - Le Moulon, 91190, Gif-sur-Yvette, France
Référent : AgroParisTech

Thèse présentée et soutenue à Gif-sur-Yvette, le 20 Janvier 2020, par

Antoine ALLIER

Composition du Jury

Christine DILLMANN Professeur, Université Paris-Sud	Présidente
Jean-Christophe GLASZMANN Directeur de recherche, CIRAD	Rapporteur
Gregor GORJANC Associate professor, Roslin Institute	Rapporteur
Sophie BOUCHET Chargée de recherche, INRAE	Examinatrice
Leopoldo SANCHEZ-RODRIGUEZ Directeur de recherche, INRAE	Examinateur
Alain CHARCOSSET Directeur de recherche, INRAE	Directeur de thèse
Christina LEHERMEIER Responsable Statistical Genetics Unit, RAGT2n	Co-encadrante
Hélène PASCAL Sélectionneur maïs, RAGT2n	Invitée
Gwendal RESTOUX Chargé de recherche, INRAE	Invité

Thèse de doctorat

NNT : 2020UPASA002

Acknowledgements

Ce manuscrit est l'aboutissement de trois années de thèse. Trente-six mois, c'est le temps requis pour qu'un ingénieur se développe, s'épanouisse et devienne docteur. Comme pour toute ''espèce'' cultivée, rendement et qualité sont déterminés par de nombreux facteurs. A commencer par l'étudiant, facteur que nous supposerons fixé par la suite. La marge de manœuvre résulte, donc, de l'interaction de nombreux facteurs environnementaux. Je vais, maintenant, remercier toutes les personnes ayant contribué à ces interactions et qui ont rendu ce travail de thèse possible.

Tout d'abord, je souhaite remercier mes co-directeurs.ices et co-encadrants.es de thèse de m'avoir fait confiance, de m'avoir accompagné et transmis sur le plan scientifique mais aussi humain durant ces trois années.
Je tiens à remercier Alain Charcosset de m'avoir accompagné dans ma décision de commencer cette thèse. Merci pour ta contribution au projet de thèse et merci de m'avoir fait confiance et laissé une grande liberté. Merci aussi pour tes nombreuses suggestions qui ont orienté mes travaux et merci de m'avoir fait bénéficier de ta culture générale partagée gracieusement autour d'un café ou d'un verre de vin. Merci Laurence Moreau pour ton implication, tes remarques et corrections qui ont largement contribué à améliorer la qualité de mes travaux. Merci également d'avoir ouvert mon esprit scientifique à d'autres espèces végétales et animales lors de réunions des méta-programmes R2D2 et GDivSelGen. Merci à Alain et Laurence de m'avoir hébergé pendant trente mois au Moulon dans une équipe dynamique et très sympathique. Ce fut un réel plaisir de travailler avec vous.
Je souhaiterais également remercier Simon Teyssèdre de m'avoir accepté en thèse CIFRE chez RAGT2n. Merci pour ton accompagnement scientifique durant ces trois années, tu m'as énormément appris en génétique quantitative et informatique. Merci à Christina Lehermeier qui nous a rejoint en cours de thèse. Je tiens à te remercier pour tout ce que tu m'as transmis et ta contribution à mes travaux. Für alles möchte ich dir bedanken ! Merci à Simon et Christina pour leur disponibilité lors de mes séjours à Rodez, incluant aussi de nombreux trajets pendant les six mois cumulés.

Le succès d'Alain, Laurence, Simon et Christina à me tuteurer et cultiver mon esprit scientifique durant ces trois années est aussi dû à un environnement global favorable. Je souhaiterais donc remercier toutes les personnes qui, directement ou indirectement, ont rendu les environnements ruthénois et moulonien favorables à l'aboutissement de cette thèse.
Un grand merci au personnel du Moulon de m'avoir accueilli au sein de votre village de gaulois où, au milieu des chantiers, il y fait bon vivre. J'ai essayé de me mettre au basket mais, sans grand succès ! Je suis donc passé du grand au petit ballon de rouge : le club œnologique m'a mieux réussi ! Merci à Cyril Bauland de nous transmettre ta passion. Merci à toute l'équipe GQMS pour votre gentillesse, nos discussions et votre participation à la récolte des données analysées. Merci pour nos pique-niques et particulièrement à Valérie Combes pour ses excellents gâteaux. Comment passer à côté de mes trois acolytes ? Adama Seye, Clément Mabire et Simon Rio, j'ai passé avec vous de supers moments scientifiques, festifs et même ''scientifestifs'' ! Mes premières deux années et demi de thèse ont été un réel plaisir à vos côtés. Merci à Clément, au club de rugby de l'université d'Orsay et à l'équipe de touch-rugby du CEA pour nos entrainements au ballon ovale. Merci aussi aux futurs doctorants GQMS :

Alizarine, Aurélien et Dimitri de m'avoir supporté lors des derniers mois de rédaction. Je vous souhaite le meilleur pour votre thèse.
Un tout grand merci au personnel RAGT2n de m'avoir accueilli durant six mois et votre participation à la récolte des données que j'ai analysées. Je souhaite tout particulièrement remercier Bruno Claustres, Philippe Dufour et Sébastien Chatre pour votre implication dans la gestion administrative de ma thèse, Stéphane Melkior, Stéphane Maltese et les autres sélectionneurs maïs pour nos échanges constructifs, Francoise Fabre et Michel Romestant pour les données de génotypage ainsi que nos discussions. Merci aussi à Brice Lascourreges et Emma Sabatier pour votre réactivité et votre aide administrative lors de mes déplacements. Un remerciement particulier à *@qfazill+* pour tes astuces en linux, expressions régulières, développement d'application Shiny et de m'avoir initié au palet vendéen ainsi qu'à la pêche. Merci à *@aallard* pour nos échanges et l'animation de mes soirées et week-ends ruthénois.

Un esprit scientifique qui se développe bien est un esprit scientifique bien préparé. Merci aux personnes qui m'ont formé avant ma thèse, incluant mes professeurs de classe préparatoire pour leur excellent enseignement et la méthodologie de travail qu'ils m'ont transmise. Merci à Julie Fievet et Philippe Brabant pour votre formation en amélioration des plantes. Je souhaiterais de même remercier les personnes qui m'ont fait confiance lors d'expériences professionnelles riches : Werner Beyer (KWS), Milagros Garcia et Eric Mandrou (Euralis) qui ont confirmé mon intérêt pour la génétique quantitative et m'ont orienté vers la thèse. Merci à Jean-Michel Elsen, John Hickey et Tristan Mary-Huard pour vos retours et recommandations qui ont activement participé au bon déroulement de ma thèse.

Un tout grand merci à Jean-Christophe Glaszmann et Gregor Gorjanc d'avoir accepté d'examiner ce manuscrit et à tous les membres du jury pour votre participation à son évaluation : Christine Dillmann, Sophie Bouchet, Leopoldo Sanchez-Rodriguez, Gwendal Restoux et Hélène Pascal.

Enfin un grand merci à mes parents de m'avoir donné les clefs intellectuelles et humaines afin de suivre ces études. Merci à mes sœurs pour leur soutien indéfectible. Enfin merci à Lola avec qui j'ai eu la joie de partager ce parcours. Nous passerons cette étape même jour même heure à 333 km de distance. Merci pour ton soutien et tous les bons moments passés à Rennes et à Paris qui ont été autant de repos bienvenus.

Contents

List of abbreviations

BLUE	Best Linear Unbiased Estimation
BLUP	Best Linear Unbiased Prediction
CD	Coefficient of Determination
CRISPR	Clustered Regulatory Interspaced Short Palindromic Repeats
CSI	Cross Selection Index
DE	Differential Evolutionary algorithm
DH	Doubled Haploid
DNA	DeoxyriboNucleic Acid
FAO	Food and Agriculture Organization
GBS	Genotyping By Sequencing
GCA	General Combining Ability
GEBV	Genomewide Estimated Breeding Value
GEM	Germplasm Enhancement of Maize
GP	Genomic Prediction
GS	Genomic Selection
He	Excpected Heterozygosity
HEBV	Haplotypic Estimated Breeding Value
HOPE	Hierarchical Open-ended Population Enrichment
IBD	Identity By Descent
IBS	Identity By State
IID	Identically and Independently Distributed
INRA	Institut National de la Recherche Agronomique
ISSS	Iowa Stiff Stalk Synthetics
LAMP	Latin American Maize Project
LD	Linkage Disequilibrium
LE	Linkage Equilibrium
MAF	Minor Allele Frequency
MAS	Marker Assisted Selection
MCMC	Markov Chain Monte Carlo
MRD	Modified Roger's Distance
Ne	Effective population size
OCS	Optimal Cross Selection
OHV	Optimal Haploid Value
OPV	Optimal Population Value
OPVs	Open Pollinated Varieties
PAGE	Promotion of Alleles by Genome Editing
PC	Parental Contribution
PCV	Predicted Cross Value
PM	Parental Mean
PMV	Posterior Mean Variance
PP	Prediction Population
PVPA	Plant Variety Protection Act
QTL	Quantitative Trait Locus
REML	Restricted Maximum Likelihood
RIL	Recombinant Inbred Line
ROHe	Run Of Expected Homozygosity
RRS	Recurrent Reciprocal Selection
SCA	Specific Combining Ability
SeeD	Seed of Discovery
SNP	Single Nucleotide Polymorphism
TALEN	Transcription Activator-Like Effector Nuclease
TBV	True Breeding Value
TP	Training Population
TS	Training Set
UC	Usefulness Criterion
UCPC	Usefulness Criterion Parental Contributions
UPOV	International Union for the Protection of New Varieties of Plants
USDA	United States Department of Agriculture
VPM	Variance of Posterior Means
ZNF	Zinc Finger Nuclease

Résumé (Français)

Une sélection efficace et durable repose sur un compromis entre efforts à court terme afin de proposer aux agriculteurs des variétés compétitives, et le maintien d'une base génétique large garantissant des variétés futures qui répondront aux défis climatiques, biologiques et sociétaux de demain. Les avancées du génotypage haut débit ont ouvert de nouvelles perspectives de sélection pour les caractères quantitatifs, telles que la prédiction génomique de performances individuelles, la prédiction de l'intérêt de plans de croisements, ainsi que la gestion de la diversité. L'objectif de cette thèse est de contribuer au développement de méthodologies et schémas de sélection efficaces et durables. Cela inclue l'évaluation de la diversité génétique des populations élites, sa conversion efficace en gain génétique à court et long termes, ainsi que l'identification de sources de variabilité génétique d'intérêt et leur introduction dans les populations de sélection.

Nous proposons tout d'abord d'exploiter des séries temporelles de données phénotypiques et génotypiques afin d'évaluer l'effet de la sélection sur la diversité génétique des populations élites ainsi que leur réponse attendue à la sélection. En **Chapitre 1**, nous proposons trois séries d'indicateurs : phénotypique, génotypique et génomique. Le fondement théorique de ces indicateurs est tout d'abord présenté. Ils sont ensuite appliqués à un programme de sélection maïs grain portant sur les groupes hétérotiques cornés et dentés. Un gain génétique significatif est observé sur dix ans dans les populations ''cornée'' et ''dentée'' en sélection et est accompagné d'une perte de variance génétique additive en absence d'introductions de matériel externe dans la population ''denté''. Une perte significative de diversité génétique ainsi que des régions à très faible diversité dans les régions péri-centromériques sont aussi observées dans ce groupe. Enfin, il est estimé que la répulsion entre locus causaux capture 24% de la variance génique additive totale chez les dentés, soit 4,9% de la réponse potentielle maximale à la sélection. Cette proportion varie entre chromosomes ce qui permet de suggérer différentes stratégies de gestion et d'amélioration de la réponse à la sélection selon les chromosomes. Ces indicateurs sont faciles à implémenter et permettent d'exploiter, à moindre coût, les données phénotypiques et génotypiques stockées dans des bases de données sur plusieurs générations de sélection afin d'aider les sélectionneurs dans leurs décisions stratégiques.

Par la suite, nous nous sommes intéressés à la gestion de la diversité génétique afin d'optimiser sa conversion en gain génétique à court terme sans compromettre le gain génétique à long terme. La sélection du plan de croisement qui génère des descendants performants et maintient suffisamment de diversité est un facteur clef du succès à court et long termes des programmes de sélection récurrente. L'identification du croisement maximisant la probabilité de sélectionner une descendance meilleure que les parents de départ repose sur la prédiction de la distribution d'un caractère quantitatif dans la descendance du croisement. Cette approche est communément appelée critère d'utilité et prend en compte la complémentarité entre parents, i.e. la ségrégation mendélienne dans la descendance, pour le caractère quantitatif considéré. En **Chapitre 3**, le modèle prédictif de la distribution d'un caractère quantitatif dans une famille biparentale est étendu au cas des familles multi-parentales. Une approche multi-caractères est ensuite proposée, considérant les performances agronomiques et les contributions parentales comme des caractères quantitatifs corrélés et normalement distribués. Cette approche dénommée critère d'utilité et contributions parentales (UCPC) permet de prédire la performance moyenne et la diversité attendues dans la fraction sélectionnée de la descendance d'un croisement. L'UCPC peut être utilisé afin d'étendre la sélection optimale de plan de croisements (OCS) qui a pour but de maximiser le gain génétique tout en limitant la perte de diversité. En **Chapitre 4**, nous comparons différents plans de croisements par simulation. Il est tout d'abord observé qu'une sélection des croisements basée sur le critère d'utilité maximise le gain à court et long termes comparativement à une sélection basée sur la moyenne des performances parentales sans prise en compte de la ségrégation attendue de leur descendance. Ensuite, nous montrons que les approches de croisement optimales (OCS) sont plus performantes à long terme mais au prix d'une pénalité à court terme comparativement au critère d'utilité. Finalement, l'OCS basée sur l'UCPC convertit plus efficacement la diversité génétique en gain à court et long termes que l'OCS. Ainsi, la sélection de croisement optimale basée sur l'UCPC aide les sélectionneurs dans leur choix de plan de croisements pour satisfaire leurs objectifs à court et long termes.

Une base génétique étroite des populations élites compromet le gain génétique à long terme. De ce fait, une stratégie d'élargissement de leur base génétique sans compromettre le gain à court terme est nécessaire. De nombreuses sources de diversité peuvent être considérées mais toutes ne peuvent être évaluées. En **Chapitre 2**, différents critères prédictifs sont passés en revue et comparés afin d'évaluer l'utilité de ressources génétiques pour enrichir un pool élite. Ces critères évaluent la complémentarité entre ressources génétiques et lignée élite receveuse afin d'assurer l'apport de nouveaux allèles ou haplotypes favorables absents de la population élite. Les critères proposés s'appuient sur les effets aux marqueurs estimés dans un panel collaboratif constitué de lignées de diversité publiques et de lignées élites privées (panel denté issu du projet « Amaizing »). La qualité

prédictive obtenue par validation croisée sur le panel collaboratif ainsi que la qualité prédictive non nulle obtenue sur une large population élite montre l'intérêt d'utiliser ces effets à des fins d'identification de ressources génétiques pour l'élargissement de la base génétique élite. Enfin, dans le **Chapitre 5**, nous proposons d'utiliser l'OCS basée sur l'UCPC afin d'identifier le croisement optimal entre ressources génétiques et lignées élites en fonction des caractéristiques d'originalité et de performance des ressources génétiques. Nous proposons d'améliorer les ressources génétiques (pre-breeding), puis de connecter les ressources génétiques améliorées au matériel élite (bridging) avant de les introduire dans la population en sélection. Par simulations, nous montrons l'intérêt de réaliser des introductions récurrentes de ressources génétiques préalablement améliorées afin de maximiser le gain génétique tout en maintenant la diversité constante dans la population élite. De même, nous montrons l'importance de la composition de la population utilisée pour calibrer le modèle de sélection génomique utilisé lors de l'introduction des ressources génétique dans la population élite. Nous préconisons de considérer une population de référence constituée de lignées élites et de la descendance de croisement entre lignées élites et lignées issues de ressources génétiques. Ce dernier chapitre fournit des recommandations quant à l'exploitation de la variabilité polygénique présente dans les ressources génétiques afin d'enrichir la base génétique d'une population élite.

L'ensemble de ces travaux ainsi que les récentes études cités au long de ce manuscrit ouvrent de nouvelles perspectives pour la gestion de la diversité génétique au sein de programmes de sélection compétitifs et durables.

General introduction

Crop adaptation to human needs, i.e. crop improvement, is as ancient as agriculture itself (app. 10,000 years ago, Doebley *et al.* 2006). Crop improvement, like natural evolution, occurs through the selection operating on the genetic variability of plant populations (Lush 1937; Simmonds 1962). Both, natural evolution and early agricultural practices have left their signatures and shaped the genetic diversity of modern crops. Human selection initially carried out by farmers has been recently, for main crops and industrialized countries, optimized and structured into variety improvement by breeders and production by farmers (e.g. first French "seed dealers" in the mid-17th century). In this context, different levels of diversity can be distinguished for each crop: (i) the overall crop diversity stored in *ex situ* collections, (ii) the diversity of modern crop breeding populations (i.e. intra-breeding program) and (iii) the diversity of cultivated varieties delivered by breeders to farmers (Figure 1). At farmers' level, the diversity of varieties and crops contribute to the agroecosystem resilience to biotic and abiotic perturbations (Vandermeer *et al.* 1998; Malézieux *et al.* 2009). Thus, the management of genetic diversity at each level is of critical importance in a context of climate change characterized by an increased frequency of unpredictable extreme temperatures, drought, pests and plant pathogen outbreaks (McCouch *et al.* 2013). In the following, this dissertation will focus mostly on the second level of diversity, i.e. diversity within breeding populations that determinates the diversity available to breeders to develop new varieties (Figure 1).

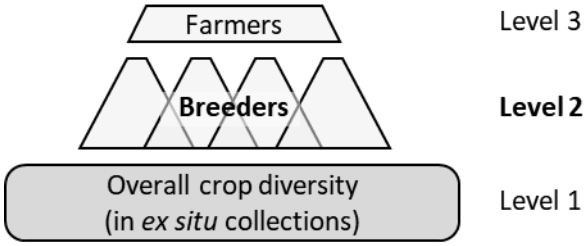


Figure 1 Diagram illustrating the three hierarchical levels of crop diversity for major crops in industrialized countries. This thesis focused on the diversity within breeding populations (**Level 2**).

Importance of genetic diversity for crop improvement

Domestication and improvement shaped crop genetic diversity

During their evolutionary history, crops have experienced different genetic bottlenecks through selection and drift during domestication and migrations (Spillane and Gepts 2001). Such events explain the reduction of current genetic diversity in main crops compared to the wild relatives and traditional varieties referred to as landraces (Ladizinsky 1985; Doebley *et al.* 2006). Artificial selection by farmers and modern plant breeders yielded major improvement in most crops to sustain humanity development but also reduced the genetic variability (Simmonds 1962; Cooper *et al.* 2001; Fu 2006, 2015). For instance, cultivated barley (Brown and Clegg 1983; Petersen *et al.* 1994), soybean (Doyle 1988; Hyten *et al.* 2006; Han *et al.* 2016), chickpea (Cooper *et al.* 2001), peanut (Fonceka *et al.* 2012) and wheat (Charmet 2011) show a narrow genetic base because of bottlenecks at domestication and migration. Other crops such as maize present a narrow genetic base arising from bottlenecks during modern breeding but contain a much larger available diversity in older germplasm (Tallury and Goodman 2001).

The loss of genetic variability in closed and finite selected populations is due to genetic drift induced by selection of a limited number of individuals. Also, directional selection for some agronomic traits (e.g. yield, quality, diseases tolerance) favors a favorable allele, respectively disfavors an unfavorable allele, at quantitative trait loci (QTLs) underlying the selected traits. As a result of selection, the allele frequency shifts in one direction yielding a reduction of deoxyribonucleic acid (DNA) sequence diversity at the QTLs and neighboring regions by linkage drag (Maynard-Smith and Haigh 1974). Alternatively, a balancing-stabilizing selection (e.g. selection for an optimal precocity), maintains multiple alleles in the breeding population and elevates sequence diversity at the QTLs and surrounding regions.
In practice, the impact of selection on crop genetic diversity at the farmers' level, also referred to as diversity erosion (Wouw *et al.* 2010), is difficult to observe. For instance, Fu (2006) reviewed 23 studies released from 2000 to 2005 evaluating the impact of modern plant improvement on genetic diversity of agricultural crops, such as maize (Duvick *et al.* 2004; Clerc *et al.* 2005; Reif *et al.* 2005b), wheat (Reif *et al.* 2005a; Roussel *et al.* 2005; Fu *et al.* 2006), barley (Koebner *et al.* 2003) and oat (Fu *et al.* 2004). This review revealed different impacts of modern crop improvement on elite germplasm. In general, the genomewide reduction of crop genetic diversity over time was minor, but allelic reduction at individual chromosomal segments was substantial. Only few studies focused on the impact of long-term selection on genetic diversity at the level of a given breeding program (e.g. in maize, Labate *et al.* 1999; Feng *et al.* 2006; Fischer *et al.* 2008; Van Inghelandt *et al.* 2010; Gerke *et al.* 2015, in soybean Bruce *et al.* 2019). The authors observed either significant reductions of genetic diversity or complex changes in genetic diversity due to large open breeding systems, i.e. with introductions of new extrinsic allelic variation (Feng *et al.* 2006; Bruce *et al.* 2019). Since every breeding population is subject to different breeding strategies, additional studies of the evolution of genetic diversity within commercial breeding programs and consequences on genetic improvement are required to drive an empirical consensus on good breeding practices.

Genetic diversity a cornerstone for crop improvement

The relationship between the additive genetic variation and the expected response to selection is known as the "breeder's equation" (Lush 1937). Assuming an infinite breeding population and a normally distributed targeted trait, the expected change in mean performance ($\Delta\mu$) per generation is proportional to the selection intensity (i), the selection accuracy (h) and the population additive genetic standard deviation of the targeted trait (σ_A):

$$\Delta\mu = ih\sigma_A, \text{ (Eq. 1)}$$

where the selection accuracy (h) is defined as the correlation between the value used for selection and the additive genetic value for the targeted trait. Equation 1 states that in absence of mutation and epistasis, the total response to selection is limited by the initial standing additive variation (σ^2_{A}, the variance of additive genetic values which corresponds to the sum of the additive diversity at causal loci and the additive covariances between causal loci, Bulmer 1971; Lynch and Walsh 1998; Gianola *et al.* 2009). Larger initial σ^2_{A} in the breeding population yields higher expected response to selection per generation.
Two parameters are commonly used to characterize the level of diversity in selected populations. The first one is the effective population size (Ne, Fischer 1930; Wright 1931), which refers to the number of breeding individuals in an idealized panmictic population with absence of selection that would show the same amount of genetic diversity as the population at hand. The second one is the expected

heterozygosity in the idealized population (He, Nei 1973). For biallelic loci, the expected heterozygosity in a panmictic population and no selection is $He = \frac{1}{m}\sum_{j=1}^{m} 2p_j(1-p_j)$, with p_j the frequency of the reference allele at locus $j \in [\![1, m]\!]$. The effective population size (Ne) can be estimated from changes in frequency of heterozygotes in the panmictic population assuming only drift: $He_{t+1} = He_t\ (1 - 1/2N_e)$ (Falconer and Mackay 1996). Thus, both expected heterozygosity (He) and effective population size (Ne) are related concepts.

In a long-term perspective, large and diverse populations show a greater efficiency of selection (Fischer 1930, p. 102; Weber and Diggins 1990). The effect of Ne on potential maximal response to selection is well known in quantitative genetic literature (Robertson 1960). Under the assumptions of an infinitesimal model (Fisher 1918), i.e. many locus of small effects underlying the trait, absence of mutation, a selection intensity i, an accuracy h, a population with effective size Ne and additive genetic standard deviation σ_A, the maximum potential response in long-term is:

$$2Ne\ ih\sigma_A. \text{ (Eq. 2)}$$

The maximum potential response to selection reduces to $2Ne\Delta\mu$ with $\Delta\mu$ being the expected response to selection in the first generation as defined in Eq. 1. Thus, a first advantage of a larger effective population size is to reduce the loss of initial genetic variance by genetic drift resulting in an increased selection limit. A second advantage is the greater accumulation of genetic variation by recombination events and mutations. Hill (1982a; b) derived that if new mutations are steadily accumulated and generate an additional variance σ_M^2 per generation, then $2Ne\ i\ \sigma_M^2/\sigma_P$ is the eventual additional response rate per generation, with σ_P being the phenotypic standard deviation. More recently, Barton extended this work including epistasis (2017).
While the expected response to selection is proportional to the selection intensity i (Eq.1, 2), the effective population size Ne is inversely proportional to the square of the selection intensity i^2 (Robertson 1961; Wray and Thompson 1990; Sanchez *et al.* 2006; Woolliams *et al.* 2015). Consequently, maximizing the selection intensity to maximize the short-term response to selection will inevitably reduce the effective population size and long-term response to selection (Eq. 2). This highlights the inherent dilemma between the genetic diversity and the genetic gain and opens the scope for optimization.

As expressed in Lush (1937) and Robertson (1960), a reduced genetic diversity in breeding populations might induce yield plateau or substantially increase breeding efforts and investments to keep constant rates of genetic gain. A reduced genetic diversity in breeding populations might also induce a reduced diversity in fields limiting the ability to overcome biotic and abiotic stresses, or even yielding crop failure in a changing environment (McCouch *et al.* 2013). One of the disastrous evidence is the Irish potato famine in the 1840s, caused by the homogenous sensitivity of cultivated varieties to late blight. More recently, the southern leaf blight epidemic in the US maize crop in 1969-1970 induced 15% losses caused by the use of the same cytoplasmic DNA male sterility in developed maize varieties which were uniformly susceptible to a race of the fungus (Ullstrup 1972; Bruns 2017). Consequently, there is a continuing need to balance improvement and diversity in crop breeding through an optimized management of intrinsic (i.e. internal to the breeding population) genetic variability and enrichment in new variability from different extrinsic (i.e. external to the breeding population) genetic resources

to increase breeding ceiling and reduce the genetic susceptibility to rising and yet unknown biotic and abiotic stresses.

Managing and broadening the genetic base of breeding programs

It is generally recognized in species suffering strong inbreeding depression and where the breeding population is also the production population (e.g. animal breeding) that one cannot simply select and mate the best individuals without also taking into account the degree of relatedness among them to limit consanguinity and the impact of deleterious alleles causing inbreeding depression. The identification of the mating plan that maximizes the genetic merit in the next generation while constraining the average relationship between parents involves the optimization of parental contributions, i.e. the fraction of genes contributed by a parent to the future generation, a concept well known in animal genetics (James and McBride 1958; Woolliams *et al.* 2015). Parental contributions have simple relationships with key parameters of population genetics. While the genetic gain is proportional to the product of individuals' contributions and deviations from population mean (Woolliams and Thompson 1994; Woolliams *et al.* 1999), the rate of inbreeding, i.e. loss of diversity, is inversely proportional to the square of individuals' contributions (Robertson 1961; Wray and Thompson 1990; Sanchez *et al.* 2006; Woolliams *et al.* 2015). Based on this theory, a mating strategy called optimal contribution selection has been investigated for decades in animal breeding (e.g. Wray and Goddard 1994; Meuwissen 1997; Kinghorn 2011), in tree breeding (e.g. Kerr *et al.* 1998; Hallander and Waldmann 2009a; b) and has been increasingly adopted in crop breeding (e.g. Akdemir and Sánchez 2016; De Beukelaer *et al.* 2017; Gorjanc *et al.* 2018; Akdemir *et al.* 2018).
There are several reasons that might explain why such considerations have been firstly developed in animal breeding and only recently adopted in crop breeding. One reason may be that major crops are inbred species (e.g. wheat, barley) and suffer little inbreeding depression or pass by a hybrid stage (e.g. maize) allowing to complement recessive sub lethal alleles. Complementarily, since most crop breeders have the possibility to broaden the genetic base of their population using different extrinsic genetic resources publically available (e.g. current and old varieties) and conserved worldwide in international gene banks and national collections (e.g. wild relatives, exotic germplasm accessions and landraces, Hammer *et al.* 2003; Commission on Genetic Resources for Food 2010), they might have underestimated the importance of intrinsic diversity management. The recent increased interest of crop breeders for intrinsic genetic diversity management might be explained by the fact that the more breeding germplasm is improved, the more expensive and time consuming becomes the introduction of extrinsic diversity.

Crop genetic resources are defined as "genetic material of actual or potential value" by the Convention on Biological Diversity (https://www.cbd.int/) and provide the basis to improve productivity, resilience and nutritional quality of crops (Wang *et al.* 2017). Although plant breeders recognize the importance of genetic resources for elite genetic base broadening, only little use has been made of it (Glaszmann *et al.* 2010; Wang *et al.* 2017). The main reason is that breeding progress continues to be made in most crops (e.g. in maize grain yield, Duvick 2005, in wheat, Tadesse *et al.* 2019) and that breeders are reluctant to compromise elite germplasm with unadapted and unimproved genetic resources (Kannenberg and Falk 1995). Consequently, there is a need for a breeding system that can efficiently broaden the genetic base of elite germplasm while not compromising the performance of released varieties. Such a system first involves the description and the understanding of the genetic diversity

present in collections and the definition of core sets of genetic resources representing the global diversity (Frankel 1984; Brown 1989). Genetic resources are characterized for adaptation traits in few locations (e.g. flowering day length, earliness, stress resistance …). Adapted genetic resources should be further extensively evaluated for agronomic traits (e.g. grain yield, quality …) and their genotype by environment interactions (GxE) before being identified as interesting for breeding purpose. The identification can be based on phenotypic evaluation of potential donors, progeny of the cross donor x elite material or considering molecular information (e.g. Bernardo 2014; Crossa *et al.* 2016; Yu *et al.* 2016). In the case of traits determined by few genes of large effect, the favorable alleles can be identified and introgressed into elite germplasm (Figure 2) following well established marker-assisted backcross procedures (e.g. Charmet *et al.* 1999; Servin *et al.* 2004; Bernardo 2016; Han *et al.* 2017). Introgressions have been successful for mono- or oligogenic traits (e.g. earliness loci in maize, Simmonds 1979; Smith and Beavis 1996, SUB1 gene in rice, Bailey-Serres *et al.* 2010). Introgressions also proved to be successful for more polygenic traits where few major causal regions have been identified. For instance, Ribaut and Ragot (2006) successfully introgressed five regions associated with maize flowering time and yield components under drought conditions. For complex traits controlled by numerous genes with small effect introgression procedures were mostly unsuccessful to broaden the genetic base of breeding populations (Simmonds 1993). Simmonds (1993) proposed a general scheme for genetic base broadening that consists in the incorporation of extrinsic polygenic variation in the breeding population. Simmonds distinguished three hierarchical steps starting from a broad population of genetic resources to the locally adapted breeding population. It starts with the pre-breeding, called base broadening in Simmonds (1993), to improve genetic resources in order to reduce the performance gap with the breeding population. Pre-breeding can be defined as the recurrent improvement of genetic resources to release donors that can be further introduced into the elite breeding population (Figure 2). For Simmonds, the pre-breeding must be kept completely independent of the breeding population until it starts to provide performing resources (Simmonds 1993). Best pre-breeding progeny are further considered for incorporation in a buffer population with some of the breeding material. This population bridges the elite breeding genetic base with the pre-breeding genetic base and this step is referred to as bridging (Figure 2). For the sake of clarity, bridging aims at limiting the negative impact of introductions on short-term varieties' performance. The best bridging individuals are further considered as breeding parents in the routine breeding program (Figure 2). Alternatively, one could suggest to skip the bridging if pre-breeding releases material that is directly competitive with elite parents (Figure 2).

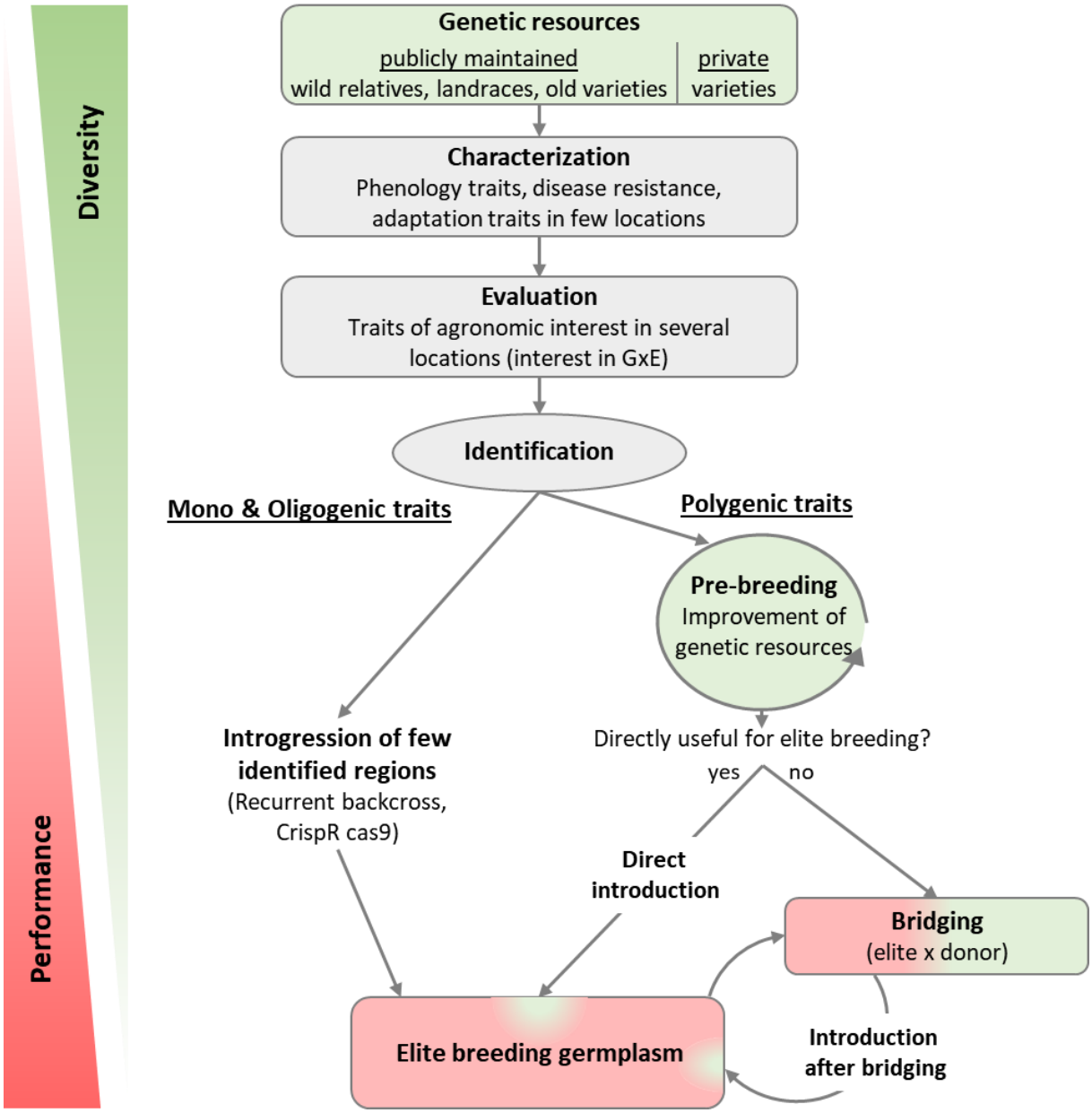


Figure 2 Diagram illustrating the difference between genetic base broadening, i.e. polygenic trait enrichment and introgression (adapted from Simmonds 1993).

A maize perspective

In this section, the maize history and modern hybrid breeding that shaped maize genetic diversity are presented. The interest of genetic base broadening in the maize context is further discussed and some maize genetic base broadening projects are shortly reviewed.

Maize domestication and adaptation shaped the maize genetic diversity

Maize production exceeded 1.3 billion tons on about 240 million ha worldwide in 2017, which makes maize the first crop before rice in terms of production (nearly a billion tons) (Food and Agriculture Organization, FAO 2019). Maize was domesticated once from its wild progenitor teosinte Zea *mays ssp. parviglumis* about 9,000 years ago in the Balsas valley of Mexico (Beadle 1939; Doebley 1990; Matsuoka *et al.* 2002). Maize domestication resulted in original maize landrace varieties further spread and adapted by Native Americans in a wide range of environmental conditions: as far as the current Canada and southern Chile (Figure 3). For instance, the American Northern Flint landraces were adapted to cold temperate regions (Brown and Anderson 1947) and are genetically divergent compared to other tropical or subtropical landraces (Doebley *et al.* 1986). About 200 years ago, Southern Dent and American Northern Flint germplasm were hybridized and gave rise to the Corn Belt Dent type adapted to the mid United States region (Doebley *et al.* 1988; Camus-Kulandaivelu *et al.* 2006). Due to day-length adaptation bottleneck, most of the tropical maize diversity is not represented in Corn Belt Dent (Goodman 1985).

The first introduction of tropical maize in south Europe is commonly attributed to Columbus in 1493 (Figure 3). European Northern Flint originated from the second introduction of pre-acclimated sources of maize from the eastern coast of North America in the north of Europe, currently Germany, Belgium and Netherlands, during the 16th century (Brandolini 1970; Rebourg *et al.* 2001, 2003; Dubreuil *et al.* 2006; Camus-Kulandaivelu *et al.* 2006). Further introductions may have occurred in Italy from South America (Brazil, Argentina) explaining the high similarity between traditional varieties of these regions (Tenaillon and Charcosset 2011). As a consequence of these introductions, European maize diversity derives from America and presents only few European specific alleles (Rebourg *et al.* 2003). Admixture events were also observed in Europe between different genetic backgrounds and led to the creation of new groups such as the broad European Flints group spanning from north to south Europe (Brandenburg *et al.* 2017).

As a result of domestication and adaptation to different growing conditions, maize exhibits a strong morphological variability among different origins. Maize is cultivated mainly for grain but also for silage in a wide range of environments, from temperate to tropical regions. As an allogamous species showing substantial inbreeding depression for grain yield (A. R. Hallauer and Miranda 1988, chapter 9), maize was historically, and is still in some regions (e.g. Bellon *et al.* 2003), cultivated in heterogeneous populations of heterozygous individuals called open-pollinated varieties (OPVs).

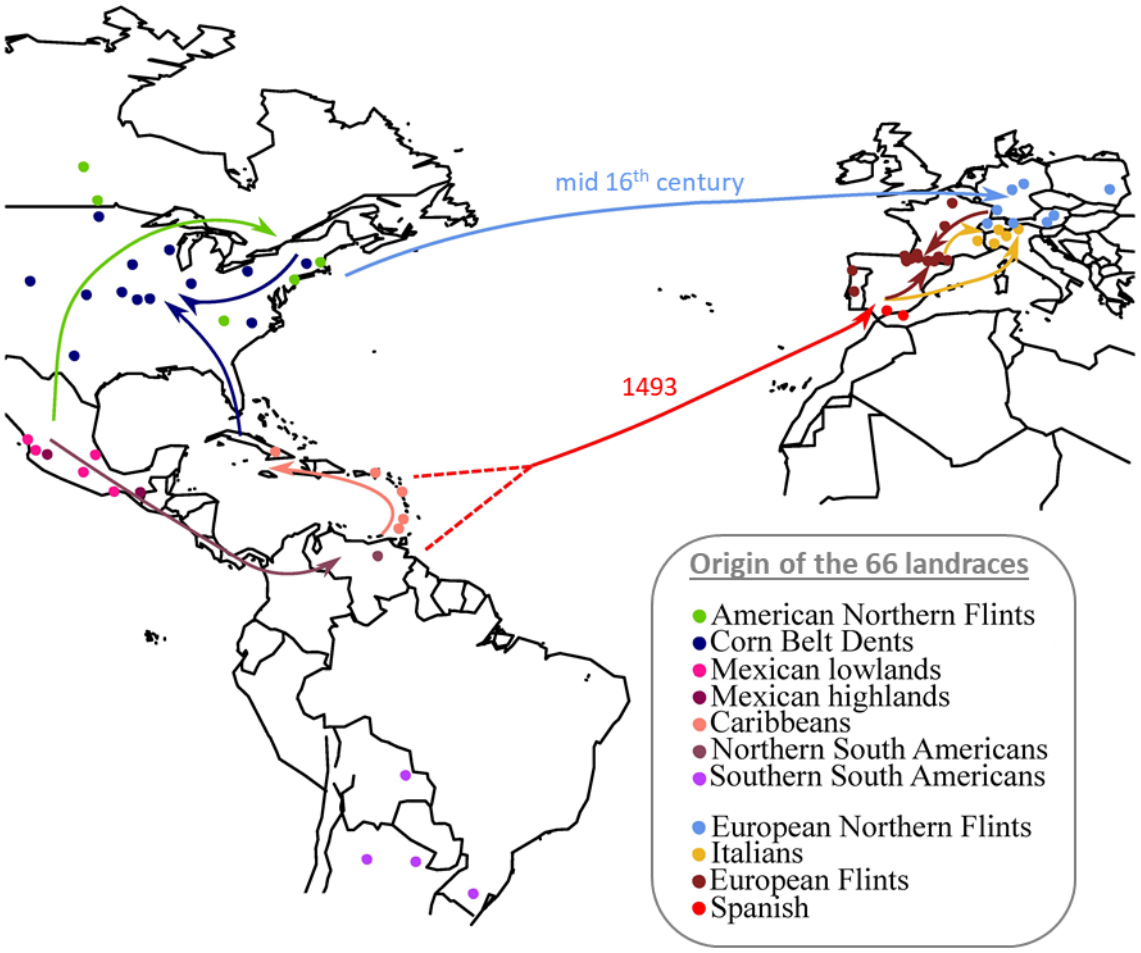


Figure 3 Maize genetic groups and diffusions pathways inferred from 66 maize landraces (adapted from Brandenburg *et al.* 2017). Points represent the 66 landraces origin and arrows the migration flux.

Modern maize breeding: hybrid breeding

Historically the OPVs of maize were the source of material used in temperate maize breeding programs. In the early 1900s, Shull (1908) proposed to "clone" the best heterozygote individual in the OPV as an hybrid between inbred parents (East 1908; Shull 1909). This revolutionized maize breeding and led to the rediscovery of the concept of hybrid vigor (Darwin 1876) further described as heterosis (Shull 1914). In the first generations, few OPVs served as source populations to derive inbred lines for use as hybrid parents. Due to strong inbreeding depression, the quantity of seeds produced by the first derived inbred lines was too small to directly used these lines as parents of commercial hybrids. And thus, first hybrids were double cross hybrids resulting from [Inbred1 x Inbred 2] x [Inbred 3 x Inbred 4] (Jones 1918). In the 1960s, with the improvement in seed quantity and quality traits, breeders switched from double cross hybrids to single cross hybrids resulting directly from Inbred1 x Inbred 2. It rapidly and completely replaced mass-selected OPVs in the United States and Europe (Anderson 1944; Troyer 1999). Hybrid breeding tremendously increased maize productivity (Figure 4). The inbred stage purges recessive deleterious alleles and increases the variance among families (Horner *et al.* 1969; Hallauer and Miranda 1988) and thereby increases the selection effectiveness.

Breeders defined and maintained distinct heterotic groups that maximized the inter-heterotic group hybrid vigor. Heterotic groups have been defined by testing different hybrid combinations. The hybrid breeding relies on the improvement of heterotic groups and the identification of the inbred parents from distinct heterotic groups that yield outstanding hybrids. Within heterotic groups, inbreds are improved in a reciprocal recurrent selection scheme (Russell and Eberhart 1975) designed to enhance the combining ability between the two heterotic groups, so that their cross will improve in performance over selection cycles. The hybrid performance is modeled as the sum of the general combining ability (GCA) of inbred parent from heterotic group 1 and of inbred parent from heterotic group 2 and specific combining ability (SCA) that is the effect specific to the hybrid combination. In a classical hybrid breeding scheme (Figure 5), within and between heterotic group breeding are distinct steps. Within heterotic groups, inbred segregating progeny of parental crosses are selected based on their GCA estimated from their evaluation in hybrid combination with one or few different inbreds representative of the opposite group (app. 1 to 3) called testers. Such evaluation is referred to as testcross evaluation. The best performing inbreds (app. 5%-10% best) are recycled as parents of next generation crosses. Additionally, these inbred lines are further evaluated for testcross performance on more testers and are further selected. In the second step, the best inbreds of both pools are crossed in an incomplete factorial to evaluate SCA and produce desirable commercial hybrids (Bernardo 1994; Technow *et al.* 2012, 2014) (Figure 5). Given that testcross means, i.e. GCAs, behave in a statistically additive manner (Hallauer and Miranda 1988), statistical dominance (SCA) is accounted for only in the incomplete factorial between both populations for commercial hybrid selection.

In the US by the 1960s, production of high-yielding hybrids in temperate conditions was largely based on inbreds from two Corn Belt Dent OPVs: the Reid Yellow Dents and Lancaster sure crops (Smith 1988). While the founders of these heterotic groups were not initially differentiated, the heterotic groups diverged genetically over time to become highly structured and isolated with a decreased diversity within groups (Heerwaarden *et al.* 2012). Today's North American Dent maize is composed of multiple heterotic groups and their nomenclature is complex and depends on the authors (Mikel and Dudley 2006). As a rule of thumb, the female, i.e. seed parent, is mainly from Iowa Stiff Stalk Synthetics origin (ISSS, that includes lines that were widely used in breeding e.g. B73, B14, B37, A632) which is predominantly derived from Reid Yellow Dents and the male, i.e. pollen parent, is mostly from the Lancaster sure crops origin (e.g. Oh43, Mo17, C103). More recently, the Iodent (e.g. PH207) used as male parent added early flowering time and cool conditions adaptation and contributed to spread maize cultivation further north (Goodman 1990). In Europe, hybrids between Corn Belt Dent and European Flint inbreds proved to combine productivity and environmental adaptation for maize cultivation in Northern Europe from West to East. Subsequent reciprocal selection of the two groups increased their differentiation and complementarity (Rincent *et al.* 2014). In southern Europe (e.g. Spain, Italy, Turkey), similar heterotic groups as in the Corn Belt are considered, resulting in ISSS Dent x non ISSS Dent hybrids.

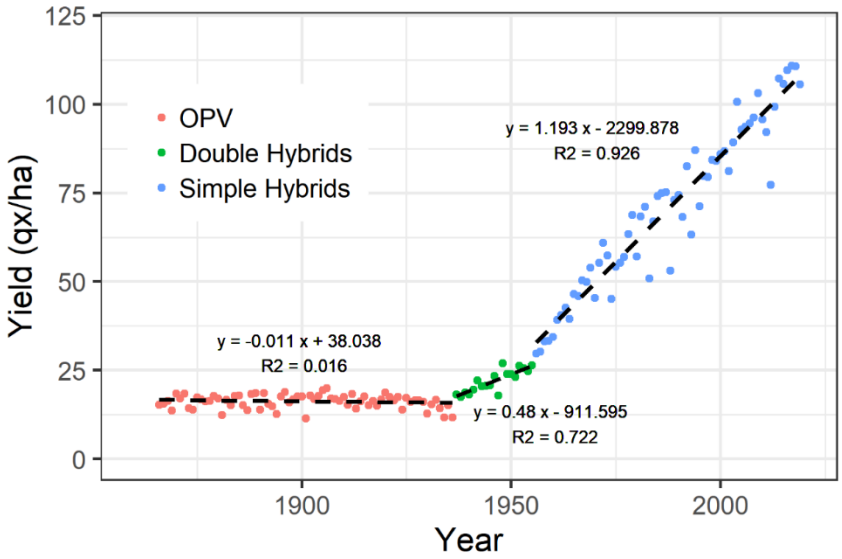


Figure 4 Maize grain yield (qx/ha) evolution in USA from 1866 to 2019 based on USDA data all states confounded (https://quickstats.nass.usda.gov/). Three linear regressions are provided for each three main eras: OPVs (1866 to 1936), Double hybrids (1937 to 1955) and Simple hybrids (1956 to 2019). Grain yield was converted from bushel/acre to qx/ha using the correspondence 1 corn bushel = 0.254 qx and 1 acre = 0.405 hectare. *(Adapted from Beckett (2017))*

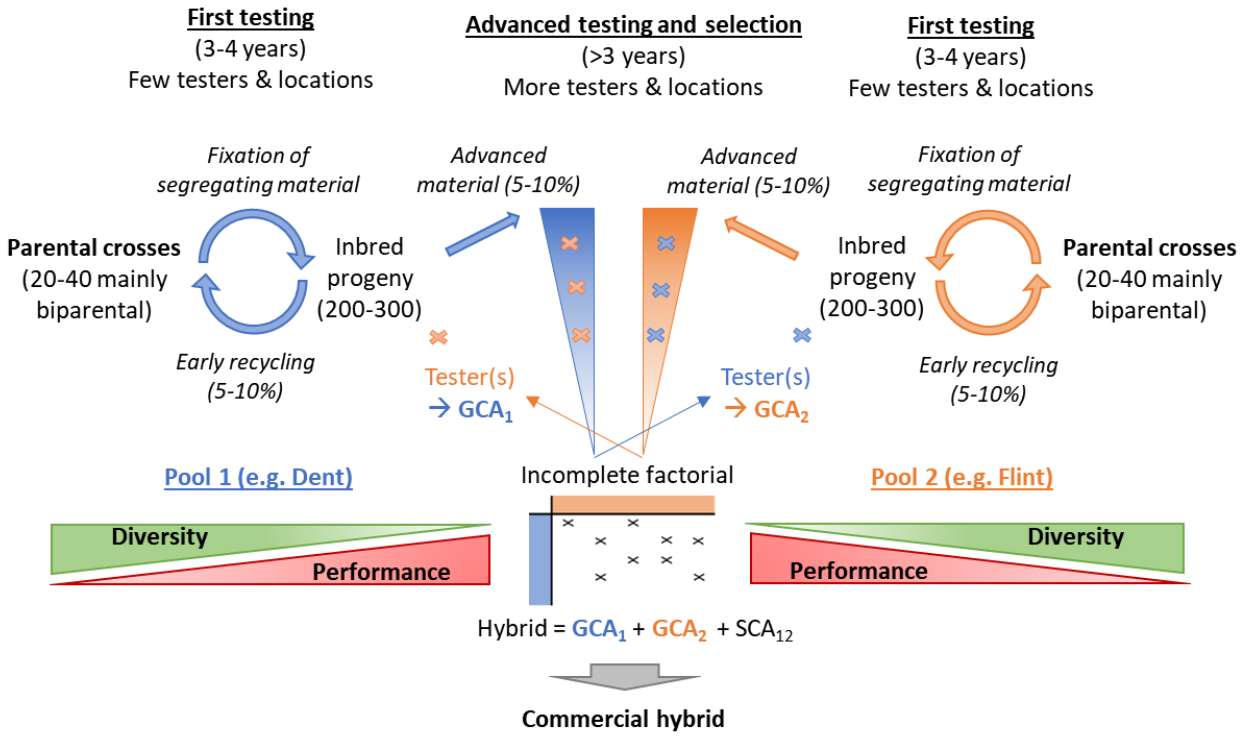


Figure 5 Schematic view of hybrid maize breeding. Plain crosses (in blue and orange) represent testcross evaluations with tester(s) of the opposite pool. Cycle length of each step in years and sample sizes are given for an averaged mid-size breeding program and do not reflect the variability between breeding programs and breeding strategies.

Broadening maize hybrid breeding programs

Despite hybrid breeding improved tremendously grain yield and quality as well as resistance to biotic and abiotic stresses at world scale (e.g. Figure 4, Duvick 2001, 2005), it also reduced elite genetic diversity. It has been observed that during the transition from landraces to hybrids, many favorable alleles have probably been lost because of their association with unfavorable alleles and/or genetic drift (Ho *et al.* 2005; Reif *et al.* 2005b; Buckler *et al.* 2006; Yamasaki *et al.* 2007). For instance, Ho *et al.* (2005) estimated that only 56% of the alleles found in the Corn Belt Dent landraces were present in a diverse set of inbred lines. US and more generally worldwide hybrid breeding is relying on the use of a very narrow elite germplasm (Goodman 1990). For instance, in the US, about three Lancaster type inbred lines (Oh43, Mo17, C103) and three ISSS type inbred lines (B73, B37, A632) and their close relatives were represented in a very high percentage (70% or more) of all U.S. hybrids (Goodman 1990). More recently, the Iodents (mainly derived from Pioneer PH207 and Dekalb/Monsanto 3IIH6 lines) took an important place in temperate non ISSS dent proprietary pedigrees (Mikel 2018). A recent high-density haplotypic analysis revealed significant haplotype sharing between maize inbred lines registered from 1976 to 1992 and key maize founders B73, Mo17 and PH207 (Coffman *et al.* 2020). Since maize hybrid breeding developed along with intellectual property rights, it also limited germplasm exchange between private programs (Goodman 1999).

Different sources of diversity can be considered to broaden the genetic base of maize breeding programs. Brown (1979) estimated that there might be 150-180 distinct “races” of maize worldwide. On a racial basis, it was indicated by Brown (1979) and Goodman (1985) that only 2% of the available germplasm was considered in temperate maize breeding and only 5% worldwide (Tallury and Goodman 2001), when excluding subsistence farming. Goodmann (1999) observed that only about 0.3% of Tropical exotic germplasm was used in US hybrid breeding in 1996. Local or exotic landraces which did not contribute to the founding material of commercial programs provide a source to broaden the genetic base of commercial breeding programs. Landraces have also been well characterized relative to elite germplasm in Europe (e.g. Dubreuil and Charcosset 1999; Rebourg *et al.* 2001; Reif *et al.* 2005b; Dubreuil *et al.* 2006; Frascaroli *et al.* 2013; Strigens *et al.* 2013) and America (e.g. Heerwaarden *et al.* 2011; Hellin *et al.* 2014). The use of reproducible libraries of doubled haploid (DH) lines from landraces has been suggested to ease genotyping, phenotyping and evaluation of the variation within landraces (Strigens *et al.* 2013; Melchinger *et al.* 2017; Böhm *et al.* 2017; Brauner *et al.* 2019; Hölker *et al.* 2019). Since maize hybrid industry is highly competitive, commercial breeders do not spend time and resources for evaluation, adaptation and improvement of non-improved landraces. Instead, commercial breeders will prefer to consider inbred lines from other than their own program (Kannenberg 2001). This includes breeding program targeting different environments and competitors’ inbreds obtained by selfing or reverse breeding from hybrids (Smith *et al.* 2008) or running out of the plant variety protection act after 20 years in the US (ex-PVPA, Mikel and Dudley 2006). Hundreds of ex-PVPA are publically released every year, which make an improved source of variation available. To broaden the genetic base of European germplasm with US inbreds is appealing. For instance, Reif *et al.* (2010) evaluated the interest to introgress US public inbreds into German European inbreds and recommended to introgress ISSS inbreds into European dents and non ISSS inbreds into European Flints.

To harness genetic variability and potential of adaptation in genetic resources, public-private collaborations that share costs between public institutes and private companies are of great interest.

In the following, some public-private maize genetic base broadening projects are listed with a focus on their contribution to the private breeding sector. Cramer and Kannenberg (1992) proposed the hierarchical open-ended population enrichment (HOPE) breeding system to release enriched maize inbreds further considered to broaden the genetic base of Canadian commercial maize breeding programs. In its last version, the HOPE system was composed of three hierarchical open-ended gene pools, i.e. the best genotypes of a basal pool were further used as parents in the superior pool, permitting the transfer of favorable alleles from genetic resources to the elite pools (Popi 1997; Kannenberg 2001). The genetic resources were introduced in the basal pool without heterotic group distinction until the introduction in the two elite pools (Popi 1997). After 20 years, only four inbreds have been released to the industry with no success story up to date. The Latin American maize project (LAMP, Pollak 1990; Salhuana *et al.* 1997; Salhuana and Pollak 2006) provided maize breeders with useful characterization and evaluation of US and Latin American tropical germplasm accessions. The germplasm enhancement of maize project (GEM, Pollak and Salhuana 2001) was a public-private collaborative effort to enhance the accessions identified as useful by LAMP with proprietary lines furnished by private partners (Pollak 2003). In practice, LAMP lines were first crossed with an elite inbred from a private partner and further crossed to a second private partner's elite line to derive a bridging germplasm carrying on average 25% of LAMP parent genome. In 2014, more than 270 temperate adapted inbreds were developed from more than 30 different exotics germplasm. Similarly, the seeds of discovery project initiated by the International Maize and Wheat Improvement Center (SeeD, Gorjanc *et al.* 2016) aims to harness favorable variation from more than four thousand landraces and to develop a bridging germplasm with on average 25% of landrace genome that would be useable for genetic base broadening in commercial maize programs. In France, the INRA/Promaïs (Gallais *et al.* 2001) project and continuation, are also examples of the interest for public-private partnership genetic base broadening projects.

Genomic selection revolutionized breeding

Marker assisted selection to genomic selection

Molecular markers refer to DNA fragments that exhibit polymorphism between individuals and that can be easily typed and used as genetic markers. In maize, different genetic markers and density have succeeded: from few multi-allelic markers such as restriction length polymorphism (RFLP), single sequence repeats (SSR) to today's commonly used single nucleotide polymorphism (SNP) that can be typed on predefined bead chips with 50k SNPs (Ganal *et al.* 2011) or 600 SNPs (Unterseer *et al.* 2014) and by sequencing (GBS, Elshire *et al.* 2011). These markers can be used on a large number of individuals to evaluate, structure and sample genetic diversity within an between ex-situ collections (Glaszmann *et al.* 2010; Mascher *et al.* 2019). These markers can also be used to monitor the genetic diversity of breeding germplasm and assist selection. The use of markers linked to QTLs, further referred to as marker assisted selection (MAS), opened new perspectives for breeding. In the 1960's, Neimann-Sorensen and Robertson (1961) considered blood groups as markers supporting selection in animals. Lande and Thompson (1990) proposed to estimate the genetic value of selection candidate by summing the estimated effects of genetic markers significantly associated with QTLs. More recently, the development of cheap high-throughput SNP genotyping and statistical developments enabled to consider a large number of genomewide markers for prediction (Whittaker *et al.* 2000; Meuwissen *et al.* 2001). This is referred to as genomic selection (GS) and this approach has been implemented in many animal and plant species over the last decades.

Genomic selection

In GS, a sample of individuals (training set, TS) is genotyped and phenotyped for a trait, before being used to train a statistical model. The statistical model is further used to predict the genetic value of genotyped individuals. Several models have been proposed (e.g. Heslot *et al.* 2012) but the most common and robust is the genomic best linear unbiased prediction model (G-BLUP) that relies on the infinitesimal model (Fisher 1918). G-BLUP considers the genomic relationship matrix between individuals to model the covariance of their genetic values (VanRaden 2008). Note that before GS, prediction of individual breeding values using BLUP with pedigree information to model genetic covariance between individuals was common in animals (Henderson 1975) and investigated in maize (Bernardo 1996a; b). A standard G-BLUP model can be written as:

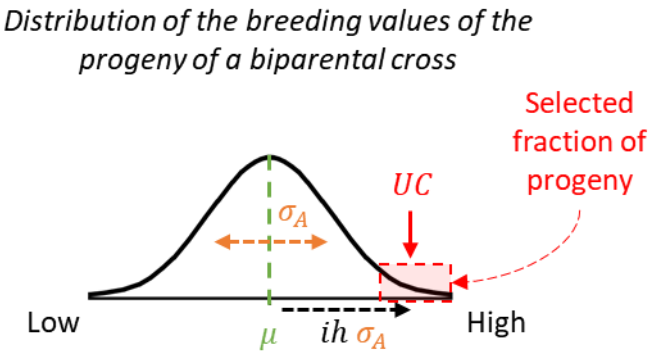
$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}, \text{ (Eq. 3)}$$

where $\boldsymbol{y}$ is the column vector of genotypes, $\boldsymbol{X}$ is the incidence matrix of fixed effects with the respective column vector effect $\boldsymbol{\beta}$ (e.g. location effect), $\boldsymbol{Z}$ is the incidence matrix of random effects, i.e. linking genotypes to genetic values, $\boldsymbol{u}$ is the column vector of genetic values with $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G}\sigma_G^2)$ and $\boldsymbol{G}$ is the genomic relationship matrix that models the covariance between individuals at markers, σ_G^2 is the genetic variance. The column vector of errors $\boldsymbol{e}$ is modeled as $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_E^2)$ with $\boldsymbol{I}$ the identity matrix. After estimation of variance components $\hat{\sigma}_G^2$ and $\hat{\sigma}_E^2$, the best linear unbiased predictor of the genetic value $\hat{u}_i$ of a given genotyped individual i is predicted using mixed model equations (Henderson 1975). Estimated marker effects can be derived out of G-BLUP model by back-solving (Wang *et al.* 2012) thanks to the equivalence with the ridge regression best linear unbiased prediction model (RR-BULP) that considers directly the matrix of allelic doses and assumes that all marker effects are drawn from the same normal distribution.

The interest of GS is commonly attributed (i) to the acceleration of selection progress by shortening generation intervals and (ii) to higher selection accuracy especially for traits difficult or costly to measure (Hayes *et al.* 2009). Different usages and implications of GS have been suggested in plant breeding (Heslot *et al.* 2015). For instance, instead of selecting progeny of parental crosses based on expensive phenotypes in multi-location replicated trials, marker information and GS models can be used to increase selection accuracy and optimize the phenotyping efforts (e.g. no more replicates or unbalanced designs). As a step further, GS can be used to predict progeny genetic values without phenotyping, which yield a gain of 3 to 5 years but also raises questions about the updating of the GS model with new phenotypes (Pszczola *et al.* 2012; Rincent *et al.* 2012; Isidro-Sanchez *et al.* 2015; Neyhart *et al.* 2017; Eynard *et al.* 2018). GS in plant breeding and particularly in maize breeding enables to generate larger biparental families and thus increases within family selection intensity. Among other applications, GS can be used to predict the interest of parental crosses based on different criteria, such as the usefulness criterion of a cross (UC, Schnell and Utz 1975) that represents the expected genetic value of the selected fraction of the progeny of the cross (Figure 6):

$$UC = \mu + ih\sigma_A, \text{ (Eq. 4)}$$

where μ is the mean genetic value of the progeny of the cross, i and h are the within family selection intensity and accuracy, respectively and σ_A^2 the within family additive genetic variance that can be



predicted for biparental crosses using information of recombination frequency and linkage disequilibrium between loci (Lehermeier *et al.* 2017b).

Figure 6 Illustration of the Eq. 4 in case of a biparental cross P1 x P2.

Genomic selection in the light of diversity management

As GS enables to shorten selection cycles and/or increase selection accuracy compared to phenotypic selection, it is expected to accelerate the loss of genetic diversity per unit of time due to rapid fixation of large effect regions. Jannink (2010) and Lin *et al.* (2016) observed by simulations that GS leaded higher loss of diversity than phenotypic selection. Experimentally, Jacobson *et al.* (2015) observed only a limited loss of genetic diversity due to genomic selection within biparental populations after one generation. However, the effect on long-term recurrent selection through both within family selection and parental cross selection is still unclear. In long-term simulations of wheat breeding, Rutkoski *et al.* (2015) observed that GS increased the loss of diversity compared to phenotypic selection. GS also tends to shrink toward the population mean the predicted genetic values of individuals with less phenotypic observations and/or less phenotypic observations on relatives in the TS and of individuals genetically distant to the TS (Habier *et al.* 2010; Pszczola *et al.* 2012). The shrinkage results in lower coefficients of determination (CD, Laloë 1993) associated with the predicted values. As a consequence, individuals with low relationship relative to the elite majority of the TS are likely predicted to be average with a small chance to be selected. Similarly, in the RR-BLUP formulation, the rare favorable allele effects are shrunk toward zero, which increases the risk of losing rare favorable alleles and consequently reduces the long-term genetic gain (Goddard 2009; Jannink 2010). Several authors suggested to up-weight rare favorable alleles to correct for shrinkage in GS model with encouraging results obtained by simulations (e.g. Goddard 2009; Jannink 2010; Sun and VanRaden 2014; Liu *et al.* 2015). However, such approaches suffer the difficulty to define appropriate up-weighting factors.

While GS raises concerns about its effect on genetic diversity erosion, it also opens new ways for intrinsic genetic diversity management and genetic base broadening. Firstly, GS models enable to estimate genomic variance components giving access to the causal diversity and the impact of linkage disequilibrium (LD) on additive genetic variance (Sorensen *et al.* 2001; Lehermeier *et al.* 2017a). Despite such decomposition can provide breeders with substantial information on the potential response to selection of a breeding population, to our knowledge, it has never been implemented in this context. Secondly, GS models might be implemented in the optimal contribution selection initially considering the pedigree information to predict the next generation merit (pedigree BLUP model) and

to constrain the pedigree relatedness among parents. Clark *et al.* (2013) observed that using genomic information for merit prediction and relatedness estimation increased optimal contribution selection performance. The optimal cross selection (OCS), an extension of the optimal contribution selection to deliver a crossing plan, has been recently adopted in plant breeding (e.g. Akdemir and Isidro-Sánchez 2016; Gorjanc *et al.* 2018; Akdemir *et al.* 2019). In previous works, OCS has been defined to balance the genetic merit and diversity in the progeny. However, as stated above, in GS plant breeding one typically has large biparental families with high within family selection intensity. Therefore, it would be likely more interesting to consider OCS that balances the genetic merit and diversity expected in the best performing fraction of each family. To our knowledge this has not yet been considered. Finally, GS models might help to characterize and identify interesting genetic resources in gene banks as suggested in Crossa *et al*. (2016) and Yu *et al.* (2016). More recently, Brauner *et al.* (2018, 2019) evaluated the predictive ability of GS models within DH lines derived from maize landraces. GS is also offering the possibility to fasten a long and expensive pre-breeding approach to harness polygenic variation in genetic resources and make it more attractive for commercial breeders (Longin and Reif 2014; Gorjanc *et al.* 2016). However, to our knowledge no simulation studies demonstrated the interest of genomic selection recurrent genetic base broadening considering pre-breeding, bridging and introductions as illustrated in Figure 2.

Objectives of this thesis

The sustainable management of genetic diversity in breeding programs is receiving increasing attention in the company RAGT2n and competitors (*personal communications*) for maize and other crops. This thesis has been articulated around five main objectives addressed in chronological order and corresponding each to a chapter of this dissertation.

1. Considering a given breeding program, how did the genetic diversity in a specific population evolve genomewide and in different genomic regions? How to release genetic variation in low diversity genomic regions?
   In **chapter 1**, we reviewed and suggested three sets of indicators based on temporal phenotypic and genotypic data to assess the past efficiency of breeding population improvement and its sustainability. We further applied the indicators on an early European grain maize program recorded from 2003 to 2016.
2. Assuming the genetic diversity is limiting, many genetic resources are accessible to breeders but cannot all be considered to broaden the elite genetic diversity. How can we identify appropriate donors for genetic base broadening of an elite population?
   In **chapter 2**, we reviewed and proposed different criteria based on estimated marker effects from GS models to select donor(s) in order to enrich elite recipient(s). To compare the different criteria, marker effects were estimated on the Amaizing Dent collaborative panel composed of 338 public Dent lines of different origins and 48 proprietary lines provided by seven companies including RAGT2n (Rio *et al.* 2019). Ten elite recipients from RAGT2n material were considered in this case study.
3. After identifying donors of diversity, how do breeders optimally cross them to elite recipients in order to maximize the expected performance and donor's polygenic contribution to progeny? Depending on the genetic and phenotypic distance of donor relative to elites is it

preferable to use biparental crosses between donor and recipient or more complex multi-parental crosses?

In **chapter 3**, we extended algebraic formulas in Lehermeier *et al.* (2017b) to predict the usefulness criterion of multi-parental crosses. We also propose to consider the parental contributions, i.e. percentage of genome in progeny inherited from a parent, as a polygenic trait in a multivariate usefulness criterion context. We validated our method by simulations.

4. Although breeders have the possibility to broaden their genetic diversity by integrating other germplasm, it requires investments and delays the genetic progress. For these reasons, an optimal management of intrinsic genetic diversity to be competitive at short-term while maintaining a long-term potential genetic gain is challenging.

   Considering a closed breeding population showing substantial genetic diversity, in **chapter 4**, we adapted the approach developed in chapter 3 for optimal cross selection (OCS) to account for the effect of within family selection on the performance and on the diversity in the next generation. We simulated 60 years of breeding and compared our strategy to OCS not accounting for within family selection.
5. Finally, in **chapter 5** we evaluated the interest of the approach developed in chapter 4 in the context of an open breeding population regularly enriched in extrinsic variability from different sources of diversity. We simulated 60 years of breeding and evaluated the interest of recurrent introductions after bridging depending on the type of donor considered. We also investigated the effect of TS diversity and composition on within family prediction accuracies and the efficiency of genetic base broadening.

The following **chapters 1**, **2**, **3** and **4** have been published in peer-reviewed journals and the edited version is provided in this manuscript. **Chapter 5** is a draft article that has not been peer-reviewed. All chapters are discussed and put into perspectives in the last section.

Cited literature

Please refer to cited literature at the end of the general discussion and perspectives section (**p. 125**).

Chapter 1 Assessment of breeding programs sustainability: application of phenotypic and genomic indicators to a North European grain maize program

This chapter has been published in the peer-reviewed journal Theoretical and Applied Genetics in 2019. The electronic version of this article is available on the publisher website:

https://link.springer.com/article/10.1007/s00122-019-03280-w

Theoretical and Applied Genetics (2019) 132:1321–1334
https://doi.org/10.1007/s00122-019-03280-w

ORIGINAL ARTICLE

Assessment of breeding programs sustainability: application of phenotypic and genomic indicators to a North European grain maize program

Antoine Allier[1,2] · Simon Teyssèdre[2] · Christina Lehermeier[2] · Bruno Claustres[2] · Stéphane Maltese[2] · Stéphane Melkior[2] · Laurence Moreau[1] · Alain Charcosset[1]

Received: 25 September 2018 / Accepted: 7 January 2019 / Published online: 21 January 2019

Abstract

Key message **We review and propose easily implemented and affordable indicators to assess the genetic diversity and the potential of a breeding population and propose solutions for its long-term management.**

Abstract Successful plant breeding programs rely on balanced efforts between short-term goals to develop competitive cultivars and long-term goals to improve and maintain diversity in the genetic pool. Indicators of the sustainability of response to selection in breeding pools are of key importance in this context. We reviewed and proposed sets of indicators based on temporal phenotypic and genotypic data and applied them on an early maize grain program implying two breeding pools (Dent and Flint) selected in a reciprocal manner. Both breeding populations showed a significant positive genetic gain summing up to 1.43 qx/ha/year but contrasted evolutions of genetic variance. Advances in high-throughput genotyping permitted the identification of regions of low diversity, mainly localized in pericentromeric regions. Observed changes in genetic diversity were multiple, reflecting a complex breeding system. We estimated the impact of linkage disequilibrium (LD) and of allelic diversity on the additive genetic variance at a genome-wide and chromosome-wide scale. Consistently with theoretical expectation under directional selection, we found a negative contribution of LD to genetic variance, which was unevenly distributed between chromosomes. This suggests different chromosome selection histories and underlines the interest to recombine specific chromosome regions. All three sets of indicators valorize in house data and are easy to implement in the era of genomic selection in every breeding program.

Introduction

Successful plant breeding implies meeting short-term goals to develop competitive cultivars, while maintaining diversity in the genetic pool to meet long-term goals. Response to selection per breeding cycle is determined by the selection intensity, the selection accuracy and the additive genetic variance of the trait (Lush 1937). Hence, a characterization of the additive genetic variance and its components in breeding populations is needed. In quantitative genetics theory, selection is expected to modify the additive genetic variance by changing allele frequency at quantitative trait loci (QTLs) and by modifying covariances between QTLs (linkage disequilibrium, LD) (Bulmer 1971). Directional selection increases the frequency of favorable alleles at selected QTLs generating a "hitch-hiking" effect at linked loci, leading ultimately to a local reduction of genetic variation (Smith and Haigh 1974). In addition, selection in a population of limited effective size leads to genetic drift, that is also expected to reduce the level of additive genetic variation (Wright 1931; Falconer and Mackay 1996). A recent simulation study (Gerke et al. 2015) suggested that genetic drift is the main force affecting evolution of genetic variation. Furthermore, directional selection induces linkage

Communicated by Benjamin Stich.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00122-019-03280-w) contains supplementary material, which is available to authorized users.

✉ Alain Charcosset
alain.charcosset@inra.fr

[1] GQE - Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France

[2] RAGT2n, Genetics and Analytics Unit, 12510 Druelle, France

disequilibrium that diminishes the additive genetic variance in the short term (i.e., repulsion between QTLs). This has been well described in quantitative genetic literature as the Bulmer effect (Bulmer 1971, 1980; Lynch and Walsh 1999).

In practice, most breeding schemes can be viewed as recurrent selection (RS) schemes in the sense that selected individuals at a given generation are intercrossed to produce the next generations. RS programs aim at increasing favorable allele frequency while maintaining genetic variability along generations to ensure long-term genetic gains (Hallauer and Darrah 1985). In species showing strong heterotic effects (e.g., maize), hybrid breeding raises additional questions for diversity management. In this case, the objective is to develop inbred lines that have the ability to generate good hybrids when making appropriate crosses. As a consequence, a recurrent reciprocal selection (RRS) firstly proposed by Comstock et al. (1949) is implemented. Its objective is to select in parallel within two pools (namely, heterotic groups) designed to yield cross-group highly performant hybrids. In this context, a trade-off exists between the complementarity among heterotic groups and the diversity within groups (Duvick et al. 2004).

In public and private breeding programs, phenotypic and genotypic data are stored over time and available for several cycles of selection. These data can be used to get a proper understanding of phenotypic and genomic changes associated with past selection, which can be useful to best orientate next breeding cycles. Phenotypic data over time allow to estimate jointly the evolution of the genetic performance of breeding pools. The evolution of genetic performance for a given trait over several selection cycles is called the genetic trend. In case of linear genetic trend, the genetic gain is estimated as a linear regression of the mean genetic performance on year (Eberhart 1964; Rutkoski 2018). Genetic trends in grain yield among commercial maize hybrids have been estimated in different environments and periods (Duvick 1984; Russell 1991; Duvick et al. 2004; Fischer et al. 2008). Nonetheless, few studies aimed to estimate the realized genetic gain in complementary breeding pools, which is of interest for breeders to better allocate resources among pools and monitor their effectiveness (Rutkoski 2018). In addition to the estimation of genetic gain, phenotypic data over time allow estimating the evolution of the additive genetic variance at each breeding cycle. Several studies have estimated the evolution of the additive genetic variance in breeding programs (Fischer et al. 2008) or selection experiments (Betrán and Hallauer 1996; Falke et al. 2007a) but none in private breeding programs.

Genetic markers based on deoxyribonucleic acid (DNA) polymorphism (e.g., restriction fragment length polymorphism, RFLP; single sequence repeat, SSR or single nucleotide polymorphism, SNP) have become more and more accessible in most crops. When available along generations, genotypic data make it possible to follow the evolution of the genetic diversity and the breeding pools structuring along breeding cycles. Labate et al. (1999) followed the genetic diversity for 13 generations of breeding in the Iowa's recurrent reciprocal (RR) maize breeding program using RFLP markers. Genetic diversity evolution in private maize breeding programs has been analyzed using low-density SSR markers (Feng et al. 2006). More recently, using high-throughput genotyping information, i.e., SNP markers, Gerke et al. (2015) observed a significant loss of genetic diversity along with a differentiation of genetic pools in Iowa's RRS program.

Using both genotypic and phenotypic information enables one to jointly estimate components of genetic variance, accounting or not for covariance between QTLs (LD) (Lehermeier et al. 2017). The effect of LD on additive genetic variance has been mostly investigated using computer simulations (Hospital and Chevalet 1996). In Falke et al. (2007b), authors followed for several selection cycles in biparental populations the genetic variance and the LD between favorable alleles at detected QTLs using SSR markers (Falke et al. 2007a). High-throughput genotyping information with genomic selection models makes it possible to extend this approach at the genome-wide level and to have a better estimation of LD effect on additive genetic variance. A Markov chain Monte Carlo (MCMC) approach coupled with a genome-wide regression has been proposed to decompose the additive genetic variation into (1) a component assuming linkage equilibrium (LE), called the additive genic variance, and (2) a deviation term accounting for covariance between QTLs (LD) (Sorensen et al. 2001; de los Campos et al. 2015; Lehermeier et al. 2017). So far it has not been applied on a private breeding program for diversity management purposes.

We investigated a North European early grain maize hybrid breeding program as a case study to assess the use of three sets of indicators to understand past selection events and derive recommendations to manage breeding pools regarding selection potential and genetic diversity. Firstly, we applied two classical indicators based on phenotypic and genotypic data, further referred as *phenotypic indicators* and *genotypic indicators*, respectively, which inform about past efficiency and future short-term tendencies. In *phenotypic indicators*, we jointly analyzed the realized genetic gain and the additive genetic variance at each generation. In *genotypic indicators*, we followed the genome-wide and local genetic diversity over time in complementary breeding pools. Finally, in *genomic indicators*, we estimated the contribution of each chromosome to the genetic variance and the portion of it that is masked by selection induced LD between QTLs. Based on results of *genomic indicators*, we suggested management strategies to improve future response to selection. We start with a theoretical background section that reviews the expected links between the different indicators, including

Theoretical and Applied Genetics (2019) 132:1321–1334

considerations on Bulmer's effect and distinction between genetic and genic variance.

Theoretical background for proposed indicators

Let us consider a classical linear model where phenotypes $y_i, i \in [1,n]$ are determined by the genotype at Q quantitative trait loci (QTL) $x_i = \{x_{ij}\}, j \in [1,Q]$ with $x_{ij} \in \{0;1;2\}$ being the reference allele content for individual i at QTL j. The n-dimensional vector of genotypes at QTL j, $x_j = \{x_{ij}\}, i \in [1,n]$ can be further centered by the frequency of the reference allele at QTL $j(p_j)$ as $z_j = x_j - 2p_j$. Then the phenotype of individual i can be modeled as:

$$y_i = \mu + \sum_{j=1}^{Q} z_{ij}\beta_j + \epsilon_i, \tag{1}$$

where μ is the intercept, β_j is the allele substitution effect of a reference allele at a biallelic QTL j, $\sum_{j=1}^{Q} z_{ij}\beta_j$ is the genetic value of the individual i, ϵ_i is the environmental residual term assumed to be independent draws from $N(0,\sigma_\epsilon^2), \forall i \in [1,n]$. Assuming that allele substitution effects at QTLs are known, absence of dominance and epistasis and orthogonality between additive genetic and environmental effects, Eq. 1 leads to the following decomposition of phenotypic variance:

$$\mathrm{Var}(y_i) = \mathrm{Var}\left(\sum_{j=1}^{Q} z_{ij}\beta_j\right) + \mathrm{Var}(\epsilon_i) = \sigma_\mathrm{A}^2 + \sigma_\epsilon^2, \tag{2}$$

where σ_A^2 is the additive genetic variance that can be further decomposed into (Lynch and Walsh 1999; de los Campos et al. 2015):

$$\sigma_A^2 = \sigma_a^2 + d, \tag{3}$$

where $\sigma_a^2 = \sum_{j=1}^{Q} \mathrm{Var}(z_{ij})\beta_j^2$ is the additive genic variance and $d = 2\sum_{j=1}^{Q}\sum_{j'>j}^{Q} \mathrm{Cov}(z_{ij}, z_{ij'})\beta_j\beta_{j'}$ depends on the covariances between QTLs. Note that $\mathrm{Var}(z_{ij})$ stands for the variance of allelic contents at QTL j and $\mathrm{Cov}(z_{ij}, z_{ij'})$ for the covariance between QTL j and j'. The additive genic variance (σ_a^2) is the genetic variance expected in the absence of gametic phase disequilibrium (linkage equilibrium, LE).

At a given generation t, the expected response to selection, i.e., genetic gain, in the next generation is determined by Lush (1937):

$$\mu(t+1) - \mu(t) = ih(t)\sqrt{\sigma_A^2(t)}, \tag{4}$$

where $\mu(t+1)$ is the expected mean genetic performance in the next generation, $\mu(t)$ is the mean genetic performance of generation t, i is the selection intensity, $h(t)$ is the selection accuracy and $\sigma_\mathrm{A}^2(t)$ is the additive genetic variance in generation t. In order to assess the past history of a breeding population and the short-term expected evolution assuming no changes in breeding practices, it is interesting to estimate the mean and the additive genetic variance for several cycles until the current generation (*phenotypic indicators*).

Considering a set of inbred lines and assuming that genotypes at QTL are independent (LE), we can express the expected additive genic variance as a function of genetic diversity. We define the genetic diversity for a set of inbred lines as the expected heterozygosity averaged on Q biallelic QTLs (He, Nei 1978) $\bar{H}_\mathrm{e} = \frac{1}{Q}\sum_{j=1}^{Q} 2p_j(1-p_j)$. In a population of inbred lines the variance is doubled compared to a population in Hardy–Weinberg equilibrium, so the variance at locus j $\mathrm{Var}(z_{ij}) = 4p_j(1-p_j)$. The genic variance that does not take into account covariances between QTL can be expressed as $\sigma_\mathrm{a}^2 = \sum_{j=1}^{Q} \mathrm{Var}(z_{ij})\beta_j^2 = \sum_{j=1}^{Q} 4p_j(1-p_j)\beta_j^2$.

In practice, allele substitution effects are unknown and commonly estimated in a linear regression model assuming them to be random draws from a normal distribution $\beta \sim N\left(0, \sigma_\beta^2 I\right)$ (Meuwissen et al. 2001; VanRaden 2008; Gianola et al. 2009). Following this framework, the expected value for the additive genic variance can be derived as (Gianola et al. 2009):

$$\mathrm{E}[\sigma_\mathrm{a}^2] = E\left[\sum_{j=1}^{Q} 4p_j(1-p_j)\beta_j^2\right] = \sigma_\beta^2 \sum_{j=1}^{Q} 4p_j(1-p_j) = 2\sigma_\beta^2 Q\bar{H}_\mathrm{e}, \tag{5}$$

where $\bar{H}_\mathrm{e}$ is the genetic diversity averaged on Q biallelic QTLs defined previously and σ_β^2 the variance of allele substitution effect distribution. Assuming complete LE between QTLs, i.e., the covariance term (d) in Eq. 3 is null, the additive genetic variance is equal to the additive genic variance. Under this assumption, genome-wide genetic diversity at causal loci (QTLs) is expected to be a proxy of the additive genetic variance and therefore to the expected response to selection. Consequently, it is of interest to evaluate the past evolution of genetic diversity (*genotypic indicators*) jointly with that of additive genetic variance.

The LE assumption between QTLs is not realistic when there is selection, genetic drift or introgression (Gianola et al. 2009) and the covariance component of the additive genetic variance should be accounted for. The contribution of LD to the additive genetic variance is well known in quantitative genetics (Hill and Robertson 1966; Bulmer 1971; Avery and Hill 1977; Gianola et al. 2009). The LD is built up as the result of two opposite forces: selection that generates LD between QTLs and recombination that tends to break LD in every subsequent generation by recombination events until an equilibrium is reached. A positive LD ($d > 0$) implies that genetic variance is created by positive

covariance between QTLs (i.e., coupling) and occurs under a structuration of the population into subpopulations of contrasted means (i.e., likely due to divergent selection or recent admixture) (Felsenstein 1965). For instance, a positive d value was observed by Lehermeier et al. (2017) in an Arabidopsis panel evaluated for flowering time, showing strong evidence for diversifying selection. A negative LD ($d<0$) implies that additive genic variance is partly hidden by negative covariance between QTLs (i.e., repulsion), as occurs under directional and stabilizing selection. In this case, there is an interest to generate additive genetic variance (σ^2_A) by recombination. Efficiency of recombination to break repulsion between QTLs will depend on the linkage between QTLs (Lynch and Walsh 1999, chapter 16 and 26; Hospital and Chevalet 1996) and the frequency of alternative haplotypes. Consequently, a proper estimation of additive genetic variance components (Eq. 3) jointly with an analysis of genetic diversity (He) allows one to evaluate the interest of recombination to unlock the potential response to selection (*genomic indicators*).

Materials and methods

Genetic material

We worked on a subset of a North European early grain maize private breeding program from RAGT2n. This program was organized around a Dent–Flint heterotic pattern and aimed at improving grain yield (quintal per hectare, qx/ha) performance and stability while keeping grain maturity constant. Both groups were evaluated in the same network of testing locations in the north of France and Germany from 2006 to 2016. In brief, each year new segregating recombinant inbred lines (RILs: F4 to more advanced inbreds) or doubled haploid (DH) lines were evaluated 3 years after the biparental cross was made (namely, breeding start). Their hybrid progeny with one representative line from the complementary group (tester) was evaluated for 1 year in four to six locations. Best performing lines (F6 and inbred lines or DH lines) were then tested more extensively with several testers and for several years. Lines considered in this study were evaluated for 1–11 years on an average of 1.3 different testers. Testers were running on average for 5 years, allowing some bridges between years for hybrid value decomposition into its general combining ability (GCA) and specific combining ability (SCA) components. To follow temporal evolution of proposed indicators, we considered cohorts that consisted in lines derived from breeding starts realized at a given year within one heterotic group. We considered eleven cohorts from 2003 to 2013 with an average of 316 and 272 lines in the Dent and Flint cohorts, respectively. The average generation interval, i.e., average number of years between the creation of an inbred and that of its parents, was 5 years. Commercial checks and most advanced material were evaluated during several years, which makes it possible to obtain estimates of mean performances and variance components over cohorts (*phenotypic indicators*).

For genetic diversity analysis (*genotypic indicators*), only parental inbred lines of breeding start crosses were considered. Parental inbred lines contributed on average to two cohorts, leading to some overlap between cohorts in terms of parents. Every parental inbred line was genotyped with the MaizeSNP50 Illumina ® BeadChip (Ganal et al. 2011). Only Panzea markers designed from 27 diverse founder lines were used for diversity analysis to reduce SNP discovery ascertainment bias (Gore et al. 2009). Further, only markers that mapped on the B73 refgenV4 genome (Jiao et al. 2017), presenting a call rate ≥ 0.9, and a heterozygosity level ≤ 0.15 were kept, resulting in 28,803 genome-wide SNPs used for the analyses. Lines exhibiting a call rate ≥ 0.8 and heterozygosity level ≤ 0.10 were considered, representing, respectively, 214 Dent and 111 Flint parental inbred lines. Missing marker values were imputed using Beagle v4 (Browning and Browning 2007).

The evaluation of the impact of LD and genetic diversity on genetic variance (*genomic indicators*) was not applied on temporal data but only on the six last Dent cohorts representing 1809 RILs or DH lines evaluated on Flint testers from 2011 to 2016 (i.e., cohorts 2008–2013) for which marker data were available. These 1809 lines were genotyped with a low-density array and imputed on 28,803 SNPs based on parental genotypes using Beagle v4 (Browning and Browning 2007).

Phenotypic indicators: genetic gain and additive genetic variance evolution

Linear genetic gain and intra-cohort additive genetic variance were estimated in each heterotic group using Eq. 6 fitted by R-ASReml (Butler et al. 2009; R Core Team 2017):

$$Y_{icjer} = \mu + E_e + (1-\delta_{ij})C_{ij} + \delta_{ij}(\Gamma_c + \alpha_{1ci} + \alpha_{2j} + \theta_{12ij}) + \epsilon_{icjer}, \quad (6)$$

where Y_{icjer} is the phenotype of the hybrid between line i of cohort $c \in [1, 13]$ and tester j evaluated in environment e (Location × Year) and repetition r. μ is the intercept, E_e is the environment e fixed effect, δ_{ij} is a dummy variable equal to zero for checks and to one otherwise, C_{ij} is the corresponding check hybrid fixed effect, Γ_c a fixed effect that differently accounts for the cohort c performance in Model 1a and Model 1b (detailed hereinafter), $\alpha_{1ci} \sim N(O, \sigma^2_c)$ is the GCA random effect of the tested line i within cohort c considered as being independently distributed with an intra-cohort specific variance σ^2_c, α_{2j} is the tester j GCA

fixed effect, $\theta_{12ij} \sim N(O, \sigma^2_\theta)$ is the hybrid ij SCA random effect. Finally, $\epsilon_{icjer} \sim N(O, \sigma^2_\epsilon)$ is the random residual error assumed to be independent and identically distributed (IID). In Model 1a, $\Gamma_c = a\gamma_c$ with a the linear regression slope of performances on cohort's year γ_c considered as numeric variable. The estimated regression slope a was used to assess linear genetic gain and its significance was estimated using a Wald test conditionally to other fixed effects. To estimate the intra-cohort GCA variance, i.e., Model 1b, Γ_c is defined as the fixed effect of cohort c with cohort considered as categorical variable. The evolution of intra-cohort GCA variance was estimated assuming a linear regression of σ^2_c on cohort's year and linear slope parameter significance was estimated using a t test. Further, the annual genetic gain in units of initial genetic standard deviation was estimated using Model 1a after scaling Y_{icjer} by the intra-cohort genetic standard deviation of the first cohort $\sigma_{c=1}$ (i.e., cohort 2003) obtained in Model 1b.

Genotypic indicators: Patterns of diversity and differentiation between pools

Due to limited number of parental lines per cohort, cohorts were merged into 5 years periods that corresponded approximately to the interval between two generations of inbreds. We considered periods as sliding windows with a 1-year increment yielding 36–113 parental lines per period with an average of 66 lines in each pool. Genetic diversity was assessed by the expected heterozygosity (He) at each locus $j \in [1, m]$ as $\mathrm{He}_j = 2p_j(1 - p_j)$ (Nei 1978), where p_j is the frequency of the allele whose homozygous genotype is coded as 2 at locus j and $m = 28{,}803$ the total number of loci. He was averaged chromosome-wide and genome-wide and significance of the difference between two period means was assessed using a paired t test.

Furthermore, the minor allele frequency (MAF) at locus j was defined as $\mathrm{MAF}_j = \min\{p_j; 1 - p_j\}$. MAF was used to detect nearly fixed chunks of SNPs, defined as regions of at least ten successive SNPs presenting a MAF < 0.05 and covering more than 0.5 Mb. Such regions define chunks of successive SNPs where inbred parental lines are expected to be identical by descent (IBD), so that their within group hybrid progeny is expected to be homozygous. We referred to them as runs of expected homozygosity (ROHe), by analogy to the concept of runs of homozygosity increasingly used in animal genetics (MacLeod et al. 2009; Peripolli et al. 2017). To our knowledge, there is no reference of its use in hybrid plant breeding.

Differentiation between Flint and Dent pools was estimated along the genome using a F_{ST} index according to Nei's definition (Nei 1975). In pairwise comparison between two pools of sizes n_1 and n_2, and with allele frequencies p_{1j} and p_{2j} at a given locus j, the F_{STj} was estimated and averaged genome wide as:

$$F_{ST} = \frac{1}{m}\sum_{j=1}^{m} F_{ST_j} = \frac{1}{m}\sum_{j=1}^{m} \frac{n_1 n_2 (p_{1j} - p_{2j})^2}{(n_1 p_{1j} + n_2 p_{2j})(n_1 + n_2 - n_1 p_{1j} - n_2 p_{2j})}.$$

Genomic indicators: estimation of linkage disequilibrium contribution to additive genetic variance

We implemented a Bayesian chromosome partitioning on the four last Dent cohorts in two stages. In the first stage, we computed best linear unbiased estimators (BLUEs) of the 1809 tested lines general combining ability (GCA) using R-ASReml (Butler et al. 2009; R Core Team 2017) with the following model:

$$\begin{aligned} Y_{ijlyr} = {} & \mu + E_{ly} + (1 - \delta_{ij})\, C_{ij} \\ & + \delta_{ij}(\alpha_{1i} + \alpha_{2j} + \theta_{12ij} + \theta Y_{1iy} + \theta Y_{2jy} + \theta Y_{12ijy}) \\ & + \epsilon_{ijlyr}, \end{aligned} \quad (7)$$

where Y_{ijlyr} is the phenotype of the hybrid between line i and tester j evaluated in environment ly (Location $l \times$ Year y) and repetition r, μ is the intercept, E_{ly} is the environment ly fixed effect, δ_{ij} is a dummy variable equal to zero for checks and to one otherwise, C_{ij} is the corresponding check hybrid performance fixed effect, α_{1i} is the tested line i GCA fixed effect, α_{2j} is the tester j GCA fixed effect, $\theta_{12ij} \sim N(O, \sigma^2_\theta)$ is the hybrid ij SCA random effect. $\theta Y_{1iy} \sim N(0, \sigma^2_{\text{GCA 1×Year}})$ and $\theta Y_{2jy} \sim N(0, \sigma^2_{\text{GCA 2×Year}})$ are the tested line i and tester j GCAs by Year y random interaction effects, and $\theta Y_{12ijy} \sim N(0, \sigma^2_{\text{SCA×Year}})$ is the hybrid ij SCA by Year y random interaction effect. Finally, $\epsilon_{ijlyr} \sim N(0, \sigma^2_c)$ is the random residual term. Every random effect was assumed IID.

In a second stage, to estimate the proportion of variance explained by each chromosome, accounting or not for the covariance between QTLs, we partitioned the variance of estimated GCA (BLUEs) across chromosomes. Genome-wide markers were partitioned into two sets corresponding, respectively, to markers mapped on one chromosome and markers mapped on the nine others as suggested by Jensen, Su, and Madsen (2012)

$$\hat{\alpha}_{1i} = \mu + \sum_{j \in Q_1} z_{ij}\beta_j^{(Q_1)} + \sum_{j \in Q_2} z_{ij}\beta_j^{(Q_2)} + \epsilon_i, \quad (8)$$

where $\hat{\alpha}_{1i}$ is the GCA of Dent line i estimated in Eq. 7, $z_{ij} = x_{ij} - 2p_j$ is the centered allele content of line i at locus j, where x_{ij} is the genotype of line i at locus j coded as 0, 1, 2 and p_j is the frequency of the reference allele (as defined for *genotypic indicators*). $\beta_j^{(Q_1)}$ and $\beta_j^{(Q_2)}$ are the allele substitution effect of marker j in the partition 1 (Q_1) and 2 (Q_2) of the markers, respectively. Marker effects are treated as independent draws from a normal distribution with null mean

and partition-specific marker effect variances $\sigma^2_{\beta(Q_1)}$ and $\sigma^2_{\beta(Q_2)}$. $\epsilon \sim N(0, \sigma^2_\epsilon I)$ is the random vector of residual terms assumed IID. We implemented Eq. 8 in a Bayesian MCMC setting (Sorensen et al. 2001; Lehermeier et al. 2017), using the Bayesian ridge regression from the R function BGLR (Pérez and de los Campos 2014). We considered default parameters implemented in BGLR for the number of degrees of freedom ($df=5$) of the scaled inverse-χ^2 prior distributions. The scale parameters were set to correspond to a prior heritability of each partition as estimated beforehand by solving Eq. 8 using R-ASReml (Butler et al. 2009). The variances and covariance of genetic values of partitions Q_1 and Q_2 estimated at each MCMC post burn-in sample were averaged and used to infer the additive genetic variances (σ^2_A), referred as $\hat{\sigma}^2_{A(Q_1)}$ and $\hat{\sigma}^2_{A(Q_2)}$ and to infer the additive genetic covariance between Q_1 and Q_2 $\hat{\sigma}_{A(Q_1,Q_2)}$ [method M2, defined in Lehermeier et al. (2017)]. Additive genic variances of Q_1 and Q_2 assuming LE between QTLs (σ^2_a) were estimated using the posterior mean of marker effects variances $\hat{\sigma}^2_{\beta(Q_1)}$ and $\hat{\sigma}^2_{\beta(Q_2)}$, respectively [method M1, defined in Lehermeier et al. (2017)]. Additive genic variances were referred as $\hat{\sigma}^2_{a(Q_1)}$ and $\hat{\sigma}^2_{a(Q_2)}$, respectively. We used a total of 40,000 iterations where the first 15,000 iterations were discarded as burn-in. Every fifth sample was kept leading to 5000 samples used to estimate posterior mean and standard deviation of variances $\hat{\sigma}^2_a$, $\hat{\sigma}^2_A$ and the covariance $\hat{\sigma}_{A(Q_1,Q_2)}$. Following Eq. 3 and using estimators described above, we approximated for a given partition Q_1 the amount of genic variance captured by negative covariance between QTLs by the ratio of additive genetic variance $\hat{\sigma}^2_{A(Q_1)}$ on additive genic variance $\hat{\sigma}^2_{a(Q_1)}$. This ratio was computed at each of the 5,000 MCMC samples and based on these the posterior mean, namely $\widehat{\sigma^2_{A(Q_1)}/\sigma^2_{a(Q_1)}}$, and the posterior standard deviation were estimated.

Results

Phenotypic indicators: joint evolution of genetic performance and additive genetic variance

The genetic gain in trait units was highly significant ($p<10^{-4}$) in both pools and was higher in the Flint pool (0.85 ± 0.08 qx/ha/year) than in the Dent pool (0.58 ± 0.07 qx/ha/year) (Table 1). After scaling on the first (i.e., cohort 2003) intra-cohort genetic standard deviation, the genetic gain in standard deviation units was still higher in the Flint pool (0.16 ± 0.02 sd/year) than in the Dent pool (0.13 ± 0.02 sd/year) (Table 1). The variance of intra-cohort GCA estimated in Model 1b showed no significant evolution over time in the Flint pool but its reduction was significant at a 10% risk level in the Dent pool (Tables 1, S1). The intra-cohort additive genetic variance averaged over cohorts was higher in the Flint pool (18.94 ± 11.06) than in the Dent pool (9.68 ± 5.83) (Table 1).

Genotypic indicators: changes in genetic diversity over time

Genome-wide genetic diversity was higher on average in the Flint (0.276) than in the Dent pool (0.147). Genetic diversity was nearly stable over time in the Flint pool while it was depleted in the Dent pool (from 0.160 in 2003–2007 to 0.136 in 2012–2016, Fig. 1, Table S2). This evolution of genetic diversity was associated with that of the differentiation between heterotic groups (Fst), which raised from 0.156 in 2003–2007 to 0.178 in 2012–2016 (Fig. 1, Table S2). Considering only two extreme non-overlapping 7 years periods, 2003–2009 and 2010–2016, we observed different patterns for the evolution of genetic diversity between chromosomes and heterotic groups (Fig. 2, Table S3). For sake of simplicity, we focused on three chromosomes (3, 4, 6) showing different behaviors. Chromosome 3, one of the most diverse chromosomes in the Dent pool, showed a significant increase in genetic diversity in the Dent pool (from 0.209 to 0.224) as well as a smaller, but still significant, increase in the Flint pool (from 0.290 to 0.296) (Table S3). The other

Table 1 Genetic gain between two successive cohorts (Model 1a, Eq. 6) and intra-cohort additive genetic variance (GCA) evolution (Model 1b, Eq. 6) of material generated from 2003 to 2013

Group	Genetic gain in qx/ha/year (±se)	Genetic gain in units of genetic standard deviation/year (±se)	Slope on intra-cohort GCA variance (±se)	Mean of intra-cohort GCA variance (±se)
Dent	0.58*** (±0.07)	0.13*** (±0.02)	−1.00° (±0.48)	9.68 (±5.83)
Flint	0.85*** (±0.08)	0.16*** (±0.02)	0.14^{ns} (±1.11)	18.94 (±11.06)

Annual genetic gains are expressed in qx/ha/year and in units of genetic standard deviation in the first cohort 2003. Standard error of estimates are given as (±se)

p values significance: ***$<10^{-4}$; **<0.001; *<0.01; °°<0.05; °<0.1; $^{ns}<1$

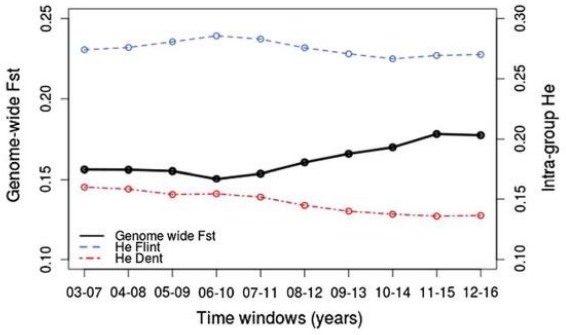


Fig. 1 Evolution of intra-group genetic diversity (He, right axis) and differentiation between heterotic groups (Fst, left axis) measured on 5 years sliding windows with 1 year increment. 03-07 stands for 2003–2007

two chromosomes presented specific patterns of diversity (Fig. 2, top panel). Chromosome 4 showed a low diversity in the pericentromeric region in the Dent but not in the Flint pool. This large region was enriched in new ROHe in the second period (Fig. 2, bottom panel). In this second period, ROHe covered about 43% of chromosome 4 physical length compared to 10% in the first period (Table S4). While global genetic diversity on chromosome 6 in Dent was significantly reduced between periods (0.115–0.105), it showed a local increase in frequency of originally rare alleles in the pericentromeric region (40–90 Mb). On the contrary, in the Flint pool, chromosome 6 did not show a significant evolution of diversity (0.311–0.312).

Genomic indicators: chromosome partitioning of additive variances and genetic diversity

The total variance of grain yield GCA in the analyzed Dent material was 55.111 (Table S5). Estimated additive genic variance ($\hat{\sigma}^2_a$, M1) and genetic variance ($\hat{\sigma}^2_A$, M2) using Eq. 8 with a genome-wide marker set were, respectively, 27.399 ± 3.864 and 20.599 ± 1.459 (Table S5). The genome-wide ratio $\widehat{\sigma^2_A/\sigma^2_a}$ (0.761 ± 0.079) indicated that repulsion

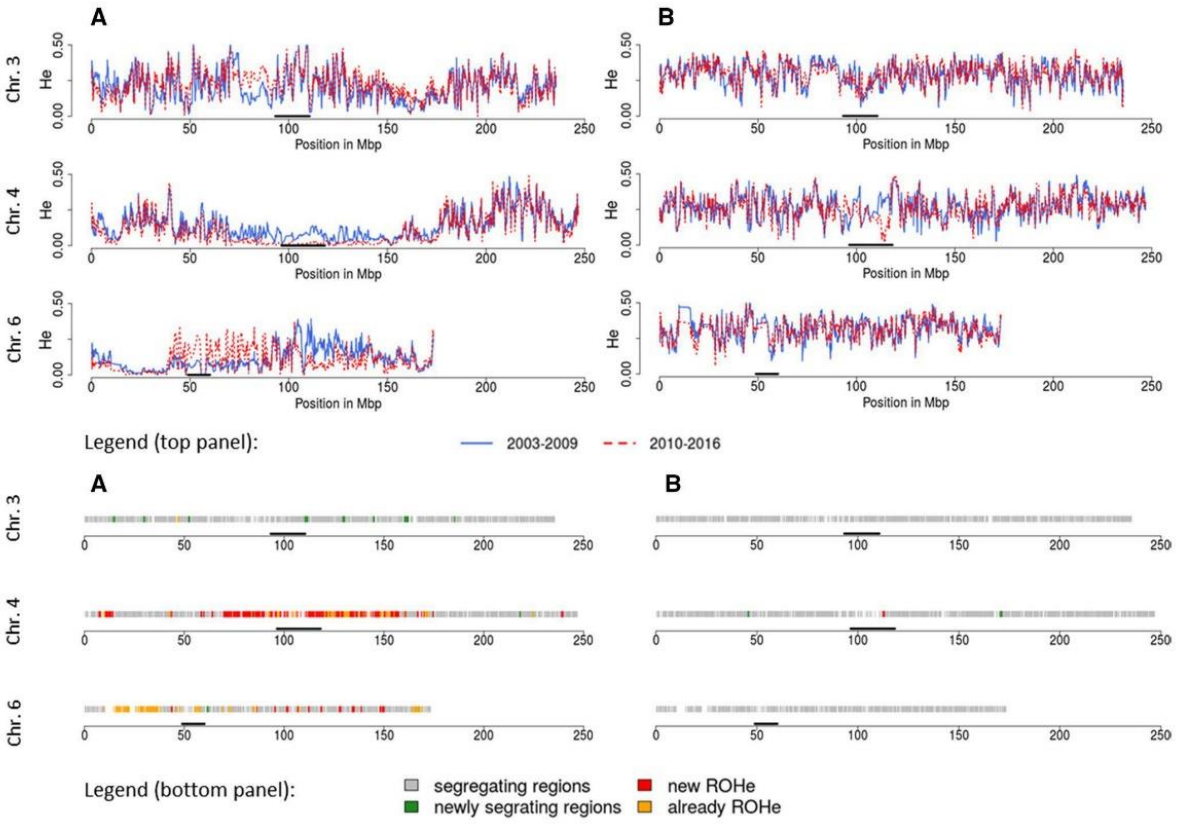


Fig. 2 Genetic diversity (top panel) and distribution of ROHe (bottom panel) along physical map. Top panel: Genetic diversity in Dent pool (**a**) and in Flint pool (**b**) for chromosomes 3, 4, 6. Genetic diversity 2003–2009 in blue full line and 2010–2016 in red dotted line. Centromeres are marked in bold on the abscissa. Bottom panel: Evolution of ROHe in Dent pool (**a**) and in Flint pool (**b**) for chromosomes 3, 4, 6. Regions are colored regarding their evolution between 2003–2009 and 2010–2016

between QTLs captured 23.9% of the additive genic variance. In chromosome partitioning, the estimated additive genic variance ($\hat{\sigma}^2_a$) was higher than the estimated additive genetic variance ($\hat{\sigma}^2_A$) for all chromosomes (Table 2). We also observed substantial differences in the contribution of chromosomes to the total variance. For instance, regarding $\hat{\sigma}^2_a$, contribution of chromosome 3 ($\hat{\sigma}^2_a = 9.147 \pm 2.820$) was about eight times superior to that of chromosome 4 ($\hat{\sigma}^2_a = 1.151 \pm 0.494$). Considering $\hat{\sigma}^2_A$, contribution of chromosome 3 ($\hat{\sigma}^2_A = 4.533 \pm 0.995$) was approximately six times that of chromosome 4 ($\hat{\sigma}^2_A = 0.707 \pm 0.292$). In accordance with the theoretical link between genetic diversity and additive genic variance (Eq. 5), a significant tendency at risk level 10% ($R^2 = 0.306$, $p = 0.097$) was observed between chromosome He and $\hat{\sigma}^2_a$ explained by one chromosome, after correction for chromosome length (Fig. S1). A significant relationship at risk level 5% was observed for the additive genetic variance $\hat{\sigma}^2_A$ ($R^2 = 0.427$, $p = 0.040$). The genetic covariances $\hat{\sigma}_{A(Q_1,Q_2)}$ (method M2) between one chromosome (partition Q_1) and the rest of the genome (partition Q_2) ranged from -0.593 ± 0.682 (between chromosome 2 and Q_2) to 0.810 ± 0.442 (between chromosome 3 and Q_2) (Table 2). The ratio $\widehat{\sigma^2_A/\sigma^2_a}$ varied over chromosomes with an average value of 0.597 and an average standard deviation of 0.166 (Table 2). We observed different extreme chromosomes for $\widehat{\sigma^2_A/\sigma^2_a}$ and He. Chromosome 10 showed a low diversity (He = 0.064) and a low $\widehat{\sigma^2_A/\sigma^2_a}$ (0.372 ± 0.116). Chromosome 8 also showed a low diversity (He = 0.063) but a higher $\widehat{\sigma^2_A/\sigma^2_a}$ (0.831 ± 0.300). For more diverse chromosomes, chromosome 5 (He = 0.119) showed a low $\widehat{\sigma^2_A/\sigma^2_a}$ (0.523 ± 0.120) while chromosome 2 (He = 0.119) showed a higher $\widehat{\sigma^2_A/\sigma^2_a}$ (0.647 ± 0.145).

Discussion

Here, we discuss the main results for each indicator, propose generalized interpretation grids and suggest decision guidelines.

Phenotypic indicators: joint evolution of genetic performance and additive genetic variance

Assuming complete additivity, estimations of genetic gain in both breeding pools led to a total annual genetic gain at the hybrid level of 1.43 qx/ha/year. As a matter of comparison, Duvick (1984) estimated on released commercial hybrids a genetic gain of 1.12 qx/ha/year from 1955 to 1980. More recently, Duvick et al. (2004) estimated on the period 1930–2001 a total genetic gain of 0.77 qx/ha/year. In Europe, Fischer et al. (2008) observed over 29 years of European Flint × Dent breeding program at the University of Hohenheim a significant evolution of mean hybrid performance of 1.70 qx/ha/year.

In this study, the additive genetic variance showed a reduction (significant at a 10% risk level) in the Dent pool but not in the Flint pool. The absence of a monotonic trend toward a reduction of variance in the Flint pool is due to genetic diversity introgression in 2011 and 2013 Flint cohorts (Table S1). We decided to keep these years in the evaluation since they reflected the recent breeding strategy in the Flint pool, which contrasts with that of the Dent pool. The lower mean intra-cohort additive genetic variance in Dent (9.68 ± 5.83) compared to Flint (18.94 ± 11.06) may explain the lower genetic gain observed in Dent, since similar selection intensities were applied in both pools. In the Hohenheim University breeding program, Fischer et al. (2008) observed neither a significant temporal evolution of the GCA variance in the Flint or Dent pools, nor significantly different mean variances between pools. These findings reflect differences in breeding strategies and objectives between the Hohenheim University breeding program and the one analyzed in the present study, the Hohenheim University germplasm being regularly enriched with new germplasm (Fischer et al. 2008).

In this study, the Flint pool showed an increase in genetic performance and a stable non-zero additive genetic variance over one decade (Table 1). This situation reveals a sustainable breeding population with promising expected response to selection. The introgression events in the Flint population in 2011 and 2013 likely contributed to maintain the genetic variance and long-term selection response. It can be advised for this Flint population to keep managing existing genetic diversity and/or introduce new original favorable alleles to maintain long-term genetic gain while maximizing short-term genetic gain. On the contrary, the Dent pool, showed an increase in genetic performance and a reduction of additive genetic variance over one decade (Table 1). Such a situation will potentially lead to a depletion of the expected response to selection and therefore genetic gain in a mid-long-term future. Identification of relevant sources of diversity to be introduced in the breeding pool may mitigate this risk. This identification should consider the potential increase in genetic diversity, the level of performance of introduced accessions, as well as the maintenance of the complementarity between heterotic groups.

The reduction in the additive genetic variance is inherent to directional selection but its conversion into short and long-term genetic gain can be optimized. Several approaches to optimize parental cross-designs have been suggested to this end. For instance, optimum contribution selection approaches (Brisbane and Gibson 1995; Meuwissen 1997; Woolliams et al. 2015; Akdemir and Sánchez 2016; Gorjanc et al. 2018) aim to optimize the contribution of parents to

Table 2 Posterior means (± posterior standard deviation) of additive genomic variances for Grain Yield (qx/ha) of the ten maize chromosomes estimated in Eq. 8 with method M1 ($\hat{\sigma}^2_a$, assuming LE) or method M2 ($\hat{\sigma}^2_A$, accounting for LD)

Chr	He	Length (Mbp)	*M1*			*M2*				Ratio (± sd)
			$\hat{\sigma}^2_a$ (± sd)		Total (± sd)	$\hat{\sigma}^2_A$ (± sd)		$\hat{\sigma}_{A(Q_1,Q_2)}$ (± sd)	Total (± sd)	$\frac{\widehat{\sigma^2_{A(Q_1)}}}{\sigma^2_{a(Q_1)}}$
			Chr. (Q_1)	Others (Q_2)		Chr. (Q_1)	Others (Q_2)			
1	0.080	305.8	7.141 (±2.510)	21.552 (±3.598)	63.194 (±3.863)	3.688 (±1.005)	16.207 (±1.746)	0.437 (±0.629)	55.270 (±1.689)	0.543 (±0.133)
2	0.119	244.3	4.292 (±1.454)	23.565 (±3.839)	62.382 (±3.619)	2.668 (±0.747)	19.240 (±1.929)	−0.593 (±0.682)	55.246 (±1.694)	0.647 (±0.145)
3	0.141	234.9	9.147 (±2.820)	19.618 (±3.524)	63.334 (±3.757)	4.533 (±0.995)	14.630 (±1.701)	0.810 (±0.442)	55.352 (±1.702)	0.518 (±0.112)
4	0.066	246.6	1.151 (±0.494)	27.138 (±3.970)	62.753 (±3.751)	0.707 (±0.292)	19.899 (±1.763)	0.128 (±0.475)	55.326 (±1.700)	0.642 (±0.198)
5	0.119	221.9	4.886 (±1.808)	23.625 (±3.686)	62.971 (±3.607)	2.430 (±0.696)	18.397 (±1.976)	0.015 (±0.604)	55.317 (±1.685)	0.523 (±0.120)
6	0.047	173.2	3.059 (±1.173)	25.246 (±3.731)	62.727 (±3.639)	1.870 (±0.622)	17.573 (±1.638)	0.723 (±0.374)	55.309 (±1.700)	0.644 (±0.183)
7	0.079	179.2	2.323 (±1.051)	25.900 (±3.836)	62.701 (±3.596)	1.286 (±0.474)	18.699 (±1.639)	0.427 (±0.367)	55.319 (±1.703)	0.591 (±0.180)
8	0.063	180.5	1.824 (±0.757)	27.222 (±4.103)	63.389 (±3.838)	1.388 (±0.425)	19.018 (±1.637)	0.319 (±0.289)	55.388 (±1.730)	0.831 (±0.300)
9	0.090	159.5	2.710 (±1.078)	25.819 (±4.125)	62.974 (±3.828)	1.694 (±0.542)	18.265 (±1.795)	0.472 (±0.460)	55.348 (±1.726)	0.659 (±0.176)
10	0.064	149.6	4.422 (±1.991)	24.847 (±3.975)	63.716 (±4.028)	1.530 (±0.542)	18.387 (±1.743)	0.449 (±0.400)	55.262 (±1.692)	0.372 (±0.116)

Total M2 variance accounts for additive variance explained by partitions (Q_1 and Q_2), residual variance and covariance between partitions ($\hat{\sigma}_{A(Q_1,Q_2)}$) while total M1 variance accounts only for additive and residual variances. Genetic diversity is the averaged He for each chromosome in the 1809 candidate lines. Physical length of chromosomes is given as last minus first position of considered markers in Mbp (B73 refgenV4 genome, Jiao et al. 2017)

the next generation to minimize the loss of diversity (i.e., inbreeding) while maximizing the short-term genetic gain. In case of multi-objective selection, maintaining an appropriate level of diversity is even more critical (1) to ensure a long-term genetic gain on several traits and (2) to be able to select for emergent traits in changing environments. In this context Akdemir et al. (2018) extended the optimal contribution to multi-objective selection.

Genotypic indicators: changes in genetic diversity over time

Genome-wide markers make it possible to estimate genetic diversity and therefore bring complementary elements to the evolution of additive genetic variance. For practical reasons we worked on the parental lines of the crosses involved in each cohort but the analysis could be easily extended to progeny if such genotypic data is available. The lower genome-wide genetic diversity in the Dent pool compared to the Flint pool is in contradiction with previous observations on maize panels assembling diversity from different heterotic groups (Rincent et al. 2014). This can be explained by the fact that the Dent pool analyzed here belongs to a specific segment of Dent lines which is complementary to Flint lines. It is known that the elite Dent germplasm that has been introduced into Europe since the 1950s until today to complement Flint germplasm presents a reduced genetic basis compared to that of the total Dent gene pool that also includes heterotic groups used in the corn belt and Southern Europe (Inghelandt et al. 2010; personal communication Alain Charcosset). The steady temporal decrease in genome-wide genetic diversity within the Dent pool of parental lines is consistent with the reduction of GCA variance in the progeny (*phenotypic indicators*). Symmetrically, in the Flint pool no significant tendency toward depletion in genome-wide genetic diversity was observed. The introgression of external diversity sources in cohorts 2011 and 2013 likely contributed to this steady genetic diversity. The corresponding increase in differentiation between Dent and Flint pools is the result of reciprocal selection that increases the differentiation between pools (Feng et al. 2006; Gerke et al. 2015) and minimizes the variation of SCA relative to that of GCA (Reif et al. 2007; Fischer et al. 2009).

In addition to global trends, we observed that selection increased the frequency of initially rare alleles in some regions (e.g., chromosome 6 in Dent) and fixed alleles in other regions (e.g., chromosome 4 in Dent). Regions with an initially low diversity mainly presented a further decrease in diversity in the next generations. As previously observed by Gerke et al. (2015) in maize, such regions were mainly located in low recombination pericentromeric regions (e.g., chromosome 4 in Dent, Fig. S2). We can suggest several nonexclusive forces leading to regions with such long ROHe, which correspond to identical by descent (IBD) haplotypes. Founder effects, known to be important in the European Dent pool, can be a first explanation. Some ROHe might be due to selection induced linkage drag in regions surrounding causal genes. This assumption is supported by the ongoing extension of low diversity regions, e.g., genome-wide average length of ROHe increased from 1.45 Mb (2003–2009) to 2.46 Mb (2010–2016) in the Dent pool (Table S4). A complementary explanation is an intense selection along with a reduced effective population size that induces genetic drift. It is difficult to distinguish the direct effect of selection and its indirect effect through induced genetic drift. Based on simulation results, Gerke et al. (2015) suggested that changes in allele frequencies were mainly due to genetic drift rather than selection on causal loci. This underlines the importance of increasing the effective population size (i.e., He) to avoid the random loss of favorable alleles for both main and secondary traits (e.g., emergent stress tolerance). In theory, the effective population size can be estimated from the rate of LD decay and recombination rate assuming selective neutrality, close and panmictic populations (Weir and Hill 1980; Hill 1981; Tenesa et al. 2007; Waples and England 2011). In breeding populations, these assumptions are likely violated, which limits the use of such an estimator. However, a reduced effective population size is expected to be associated with a reduced genetic diversity (He) and higher linkage disequilibrium (Waples and England 2011). For instance, Truntzler et al. (2012) observed that LD was higher in a population of private maize dent lines compared to a more diverse dent panel. Only few ROHe were observed in Flint (Table S4) and few were overlapping with ROHe observed in Dent (0.29%). This suggests that, in a given region fixed in Dent, several complementary haplotypes are segregating in Flint, which maintains genetic variations. The identification of large regions with low genetic diversity in the Dent pool raises the question of their enrichment by external variability to select for favorable recombinants.

Genomic indicators: chromosome partitioning of additive variance(s) and genetic diversity

Beyond estimating separately additive genetic variance and genetic diversity, molecular markers enable to estimate components of the additive genetic variance (Eq. 3). The proportion of additive variance of a trait explained by regression on markers has been mostly estimated using REML or a Bayesian setting (Yang et al. 2010). However, as stressed in de los Campos et al. (2015), the traditional definition of genomic variance does not account explicitly for the contribution of LD to genetic variance. They proposed an alternative MCMC approach to estimate both additive genic (σ^2_a) and additive genetic (σ^2_A) variances explained by each chromosome. Chromosome partitioning showed substantial

Theoretical and Applied Genetics (2019) 132:1321–1334

variations in the contribution of chromosomes to the additive genic and genetic variances. It revealed also a strong variation among chromosomes for the proportion of additive genic variance captured by LD between QTLs (Table 2). Further, in Eq. 8 the estimated additive genetic covariance $\hat{\sigma}_{A(Q_1,Q_2)}$ between partitions Q_1 (i.e., tested chromosome) and Q_2 (i.e., rest of the genome) were small and mostly positive (Table 2). These values suggested that the total M2 genetic variance, which determines the potential response to selection, was partly due to positive covariance, i.e., coupling between chromosomes. It means that on average individuals that were performant on one chromosome were also performant on the rest of the genome, except for chromosome 2 that showed a negative covariance with the rest of the genome. Notice that these results should be interpreted with care due to relatively high standard deviations of the posterior distributions.

Following Eq. 3, in the case of directional and stabilizing selection, a low σ^2_A/σ^2_a ratio (bounded by 0) means that a large amount of genic variance is hidden due to repulsion phase between QTLs. On the contrary, a high σ^2_A/σ^2_a ratio (bounded by 1 under directional selection) means that genetic variance reflects directly genic variance. As breeders only have a grip on additive genetic variance, it is interesting to evaluate how much additive genic variance is currently hidden by negative covariances and could be available under linkage equilibrium. Considering sequentially the additive genic variance (σ^2_a), and the ratio (σ^2_A/σ^2_a), we considered three general situations (namely A to C, Table 3). If a chromosome shows a high genic variance compared to others (situation A and B), disregarding any other information, we can conclude that long-term response to selection should not be compromised. Considering in addition the information provided by σ^2_A/σ^2_a, we can distinguish two cases. When σ^2_A/σ^2_a is high (situation A), the genetic variance is expected to be close to its maximum and there is no interest to favor recombination. Selection pressure can likely be increased on such a chromosome. Assuming σ^2_A/σ^2_a is low (situation B), which can occur for a chromosome subjected to recent selection, there is a potential for increasing genetic variance through recombination within the population. It can be achieved by accounting for complementarity between parents when defining the crossing design at each cycle. Finally, a chromosome showing a low genic variance, whatever its σ^2_A/σ^2_a ratio, has little potential for response to selection (situation C). Either maximum performance has been reached or new favorable alleles should be introduced to unlock response to selection on such a chromosome. Considering the illustration data set (Fig. 3), the *genomic indicators* underlines the interest to enhance recombination on chromosome 10 and to broaden the genetic diversity at causal loci in chromosomes 4 and 8 in the Dent pool. This observation on chromosome 4 is consistent with results from *genotypic indicators*.

The additive genic variance (σ^2_a) has to be estimated using a whole-genome regression model. Under the infinitesimal model assumption, i.e., assuming that every locus has a small effect, σ^2_a can be approximated by the genetic diversity (He) (Eq. 5). In practice, we observed a low but significant ($R^2 = 0.306$, $p = 0.097$) relationship between chromosome additive genic variance σ^2_a corrected for chromosome length and its genetic diversity He (Fig. S1). This low relationship can be explained by the deviation from the infinitesimal

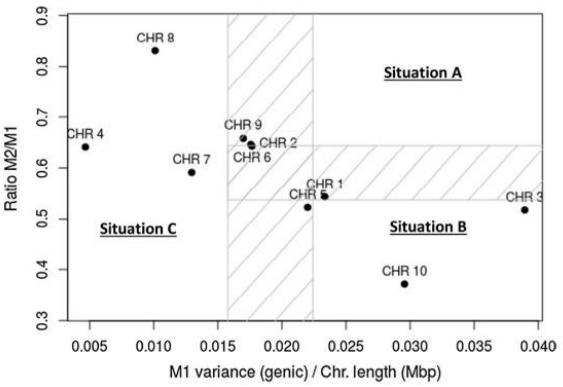


Fig. 3 Ratio $\widehat{\sigma^2_{A(Q_1)}/\sigma^2_{a(Q_1)}}$ and genic variance $\widehat{\sigma^2_{a(Q_1)}}$ corrected per chromosome length (in Mega basepairs, Mbp) of ten chromosomes in the Dent last cohorts (1809 lines). Values between the quantiles 0.3 and 0.7 are cross-hatched on both axes to distinguish the three situations in Table 3

Table 3 Decision tree using additive genic variance σ^2_a and ratio σ^2_A/σ^2_a

σ^2_a (diversity at causal loci)	σ^2_A/σ^2_a	Situation	Proposed strategy
High	High (few genic variance hidden by repulsion)	A	Increase or maintain current levels of selection intensity
	Low (lot of genic variance hidden by repulsion)	B	Potential gain to favor recombination within breeding pool
Low	High or low	C	Introduce external genetic diversity

Table 4 Summary of sets of indicators, required data and usage

Indicator set	Data required	Usage
Phenotypic indicators		
Joint temporal evolution of genetic performance and additive genetic variance	Phenotypic data over several cohorts	Assess the sustainability of a breeding population and future potential response to selection Enable to allocate efforts between populations
Genotypic indicators		
Genetic diversity and structuration between heterotic groups	Genotypic data over several cohorts (parents or progeny)	Assess the past evolution of genetic diversity Define regions with critical lack of diversity (ROHe) Allocate efforts between populations and genomic regions
Genomic indicators		
Chromosome partitioning of variance, estimation of selection induced LD contribution to genetic variance	Phenotypic and genotypic data (can be applied over several cohorts)	Define the optimal way to unlock potential response to selection per chromosome - Increase selection intensity - Introgress genetic diversity - And/or enhance recombination

model, resulting in the fact that the variance at neutral loci does not exactly reflect the variance at causal loci. Note that both the low He (genetic diversity) and the low σ^2_a (diversity at causal loci) for chromosomes 4 and 8 in the Dent last cohorts converge toward the importance to increase genetic diversity of these chromosomes.

In practice, selection of favorable recombinants on specific chromosomes while maintaining favorable configurations on others is not straightforward. Different approaches can be considered, such as planning crosses based on the molecular dissimilarity in the regions of interest, applying foreground selection for the chromosome(s) where targeted recombination events are desired and background selection to maintain a uniform genetic background on remaining chromosomes (Bernardo 2014, 2017). More recently, the promise of new breeding technics such as genome editing allow targeting and/or designing recombination events to reveal hidden genic variance and increase genetic gain as simulated in Gonen et al. (2017) or Bernardo (2017).

In this study, we partitioned the genome into chromosomes but Eq. 8 is defined in a general case. In theory, it is possible to consider finer regions (Speed and Balding 2014; Gusev et al. 2014). However, it might be computationally intensive and potential non-orthogonality between partitions might yield unreliable variance estimates. Also, we focused on the last Dent cohorts for which all candidates were genotyped. As this generation showed the lowest additive genetic variance, evaluating the temporal evolution of the LD effect on additive genetic variance would be interesting. For instance, using REML estimates of additive genetic variance estimated in Flachenecker et al. (2006) and LD between favorable alleles at QTLs, Falke et al. (2007b) did not observe a long-term reduction of the additive genetic variance due to negative LD. For the authors these results supported the hypothesis that in maize recurrent selection programs, LD generated by high intensity of selection is not a limiting factor if an efficient recombination procedure is employed during mate allocation.

Conclusion

Temporal series of phenotypic and genotypic data available in breeding programs can be used to analyze the phenotypic and genomic changes associated with past selection. This provides a basis for managing the trade-off between genetic diversity and short-term genetic gain. In this study, we proposed to harness this information through three indicators summarized in Table 4. The joint analysis of achieved genetic gain and evolution of additive genetic variance (*phenotypic indicators*) assesses the past efficiency of selection and the sustainability of the breeding population. Complementarily, the analysis of genetic diversity evolution along the genome (*genotypic indicators*) can reveal local genomic patterns indicative of past breeding strategies. The proposed definition of ROHe enables to draw breeders' attention on populations likely requiring genetic diversity introgression and to define regions that might gain to be enriched in diversity. Finally, using molecular markers in a MCMC approach (*genomic indicators*) enables to estimate to which extent the additive genetic variance is affected by selection induced linkage disequilibrium. It allows to propose fine scale strategies to manage genetic diversity and to unlock the potential response to selection (Bernardo 2017; Gonen et al. 2017). Advances in genomic technologies promise to make it feasible, affordable and highly beneficial. Proposed indicators can easily be extended to other breeding programs and

species, which are likely to show different results depending on past choices and long-term strategies.

Author contribution statement AC, LM and ST supervised the study. AA realized the study and prepared the manuscript. CL helped to review the manuscript. BC, SMa, SMe helped for raw data production and extraction. All authors read and approved the manuscript.

Acknowledgements The authors thank the experimental staff at RAGT2n for managing field experiments and data extractions. This research was funded by RAGT2n and the ANRT CIFRE Grant No. 2016/1281 for AA.

Data availability The datasets analyzed in this study are not publicly available.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest. The experiments reported in this study comply with the current laws in Europe.

References

Akdemir D, Sánchez JI (2016) Efficient breeding by genomic mating. Front Genet 7:210

Akdemir D, Beavis W, Fritsche-Neto R, Singh AK, Isidro-Sánchez J (2018) Multi-objective optimized genomic breeding strategies for sustainable food improvement. Heredity 1:17

Avery PJ, Hill WG (1977) Variability in genetic parameters among small populations. Genet Res 29:193–213

Bernardo R (2014) Essentials of plant breeding. Stemma Press, Woodbury

Bernardo R (2017) Prospective targeted recombination and genetic gains for quantitative traits in maize. Plant Genome 10:2

Betrán FJ, Hallauer AR (1996) Characterization of interpopulation genetic variability in three hybrid maize populations. J Hered 87:319–328

Brisbane JR, Gibson JP (1995) Balancing selection response and rate of inbreeding by including genetic relationships in selection decisions. Theor Appl Genet 91:421–431

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084–1097

Bulmer M (1971) The stability of equilibria under selection. Heredity 27:157–162

Bulmer M (1980) The mathematical theory of quantitative genetics. Oxford University Press, New York

Butler D, Cullis B, Gilmour A, Gogel B (2009) {ASReml}-R reference manual

Comstock RE, Robinson HF, Harvey PH (1949) A breeding procedure designed to make maximum use of both general and specific combining ability. Agron J 41:360–367

de los Campos G, Sorensen D, Gianola D (2015) Genomic heritability: what is it? PLoS Genet 11:e1005048

Duvick DN (1984) Chapter 2, genetic contributions to yield gains of US Hybrid Maize, 1930 to 1980. In: Genetic contributions to yield gains of five major crop plants, ASA, CSSA, 677 South Segoe Road

Duvick DN, Smith JSC, Cooper M (2004) Long-term selection in a commercial hybrid maize breeding program. Plant Breed. Rev. J Janick Ed Vol 24 Part 2 Long Term Sel. Crops Anim. Bact., Wiley, New York, pp 109–151

Eberhart SA (1964) Least squares method for comparing progress among recurrent selection methods 1. Crop Sci 4:230–231

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longmans Green, Harlow, Essex, UK

Falke KC, Flachenecker C, Melchinger AE, Piepho H-P, Maurer HP et al (2007a) Temporal changes in allele frequencies in two European F(2) flint maize populations under modified recurrent full-sib selection. TAG Theor Appl Genet Theor Angew Genet 114:765–776

Falke KC, Maurer HP, Melchinger AE, Piepho H-P, Flachenecker C et al (2007b) Linkage disequilibrium in two European F(2) flint maize populations under modified recurrent full-sib selection. TAG Theor Appl Genet Theor Angew Genet 115:289–297

Felsenstein J (1965) The effect of linkage on directional selection. Genetics 52(2):349–363

Feng L, Sebastian S, Smith S, Cooper M (2006) Temporal trends in SSR allele frequencies associated with long-term selection for yield in maize. Maydica 51:293–300

Fischer S, Möhring J, Schön CC, Piepho H-P, Klein D et al (2008) Trends in genetic variance components during 30 years of hybrid maize breeding at the University of Hohenheim. Plant Breed 127:446–451

Fischer S, Möhring J, Maurer HP, Piepho H-P, Thiemt E-M et al (2009) Impact of genetic divergence on the ratio of variance due to specific vs. general combining ability in winter triticale. Crop Sci 49:2119–2122

Flachenecker C, Frisch M, Falke KC, Melchinger AE (2006) Trends in population parameters and best linear unbiased prediction of progeny performance in a European F2 maize population under modified recurrent full-sib selection. Theor Appl Genet 112:483–491

Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES 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

Gerke JP, Edwards JW, Guill KE, Ross-Ibarra J, McMullen MD (2015) The genomic impacts of drift and selection for hybrid performance in maize. Genetics 201:1201–1211

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

Gonen S, Battagin M, Johnston SE, Gorjanc G, Hickey JM (2017) The potential of shifting recombination hotspots to increase genetic gain in livestock breeding. Genet Sel Evol 49:55

Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES et al (2009) A first-generation haplotype map of maize. Science 326:1115–1117

Gorjanc G, Gaynor RC, Hickey JM (2018) Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. Theor Appl Genet 131:1953–1966

Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ et al (2014) Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet 95:535–552

Hallauer AR, Darrah LL (1985) Compendium of recurrent selection methods and their application. Crit Rev Plant Sci 3:1–33

Hill WG (1981) Estimation of effective population size from data on linkage disequilibrium 1. Genet Res 38:209–216

Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. Genet Res 8:269–294

Hospital F, Chevalet C (1996) Interactions of selection, linkage and drift in the dynamics of polygenic characters. Genet Res 67:77–87

Inghelandt DV, Melchinger AE, Lebreton C, Stich B (2010) Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor Appl Genet 120:1289–1299

Jensen J, Su G, Madsen P (2012) Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. BMC Genet 13:44

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC et al (2017) Improved maize reference genome with single-molecule technologies. Nature 546:524–527

Labate JA, Lamkey KR, Lee M, Woodman WL (1999) Temporal changes in allele frequencies in two reciprocally selected maize populations. Theor Appl Genet 99:1166–1178

Lehermeier C, de los Campos G, Wimmer V, Schön C-C (2017) Genomic variance estimates: with or without disequilibrium covariances? J Anim Breed Genet 134:232–241

Lush JL (1937) Animal breeding plans. Iowa State College Press, Iowa

Lynch M, Walsh B (1999) Evolution and selection of quantitative traits, Sunderland

MacLeod IM, Meuwissen THE, Hayes BJ, Goddard ME (2009) A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. Genet Res 91:413–426

Meuwissen TH (1997) Maximizing the response of selection with a predefined rate of inbreeding. J Anim Sci 75:934–940

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Nei M (1975) Molecular population genetics and evolution. Front Biol 40:I-288

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583–590

Pérez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495

Peripolli E, Munari DP, Silva MVGB, Lima ALF, Irgang R et al (2017) Runs of homozygosity: current knowledge and applications in livestock. Anim Genet 48:255–271

R Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Reif JC, Gumpert F-M, Fischer S, Melchinger AE (2007) Impact of interpopulation divergence on additive and dominance variance in hybrid populations. Genetics 176:1931–1934

Rincent R, Nicolas S, Bouchet S, Altmann T, Brunel D et al (2014) Dent and Flint maize diversity panels reveal important genetic potential for increasing biomass production. TAG Theor Appl Genet Theor Angew Genet 127:2313–2331

Russell WA (1991) Genetic improvement of maize yields. Adv Agron 46:245–298

Rutkoski J (2018) Estimation of realized rates of genetic gain and indicators for breeding program assessment. bioRxiv 409342

Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23:23–35

Sorensen D, Fernando R, Gianola D (2001) Inferring the trajectory of genetic variance in the course of artificial selection. Genet Res 77:83–94

Speed D, Balding DJ (2014) MultiBLUP: improved SNP-based prediction for complex traits. Genome Res 24:1550–1557

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM et al (2007) Recent human effective population size estimated from linkage disequilibrium. Genome Res 17:520–526

Truntzler M, Ranc N, Sawkins MC, Nicolas S, Manicacci D et al (2012) Diversity and linkage disequilibrium features in a composite public/private dent maize panel: consequences for association genetics as evaluated from a case study using flowering time. Theor Appl Genet 125:731–747

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Waples RS, England PR (2011) Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. Genetics 189:633–644

Weir BS, Hill WG (1980) Effect of mating structure on variation in linkage disequilibrium. Genetics 95:477–488

Woolliams JA, Berg P, Dagnachew BS, Meuwissen THE (2015) Genetic contributions and their optimization. J Anim Breed Genet 132:89–99

Wright S (1931) Evolution in mendelian populations. Genetics 16:97–159

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK et al (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569

Chapter 2 Genomic prediction with a maize collaborative panel: identification of genetic resources to enrich elite breeding programs

This chapter has been published in the peer-reviewed journal Theoretical and Applied Genetics in 2020. The electronic version of this article is available on the publisher website:

https://link.springer.com/article/10.1007/s00122-019-03451-9

Theoretical and Applied Genetics (2020) 133:201–215
https://doi.org/10.1007/s00122-019-03451-9

ORIGINAL ARTICLE

Genomic prediction with a maize collaborative panel: identification of genetic resources to enrich elite breeding programs

Antoine Allier[1,2] · Simon Teyssèdre[2] · Christina Lehermeier[2] · Alain Charcosset[1] · Laurence Moreau[1]

Received: 20 May 2019 / Accepted: 28 September 2019 / Published online: 8 October 2019

Abstract

Key message **Collaborative diversity panels and genomic prediction seem relevant to identify and harness genetic resources for polygenic trait-specific enrichment of elite germplasms.**

Abstract In plant breeding, genetic diversity is important to maintain the pace of genetic gain and the ability to respond to new challenges in a context of climatic and social expectation changes. Many genetic resources are accessible to breeders but cannot all be considered for broadening the genetic diversity of elite germplasm. This study presents the use of genomic predictions trained on a collaborative diversity panel, which assembles genetic resources and elite lines, to identify resources to enrich an elite germplasm. A maize collaborative panel (386 lines) was considered to estimate genome-wide marker effects. Relevant predictive abilities (0.40–0.55) were observed on a large population of private elite materials, which supported the interest of such a collaborative panel for diversity management perspectives. Grain-yield estimated marker effects were used to select a donor that best complements an elite recipient at individual loci or haplotype segments, or that is expected to give the best-performing progeny with the elite. Among existing and new criteria that were compared, some gave more weight to the donor–elite complementarity than to the donor value, and appeared more adapted to long-term objective. We extended this approach to the selection of a set of donors complementing an elite population. We defined a crossing plan between identified donors and elite recipients. Our results illustrated how collaborative projects based on diversity panels including both public resources and elite germplasm can contribute to a better characterization of genetic resources in view of their use to enrich elite germplasm.

Introduction

Successful plant-breeding programs rely on balanced efforts between short-term goals to develop competitive cultivars and the maintenance of a broad genetic pool to guarantee long-term progress. In practice, as newer lines and varieties have been mainly derived from intercrosses of existing elite lines, genetic improvement has been accompanied by a narrowing of the elite germplasm genetic diversity in several crops (Jenkins 1978; Mikel and Dudley 2006; Rauf et al. 2010). For instance, Allier et al. (2019a) observed a significant reduction in the genetic diversity over time in a private maize breeding pool with chromosome regions showing critical lack of genetic variation. These results supported previous observations in different maize breeding programs (Feng et al. 2006; Gerke et al. 2015). Low genetic diversity may restrict breeding potential to address new constraints related to climate change and changes in agronomical practices to respond to social demands (e.g., low input farming Fess et al. 2011 for a review). In this context, harnessing appropriate genetic resources to enhance the potential of breeding programs is a key factor of long-term success.

A broad range of candidate genetic resources are available to plant breeders. The following considerations are focused on maize but can be generalized to many other plant species. A maize breeder working for a given climatic zone has access to a large amount of potential genetic resources that

Communicated by Benjamin Stich.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00122-019-03451-9) contains supplementary material, which is available to authorized users.

✉ Laurence Moreau
laurence.moreau@inra.fr

[1] GQE - Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France

[2] RAGT2n, Genetics and Analytics Unit, 12510 Druelle, France

require more or less improvement for adaptation or performance, i.e., pre-breeding, before being integrated into the elite germplasm of interest. Considering first the closest to this elite germplasm, original favorable alleles may be found in elite breeding programs of the company which target other environments, in competitor lines obtained from commercial exchanges or in material derived from competitor hybrid cultivars using reverse breeding (Smith et al. 2008). Secondly, genetic variability can be accessed in US commercial inbred lines with expired plant variety protection act (ex-PVPA). Ex-PVPAs show about 20 years of genetic performance gap compared to current elite material but have been identified to present interesting genetic variability absent from current elite material (Mikel and Dudley 2006; Nelson et al. 2008; Kurtz et al. 2016). Thirdly, genetic variability can be accessed among and within landraces and the first inbred lines that have been created in early steps of hybrid breeding. Such genetic resources are maintained and are accessible in collections. Several studies characterized the diversity available in maize landraces and derived lines (Rebourg et al. 2001; Gauthier et al. 2002; Gouesnard et al. 2017). Despite a strong genetic performance gap accumulated since the beginning of hybrid breeding 60 years ago, landraces have been shown to represent interesting sources of adaptation alleles that can be harnessed using doubled-haploid technology (Strigens et al. 2013; Hellin et al. 2014; Melchinger et al. 2017; Böhm et al. 2017; Mayer et al. 2017; Brauner et al. 2019). Finally, breeders can exploit the genetic variability present in exotic material and derived inbred lines (Pollak and Salhuana 2001; Salhuana and Pollak 2006). This last source of original alleles has been depicted in several studies (e.g., Warburton et al. 2005; Wu et al. 2016) but requires more pre-breeding investments for improvement and adaptation to temperate conditions.

Regarding this large number of candidate materials, methods to predict the most interesting genetic resources to enrich elite recipient(s) in new favorable alleles are needed. In case of traits determined by major genes, causal variants can be identified using genome-wide association studies (e.g., Millet et al. 2016) and candidate genetic resources carrying the favorable allele can be further introgressed into elite material using backcross or gene pyramiding approaches (Servin et al. 2004; Han et al. 2017). For polygenic traits enrichment, i.e., the enrichment for a trait determined by a large number of small effect loci, the identification of donors is more complex and should account for genome-wide variants. Genomic prediction (GP) has been widely implemented to complement the expensive phenotyping of candidates in elite breeding (Heslot et al. 2015; Crossa et al. 2017). In genomic prediction, genome-wide molecular marker effects are estimated using both phenotypes and genotypes on a training population (TP) and are used to predict the performances of genotyped individuals of the prediction population (PP) (Whittaker et al. 2000; Meuwissen et al. 2001). It has been also recently shown that GP trained across a broad diversity is promising to mine natural variation present in gene banks: *Triticum aestivum* L. (Crossa et al. 2016) and *Sorghum bicolor* L. (Yu et al. 2016) but also in DH libraries from European flint maize landraces (Brauner et al. 2018).

The interest of genetic resources for polygenic trait enrichment of an elite germplasm depends on the recipient elite material considered. Therefore, one appealing strategy is to calibrate prediction models on a population assembling both types of material: genetic resources and elite material, further referred to as collaborative diversity panel. The public diversity component may include founders of breeding pools, elite material recently released into public domain (ex-PVPA) and public breeding material, whereas the proprietary elite component would come from different private partners' elite breeding programs. The interest of using such collaborative diversity panels relies on a shared investment between partners, a broader diversity covered and potentially enabling more accurate phenotyping for traits difficult or expensive to measure in fields and that cannot be evaluated in routine in breeding populations. Furthermore, the collaborative combination of facilities and expertise improves the identification and spreading of traits. To our knowledge, no study investigated the interest of genomic prediction models trained on collaborative diversity panels to identify genetic resources to enrich an elite germplasm.

The identification of genetic resource(s) to complement and enrich elite recipient(s) can be grounded on different approaches proposed to select inbred parents based on parental performance and parental complementarity assessment. For a given pair of parents, Dudley (1984, 1987) proposed to subdivide biallelic quantitative trait loci (QTLs) into four genotypic classes (I, J, K, and L) and count the number of QTLs in each class (Table 1). Class I and class L correspond to QTLs where donor and recipient both carry the favorable and unfavorable allele, respectively. Class J and class K correspond to QTLs where parents are polymorphic. In the following, I, J, K, and L refer to the proportion of QTLs in these classes. When considering a donor × elite recipient cross, K (respectively J) is the proportion of QTLs for which the donor carries the favorable (respectively unfavorable)

Table 1 Classes of loci according to Dudley (1984) and Bernardo (2014). For each biallelic locus the favorable allele is denoted + and the unfavorable allele –

Class of loci	Recipient (Inbred)	Donor (Inbred)
I	+/+	+/+
J	+/+	–/–
K	–/–	+/+
L	–/–	–/–

allele and the recipient the alternative allele. In this context, one is seeking the donor maximizing K, while minimizing J. Initially, Dudley (1984, 1987) used phenotypic evaluations of single cross and per se values of parents to measure heterosis as an indicator of the complementarity between parents to estimate I, J, K and L. With high density genotyping and assuming that each locus is in linkage with QTLs, genome-wide estimated marker effects can replace phenotypic evaluations to predict I, J, K, and L as done by Bernardo (2014).

However, individual estimated marker effects do not directly reflect the effect of QTLs. In addition, the limited number of recombination events while deriving the progeny from donor × elite recipient crosses, i.e., haplotype block inheritance, needs to be accounted for when evaluating the complementarity between parents. It has been suggested to integrate estimated marker effects on predefined haplotype segments to identify the most complementary parents using the optimal haploid value (OHV, Daetwyler et al. 2015). The OHV of a biparental cross aims at predicting the best doubled-haploid progeny that can be produced from this cross. Goiffon et al. (2017) generalized the OHV to a population, namely optimal population value (OPV), that predicts the performance of the best possible doubled-haploid progeny produced by a given population after an infinite number of generations. In a donor identification context, OHV can be considered to identify the donor that complements at best an elite recipient and OPV to identify donor(s) that complement at best an elite recipient population. However, neither OHV nor OPV accounts for the recombination rate in donor × elite recipient cross(es).

The interest of a donor to complement an elite recipient can be further evaluated using the usefulness criterion (UC, Schnell and Utz 1975), that corresponds to the expected genetic gain from the cross between the elite recipient and the donor. The UC of a cross is defined as a sum of the expected mean progeny performance μ and the expected genetic gain, considering that the best progeny is selected with an intensity i, an accuracy h, for a trait-specific genetic standard deviation in progeny σ, resulting in $UC = \mu + ih\sigma$. While μ can be easily estimated as the mean of parental performances, the estimation of progeny variance (σ) is not straightforward. Bernardo et al. (2006) estimated the genetic variance using either phenotypic data or QTL detection results. Nonetheless, results based on QTL detection were not convincing, likely illustrating the limits of progeny variance estimates based on QTL detection approaches. To overcome this limitation, genome-wide estimated marker effects can be used to estimate the progeny variance. Bernardo (2014) and Mohammadi et al. (2015) proposed to use genome-wide estimated marker effects and stochastic simulations of progeny to estimate progeny variance. Recently, Lehermeier et al. (2017b) proposed an efficient algebraic formula accounting for the recombination rate and linkage disequilibrium between QTLs in parental lines to predict progeny mean and variance for biparental crosses. This was further extended to heterozygote parents and crosses implying up to four parents (Allier et al. 2019b).

In this study, we considered genome-wide marker effects estimated on the maize "Amaizing" dent diversity panel showing a continuum from old accessions to elite material (Rio et al. 2019) to help the identification of donor(s)

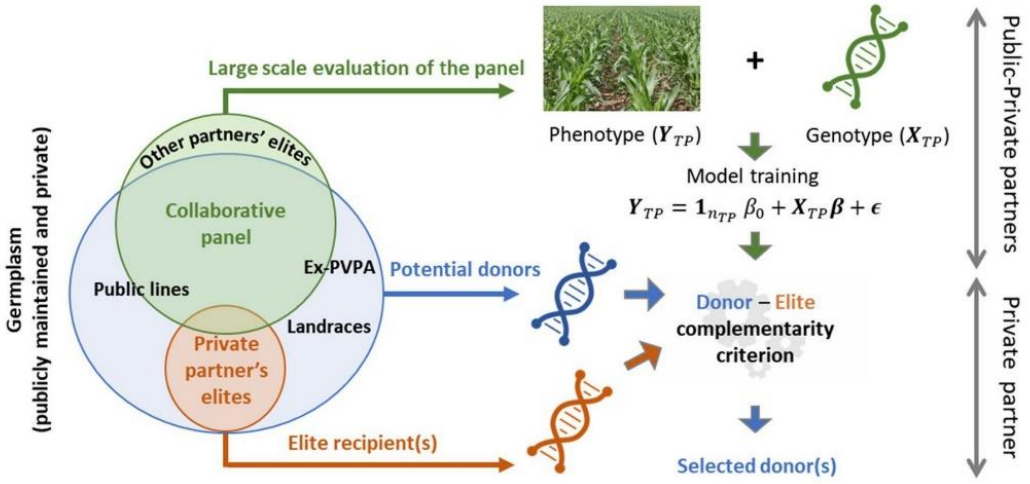


Fig. 1 Illustration of the suggested strategy to predict the interest of potential donor(s) to complement elite recipient(s). Collaboration between public and several private partners is presented on the top part and internal application by a given private partner is presented on the bottom part. The model training equation refers to Eq. 1. In this study, we considered as potential donors the lines evaluated in the collaborative panel, excluding the other partners' elites

to enrich an elite germplasm for grain yield (Fig. 1). We first evaluated the predictive ability of the estimated marker effects on a private elite breeding material covering 13 years of breeding. We then applied different criteria based on the estimated marker effects to evaluate the interest of candidate donors from the "Amaizing" dent panel. These criteria account differently for candidate donor performance and complementarity to the elite recipient. After considering the selection of a single donor to enrich a given elite recipient, we extended the approach to the identification of a set of donors to enrich an elite population. We further proposed a crossing plan between identified donors and elite recipients to maximize the expected short-term genetic gain considering doubled-haploid progeny. These objectives were addressed and illustrated sequentially considering as elite material representative Iodent maize inbred lines from a private breeding program (RAGT2n).

Materials and methods

Collaborative panel

We worked with the "Amaizing" dent panel presented in Rio et al. (2019), composed of 389 dent maize lines genotyped with the MaizeSNP50 Illumina® BeadChip (Ganal et al. 2011). After quality control (line call rate ≥ 0.8, line heterozygosity rate ≤ 0.1, marker call rate ≥ 0.9 and marker heterozygosity rate ≤ 0.15), 386 dent lines genotyped for 40,478 single nucleotide polymorphism markers (SNPs) were considered. The positions of the markers on a genetic map were obtained by predicting genetic positions from physical positions (Jiao et al. 2017) using a spline-smoothing interpolating procedure described in Bauer et al. (2013) and the consensus dent genetic map of Giraud et al. (2014). In the following, for all genetic material considered, the heterozygous marker genotypes were set as missing genotypes and all missing genotypes were imputed using Beagle v4 (Browning and Browning 2007, 2016). These 386 lines (training population, TP) were evaluated for test-cross performances on a single Flint tester (UH007) in seven locations for one year. In this study we considered Lsmeans for each line over the seven locations for grain yield standardized at 15% of grain moisture (qx/ha), male flowering time and grain moisture (Rio et al. 2019). This panel showed several interesting properties regarding our objectives. The 386 dent lines were from different origins corresponding to three main dent heterotic groups: 57 Iodent, 82 Stiff Stalk and 199 Other dent public lines and 27 Iodent, 16 Stiff Stalk and 5 Other dent private elite lines (Rio et al. 2019). Good prediction accuracies for within- and between-group genomic predictions were obtained (Rio et al. 2019). The panel includes 338 lines of public origin consisting (1) in founders of maize dent groups (Iodent, Stiff Stalk, and Other dent) (2) lines derived from landraces (3) elite material released into public domain (ex-PVPA) and (4) breeding material derived out of them by public institutes (principally CSIC, CIAM, Hohenheim University and INRA). Ex-PVPA and derived material ensured a continuum toward the current private elite germplasm represented by 48 lines provided by seven breeding companies and also included in the panel.

Genomic prediction model

In order to estimate unbiased estimator of progeny variance (PMV; posterior mean variance) presented in Lehermeier et al. (2017a,b), phenotypes and genotypes of the TP were used to estimate genome-wide marker effects in a Bayesian Ridge Regression implemented in BGLR (Pérez and de los Campos 2014):

$$\boldsymbol{Y}_{\mathrm{TP}} = 1_{n_{\mathrm{TP}}}\beta_0 + \boldsymbol{X}_{\mathrm{TP}}\boldsymbol{\beta} + \epsilon, \quad (1)$$

where $\boldsymbol{Y}_{\mathrm{TP}}$ is the n_{TP}-dimensional vector of phenotypes in the TP with $n_{\mathrm{TP}} = 386$, $1_{n_{\mathrm{TP}}}$ is a n_{TP}-dimensional vector of ones, β_0 is the TP mean performance, $\boldsymbol{X}_{\mathrm{TP}}$ is the $[n_{\mathrm{TP}} \times M]$-dimensional matrix of reference allele counts of the TP coded in 0 or 2, $\boldsymbol{\beta}$ is a M-dimensional vector of random marker effects and ϵ the n_{TP}-dimensional vector of random residuals. In this Bayesian model, identical and independent prior Gaussian distributions were assigned to marker effects $N(0, \boldsymbol{I}\sigma^2_\beta)$ and residual terms $N(0, \boldsymbol{I}\sigma^2_\epsilon)$. Scaled inverse χ^2 distributions were assigned to the marker effects and residual variances (σ^2_β and σ^2_ϵ respectively). Hyperparameters for the scaled inverse χ^2 prior distributions were defined according to default settings in BGLR, resulting in sparsely informative priors. Samples from the posterior distributions were generated from a Markov chain Monte Carlo (MCMC) algorithm implemented in BGLR. We used 20,000 iterations where the first 5000 were discarded as burn-in. One-fifth of the samples were kept for posterior inference resulting in a total of $S = 3000$ samples. The M-dimensional vector of posterior mean marker effects ($\hat{\boldsymbol{\beta}}$) was derived as the mean of marker effects estimated in all samples of the thinned post burn-in MCMC chain as: $\hat{\boldsymbol{\beta}} = S^{-1}\hat{\boldsymbol{\beta}}^{(S)}1_S$, where, $\hat{\boldsymbol{\beta}}^{(S)}$ is a $[M \times S]$-dimensional matrix of marker effect samples with the mth row representing marker m ($m \in [1, M]$) and the sth column the sample s ($s \in [1, S]$). After fitting the model using the whole TP (386 lines), we considered only the 338 lines of public origin as candidate donors.

Elite material

To evaluate the relevance of the model and the TP to predict the interest of a donor relatively to an elite population, we firstly evaluated its predictive ability in an elite population of 594 inbred lines (prediction population, PP). The 594 inbred

lines were generated between 2004 and 2016 in an early Iodent grain maize breeding program (RAGT2n). Lines were evaluated on private Flint testers for grain yield standardized at 15% of grain moisture (qx/ha), male flowering time and grain moisture (more details in File S1). From these data, we estimated best linear unbiased estimators of their general combining ability (GCA) (more details in File S1). These lines were genotyped with the MaizeSNP50 Illumina® BeadChip and the same set of SNPs as for the calibration set was kept after imputation of missing values. Posterior mean marker effects obtained from Eq. 1 were used to predict the genomic estimated breeding values (GEBVs) of the PP. The predictive ability was defined as the correlation between the predicted GEBVs and estimated GCA for individuals in the PP (more details in File S1). As an illustrative elite population to be complemented by donors, we considered a total of 10 elite Iodent lines (named E1 to E10) from the same breeding pool as the PP.

Criteria to select a single donor to enrich a unique elite recipient

We compared different criteria for selecting a donor among a population of candidate donors, which will most likely complement a given elite recipient for a polygenic trait. Following Dudley (1987) and Bernardo (2014), we considered the genotypic information and the sign of posterior mean marker effects to estimate the proportions of QTLs I, J, K, and L (Table 1) for each donor × elite recipient cross. We ranked the donors depending on the fraction of new favorable alleles initially absent from the elite recipient (criterion K). In addition, we also considered the risk of bringing unfavorable alleles using the ratio K/J.

To account for differences in marker effect estimates and prevent consequences of the inaccuracy in estimations, one can integrate individual marker effects across haplotype segments. This enables to evaluate the complementarity of parents on predefined haplotype segments instead of single loci. The posterior mean marker effects are summed across haplotypes to define the haplotypic estimated breeding value matrix (HEBV):

$$\boldsymbol{HEBV} = \left(\boldsymbol{X} \circ 1_N \hat{\boldsymbol{\beta}}^T\right)\boldsymbol{Z}, \tag{2}$$

where $\boldsymbol{X}$ is the $[N \text{x} M]$-dimensional genotyping matrix coded in 0 or 2 of the candidate donors and elite recipients, 1_N is an N-dimensional vector of ones, $\hat{\boldsymbol{\beta}}$ is the M-dimensional vector of marker effects estimated in Eq. 1, $\circ$ denotes the entry-wise product. $\boldsymbol{Z}$ is a $[M \times nH]$-dimensional design matrix of 0 and 1, where nH is the total number of haplotype segments considered. Elements $Z(j,h)$, $\forall j \in [1,M], h \in [1,nH]$ indicate whether locus j is in haplotype segment h ($Z(j,h) = 1$) or not ($Z(j,h) = 0$). It results that HEBV is an $[N \times nH]$-dimensional matrix with N lines in rows and nH haplotype segments in columns. Elements of HEBV are the estimated genetic effects for each haplotype segment in each individual. In OHV defined by Daetwyler et al. (2015), nSeg non-overlapping continuous segments per chromosome are considered in the design matrix $\boldsymbol{Z}$ resulting in HEBV^{OHV} (Eq. 2). The authors suggested to consider a relatively small number of segments per chromosome (nSeg = 1 to 3) to reflect the small number of recombination events occurring per meiosis. However, when selecting donor × elite recipient crosses, one is interested in evaluating the complementarity between parents at a finer scale. Furthermore, recombination events likely take place genome-wide rather than at fixed locations defined by nSeg parameter and chromosome size. We considered different nSeg in OHV (nSeg = 2 and nSeg = 100). We also extended the OHV to the criterion H considering overlapping segments in the design matrix $\boldsymbol{Z}$ resulting in HEBV^{H} (Eq. 2). In the illustration, we considered haplotype segments of 100 SNPs with a 20 SNPs increment. The OHV and H of the cross between an inbred donor and an inbred elite recipient are defined as:

$$\lambda \sum_{h=1}^{nH} \max\left\{\text{HEBV}_{(\text{donor},h)}, \text{HEBV}_{(\text{elite},h)}\right\}, \tag{3}$$

where HEBV stands for the matrix HEBV^{OHV} or matrix HEBV^{H} for OHV or H criterion, respectively. The scaling parameter λ = number of loci per segment/number of loci increment aims at correcting for the overlapping segments and that on average a locus is taken into account several times. Note that $\lambda = 1$ in case of OHV. With this, criteria H and OHV are comparably scaled and represent an expectation of the performance of the best doubled-haploid line that can be derived from the cross donor × elite recipient when the number of progeny is sufficiently large and recombination events take place as assumed for the construction of the HEBV matrix. The visualization of the HEBV^{H} matrix along the genome for a candidate donor and an elite recipient also enables to identify chromosomic regions where the donor is more or less performant than the elite (as illustrated in Fig. 2, with $H(+)$ and $H(-)$ regions, respectively). $H(+)$ regions can be further considered to target specific recombination events in progeny of the cross donor × recipient (as proposed by Bernardo 2017). Similarity, $H(-)$ regions, i.e., risk of introgressing unfavorable alleles, can be considered as a secondary criterion to distinguish candidate donors showing similar H values.

Furthermore, we can account for recombination rate (i.e., genetic position of markers) to evaluate the interest of a donor using the usefulness criterion (UC) of the cross donor × elite recipient. Considering the posterior mean estimated marker effects ($\hat{\boldsymbol{\beta}}$), no dominance and no epistasis, the progeny mean can be estimated as the mean of

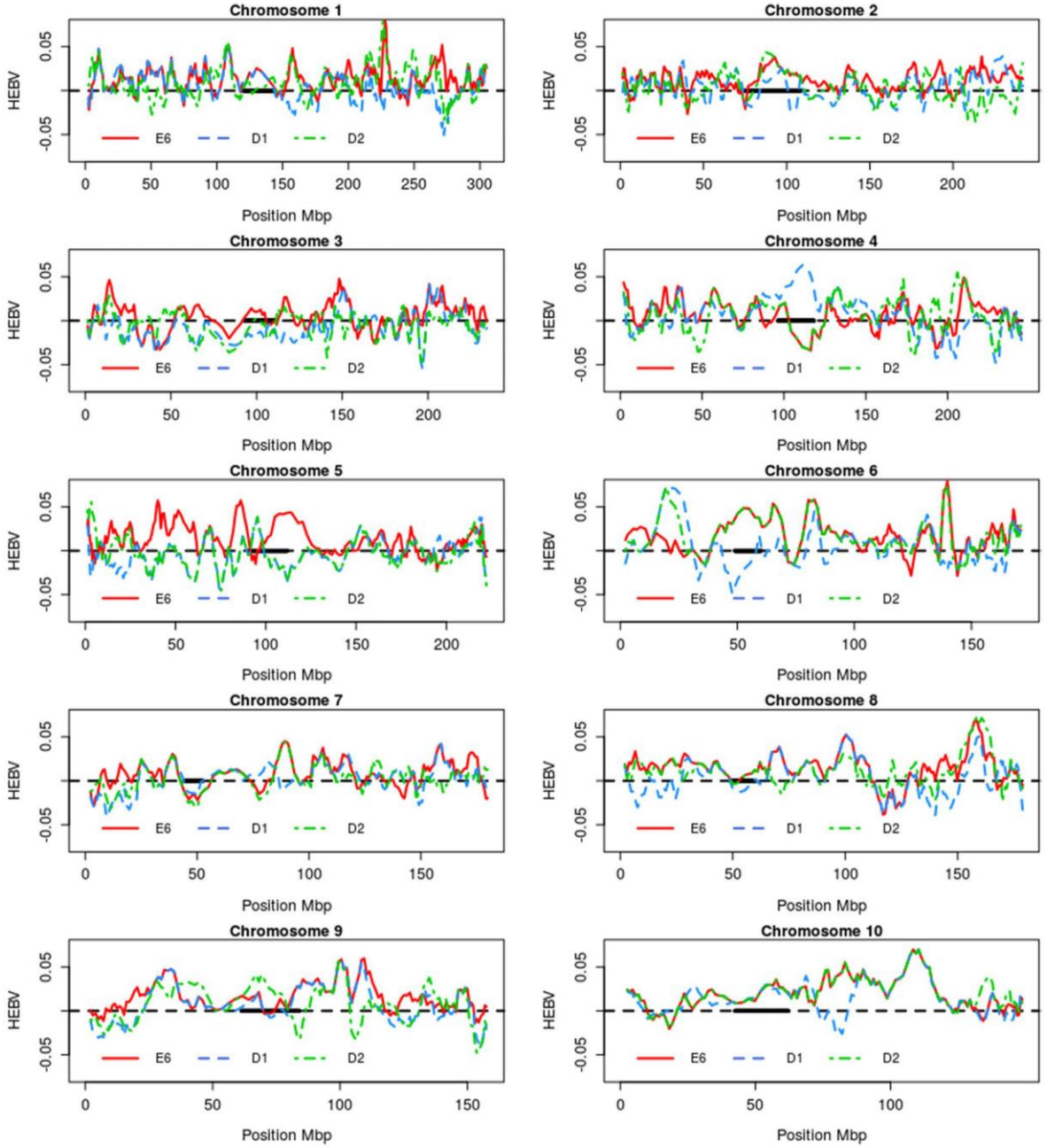


Fig. 2 Visualization of the haploid estimated breeding value (HEBV) with overlapping segments (100 SNP with 20 SNP increment) along the genome considering elite line E6 and two donors (D1 and D2). The *x*-axis represents the mean physical position of each haplotype segments in Mbp. The centromere regions are represented in bold on the *x*-axis. This illustration highlights genomic regions where the donor was more performant than the elite recipient, called *H*(+) (e.g., centromeric region of chromosome 4 for D1), its opposite *H*(−) (e.g., chromosome 8: 0–30 Mbp for D1) and regions where both elite and donor showed the same HEBV (e.g., centromeric region of chromosome 4 for D2)

the parental breeding values $\hat{\mu} = \frac{1}{2}\left(\boldsymbol{x}_d^T\hat{\boldsymbol{\beta}} + \boldsymbol{x}_r^T\hat{\boldsymbol{\beta}}\right)$, where $\boldsymbol{x}_d$ and $\boldsymbol{x}_r$ are the marker genotypes of donor and elite recipient, respectively. Using estimated marker effects in each MCMC sample $s \in [1, S]$, the posterior mean progeny variance accounting for linkage disequilibrium between markers can be derived as proposed by Lehermeier et al. (2017b): $\hat{\sigma}^2 = \frac{1}{S}\sum_{s=1}^{S}\hat{\boldsymbol{\beta}}^{(s)T}\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}^{(s)}$. The genotypic covariance matrix between loci genotypes in progeny $\boldsymbol{\Sigma}$ is defined for a population of DH lines derived from the F1 plant following Lehermeier et al. (2017b). This matrix relies on recombination rates between markers that were estimated using the Haldane function (Haldane 1919) and the genotype of parental lines. The estimated UC of all donor × elite recipient crosses was computed considering $h = 1$ (i.e., assuming within family selection accuracy of one) and different selection intensities (i):

$$\mathrm{UC} = \hat{\mu} + ih\hat{\sigma}. \quad (4)$$

A selection of the 5% most performant progeny ($i = 2.06$) was considered to represent a common selection intensity (UC_1). In addition, to give more weight to the progeny variance we considered an extreme selection of the 10^{-8}% most performant progeny ($i = 5.78$) corresponding to UC_2 that reflected a selection limit. We also considered the progeny variance ($\hat{\sigma}^2$) as a criterion to identify donors, further referred to as VarG.

Finally, we evaluated the ranking of 57 Iodent candidate donors to enrich the elite recipient depending on the criteria used: genotypic classes of loci (K and K/J), H, OHV (nSeg = 2 or 100), usefulness criteria UC_1, UC_2, and VarG. We considered two simpler criteria as benchmark: the GEBV of donors ($\boldsymbol{x}_d^T\hat{\boldsymbol{\beta}}$), that selects donors based on their predicted values, and the pairwise modified Roger's distance between donors and the elite recipient (MRD, Wright 1978), that selects donors based on their originality compared to the elite line.

Selection of a set of donors to enrich an elite population

Instead of a single elite recipient, breeders might want to enrich a whole elite population with new favorable alleles. In this context, we accounted for within elite lines variation to favor the selection of donors carrying favorable alleles not present yet in the population of elite lines. We built an iterative forward approach to identify a set of interesting donors from a population PopD of k candidate inbred donors (P_{21} to P_{2k}) complementing with favorable alleles a population PopE of n recipient inbred lines (P_{11} to P_{1n}). First we computed HEBV^H for all donors and elite recipients. Similarly as OHV (Daetwyler et al. 2015) has been extended to a population with the OPV criterion (Goiffon et al. 2017), we extended the criterion H (Eq. 3) to consider a population of elite recipients. For each donor in PopD, we computed:

$$H = \lambda \sum_{h=1}^{nH} \max\left\{\mathrm{HEBV}^H_{(\mathrm{donor},h)}, \max_{\mathrm{elite}\in\mathrm{PopE}}\left\{\mathrm{HEBV}^H_{(\mathrm{elite},h)}\right\}\right\}, \quad (5)$$

where λ is the scaling parameter defined in Eq. 3. In this context, H reflects the maximum doubled-haploid line performance expected after several generations of intercross and selection of the donor and elite lines in PopE. The donor maximizing H was included in PopE for the next iteration. We iterated this step to identify a set of donors complementing elite recipients with most of the favorable haplotype segments available. The number of donors in the set was determined considering the relative gain of introducing a new donor, into the population PopE composed of elites and previously selected donors. This rationale aimed at balancing the gain and the costs of selecting an additional donor. Finally, the usefulness criterion (UC_1) of all possible crosses between the selected donors and the elite lines $\{P_{11}, P_{12}, \ldots, P_{1n}\}$ was predicted. We selected for each donor the donor × elite recipient cross maximizing UC_1. An exemplary R script (R Core Team 2017) for the forward section of donors and UC computation is provided in File S2. To illustrate defined criteria, we considered the 57 publicly available Iodent as diverse candidate donors and 10 private elite Iodent recipients (E1 to E10). When considering a single elite recipient, results are shown for the elite line E6 that is representative of the ten private lines (E1 to E10).

Results

Relevance of the predictive model for the elite lines

We observed a positive correlation r = 0.404 between the performances predicted using marker effects estimated on the collaborative panel and the observed performances of 594 RAGT2n elite lines for grain yield (GY, Fig. 3a). The estimated marker effects predicted partly the realized genetic improvement over the 13 years considered. When focusing on lines derived a same year, i.e., on average 46 lines for grain yield, the predictive ability became very variable across years for all traits (e.g., $r = -0.062$ to $r = 0.722$ for GY, with a mean value 0.305, File S1 Table S1). The predictive ability of RAGT2n elite lines was slightly lower when considering the training population without the 48 elite private lines ($r = 0.377$, File S1 Table S1). For traits under stabilizing selection such as male flowering time (MF) or grain moisture (GM), the predictions were more accurate with $r = 0.495$ and $r = 0.550$ (Fig. 3b, c), respectively.

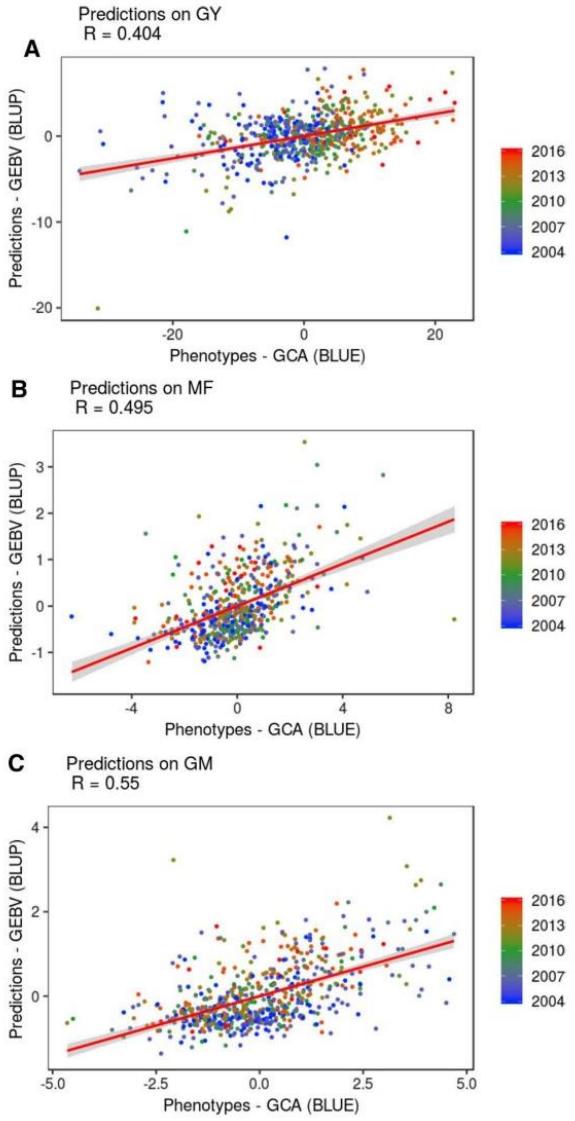


Fig. 3 Prediction of GCA of elite RAGT Iodent lines using marker effects estimated on the "Amaizing" dent panel (386 dent lines: 338 public and 48 private partner lines). Each dot corresponds to one elite line and the color corresponds to the first testing year of the line. **a** Grain yield (594 lines), **b** male flowering time (539 lines), **c** grain moisture (594 lines). Notice that both GEBV BLUPs and GCA BLUEs were centered

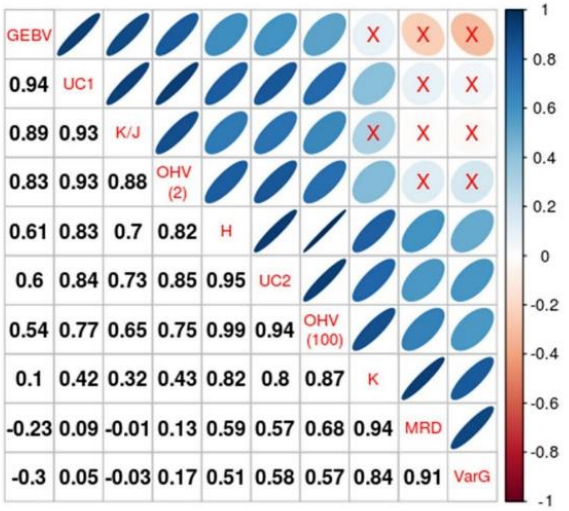


Fig. 4 Linear Pearson correlations (r) between different criteria. Criteria are ordered according to their correlation with the GEBV of the donor. Nonsignificant correlations at risk level 1% (i.e., $\alpha \geq 0.01$) are marked with a red cross. K is the percentage of loci in class K and K/J the ratio of percentage of loci in class K over those in class J. OHV (2) and OHV (100) are the optimal haploid value (Eq. 3) of the cross donor×E6 considering two (nSeg = 2) and 100 segments (nSeg = 100) per chromosome, respectively. H criterion is defined in Eq. 3. UC_1 and UC_2 are the usefulness criterion (Eq. 4) of the cross donor×E6 for selection intensity $i = 2.06$ and $i = 5.78$, respectively. VarG is the progeny variance of the cross donor×E6 and MRD is the modified Roger's distance E6-donor

Criteria to select a single donor to enrich a unique elite recipient

The 10 different criteria ranked the Iodent candidate donors in different ways relatively to their interest to enrich the Iodent elite recipient E6. The candidate donor performance (GEBV) and the complementarity between candidate donor and E6 at "neutral" loci (MRD), were negatively correlated ($r = -0.23$, nonsignificant at 1% risk level, i.e., $\alpha \geq 0.01$) (Fig. 4). We observed a gradient from criteria ranking candidate donors similarly as donor GEBV to criteria more correlated with the MRD between donor and E6. For instance, the ratio K/J that estimated the ratio of new favorable alleles over new unfavorable alleles brought by the donor relatively to E6, the optimal haploid value (OHV) (nSeg = 2), and the usefulness criterion with 5% selection pressure (UC_1) of the donor×E6 crosses was significantly positively correlated with the donor GEBV ($r = 0.89$, $r = 0.83$ and $r = 0.94$, respectively). Interestingly, the haplotypic criterion H, the OHV (nSeg = 100) and the usefulness criterion with 10^{-8}% selection pressure (UC_2) showed lower but still positive correlations with almost all criteria and seemed to account for both donor performance and distance with the elite recipient (Fig. 4). When increasing the number of segments (nSeg) in OHV from 2 to 100, i.e., considering more recombination events, the correlation between OHV and donor GEBV decreased ($r = 0.83$ to $r = 0.54$), whereas the correlation between OHV and MRD donor–E6 increased ($r = 0.13$ to $r = 0.68$). OHV (nSeg = 100) also converged toward H ($r = 0.99$) and UC_2 ($r = 0.94$) which assumed that somehow

Table 2 Summary of the top five identified donors for each criterion and corresponding criterion values

	GEBV[a] (qx/ha)	UC_1[a] (qx/ha)	K/J	OHV (2)[a] (qx/ha)	H[a] (qx/ha)	UC_2[a] (qx/ha)	OHV (100)[a] (qx/ha)	K	MRD	VarG (qx^2/ha^2)
#1	**HMV5301**	**HMV5301**	**HMV5301**	**NQ508**	**HMV5502**	**HMV5502**	**HMV5502**	**UH_P072**	**UH_P072**	**UH_P072**
	(14.679)	(19.249)	(1.128)	(15.931)	(23.418)	(35.621)	(26.828)	(0.136)	(0.560)	(22.120)
#2	Lo1242	**HMV5502**	**HMV5502**	Lo1180	**NQ508**	Lo1180	**NQ508**	F1819	F1819	Lo1180
	(12.161)	(19.108)	(1.023)	(15.885)	(22.212)	(34.804)	(25.340)	(0.134)	(0.541)	(20.677)
#3	PHG83	Lo1242	SGI061	**HMV5301**	Lo1242	Lo1242	Lo1242	PHG71	UH_P017	UH_P017
	(12.012)	(19.048)	(1.014)	(15.704)	(21.940)	(34.181)	(25.330	(0.134)	(0.536)	(19.764)
#4	**NQ508**	PHG83	**NQ508**	**HMV5502**	UH_P084	B103	Lo1180	**HMV5502**	PHG71	**HMV5502**
	(11.739)	(18.239)	(1.000)	(15.38)	(21.723)	(32.556)	(24.719)	(0.133)	(0.532)	(19.731)
#5	SGI061	**NQ508**	PHG83	Lo1242	Lo1180	**HMV5301**	PHG71	Lo1242	UH_P075	UH_P089
	(11.174)	(18.184)	(0.991)	(15.197)	(21.671)	(32.475)	(24.679)	(0.131)	(0.531)	(18.661)

Donors highlighted in bold are the best donors for at least one criterion

[a]These criteria are centered on training population (TP) mean performance $\hat{\beta}_0 = 83.299$ qx/ha

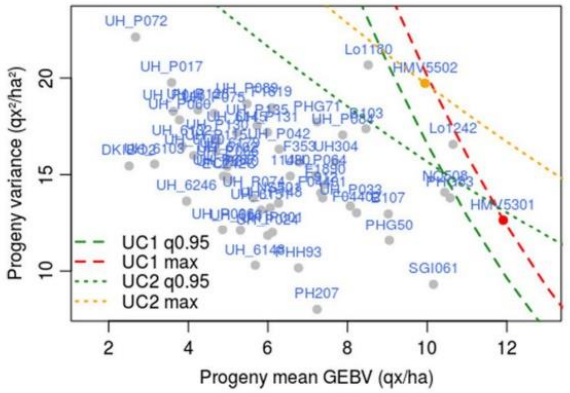


Fig. 5 Expected progeny variance (VarG, in qx^2/ha^2) and progeny mean GEBV (in qx/ha) of crosses donor×E6. Dotted lines represent isoclines for usefulness criteria UC_1 and UC_2 regarding the 95% quantile and the observed maximum. UC_1 corresponds to the expected mean performance of the 5% most performant progeny, which represents a common selection intensity. UC_2 corresponds to the expected mean performance of the 10^{-8}% most performant progeny, which reflects a selection limit. The red dot corresponds to the donor maximizing UC_1 and orange dot to the donor maximizing UC_2

the number of recombination events approached infinity (Fig. 4). Finally, the proportion of loci K and the predicted variance in progeny of the cross donor–E6 (VarG) were significantly positively correlated with the originality of the donor: MRD donor–E6 ($r = 0.94$ and $r = 0.91$, respectively).

When considering the best donor selected, we observed that some criteria identified the same donor (Table 2), consistent with the observed correlations across the 57 candidate donors. On the one side, GEBV, UC_1, and K/J selected the donor HMV5301 (GEBV = 14.679 qx/ha, $UC_1 = 19.249$ qx/ha, K/J = 1.128, Table 2). The criterion OHV (nSeg = 2) selected the candidate donor NQ508 (OHV (nSeg = 2) = 15.931 qx/ha, Table 2), and HMV5301 was considered as the third best donor. The criteria H, UC_2, and OHV (nSeg = 100) selected the donor HMV5502 ($H = 23.418$ qx/ha, $UC_2 = 35.621$ qx/ha, OHV (nSeg = 100) = 26.828 qx/ha, Table 2). The selection of HMV5502 by UC_2 instead of HMV5301 by UC_1 was explained by the fact that when crossed to E6 (GEBV = 9.143 qx/ha), HMV5502 (GEBV = 10.748 qx/ha) generated a lower progeny mean than HMV5301 (GEBV = 14.679 qx/ha) but generated more progeny variance (VarG = 19.731 qx^2/ha^2 and 12.656 qx^2/ha^2, respectively) (Fig. 5, Table 2). The criteria K, MRD and VarG yielded the identification of the most distant donor UH_P072 ($K = 0.136$, MRD = 0.560, VarG = 22.120 qx^2/ha^2, Table 2).

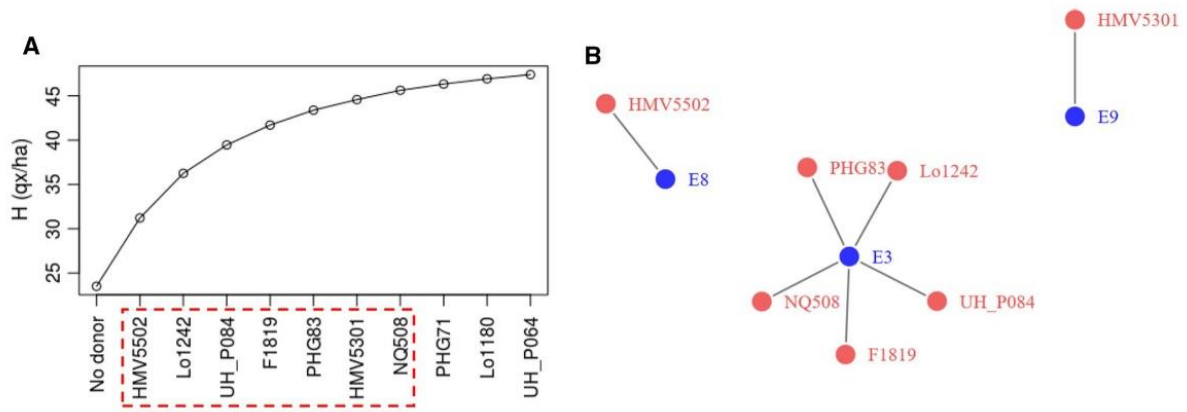


Fig. 6 Results of the stepwise selection of a set of donors using the criterion *H*. **a** Increase of *H* (Eq. 5) when selecting donors. The seven donors selected are marked in the red box. **b** Crossing plan that maximizes the UC_1 (Eq. 4) of crosses between each donor (red) with an elite recipient (blue). Gray links represent the biparental donor × elite recipient crosses

Selection of a set of donors to enrich an elite population

Considering the complementarity within elite recipients E1 to E10 during the forward selection of donors, we built up an ideotype population carrying most favorable haplotypes. The increase in the criterion *H*, i.e., the interest of selecting an additional donor, plateaued at around seven donors (Fig. 6a) and identified the donors HMV5502, Lo1242, UH_P084, F1819, PHG83, HMV5301, and NQ508. These donors were already identified as part of the five best donors using most of previously defined criteria to enrich E6 (Table 2). For instance, the candidate HMV5502 was selected by eight criteria, Lo1242 by seven criteria, NQ508 by six criteria and HMV5301 by five criteria (Table 2). Alternatively, some selected donors were rarely selected to enrich E6, such as F1819 selected by criteria K and MRD and UH_P084 selected only by the criterion *H* (Table 2). Note that donors maximizing the different criteria to enrich the recipient E6 were also selected in the forward approach, except UH_P072 that maximized criteria related to its genetic distance to E6 (Table 2). We further determined for each identified donor the biparental cross with an elite recipient that maximized the short-term genetic gain UC_1. The resulting crossing plan involved three elite lines with an intensive use of the best-performing elite line E3 in five out of seven crosses (Fig. 6b).

Discussion

Interest of collaborative diversity panels

We estimated marker effects across the "Amaizing" maize dent panel to identify genetic resources to enrich an elite germplasm in new favorable alleles for a polygenic trait. This approach relied on estimated marker effects that are assumed to be predictive for genetic resources and to have a certain predictive ability within the elite germplasm considered. It has been shown by cross-validations that marker effects estimated in this panel on a mixture of several dent groups predicted accurately individuals from one specific group (Rio et al. 2019). The same study also showed that prediction models trained over materials developed by public institutes could predict variation across the set of private elite lines obtained by different partner companies. In the present study, we observed that estimated marker effects were able to predict main differences in a larger serie of 594 RAGT elite Iodent lines covering 13 years of breeding for grain yield ($r=0.404$, Fig. 3a). Note that the observed predictive ability was quite high considering (1) the low heritability in the PP (GY, $\sqrt{h^2}=0.347$; MF, $\sqrt{h^2}=0.519$; GM, $\sqrt{h^2}=0.681$, File S1 Table S1) and (2) that the tester and locations used to evaluate the TP differed from those used to evaluate the PP. When considering only lines generated the same year, i.e., on average 46 lines for grain yield, predictive ability became very variable across years for all traits (e.g., GY, $r=-0.062$ to 0.722, File S1 Table S1). When excluding the elite private material (48 lines) from the training population, we observed only a slight loss of accuracy for grain yield ($r=0.404$ to 0.377, File S1 Table S1) and small changes for male flowering time ($r=0.495$ to 0.509, File S1 Table S1) and grain moisture ($r=0.550$ to 0.541, File S1 Table S1). These results may be explained by (1) the broad diversity covered by the public lines including lines directly derived from landraces, old elite material (ex-PVPA) and public breeding material and (2) the small fraction of elite private material in the panel (48 out of 386 lines). Despite predictions do not seem accurate enough for breeding perspectives within a given year, the "Amaizing" dent panel showed a stability

in prediction efficiency over larger time trends and material origins that appears promising to address the identification of donors complementary to elite material.

Criteria to select a single donor to enrich a unique elite recipient

The interest of a donor to enrich a specific recipient elite line relies on a balance between its genetic value, which determines the expected mean performance of the progeny, and its originality at QTLs, which contributes to the expected long-term genetic progress. Several criteria have been proposed to identify crosses between inbred lines based on the complementarity between parents at individual loci (Dudley 1984, 1987; Bernardo 2014), the complementarity between parents at haplotype segments (Daetwyler et al. 2015; Goiffon et al. 2017) and based on the expected progeny distribution (Schnell and Utz 1975; Bernardo 2014; Lehermeier et al. 2017b) that we applied in the context of donor×elite recipient crosses using genome-wide estimated marker effects. We evaluated their ability to account for donor performance and originality compare to a given elite recipient and discussed their interest depending on short- or long-term objectives.

Considering the Iodent line E6 as the elite recipient, we observed a linear tendency between the proportion of loci K (favorable allele in donor but not in recipient) and J (favorable allele in recipient but not in donor) for Iodent donors (Fig. 7). This linear tendency was related to the distance between donor and recipient E6 (r(K, MRD) = 0.94, Fig. 4), highlighting that, in this case, increasing the number of original favorable alleles generally comes at the cost of an increase in unfavorable alleles. This observation

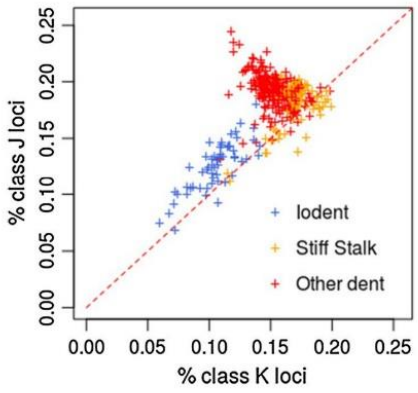


Fig. 7 Prediction of the percentage of loci in dissimilarity genotypic classes at loci K (donor +/+, elite −/−) and J (donor −/−, elite +/+) for every donor×E6 pair. All 338 dent candidate donors are presented with a color depending on their heterotic group. The bisector line is shown as red dashed line

is in line with the negative correlation between the performance of the Iodent donors (GEBV) and the genetic distance donors—E6 (MRD) likely driven by the fact that E6 is an elite line. Interestingly, this trend was less clear when considering donors from other heterotic groups than Iodent and especially the "Other dent" group (Fig. 7). Maximizing the percentage of loci in class K (i.e., the number of favorable alleles brought by the donor and not present in the elite line) yielded the selection of donors from non Iodent groups and specifically some Stiff Stalk lines appeared to be promising for improving the Iodent line E6 (Fig. 7). Furthermore, the heterotic group called "Other dent" appeared to have an excess of unfavorable alleles (class J) likely due to a large extent of old lines that predate the definition of the Iodent and Stiff Stalk groups (Rio et al. 2019). In practice, the ratio K/J was inferior to one for most of Iodent donors (Fig. 7) showing that line E6 was expected to carry more favorable alleles than most donors, consistently with the fact that E6 was more recent.

Whereas previous criteria were based on individual markers, criteria OHV and *H* addressed donor–recipient complementarity based on haplotype segments defined by splitting chromosomes into continuous segments or overlapping segments, respectively. Small haplotype segments consider a high number of recombination events, i.e., assume implicitly numerous intercross generations. In this sense, the optimal haplotype segment parameters depend on the introgression objectives, i.e., short- or long-term gain. Increasing nSeg in OHV (nSeg = 100), i.e., assuming more recombination events, accounted more for donor distance to elite recipient compared to donor performance and converged toward the criterion *H* ($r = 0.99$, Fig. 4). Note that, in criterion *H*, haplotype segments have been defined using a fixed number of markers rather than a predefined physical or genetic window size. This is justified as the MaizeSNP50 Illumina® BeadChip (Ganal et al. 2011) has been defined to reflect the density in genes, but may have to be adjusted for different genotyping tools. Haplotype segments could also account for recombination rates by considering genetic distances which would yield smaller haplotypes in recombination hotspots and larger haplotypes in recombination coldspots. Alternatively, one could focus on specific regions of interest (e.g., low diversity regions in elites as identified in Gerke et al. (2015) and Allier et al. (2019a).

One can further refine the prediction of the complementarity between a donor and an elite recipient for a polygenic trait by accounting for the recombination frequency and linkage disequilibrium between markers. This can be achieved by considering the progeny variance (VarG) in usefulness criterion (UC). When considering a selection intensity that is common for elite breeding ($i = 2.06$), UC_1 tended to select the most performant donor, which is interesting at short term,

but gives a very limited weight to the expected response to selection, i.e., progeny variance ($r(UC_1, GEBV) = 0.94$ and $r(UC_1, VarG) = 0.05$, Figs. 4, 5). On the contrary, considering an extreme selection intensity ($i = 5.78$), UC_2 accounted more for the donor × recipient progeny variance, i.e., donor originality ($r(UC_2, GEBV) = 0.60$ and $r(UC_2, VarG) = 0.58$, Figs. 4, 5). In case of UC, the higher selection intensity (i) the more importance is given to the progeny variance, i.e., somehow to longer-term gain.

In a nutshell, the comparison of tested criteria with donor performances (GEBV) and originality (MRD donor–recipient) enabled to identify three main groups of correlated criteria for the ranking of Iodent donors (Fig. 4): (1) UC_1, OHV (nSeg = 2) and the ratio K/J that correlated well with the performance of the donor (GEBV) (2) the VarG and the proportion K that correlated well with the genetic distance donor–elite recipient (MRD) and (3) the three criteria H, OHV (nSeg = 100) and UC_2 that balanced the performance and the originality of the donor. This clustering was also consistent with the best donors identified by the different criteria (Table 2) except for OHV (nSeg = 2). This illustrates the use of different criteria knowing that the optimal criteria depends on the objectives. As a rule of thumb, if one is interested in the short-term gain expected from introgressing the donor into the elite recipient one might consider OHV with few and large haplotypes per chromosomes, as suggested in Daetwyler et al. (2015), or UC with a selection intensity common in breeding (5% selected progeny in our case). On the contrary, if one is more interested in the long-term gain expected from introgressing the donor into the elite recipient, one might consider criteria accounting more for the complementarity donor–recipient such as H and OHV with smaller haplotype segments or UC with a higher selection intensity. Finally, in genetic diversity conservation program with no trait improvement objective, one might just want to maximize progeny variance (VarG) if one trait is of interest, or the MRD donor–recipient in the absence of trait-specific considerations.

Selection of a set of donors to enrich an elite population

In practice one may want to enrich a population of elite lines with new favorable alleles from different donors. To build a population mixing elites and donors in which most of the favorable alleles are segregating requires that selected donors bring new favorable alleles absent in elite material and that different donors bring different favorable alleles. The forward selection of donors based on the criterion H considers the complementarity between donors to complement at best the elite population. Using the forward selection of donors based on the extension of H criterion to populations (Eq. 5), we identified a set of seven donors. Some of these selected donors were also selected to enrich the elite line E6 (HMV5502, HMV5301, NQ508, Table 2) and others (Lo1242, UH_P084, F1819, PHG83) were in the top five donors for at least one criteria (Table 2). These results suggest that the forward selection approach identified donors covering different criteria previously described considering the elite E6, which was representative of the elite population E1 to E10. We further proposed a crossing plan between the identified donors and elite recipients. This proposal considers only one generation of selection in biparental crosses, which is likely not optimal in view of longer-term objectives to pyramid at best the favorable alleles. In case of highly polygenic traits considered in this study, the extension of UC to complex multi-parental crosses, such as suggested in Allier et al. (2019b) for four-way crosses, might be of interest but its extension to more complex crosses was not in the scope of this study. In case of a limited number of favorable alleles, Han et al. (2017) proposed the predicted cross value (PCV) to identify the biparental cross maximizing the likelihood of pyramiding major favorable alleles from a donor into a given recipient. As suggested by the authors, the extension of PCV to multiple donors would be challenging but useful here.

Practical implementation in maize breeding

In maize hybrid breeding, lines are generally evaluated first by their test-cross performance, which is the performance of the hybrid between the line and a tester line from a complementary heterotic group. In a second step, performance evaluation implies to use several testers from the complementary heterotic group to differentiate the general combining abilities (GCAs) of the tested lines with the testers from the specific combining ability (SCA) that is peculiar to the interaction line × tester. In the "Amaizing" dent diversity panel, donors have been tested on a single Flint tester which does not allow to separate GCA and SCA for donors. As donor alleles are evaluated in combination with alleles of the tester, the tester should reflect the heterotic pattern designed in the elite breeding program. Otherwise, the interest of identified donors might be biased with respect to the breeding objective. In particular, due to inbreeding depression, the use of a tester related to part of the materials that are evaluated may lead to an underestimation of their potential for producing hybrids with unrelated materials (Larièpe et al. 2017). In this study, this risk can be considered as minor because the Flint–Iodent pattern is common for early maize breeding in Europe and UH007 is highly distant from all tested lines. The expansion of the identification of donors to all dent groups suggested the interest of Stiff Stalks to enrich and complement Iodent lines evaluated against a Flint tester. However, the introgression of genomic regions from a

third heterotic group (here Stiff Stalk) into one (here Iodent) of the two heterotic groups usually considered in this climatic zone (here Flint–Iodent) can complicate the future exchanges of genetic material within the breeding programs addressing warmer climatic zones and which work with the classical Stiff Stalk—Iodent heterotic pattern. Consequently, we advise to consider an additional constraint on donor local origin characteristics depending on private partner heterotic group management strategies.

We should also warn at this stage that despite all material in the "Amaizing" panel was screened for early to mid-flowering time, part of the interest of some donors for grain yield may be associated with late flowering, i.e., long growth cycle leading to high grain yield (e.g., HMV5502 and HMV5301 that outperform E6). This could be accounted for by correcting the grain yield by the grain moisture at harvest. We presented an approach based on a single trait of economic interest, but other donor traits or characteristics can be accounted for. For instance, selected donors are also likely bringing some agronomic flaws such as root or stalk lodging (Oyervides-Garcia et al. 1985) or, on the contrary, secondary interesting traits such as drought tolerance (Millet et al. 2016), quality or biotic stress resistance. If phenotypes are available for these traits we can consider trait indexes or to use the multivariate formulation of the usefulness criterion.

Further investigations on collaborative diversity panels

Genomic prediction models calibrated on the "Amaizing" dent collaborative diversity panel showed promising predictive ability on the elite material considered, supporting the pertinence of comparing different criteria of donor selection based on genomic predictions. Further evaluation of the predictive ability would be interesting on other elite material, along with the evaluation of the accuracy of described criteria to rank donors. On the donor side, not all available genetic resources can be evaluated within a training panel and new candidates (e.g., ex-PVPA, Kurtz et al. 2016) are released every year with only little available information, genomic prediction could be helpful to predict the missing information for these newly released lines to help breeders identifying the most appropriate ones. This strategy is in line with that described by Yu et al. (2016) and Crossa et al. (2016) to better harness genetic diversity contained in large gene banks. For instance, it would be of interest to predict the flowering time to target the right testing environment and recipient breeding pool. However, the predictive ability for these external unphenotyped genetic resources also needs to be further studied.

Our results suggested the interest of collaborative panels covering a continuum of the genetic diversity from founders of main breeding pools to current elite private material for genetic resource identification and introgression into elite germplasm. We believe that the collaboration among academic research centers, genetic collections and private breeders can help harnessing genetic resources to optimize and fasten the response to new agricultural and societal challenges in several species. Collaborative projects in maize and other species are good opportunities to efficiently investigate these assumptions.

Acknowledgements The authors thank the experimental staff at RAGT2n for elite material data extractions and Amaizing project members for the dent collaborative panel. We thank Cyril Bauland and Carine Palaffre (INRA Saint-Martin de Hinx) for the panel assembly and the coordination of seed production, private partners of the Amaizing project for field trials. We also thank Pierre Dubreuil and Simon Rio for the assembly and the analysis of phenotypic data. We thank Valerie Combes, Delphine Madur, and Stephane Nicolas for DNA extraction, analysis, and assembly of genotypic data. AA was funded by RAGT2n and the ANRT CIFRE Grant No. 2016/1281.

Author contribution statement AC, LM, CL, and ST conceived and supervised the study. AA prepared the data, performed the analysis, and wrote the early version of the manuscript. All authors reviewed and approved the manuscript.

Data availability The datasets analyzed in this study are not publicly available.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Allier A, Teyssèdre S, Lehermeier C, Claustres B, Maltese S et al (2019a) Assessment of breeding programs sustainability: application of phenotypic and genomic indicators to a North European grain maize program. Theor Appl Genet 132:1321–1334

Allier A, Moreau L, Charcosset A, Teyssèdre S, Lehermeier C (2019b) Usefulness criterion and post-selection parental contributions in multi-parental crosses: application to polygenic trait introgression. G3 Genes Genomes Genet 9:1469–1479

Bauer E, Falque M, Walter H, Bauland C, Camisan C et al (2013) Intraspecific variation of recombination rate in maize. Genome Biol 14:R103

Bernardo R (2014) Genomewide selection of parental inbreds: classes of loci and virtual biparental populations. Crop Sci 54:2586–2595

Bernardo R (2017) Prospective targeted recombination and genetic gains for quantitative traits in maize. Plant Genome 10:9999

Bernardo R, Moreau L, Charcosset A (2006) Number and fitness of selected individuals in marker-assisted and phenotypic recurrent selection. Crop Sci 46:1972–1980

Böhm J, Schipprack W, Utz HF, Melchinger AE (2017) Tapping the genetic diversity of landraces in allogamous crops with doubled haploid lines: a case study from European flint maize. TAG Theor. Appl Genet Theor Angew Genet 130:861–873

Brauner PC, Müller D, Schopp P, Böhm J, Bauer E et al (2018) Genomic prediction within and among doubled-haploid libraries from maize landraces. Genetics 210:1185–1196

Brauner PC, Schipprack W, Utz HF, Bauer E, Mayer M et al (2019) Testcross performance of doubled haploid lines from European flint maize landraces is promising for broadening the genetic base of elite germplasm. Theor Appl Genet 132:1897–1908

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084–1097

Browning BL, Browning SR (2016) Genotype imputation with millions of reference samples. Am J Hum Genet 98:116–126

Crossa J, Jarquín D, Franco J, Pérez-Rodríguez P, Burgueño J (2016) Genomic prediction of gene bank wheat landraces. G3 Genes Genomes Genet 6:1819–1834

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D et al (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22:961–975

Daetwyler HD, Hayden MJ, Spangenberg GC, Hayes BJ (2015) Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. Genetics 200:1341–1348

Dudley JW (1984) A method of identifying lines for use in improving parents of a single cross. Crop Sci 24:355–357

Dudley JW (1987) Modification of methods for identifying inbred lines useful for improving parents of elite single crosses. Crop Sci 27:944–947

Feng L, Sebastian S, Smith S, Cooper M (2006) Temporal trends in SSR allele frequencies associated with long-term selection for yield in maize. Maydica 51:293–300

Fess TL, Kotcon JB, Benedito VA (2011) Crop breeding for low input agriculture: a sustainable response to feed a growing world population. Sustainability 3:1742–1772

Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES 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 6:28334

Gauthier P, Gouesnard B, Dallard J, Redaelli R, Rebourg C et al (2002) RFLP diversity and relationships among traditional European maize populations. TAG Theor Appl Genet Theor Angew Genet 105:91–99

Gerke JP, Edwards JW, Guill KE, Ross-Ibarra J, McMullen MD (2015) The genomic impacts of drift and selection for hybrid performance in maize. Genetics 201:1201–1211

Giraud H, Lehermeier C, Bauer E, Falque M, Segura V et al (2014) Linkage disequilibrium with linkage analysis of multiline crosses reveals different multiallelic QTL for hybrid performance in the flint and dent heterotic groups of maize. Genetics 198:1717–1734

Goiffon M, Kusmec A, Wang L, Hu G, Schnable PS (2017) Improving response in genomic selection with a population-based selection strategy: optimal population value selection. Genetics 206:1675–1682

Gouesnard B, Negro S, Laffray A, Glaubitz J, Melchinger A et al (2017) Genotyping-by-sequencing highlights original diversity patterns within a European collection of 1191 maize flint lines, as compared to the maize USDA genebank. TAG Theor Appl Genet Theor Angew Genet 130:2165–2189

Haldane J (1919) The combination of linkage values, and the calculation of distances between the loci of linked factors. J Genet 8:299–309

Han Y, Cameron JN, Wang L, Beavis WD (2017) The predicted cross value for genetic introgression of multiple alleles. Genetics 205:1409–1423

Hellin J, Bellon MR, Hearne SJ (2014) Maize landraces and adaptation to climate change in Mexico. J Crop Improv 28:484–501

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

Jenkins MT (1978) Maize breeding during the development and nearly years of hybrid maize. In: Walden DB (ed) Maize breeding and genetics. Wiley, New York, pp 13–28

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC et al (2017) Improved maize reference genome with single-molecule technologies. Nature 546:524–527

Kurtz B, Gardner CAC, Millard MJ, Nickson T, Smith JSC (2016) Global access to maize germplasm provided by the US National Plant Germplasm System and by US Plant Breeders. Crop Sci 56:931–941

Larièpe A, Moreau L, Laborde J, Bauland C, Mezmouk S 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

Lehermeier C, Campos G, Wimmer V, Schön C-C (2017a) Genomic variance estimates: With or without disequilibrium covariances? J Anim Breed Genet 134:232–241

Lehermeier C, Teyssèdre S, Schön C-C (2017b) Genetic gain increases by applying the usefulness criterion with improved variance prediction in selection of crosses. Genetics 207:1651–1661

Mayer M, Unterseer S, Bauer E, de Leon N, Ordas B et al (2017) Is there an optimum level of diversity in utilization of genetic resources? Theor Appl Genet 130:2283–2295

Melchinger AE, Schopp P, Müller D, Schrag TA, Bauer E et al (2017) Safeguarding our genetic resources with libraries of doubled-haploid lines. Genetics 206:1611–1619

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Mikel MA, Dudley JW (2006) Evolution of North American dent corn from public to proprietary germplasm. Crop Sci 46:1193–1205

Millet EJ, Welcker C, Kruijer W, Negro S, Coupel-Ledru A et al (2016) Genome-wide analysis of yield in Europe: Allelic effects vary with drought and heat scenarios. Plant Physiol 172:749–764

Mohammadi M, Tiede T, Smith K (2015) PopVar: A genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. Crop Sci 55:2068–2077

Nelson PT, Coles ND, Holland JB, Bubeck DM, Smith S et al (2008) Molecular characterization of maize inbreds with expired U.S. plant variety protection. Crop Sci 48:1673–1685

Oyervides-Garcia M, Hallauer AR, Cortez-Mendoza H (1985) Evaluation of improved maize populations in mexico and the U.S. corn belt. Crop Sci 25:115–120

Pérez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495

Pollak LM, Salhuana W (2001) The germplasm enhancement of maize (GEM) project: private and public sector collaboration. In: Cooper HD, Spillane C, Hodgkin T (eds) Broadening the genetic base of crop production. CABI, Wallingford, pp 319–329

R Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Rauf S, Teixeira da Silva JA, Khan AA, Naveed A (2010) Consequences of plant breeding on genetic diversity. Int J Plant Breed 4:1–21

Rebourg C, Gouesnard B, Charcosset A (2001) Large scale molecular analysis of traditional European maize populations. Relationships with morphological variation. Heredity 86:574–587

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

Salhuana W, Pollak L (2006) Latin American Maize Project (LAMP) and Germplasm Enhancement of Maize (GEM) project: generating useful breeding germplasm. Maydica 51:339–355

Schnell F, Utz H (1975) F1-Leistung und Elternwahl in der Züchtung von Selbstbefruchtern. Bericht über die Arbeitstagung der Vereinigung österreichischer Pflanzenzüchter. BAL Gumpenstein, Austria, pp 243–248

Servin B, Martin OC, Mézard M, Hospital F (2004) Toward a theory of marker-assisted gene pyramiding. Genetics 168:513–523

Smith JSC, Hussain T, Jones ES, Graham G, Podlich D 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

Strigens A, Schipprack W, Reif JC, Melchinger AE (2013) Unlocking the genetic diversity of maize landraces with doubled haploids opens new avenues for breeding. PLoS ONE 8:e57234

Warburton ML, Ribaut JM, Franco J, Crossa J, Dubreuil P et al (2005) Genetic characterization of 218 elite CIMMYT maize inbred lines using RFLP markers. Euphytica 142:97–106

Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res 75:249–252

Wright S (1978) Evolution and the genetics of populations. Volume 4: variability within and among natural populations. University of Chicago Press, Chicago

Wu Y, Vicente FS, Huang K, Dhliwayo T, Costich DE et al (2016) Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. Theor Appl Genet 129:753–765

Yu X, Li X, Guo T, Zhu C, Wu Y et al (2016) Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat Plants 2:16150

Chapter 3 Usefulness criterion and post-selection parental contributions in multi-parental crosses: application to polygenic trait introgression

This chapter has been published in the peer-reviewed journal G3: Genes, Genomes, Genetics in 2019. The electronic version of this article is open-access on the publisher website:

https://www.g3journal.org/content/9/5/1469

GENOMIC PREDICTION

Usefulness Criterion and Post-selection Parental Contributions in Multi-parental Crosses: Application to Polygenic Trait Introgression

Antoine Allier,*,† Laurence Moreau,* Alain Charcosset,* Simon Teyssèdre,†,1 and Christina Lehermeier†,1

*GQE - Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Univ. Paris-Saclay, 91190 Gif-sur-Yvette, France and †RAGT 2n, Genetics and Analytics Unit, 12510 Druelle, France

ORCID IDs: 0000-0001-6578-1715 (A.A.); 0000-0002-7195-1327 (L.M.); 0000-0001-6125-503X (A.C.); 0000-0001-7724-0887 (C.L.)

ABSTRACT Predicting the usefulness of crosses in terms of expected genetic gain and genetic diversity is of interest to secure performance in the progeny and to maintain long-term genetic gain in plant breeding. A wide range of crossing schemes are possible including large biparental crosses, backcrosses, four-way crosses, and synthetic populations. *In silico* progeny simulations together with genome-based prediction of quantitative traits can be used to guide mating decisions. However, the large number of multi-parental combinations can hinder the use of simulations in practice. Analytical solutions have been proposed recently to predict the distribution of a quantitative trait in the progeny of biparental crosses using information of recombination frequency and linkage disequilibrium between loci. Here, we extend this approach to obtain the progeny distribution of more complex crosses including two to four parents. Considering agronomic traits and parental genome contribution as jointly multivariate normally distributed traits, the usefulness criterion parental contribution (UCPC) enables to (i) evaluate the expected genetic gain for agronomic traits, and at the same time (ii) evaluate parental genome contributions to the selected fraction of progeny. We validate and illustrate UCPC in the context of multiple allele introgression from a donor into one or several elite recipients in maize (*Zea mays* L.). Recommendations regarding the interest of two-way, three-way, and backcrosses were derived depending on the donor performance. We believe that the computationally efficient UCPC approach can be useful for mate selection and allocation in many plant and animal breeding contexts.

KEYWORDS
progeny variance
parental genome contribution
genome-wide prediction
multi-parental crosses
Genomic Prediction
GenPred
Shared Data Resources

Allocation of resources is a key factor of success in plant and animal breeding. At each selection cycle, breeders are facing the choice of crosses to generate the genetic variation on which selection will act at the next generation. In case of limited genetic variation for targeted traits, the introduction of favorable alleles from donors to elite material is necessary to ensure long term genetic gain. Several approaches have been proposed to introgress superior quantitative trait locus (QTL) alleles from a donor into a recipient. In case of a single desirable allele, it can be accomplished using molecular assisted introgression (Visscher *et al.* 1996; Frisch *et al.* 1999). In case of multiple desirable alleles, gene pyramiding strategies have been proposed (Hospital and Charcosset 1997; Charmet *et al.* 1999; Servin *et al.* 2004). More recently, Han *et al.* (2017) proposed the predicted cross value (PCV) to select at each generation crosses that maximize the likelihood of pyramiding desirable alleles in their progeny. For quantitative traits implying numerous QTL with small individual effects, genomic selection has been proposed to fasten the introgression of exotic alleles into elite germplasm (Bernardo 2009) and to harness polygenic variation from genetic resources (Gorjanc *et al.* 2016) using two-way crosses or backcrosses. However, plant breeders are not only considering biparental crosses such as two-way crosses or backcrosses but also

doi: https://doi.org/10.1534/g3.119.400129
Manuscript received November 30, 2018; accepted for publication February 27, 2019; published Early Online February 28, 2019.

Supplemental material available at Figshare: https://doi.org/10.25387/g3.7405892.
[1]Corresponding authors: RAGT 2n, Genetics & Analytics Unit, 12510 Druelle, France, E-mail: clehermeier@ragt.fr, RAGT 2n, Genetics & Analytics Unit, 12510 Druelle, France, E-mail: steyssedre@ragt.fr

multi-parental crosses including three-way crosses, four-way crosses or synthetic populations (Gallais 1990; Schopp *et al.* 2017). Crosses implying several parental lines are highly interesting for breeders to exploit at best the genetic diversity underlying one or several traits. Beyond fastening the introgression of genetic resources into elite germplasm, genomic selection could be used to predict the interest of a multi-parental cross involving one or several donors and recipients. Among possible crosses, the identification of those that secure the performance in progeny and maximize the genome contribution of donors to the selected progeny is essential for increasing or maintaining genetic gain and diversity of an elite population.

The interest of a cross for a given quantitative trait can be defined using the usefulness criterion (Schnell and Utz 1975) that is determined by its expected genetic mean (μ) and genetic gain ($ih\sigma$): $UC = \mu + i\ h\ \sigma$, where σ is the progeny genetic standard deviation. The selection intensity (i) depends on the selection pressure and the selection accuracy (h) can be assumed to be one when selecting on genotypic effects (Zhong and Jannink 2007). While μ can be easily predicted for different crossing schemes by the weighted average of parental values, the difficulty to have a good prediction of progeny variance (σ^2) hindered the use of UC in favor of simpler criteria (for a recent review on different criteria, see Mohammadi *et al.* 2015). Bernardo *et al.* (2006) suggested to predict the progeny variance of a given population using genotypic data of its progenitors and quantitative trait loci (QTL) effect estimates, assuming unlinked QTL. Zhong and Jannink (2007) extended this concept to linked loci. With the availability of high-density genotyping, it has been proposed to predict the progeny variance using *in silico* simulations of progeny and genome-wide marker effects (Iwata *et al.* 2013; Bernardo 2014; Lian *et al.* 2015; Mohammadi *et al.* 2015). However, the geometrically increasing number of cross combinations possible for n parents makes the testing of all crosses computationally intensive. For instance, with only $n = 50$ potential parents, a total of $C_2^n = \frac{n(n-1)}{2} = 1,225$ genetically different two-way crosses can be formed. This number increases by a factor of n when crossing all the possible two-way crosses to the n different parents, so that $nC_2^n = 61,250$ three-way crosses and backcrosses are possible. Recently, Lehermeier *et al.* (2017b) derived algebraic formulas to predict for a single trait the genetic variance of doubled haploid (DH) or recombinant inbred line (RIL) progeny derived from two-way crosses, using information of recombination frequency and linkage disequilibrium in parental lines. These algebraic formulas have not been extended so far to multi-parental crosses, hindering the prediction of the interest of such crosses.

While the expected genetic gain (UC) is a meaningful measure of the interest of a cross for breeding, it does not account for the parental genome contributions to the selected fraction of progeny that determine the genetic diversity in the next generation. Parental genome contribution to unselected progeny has been studied for several years and is of specific interest in breeding for donor introduction and to manage long term genetic gain and inbreeding rate (Hill 1993; Bijma 2000; Woolliams *et al.* 2015). Hill (1993) derived the variance of the non-recurrent parent genome contribution to heterozygous backcross individuals in cattle. Wang and Bernardo (2000) formulated the variance of parental genome contribution to F2 and backcross plant progeny considering a finite number of loci. Frisch and Melchinger (2007) extended this approach to a continuous integration over loci and showed that a normal distribution approximated well parental genome contribution obtained from computer simulations. Also empirical data on pairs of human full-sibs confirmed that parental genome contributions, *i.e.*, additive relationship, can be considered as normally distributed around the expected value of 0.5 (Visscher *et al.* 2006; Visscher 2009). All these studies considered the parental genome contribution distribution in unselected progeny. However, to control parental contribution during polygenic traits introgression, it is of interest to predict parental genome contribution after selection for quantitative traits.

In this study, we develop a multivariate approach called usefulness criterion parental contribution (UCPC) to evaluate the interest of a multi-parental cross implying a donor line and one or several elite recipients based on the expected genetic gain (UC) and the diversity (parental contributions, PC) in the selected progeny. We extend here the rational given by Lehermeier *et al.* (2017b) for two important aspects. We address the prediction of progeny variance for multi-parental crosses implying two to four parents and we consider the parental contribution as an additional quantitative trait. The originality of this approach is that it uses derivations of the prediction of progeny variance in multi-parental crosses implying up to four parents to jointly predict (i) the performance of the next generation using the usefulness criterion and (ii) the parental contributions to the selected fraction of progeny, which to our knowledge has not been investigated so far. We illustrate the use of UCPC in the context of external genetic resources introgression into elite material considering the specific case of a unique donor that is crossed to one or several elite recipients. We address the type of multi-parental cross that should be preferred among two-way crosses, three-way crosses or backcrosses in order to maximize genetic gain while introgressing donor alleles in the elite population within one selection cycle.

MATERIALS AND METHODS

Application example: breeding context

We assumed a generic plant breeding population of fully homozygote inbred lines genotyped for biallelic single nucleotide polymorphism (SNP) markers with known positions. We considered a quantitative agronomic trait (*e.g.*, grain yield) implying p QTL with known additive effects and with positions sampled among the SNP marker positions. Further, we considered that the breeding population is an elite population that should be enriched with several alleles from a donor without *a priori* knowledge on major QTL to be introgressed. We assumed a donor line (D) has been identified and should be crossed with lines from the elite population (*e.g.*, E_1 and E_2) in order to obtain high-performing progeny that combine donor favorable alleles in a performing elite background. This donor line can vary in its performance level and its diversity relative to the elite population.

In this context, we aimed at evaluating the interest of two-way crosses (*i.e.*, D x E_1 and D x E_2), backcrosses (*i.e.*, (D x E_1) x E_1 and (D x E_2) x E_2) or three-way crosses (*i.e.* (D x E_1) x E_2 and (D x E_2) x E_1) based on (i) the mean performance of the selected progeny and (ii) the average genome contribution of the donor to the selected progeny. Considering different donor characteristics, *i.e.*, originality and performance level, we compared the interest of the multi-parental crosses listed above in order to derive guidelines for the use of the donor D. As a benchmark, we also evaluated the interest of different elite multi-parental crosses.

Usefulness Criterion Parental Contribution

In order to predict the progeny distribution of a given cross in terms of expected genetic gain and genetic diversity, we considered the agronomic trait and the parental genome contribution as jointly multivariate normally distributed traits. This enabled us to (i) evaluate the genetic

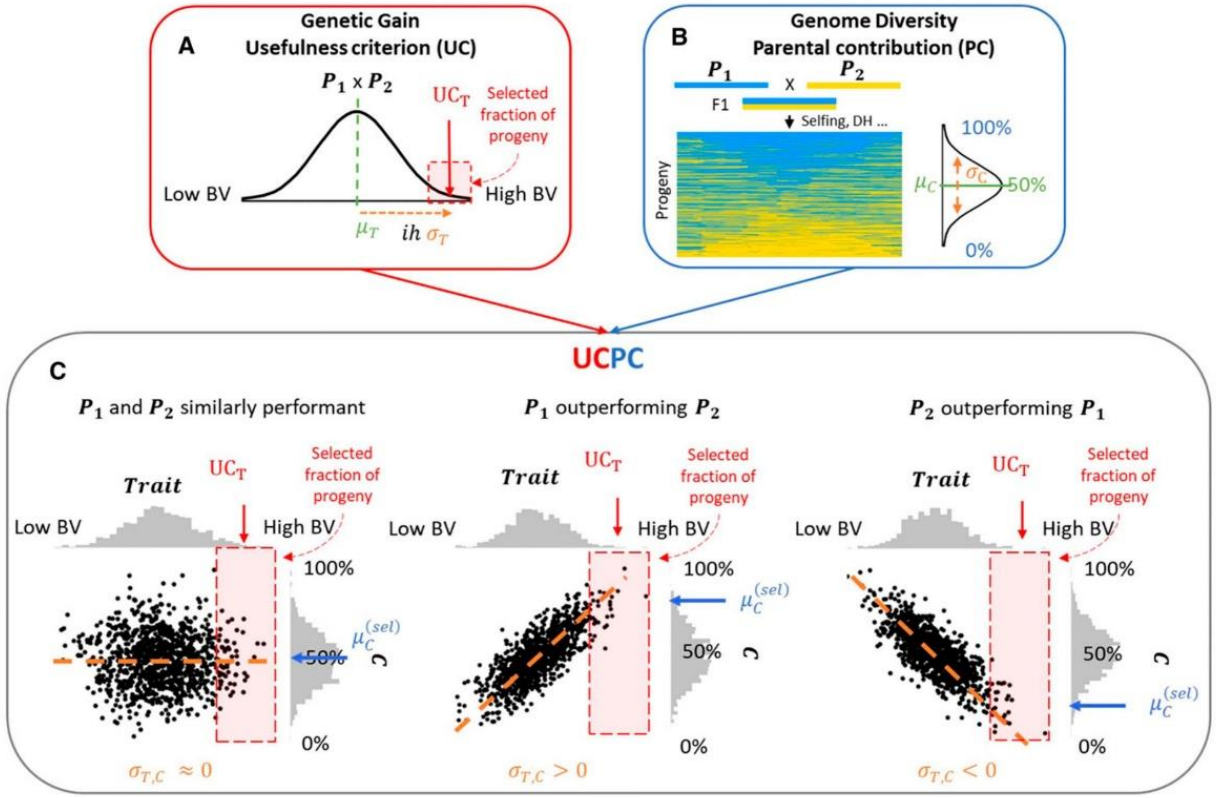


Figure 1 Illustration of Usefulness Criterion Parental Contribution (UCPC) for a two-way cross between P_1 and P_2. UCPC combines (A) the concept of usefulness criterion for an agronomic trait normally distributed ($N(\mu_T, \sigma_T)$) and (B) P_1 genome contribution considered as a normally distributed quantitative trait ($N(\mu_C, \sigma_C)$) in a multivariate approach (C). UCPC enables to predict the expected progeny performance for the trait (UC_T) and P_1 genome contribution to the selected fraction of progeny ($\mu_C^{(sel)}$) that depends on the covariance $\sigma_{T,C}$ mainly driven by the difference between P_1 and P_2 performances.

gain of the selected progeny for the agronomic trait, and to (ii) evaluate the contribution of each parental line to this selected progeny. An illustration of the concept of UCPC is given in Figure 1. In the following sections we present in more detail the theory underlying UCPC in the general case of a four-way cross.

Multi-parental crosses and genetic model: To cover diverse types of crosses, we consider a general multi-parental cross implying four fully homozygous parents (P_1, P_2, P_3 and P_4, Figure 2). Note that for this general presentation of the theory, parents can be lines from the elite population and/or considered as external donors. This four-way cross implies two initial crosses giving generations $F_1^{(1)}$ and $F_1^{(2)}$, respectively (Figure 2). A second cross between $F_1^{(1)}$ and $F_1^{(2)}$ yields the generation $F1'$ standing for pseudo F1. Two-way crosses, three-way crosses and backcrosses can be seen as specific cases of four-way crosses depending on the number of parents considered as visualized in Figure 2.

Assuming known genotypes at p QTL underlying the quantitative trait considered and biallelic markers at QTL positions, $\boldsymbol{x}_i$ denotes the p-dimensional genotype vector of parent i, with the j^{th} element coded as 1 or -1 for the genotypes AA or aa at locus j. Assuming biallelic QTL effects, a classical way to define the parental genotypes matrix would be a (4 x p)-dimensional matrix $(\boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3\ \boldsymbol{x}_4)'$. Addressing parental specific effects and following the identical by descent (IBD) genome contribution of parents to progeny requires to consider parental specific alleles. Thus, we extend the definition of parental genotypes to a multi-allelic coding:

$$\boldsymbol{X}_{Parental} = \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \boldsymbol{X}_3' \\ \boldsymbol{X}_4' \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1' & 0_p' & 0_p' & 0_p' \\ 0_p' & \boldsymbol{x}_2' & 0_p' & 0_p' \\ 0_p' & 0_p' & \boldsymbol{x}_3' & 0_p' \\ 0_p' & 0_p' & 0_p' & \boldsymbol{x}_4' \end{pmatrix},$$

with $\boldsymbol{X}_{Parental}$ a $(4\times 4p)$ dimensional matrix defining the genotype of the four parents at the $4p$ parental alleles at QTL, $\boldsymbol{X}_i$ the $4p$-dimensional vector defining the genotype of parent i and 0_p a p-dimensional vector of zeros.

We first concentrate on doubled haploid (DH) lines derived from the $F1'$ generation (DH-1), and then extend our work to DH lines generated after more selfing generations from the $F1'$ and to recombinant inbred lines (RILs) at different selfing generations, *i.e.*, partially heterozygous progeny. Absence of selection is assumed while deriving the progeny from generation $F1'$. In case of DH-1, we denote the (N x $4p$)-dimensional genotyping matrix of N progeny derived from a four-way cross (Figure 2) in a multi-allelic context as:

$$\boldsymbol{X}_{Progeny} = \left(\boldsymbol{X}_{1\,Progeny}\ \boldsymbol{X}_{2\,Progeny}\ \boldsymbol{X}_{3\,Progeny}\ \boldsymbol{X}_{4\,Progeny}\right),$$

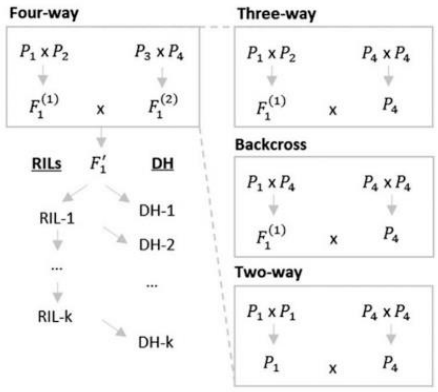


Figure 2 Illustration of four-way crosses (left) and derived crossing schemes (right). In the general case of four-way crosses, nomenclature is defined for recombinant inbred lines (RILs) after k generations of selfing (RIL-k) from pseudo F1 generation (F1′) and doubled haploid lines (DH) derived from the RIL generation k-1 (DH-k, for k > 1). RIL-1 corresponds to the pseudo F2 generation and RIL ∞ = DH ∞.

where for instance $\boldsymbol{X}_{1\,Progeny}$ is a $(N \times p)$-dimensional matrix of progeny genotypes at QTL coded -1 or 1 for alleles inherited from parent P_1 and 0 otherwise.

The multi-parental coding enables to consider $\boldsymbol{\beta}_T = (\boldsymbol{\beta}_{T1}' \boldsymbol{\beta}_{T2}' \boldsymbol{\beta}_{T3}' \boldsymbol{\beta}_{T4}')'$ a $4p$-dimensional vector of known parental specific additive effects for the agronomic trait. Thus, $\boldsymbol{X}_{Progeny}\ \boldsymbol{\beta}_T$ is the vector of progeny breeding values of the agronomic trait. As we assumed additive effects, the breeding value equals the genetic value. Assuming no parental specific effects for the agronomic trait, as in the application example considered, $\boldsymbol{\beta}_T$ reduces to $\boldsymbol{\beta}_T = (\boldsymbol{\beta}_0' \boldsymbol{\beta}_0' \boldsymbol{\beta}_0' \boldsymbol{\beta}_0')'$, where $\boldsymbol{\beta}_0$ is the vector of known QTL effects in the elite and donor populations. Furthermore, the multi-parental coding considered enables to define the effects to follow IBD parental contributions either genome-wide (namely C, $\boldsymbol{\beta}_C$) or considering only the favorable alleles (namely $C(+)$, $\boldsymbol{\beta}_{C(+)}$). In this study, we focused on the first parent (P_1) genome IBD contributions, but a generalization to every parent is straightforward. In the following, $\boldsymbol{\beta}_C$ is a $4p$-dimensional vector defined to follow P_1 genome-wide contribution and $\boldsymbol{\beta}_{C(+)}$ a $4p$-dimensional vector defined to follow P_1 genome contribution at favorable alleles. In the general case of four-way crosses $\boldsymbol{\beta}_C = \frac{1}{p}(\boldsymbol{x}_1' \ 0_p' \ 0_p' \ 0_p')' = \frac{1}{p}\ \boldsymbol{X}_1$ and $\boldsymbol{\beta}_{C(+)}$ is identical to $\boldsymbol{\beta}_C$ except that if P_1 has the unfavorable allele at QTL $q \in [1, p]$, the corresponding element of $\boldsymbol{\beta}_{C(+)}$ is null. Thus, $\boldsymbol{X}_{Progeny}\ \boldsymbol{\beta}_C$ represents the proportion of alleles in the progeny that are inherited from P_1 independently of the allele effect and $\boldsymbol{X}_{Progeny}\ \boldsymbol{\beta}_{C(+)}$ represents the proportion of alleles in the progeny that are inherited from P_1 and favorable. In the specific case of two-way crosses (*i.e.*, $P_1 = P_2$ and $P_3 = P_4$ so $\boldsymbol{x}_1 = \boldsymbol{x}_2$ and $\boldsymbol{x}_4 = \boldsymbol{x}_3$), P_1 genome-wide contribution is defined by $\boldsymbol{\beta}_C = \frac{1}{p}\ (\boldsymbol{x}_1' \ \boldsymbol{x}_1' \ 0_p' \ 0_p')'$.

Prediction of progeny mean and progeny variance: In this section we consider a generic quantitative trait defined by the $4p$-dimensional vector of parent specific additive effects $\boldsymbol{\beta} = (\boldsymbol{\beta}_1' \ \boldsymbol{\beta}_2' \ \boldsymbol{\beta}_3' \ \boldsymbol{\beta}_4')'$. The vector $\boldsymbol{\beta}$ can be replaced by $\boldsymbol{\beta}_T$, $\boldsymbol{\beta}_C$ or $\boldsymbol{\beta}_{C(+)}$ without loss of generality. In order to evaluate the performance of a four-way cross, we derive its expected progeny mean and variance. The expected progeny mean can be derived as the mean of all four parents' breeding values:

$$\mu_{Progeny} = \frac{1}{4}\ \mathbf{1}_4'\ \boldsymbol{X}_{Parental}\ \boldsymbol{\beta} \tag{1}$$

The progeny variance can be derived as:

$$\sigma^2_{Progeny} = var(\boldsymbol{X}_{Progeny}\ \boldsymbol{\beta}) = \ \boldsymbol{\beta}'\, var(\boldsymbol{X}_{Progeny})\boldsymbol{\beta} = \ \boldsymbol{\beta}'\ \boldsymbol{\Sigma}\boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{\Sigma}$ is the $(4p \times 4p)$-dimensional covariance matrix between parental alleles at QTL in progeny. The diagonal elements $\boldsymbol{\Sigma}_{jj}$ $(j \in [1, 4p])$ are equal to the variance of parental alleles in progeny. Note that off-diagonal elements $\boldsymbol{\Sigma}_{jl}$ $(j \neq l \in [1, 4p])$ correspond to the disequilibrium covariance between two parental alleles j and l at different QTL (*i.e.*, different physical positions) or at the same QTL. The linkage disequilibrium parameter in the progeny between parental alleles D_{jl} can be derived from the linkage disequilibrium parameter among the four parental lines and the recombination frequency between parental alleles in progeny (Table 1, see File S1 for derivation). In the specific case considered, *i.e.*, doubled haploid lines derived from generation $F1'$ (DH-1), this leads to the covariance entry:

$$\Sigma_{jl} = 4D_{jl} = (1 - 2c_{jl}^{(1)})\left(\Phi_{2jl} + \left(1 - 2c_{jl}^{(1)}\right)\Phi_{1jl}\right), \tag{3}$$

where $\Phi_{1jl} = \ D_{jl}^{12} + D_{jl}^{34}$ is the sum of the disequilibrium parameter between parental alleles j and l in pairs of parents implied in the first crosses and $\Phi_{2jl} = D_{jl}^{14} + D_{jl}^{13} + D_{jl}^{24} + D_{jl}^{23}$ is the sum of disequilibrium parameter between parental alleles j and l in pairs of parents indirectly implied in the second cross. D_{jl}^{12} denotes the linkage disequilibrium between parental alleles j and l in the pair of parental lines P_1 and P_2 which can be computed as $D_{jl}^{12} = \frac{1}{16}[(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)']_{jl}$. $c_{jl}^{(1)}$ is the recombination frequency between parental alleles j and l in the parental lines obtained from the absolute genetic distance d_{jl} in Morgan as $c_{jl}^{(1)} = \frac{1}{2}(1 - e^{-2d_{jl}})$ (Haldane 1919). When j and l refer to parental alleles at the same QTL, it holds $d_{jl} = c_{jl}^{(1)} = 0$. This formula given in [Equation 3] can be applied analogously in every case presented in Figure 2: three-way crosses, backcrosses and two-way crosses. See File S1 for a detailed derivation of the covariance in DH-1 progeny [Equation 3] and File S2 for an extension to DH progeny derived after selfing generations and to recombinant inbred lines at different selfing generations.

Indirect response to selection for parental contributions: We aim at predicting the full multivariate progeny distribution (mean, variance and pairwise covariances) for the agronomic trait, P_1 genome-wide contribution (C) and P_1 contribution at favorable alleles ($C(+)$). Therefore, we consider all three traits in the $(4p \times 3)$-dimensional multi-trait effect matrix $(\boldsymbol{\beta}_T\ \boldsymbol{\beta}_C\ \boldsymbol{\beta}_{C(+)})$. Similarly as for one trait, the mean performance $(\mu_T^{(0)})$ and mean genome-wide contribution of P_1 in progeny before selection $(\mu_C^{(0)})$ are derived as the mean of all four parents' breeding values for each trait [Equation 1]. As expected, $\mu_C^{(0)} = 0.25$ for four-way, three-way and backcrosses and $\mu_C^{(0)} = 0.5$ for two-way crosses. Progeny variances for all three traits are estimated using Equation 2 and pairwise covariances in progeny are estimated as:

$$\sigma_{T,\,C} = \boldsymbol{\beta}_T'\ \boldsymbol{\Sigma}\ \boldsymbol{\beta}_C = \boldsymbol{\beta}_C'\ \boldsymbol{\Sigma}\ \boldsymbol{\beta}_T, \tag{4a}$$

$$\sigma_{T,\,C(+)} = \boldsymbol{\beta}_T'\ \boldsymbol{\Sigma}\ \boldsymbol{\beta}_{C(+)} = \boldsymbol{\beta}_{C(+)}'\ \boldsymbol{\Sigma}\ \boldsymbol{\beta}_T \tag{4b}$$

Progeny means and (co)-variances before selection can be used to estimate the expected response to selection on multiple traits. For this purpose, we used the Usefulness Criterion (Schnell and Utz 1975) in a multi-trait approach as illustrated in Figure 1. Assuming

Table 1 Overview of genotypic covariance between loci j and l for different populations derived from the F1′ generation based on the disequilibrium parameter in pairs of parental lines

Population	Genotypic variance-covariance $\sum_{jl}$
DH generation k [a]	$(1-2c_{jl}^{(k)})\Phi_{2jl} + (1-2c_{jl}^{(k)} + c_{jl}^{(k-1)})(1-2c_{jl}^{(1)})\Phi_{1jl}$
RIL generation k [b]	$(1-2c_{jl}^{(k)} - (0.5(1-2c_{jl}^{(1)}))^{k})\Phi_{2jl} + (1-c_{jl}^{(k)})(1-2c_{jl}^{(1)})\Phi_{1jl}$

[a] Doubled haploid (DH) lines derived after k-1 generations of selfing ($k \in \mathbb{N}^*$, k = 1 for DH lines derived directly from F1′)
[b] Recombinant Inbred Lines (RIL) after k generations of selfing ($k \in \mathbb{N}^*$, k = 1 for pseudo F2 generation)
$\Phi_{1jl} = D_{jl}^{12} + D_{jl}^{34}$ and $\Phi_{2jl} = D_{jl}^{14} + D_{jl}^{13} + D_{jl}^{24} + D_{jl}^{23}$
$c_{jl}^{(t)} = \frac{2c_{jl}^{(1)}}{1+2c_{jl}^{(1)}}\,(1-0.5^{t}(1-2c_{jl}^{(1)})^{t})$

an intra-family selection of the progeny with the highest values for the agronomic trait with a selection intensity i and a selection accuracy of one (Figure 1A), the expected mean performance after selection $\mu_T^{(sel)}$ is defined as the usefulness criterion of the cross:

$$UC_T = \mu_T^{(sel)} = \mu_T^{(0)} + i\ \sigma_T \tag{5}$$

The correlated response to selection on P_1 genome-wide contribution ($\mu_C^{(sel)}$) and P_1 contribution at favorable alleles ($\mu_{C(+)}^{(sel)}$) are (Falconer and Mackay 1996):

$$\mu_C^{(sel)} = \mu_C^{(0)} + i\frac{\sigma_{T,C}}{\sigma_T} \tag{6a}$$

and

$$\mu_{C(+)}^{(sel)} = \mu_{C(+)}^{(0)} + i\frac{\sigma_{T,C(+)}}{\sigma_T} \tag{6b}$$

The contribution of P_1 at unfavorable alleles after selection can be derived as:

$$\begin{aligned}\mu_{C(-)}^{(sel)} &= \mu_{C(-)}^{(0)} + i\frac{\sigma_{T,C(-)}}{\sigma_T} = \mu_C^{(0)} - \mu_{C(+)}^{(0)} + i\ \frac{\sigma_{T,(C-C(+))}}{\sigma_T} \\ &= \mu_C^{(sel)} - \ \mu_{C(+)}^{(sel)}\end{aligned} \tag{6c}$$

Figure 1C illustrates, in the case of a two-way cross ($P1$ x P_2), the indirect response to selection on P_1 genome-wide contribution ($\mu_C^{(sel)}$) depending on the covariance $\sigma_{T,C}$ that is mainly driven by the difference of performance between P_1 and P_2.

Simulation experiments

We performed two simulation experiments. The aim of the simulation experiment 1 was the validation of the presented formulas for the moments of the distribution of progeny from four-way crosses. In simulation experiment 2, we investigated different crossing schemes (two-way, three-way and backcrosses) in terms of genetic gain and donor contribution.

Genetic material: We considered 57 Iodent inbred lines from the Amaizing Dent panel (Rio *et al.* 2019). Iodent defines a heterotic group that has been derived 50 to 70 years ago and that is commonly used in maize breeding (Troyer 1999; Van Inghelandt *et al.* 2012). In the following we refer to these lines as elite lines. Elite lines were genotyped with the Illumina MaizeSNP50 BeadChip (Ganal *et al.* 2011). After quality control and imputation, 40,478 high quality biallelic SNPs were retained. The genetic map was obtained by predicting genetic positions from physical positions (Jiao *et al.* 2017) using a spline-smoothing interpolating procedure described in Bauer *et al.* (2013) and the consensus dent genetic map in Giraud *et al.* (2014). We considered a quantitative agronomic trait (*e.g.*, grain yield) implying $p = 500$ QTL with known biallelic effects $\boldsymbol{\beta}_0$ sampled from $N(0_p, 0.002\boldsymbol{I}_p)$.

Simulation experiment 1: validation of UCPC: In order to validate the derivations for progeny (co)-variances and UCPC method in case of four-way crosses for DH and RIL progeny for selfing generations $k \in [1, 6]$ (Table 1), we randomly generated 100 four-way crosses out of the 57 elite lines. For each cross, a set of 500 QTL was randomly sampled among the 40,478 SNP markers across the genome to generate the agronomic trait. We also considered the first parent (*i.e.*, P_1) contributions: genome-wide (C) and at favorable alleles ($C(+)$). On one hand, we used algebraic formulas to predict the mean and (co)-variances for trait and contributions before selection within each cross (derivation). On the other hand, 50,000 DH or RIL progeny genotypes were simulated per cross at every selfing generation and the empirical mean and (co)-variances before selection were estimated (*in silico*). For *in silico* simulations, crossover positions were determined using recombination rates obtained with Haldane's function (Haldane 1919). The correlated response to selection on P_1 contributions after selecting the 5% upper fraction of progeny for the agronomic trait were either predicted using UCPC (derivation) or estimated after a threshold selection (*in silico*). The correspondence between predictors was assessed by the squared linear correlation and the mean squared difference between predicted (derivation) and empirical (*in silico*) values.

Simulation experiment 2: evaluation of different multi-parental crossing schemes between donor and elite lines: We used UCPC to address the question of the best crossing scheme between a given genetic resource (donor P_1, Figure 2), and elite lines. We identified the crossing scheme that maximized the short term expected genetic gain and evaluated donor genome contributions to the selected fraction of progeny. For this, we set up a simulation study where, at each iteration, an elite population of 25 lines was randomly sampled out of the 57 elite lines. Further, 500 QTL were sampled among monomorphic and polymorphic markers in the elite population in order to conserve the frequency of monomorphic loci observed on 40,478 SNPs in the entire elite population. At each iteration, 100 intra-elite two-way crosses, backcrosses, and three-way crosses were randomly sampled as benchmark. Their progeny mean (μ_T) and progeny standard deviation (σ_T) for the agronomic trait were predicted by Equation 1 and 2, respectively.

Within each iteration, 216 donor genotypes were constructed to cover a wide spectrum of donors in terms of performance and originality compared to the elite population. We defined three tuning parameters that reflect the proportions of six classes of QTL (Dudley 1984) defined by the polymorphism between the donor and the elite population (Table 2). All possible combinations of the three tuning parameters varying from 0 to 1 with steps of 0.2 were considered. For instance, among the favorable QTL in the elite population (classes I and J,

Table 2 Classes of quantitative trait loci (QTL) and tuning parameters considered for simulating the donors. The favorable allele at QTL is denoted (+) and the unfavorable is denoted (-). A polymorphic QTL in the elite population is denoted (+/−)

QTL classes	Elite Population	Single Donor	Tuning parameters
I	+	+	I/(I+J) [a]
J	+	-	
K	-	+	K/(K+L) [b]
L	-	-	
M	+/-	+	M/(M+N) [c]
N	+/-	-	

[a] proportion of monomorphic favorable QTL in the elite population where the donor had the favorable allele.
[b] proportion of monomorphic unfavorable QTL in the elite population where the donor had the favorable allele.
[c] proportion of polymorphic QTL in the elite population where the donor had the favorable allele.

Table 2), in the donor genome these QTL were randomly assigned to be favorable or unfavorable with probability I/(I+J) or J/(I+J), respectively. This was done similarly for all classes in Table 2. For each donor, we considered the simulated agronomic trait together with the donor genome contributions genome-wide (C) and at favorable alleles ($C(+)$). We defined the genetic gap with the elite population as the difference between donor and mean elite genetic values. The originality of the donor was defined as its mean pairwise modified Rogers distance (MRD) with elite lines.

For all possible 25 two-way crosses, 600 three-way crosses and 25 backcrosses between every donor and the elite population we predicted the progeny mean (μ) and the progeny standard deviation (σ) of each trait (Equation 1 and Equation 2) and the covariances between agronomic trait and contributions ($\sigma_{T,C}$, Equation 4a and $\sigma_{T,C(+)}$, Equation 4b). We defined the post-selection mean for the agronomic trait using Equation 5 with selection intensity i corresponding to a selection pressure of 5%. For comparison between iterations, we subsequently standardized the UC for the agronomic trait based on the elite population by $UC_T = (\mu_T^{(sel)} - \mu_{Elite})/\sigma_{Elite}$, where μ_{Elite} is the mean and σ_{Elite} the genetic standard deviation of the elite population. After selection on the agronomic trait, the correlated response on donor contributions was estimated using Equation 6 a-c. Finally, for each type of cross (two-way, three-way and backcrosses) and each donor, we identified the cross that maximized the expected genetic gain for the agronomic trait (UC_T).

Data availability

Simulations were based on genotypic maize data and genetic map deposited in File S4 at figshare. All simulations have been realized using R coding language (R Core Team 2017). Supplemental material available at Figshare: https://doi.org/10.25387/g3.7405892.

RESULTS

Simulation experiment 1: validation of UCPC

Predictions from the analytical derivations (Equation 1, 2, 4a, 4b, 5, 6a, 6b) showed a high correspondence with empirical results from *in silico* simulations for the 100 DH-1 families (DH lines after $F1'$, Figure 2). The predicted progeny variance from derivations and from *in silico* simulations (Figure 3A-C) as well as the covariances between the agronomic trait and parent contributions (Figure 3D-E) showed squared correlations above 0.96. Predicted and simulated post-selection mean of the agronomic trait as well as predicted and simulated post-selection parental genome contributions showed correlations above 0.9 (Figure 3F-H) ($R^2 = 1.000$ for Trait, $R^2 = 0.900$ for C and $R^2 = 0.946$ for $C(+)$). Validations for RIL and DH progeny derived from more selfing generations are presented in File S2.

Simulation experiment 2

Intra-elite multi-parental crosses: a benchmark: Considering only the elite population generated at each iteration, the mean average performance over 20 iterations was $\mu_{Elite} = 0.067 \pm 1.009$ and the mean elite standard deviation was $\sigma_{Elite} = 0.748 \pm 0.107$. We observed (Table 3) that intra-elite three-way crosses generated more progeny standard deviation (σ_T) (0.576 ± 0.034) than two-way crosses (0.510 ± 0.026) and backcrosses (0.442 ± 0.022). In terms of progeny mean (μ_T), differences were not significant between types of crosses. The gain in σ_T yielded a higher usefulness criterion ($UC_{T\ mean}$) with three-way crosses ($1.599\ \sigma_{Elite} \pm 0.317$) than two-way crosses ($1.461\ \sigma_{Elite} \pm 0.268$). On the contrary, when only considering the best cross in terms of gain for the agronomic trait ($UC_{T\ best}$), two-way crosses led to a higher UC ($3.115\ \sigma_{Elite} \pm 0.362$) than three-way crosses ($2.876\ \sigma_{Elite} \pm 0.420$) or backcrosses ($2.804\ \sigma_{Elite} \pm 0.377$).

Donor genome contribution in multi-parental crosses: For each simulated donor, we identified the two-way cross, three-way cross and backcross that maximized the UC for the agronomic trait (UC_T). Those crosses are denoted as best crosses in the following. We analyzed the relationship between donor contributions to the selected progeny of the best crosses and the genetic gap between the donor and the mean elite population (Figure 4). The genome-wide contribution, the contribution at favorable alleles, and the contribution at unfavorable alleles are shown in Figures 4A, 4B and 4C, respectively. For a given donor, the genome-wide donor contribution after selection was higher in the best two-way crosses than in the best three-way crosses or backcrosses. For illustrative purposes, we differentiated five cases from the worst donor carrying only unfavorable alleles at QTL (case 0) to the best donor carrying favorable alleles at all QTL (case 4). Starting from case 0, the selection tended to eliminate most of the donor genome in progeny until a lower bound (Figure 4A, 27.1% for the best two-way cross, 6.7% for the best three-way cross and 6.3% for the best backcross). Very badly performing donors (case 1; genetic gap ≤ -5), *i.e.*, carrying favorable alleles at maximum 180 QTL, had little chance to pass their favorable alleles to the selected progeny (Figure 4B, $\mu_{C(+)}^{(sel)} \leq 4.5\%$ in the best two-way cross, $\mu_{C(+)}^{(sel)} \leq 1.9\%$ in the best three-way cross and $\mu_{C(+)}^{(sel)} \leq 1.7\%$ in the best backcross). When the performance of the donor increased (case 2; $-5 <$ genetic gap ≤ 5), a higher portion of the donor genome was retained in the selected progeny (Figure 4B). With an increased number of favorable alleles (case 2), genome-wide donor contribution increased linearly with the genetic gap due to both, the selection of favorable alleles from the donor (Figure 4B) and the linkage drag with unfavorable alleles (Figure 4C). This linear trend continued until the donor had mainly favorable alleles (case 3; $5 <$ genetic gap). In case 3, we observed a linear increase of donor contribution at favorable alleles (Figure 4B). A correlated decrease of donor contribution at unfavorable alleles was observed at a nearly constant genome-wide contribution. Finally, in case 4, the genome-wide contribution was equal to an upper bound limit (Figure 4A, 72.6% for the best two-way cross, 42.9% for the best three-way cross and 43.5% for the best backcross).

Comparison of genetic gain among multi-parental crossing schemes: When the donor outperformed the elite population, the best two-way

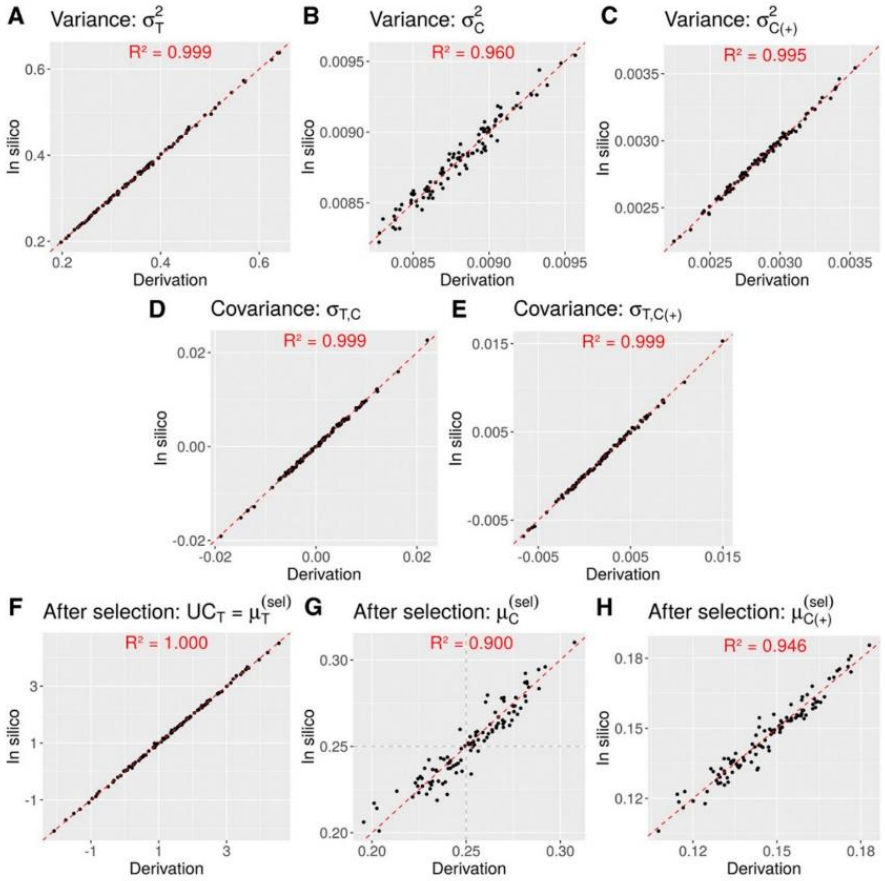


Figure 3 Comparison between predicted (derivation) and empirical (*in silico*) moments of the progeny distributions from 100 four-way crosses consisting of 50,000 DH-1 simulated progeny. Moments shown are (A) variance for the agronomic trait σ^2_T, (B) variance for the genome-wide contribution σ^2_C, (C) variance of the contribution at favorable alleles $\sigma^2_{C(+)}$, (D) covariance between agronomic trait and genome-wide contribution $\sigma_{T,C}$, (E) covariance between agronomic trait and contribution at favorable alleles $\sigma_{T,C(+)}$, (F) post-selection mean for the agronomic trait $UC_T = \mu_T^{(sel)}$, (G) post-selection mean for the genome-wide contribution $\mu_C^{(sel)}$, and (H) post-selection mean for the contribution at favorable alleles $\mu_{C(+)}^{(sel)}$. Squared correlations between predicted and empirical values are given within each plot.

cross was more likely yielding a higher genetic gain than the best three-way cross or backcross (Figure 5A). On the contrary, when the donor underperformed the elite population, the best three-way cross and backcross yielded a higher genetic gain than the best two-way cross. The higher progeny standard deviation (σ_T) in the best two-way cross compared to the best three-way cross or backcross (Figure 5B) did not compensate the loss in progeny mean (μ_T) (Figure 5C) in the best two-way cross. We observed that the type of cross maximizing the UC_T (*i.e.*, two-way cross, three-way cross or backcross) depended only on the performance of the donor, whatever the mean genetic distance with the elite population (*results not shown*). A similar comparison between three-way crosses and backcrosses showed that the best backcross yielded similar μ_T (Figure 5B) but lower σ_T than the best three-way cross (Figure 5C), especially when the donor had a genetic value close to the best elite lines. This resulted in a slightly higher expected genetic gain in three-way crosses compared to backcrosses (Figure 5A).

DISCUSSION

Usefulness criterion for quantitative traits in multi-parental crosses

Accurate predictors of progeny variance accounting for the map position of loci and linkage phase of alleles in parents have been recently derived for biparental crosses (Lehermeier *et al.* 2017b; Osthushenrich *et al.* 2017). Nonetheless, breeders might use multi-parental crosses implying more than two parents to combine best alleles segregating in the breeding population. Therefore, we extended derivations given by Lehermeier *et al.* (2017b) for two-way crosses to four-way crosses by accounting for linkage disequilibrium between pairs of parental lines. We validated the derived genetic variance of RIL and DH progeny of four-way crosses by simulations (Figure 3, File S2). As expected, the formula for four-way crosses reduces to the one given by Lehermeier *et al.* (2017b) in case of two-way crosses (File S1). The results from our simulations showed that, considering elite material only, three-way crosses generate on average more variance than two-way crosses or backcrosses, resulting in higher genetic gain (Table 3). Nevertheless, the best possible cross (*i.e.*, maximizing the expected genetic gain) was a two-way cross for most iterations (90%). This can be explained by the fact that crossing the two best elite lines generates more genetic gain than crossing them to a third less performant elite line, despite a potential gain in progeny variance. Notice that we considered only one polygenic agronomic trait but three-way crosses can be more advantageous for bringing complementary alleles for several traits. Under the formulated assumptions and with available marker effects (see discussion below), the general formula to predict mean and variance of four-way cross progeny makes it possible to identify the multi-parental cross that maximizes a given multi-trait selection objective (see discussion below) without requiring computationally intensive *in silico* simulations of progeny. The generalization to several generations of selfing for RIL progeny enables in addition to differentiate crosses releasing differently the variance in time (File S2). The presented formula for four-way crosses can also be applied to crosses

Table 3 Intra-Elite crosses predicted progeny mean (μ_T), progeny standard deviation (σ_T) and resulting expected genetic gain UC_T with a selection pressure of 5%, once averaged over all crosses ($UC_{T\ mean}$) and for the best cross identified ($UC_{T\ best}$). For all parameters the mean (± SD) over 20 iterations is given

	μ_T	σ_T	$UC_{T\ mean}$	$UC_{T\ best}$
Two-way	0.086 (± 1.016)	0.510 (± 0.026)	1.461 (± 0.268)	3.115 (± 0.362)
Three-way	0.049 (± 1.040)	0.576 (± 0.034)	1.599 (± 0.317)	2.876 (± 0.420)
Backcross	0.058 (± 1.042)	0.442 (± 0.022)	1.232 (± 0.247)	2.804 (± 0.377)

of two heterozygous parents by considering its phased genotypes as four separate parents. Doing so, our approach can be adapted for heterozygous plant varieties that are common in perennial species and for crosses with hybrids, as well as for animal breeding where the prediction of Mendelian sampling variance can be very useful for mating decisions (Bonk *et al.* 2016).

Parental contributions in multi-parental crosses under selection

Frisch and Melchinger (2007) derived the expected variance of parental contribution before selection in fully homozygote progeny accounting for linkage disequilibrium between loci assuming a biparental cross and considering only polymorphic loci. In this study, we proposed an original way to follow parental genome contribution to the selected fraction of progeny in multi-parental crosses, namely UCPC. It is grounded in a normal approximation of the probability mass function of parental contribution (Hill 1993; Frisch and Melchinger 2007) and progeny variance derivations. In the specific case of DH lines derived from two-way crosses or backcrosses and considering one chromosome of 100cM, our prediction of parental genome contribution variance converged to the one of Frisch and Melchinger (2007) when increasing the number of loci (File S3). However, the previous literature did not combine parental contributions with quantitative traits. Our original multivariate UCPC approach enables to predict the covariance between parental genome contributions and traits of economic interest. Based on multivariate selection theory, UCPC predicts the expected realized parental genome contribution after selection on traits of interest. It allows to follow parental genome contribution inheritance over generations and provides the likelihood of reaching a specific level of parental contribution while prescreening the most performing lines. Such information can guide breeders and researchers to determine the minimal number of progeny to derive from a cross between a donor and one or several elite lines so that the expected donor contribution after selection can reach a targeted value.

Predicted genome-wide donor contribution to progeny after selection was bounded to a minimum in case of the worst donor and a maximum in case of the best donor. In line with the predicted distribution of parental genome contribution before selection obtained in maize by Frisch and Melchinger (2007), these results show that in one selection cycle with a reasonable selection intensity (*e.g.*, 5%) it is unlikely to get completely rid of unfavorable parental alleles. Parental genome contribution was bounded in selected progeny due to the low probability of combining all alleles from a single parent. Note that UCPC also allows to follow the contribution of parents to progeny performance by defining a vector of effects based on parental performance marker effects. For instance, considering $(\boldsymbol{\beta}'_{T1}\ 0'_p\ 0'_p\ 0'_p)'$ enables to follow the first parent contribution to progeny performance.

Recommendations for donor by elites crosses

Using UCPC, we addressed the question of polygenic trait introgression from an inbred donor to inbred elite recipients with a focus on common plant breeding crossing schemes: two-way, three-way and backcrosses. We assumed that the objective was to derive in one selection cycle an inbred progeny that combined donor favorable alleles in a performing elite background. Such progeny can be used as parental lines for new crosses in order to quickly introgress new favorable alleles in a breeding program. Such a short term vision of genetic resource integration can be complementary to a longer term pre-breeding approach using exotic material (Bernardo 2009; Gorjanc *et al.* 2016; Yu *et al.* 2016). As expected, donors underperforming the elite population (inferior donor) yielded a higher genetic gain when complemented by two elite lines in three-way crosses or by twice an elite line in backcrosses rather than by a single elite line in two-way crosses. In this case, there is an advantage of crossing schemes involving, on average before selection, only one fourth of the donor genome instead of half of the donor genome as it would be the case for a two-way cross. On the contrary, two-way crosses were more adapted to donors outperforming the elite population. If the donor showed a similar performance level as the

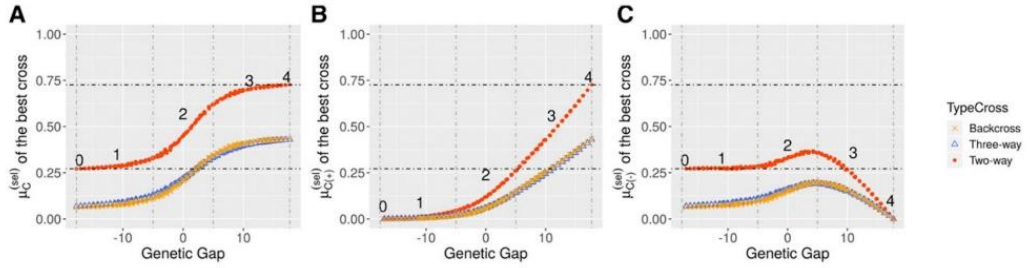


Figure 4 Donor contribution to the selected progeny of the best two-way cross (Donor*Elite), the best three-way cross ((Donor*Elite1)*Elite2) and the best backcross ((Donor*Elite1)*Elite1), depending on the genetic gap between donor line and the elite population. Each data point corresponds to the progeny of the best cross and is colored depending on the type of cross. (A) Donor genome-wide contribution after selection $\mu_C^{(sel)}$, (B) donor genome contribution at favorable alleles after selection $\mu_{C(+)}^{(sel)}$ and (C) donor genome contribution at unfavorable alleles after selection $\mu_{C(-)}^{(sel)}$. Numbers (0, 1, 2, 3, 4) correspond to illustrative cases based on genetic gap referred in the text. Illustrative cases 0 and 4 correspond to the worst and best donor respectively. Illustrative cases 1, 2, 3 are delimited by genetic gap values -5, 5 as represented by the vertical dashed lines.

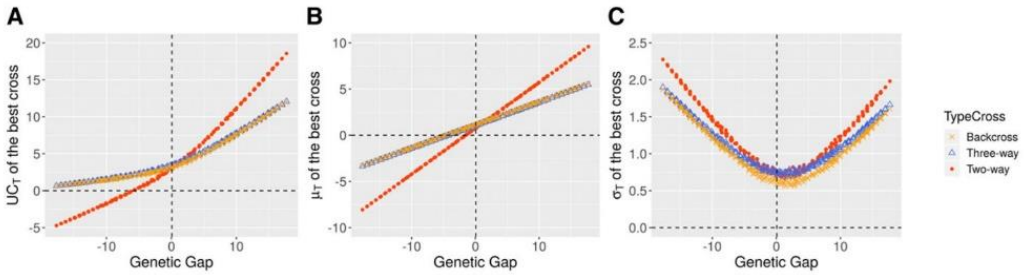


Figure 5 Comparison of the best two-way cross (Donor*Elite), the best three-way cross ((Donor*Elite1)*Elite2) and the best backcross ((Donor*Elite1)*Elite1), depending on the genetic gap (x-axis) with the elite population. Each data point corresponds to the progeny of the best cross and is colored depending of the type of cross. Comparison for the (A) expected genetic gain UC_T, (B) progeny mean (μ_T), and (C) progeny standard deviation (σ_T).

elite lines, no general rule could be drawn. In such a case, we recommend to identify the best crossing scheme by predicting every potential cross using the UCPC approach. As expected under a lower dilution of donor alleles into elite alleles in two-way crosses compared to three-way crosses or backcrosses, the predicted genome-wide donor contribution to selected progeny was higher in the best two-way cross than in the best three-way cross or the best backcross (Figure 4A).

We observed for a polygenic trait that, despite a lower competition between donor and elite favorable alleles, backcrosses were not significantly superior to three-way crosses for maintaining higher donor contribution at favorable alleles (Figure 4B). In addition, backcrosses generated less progeny variance (Figure 5C) but similar progeny mean than three-way crosses, resulting in a lower genetic gain (Figure 5A). This observation depends on the elite population considered. For instance, it might not hold if one unique elite line highly outperforms all other lines. More generally, while backcrosses only combine donor alleles with alleles of one elite parent, three-way crosses combine donor alleles with alleles of two complementary elite lines and are thus closer to material generated at the same time using two-way crosses in routine breeding. For these reasons, we suggest that three-way crosses should be preferred over backcrosses for polygenic trait introgression in elite germplasm. Our results support *a posteriori* the crossing strategy adopted in the Germplasm Enhancement of Maize project (GEM, *e.g.*, Goodman 2000). In GEM, maize exotic material has been introgressed into maize elite private lines using three-way crosses implying two different private partners. With the possibility to efficiently predict the progeny distribution of three-way crosses (UCPC), the best crossing partners can be identified to meet the targeted outcome in short time which allows to fully profit of the advantages of three-way crosses.

Multivariate selection for agronomic traits and parental contributions

We observed that badly performing donors had little chance to pass their favorable alleles to progeny selected for their agronomic trait performance. This is a consequence of the negative covariance between the performance for the trait and donor contribution in case of an inferior donor (Figure 1C). To prevent this loss of original alleles, we could account for such tension in the multivariate context, for instance by applying a truncation on donor contribution before selecting for the trait using the truncated multivariate normal theory (Horrace 2005) or vice versa. Otherwise, selection on donor contribution and the agronomic trait can be applied jointly by building a selection index, which is promising to balance short term genetic gain and long term genetic diversity (*i.e.*, selection on donor contribution) according to specific pre-breeding strategies.

More generally, the multivariate context provides the opportunity to deal with several quantitative traits on which selection is directly or indirectly applied. Further traits for which genome-wide estimated marker effects or QTL effects are available can be considered. For external genetic resource utilization, it enables to introgress secondary traits such as polygenic tolerances to biotic or abiotic stresses (*e.g.*, drought tolerance), while agronomic flaws (*e.g.*, plant lodging) can be counter-selected using threshold selection. Recently it has been shown by Akdemir *et al.* (2018) how the improvement of multiple traits can be addressed with multi-objective optimized breeding strategies.

Practical implementation of UCPC in breeding

In practice, marker effects estimated with whole-genome regression models can be used in lieu of QTL effects that are unknown. Such effects should be estimated on a proper training population mixing both elite lines and original genetic resources. Marker effects can be estimated using Bayesian Ridge Regression as suggested in Lehermeier *et al.* (2017a; b) to derive an unbiased estimator of progeny variance (PMV: posterior mean variance). In our simulation study, we considered only biallelic QTL effects. As we formulated a multi-allelic model, population-specific additive effects could be considered straightforwardly. Considering that the donor might have a different origin than the elite lines (*e.g.*, other heterotic group in hybrid crops), it might be of interest to use parental specific effects estimated by *e.g.*, multivariate QTL mapping (Giraud *et al.* 2014) or genome-wide prediction models (Lehermeier *et al.* 2015). UCPC relies on individual marker effects but the computation of the variance in the progeny accounts for collinearity among markers, *i.e.*, considers haplotype transmission. We therefore expect that inaccuracies in marker effects estimates will affect UCPC to a limited extent, but this warrant specific investigations as suggested by Müller *et al.* (2018).

Our approach is totally generic and can deal with any information on the position and the effect of QTL. However, main assumptions should be discussed at this point. We assumed known true genetic positions of QTL and no interference during crossover formation to derive recombination frequencies (Haldane 1919). In practice, the precision of recombination frequency estimates is a function of the available mapping information and the frequency of interference. Furthermore, recombination frequency might vary among the same species (Bauer *et al.* 2013) impairing the accuracy of variance prediction. To limit this risk we suggest to use a multi-parental consensus map

(*e.g.*, Giraud *et al.* 2014). The effect of genetic map inaccuracies on progeny co-variances prediction requires further investigations. Furthermore, derivations assumed no selection before developing progeny. However, selecting progeny from which to derive DH lines is likely in practice. This can involve voluntary molecular prescreening for disease resistance (*e.g.*, during selfing generations) or practical limitations (*e.g.*, originating from low DH induction rates). If the genetic correlation between those traits and the traits considered within UCPC is null, the derived progeny distribution and UC for the four-way crosses will still hold.

The derived formula for progeny mean and variance holds for mono- and oligo-genic traits, whereas the usefulness criterion underlying UCPC uses normal distribution properties. When considering traits involving a sufficient number of underlying QTL, as it is the case for most agronomic traits and parental genome contributions, this assumption of normality is likely guaranteed by the central limit theorem. If only a limited number of known major QTL should be introgressed from a donor, an allele pyramiding strategy will be more suitable (Hospital and Charcosset 1997; Charmet *et al.* 1999; Servin *et al.* 2004). Furthermore, the predicted cross value (PCV) as recently suggested by Han *et al.* (2017) can be applied in this context and could be extended to multi-parental crosses considering our derivation of progeny variance.

We presented an IBD definition of parental genome contributions using a multi-allelic approach. The multi-allelic coding yields covariance matrices that are four times larger compared to using a biallelic coding. In practice, to obtain a less computationally intensive solution, the genotyping matrix can be reduced to a bi-allelic coding which yields an identity by state (IBS) parental genome contribution that informs on the sequence similarity between one parent and progeny (see File S3). However, in such a case parental contributions do not sum up to one and it cannot be accounted for multi-allelic (*i.e.*, haplotypic) effects. For biparental crosses (*i.e.*, two-way and backcrosses), an IBS approach (File S3) considering only polymorphic markers homogeneously covering the genome can be used as an approximation of the IBD contribution.

Future research directions

UCPC is opening several future research directions. We illustrated the use of UCPC for a simple donor introgression problem but it can be extended to more complex problematics commonplace in breeding. For instance, UCPC can be applied to evaluate the interest of introgressing several donors, *e.g.*, evaluate the interest of combining alleles from two donors (D_1 and D_2) with elites (E_1 and E_2) in $(D_1 \times E_1) \times (D_2 \times E_2)$ or $(D_1 \times D_2) \times (E_1 \times E_2)$.

Mating design optimizations, *i.e.*, finding an optimized list of crosses to realize each year, accounting for a compromise between short and long term genetic gain have been investigated using two-way crosses and parental means as predictor of the expected gain and the inbreeding rate in the next generation (De Beukelaer *et al.* 2017; Gorjanc *et al.* 2018). Applying UCPC within the context of mating design optimization would enable to account for parental complementarity through the use of progeny variation, *i.e.*, within cross variance, as proposed by Shepherd and Kinghorn (1998), Akdemir and Sánchez (2016) and Müller *et al.* (2018). Furthermore, UCPC would enable to use parental contribution to the selected fraction of progeny to predict the realized inbreeding in the next generation. We conjecture that considering the realized parental genome contribution together with the usefulness criterion in UCPC is promising for mating design optimization to manage short and long term genetic gain in breeding programs. Future research will also be needed to investigate the use of multi-parental crosses in mating design optimizations. Hereby, UCPC that efficiently predicts the progeny distribution of crosses with up to four parents will represent a good starting point for further research.

Conclusions

We developed, validated and illustrated the usefulness criterion parental contribution (UCPC) that evaluates the interest of multi-parental crosses based on the expected genetic gain (UC) and the parental contributions (PC) in the next generation. UCPC allows to (i) predict the progeny variance of four-way crosses accounting for linkage disequilibrium and to (ii) follow all parental genome contributions to the selected progeny to evaluate the interest of a cross regarding an objective that is a function of the expected performance and the diversity in the selected progeny. Illustration of the use of UCPC in the context of polygenic trait introgression from a donor to elite recipients enabled to draw some major recommendations. As expected, three-way crosses and backcrosses were more adapted to donors underperforming the elite population (inferior donor) while two-way crosses were more adapted to donors outperforming the elite population. We also suggested that three-way crosses should be preferred over backcrosses for polygenic traits introgression. Furthermore, we highlighted the importance of a compromise between UC and PC in case of an inferior donor.

ACKNOWLEDGMENTS

The authors thank the Amaizing program for genotypes used in simulations. This research was funded by RAGT 2n and the ANRT CIFRE Grant n° 2016/1281 for AA.

LITERATURE CITED

Akdemir, D., and J. I. Sánchez, 2016 Efficient Breeding by Genomic Mating. Front. Genet. 7: 210. https://doi.org/10.3389/fgene.2016.00210

Akdemir, D., W. Beavis, R. Fritsche-Neto, A. K. Singh, and J. Isidro-Sánchez, 2018 Multi-objective optimized genomic breeding strategies for sustainable food improvement. Heredity. https://doi.org/10.1038/s41437-018-0147-1

Bauer, E., M. Falque, H. Walter, C. Bauland, C. Camisan *et al.*, 2013 Intraspecific variation of recombination rate in maize. Genome Biol. 14: R103. https://doi.org/10.1186/gb-2013-14-9-r103

Bernardo, R., L. Moreau, and A. Charcosset, 2006 Number and Fitness of Selected Individuals in Marker-Assisted and Phenotypic Recurrent Selection. Crop Sci. 46: 1972–1980. https://doi.org/10.2135/cropsci2006.01-0057

Bernardo, R., 2009 Genomewide Selection for Rapid Introgression of Exotic Germplasm in Maize. Crop Sci. 49: 419–425. https://doi.org/10.2135/cropsci2008.08.0452

Bernardo, R., 2014 Genomewide Selection of Parental Inbreds: Classes of Loci and Virtual Biparental Populations. Crop Sci. 54: 2586–2595. https://doi.org/10.2135/cropsci2014.01.0088

De Beukelaer, H. D., Y. Badke, V. Fack, and G. D. Meyer, 2017 Moving beyond managing realized genomic relationship in long-term genomic selection. Genetics 206: 1127–1138. https://doi.org/10.1534/genetics.116.194449

Bijma, P., 2000 Long-term genetic contributions: prediction of rates of inbreeding and genetic gain in selected populations (Doctoral dissertation). Veenendaal, The Netherlands.

Bonk, S., M. Reichelt, F. Teuscher, D. Segelke, and N. Reinsch, 2016 Mendelian sampling covariability of marker effects and genetic values. Genet. Sel. Evol. 48: 36. https://doi.org/10.1186/s12711-016-0214-0

Charmet, G., N. Robert, M. R. Perretant, G. Gay, P. Sourdille *et al.*, 1999 Marker-assisted recurrent selection for cumulating additive and interactive QTLs in recombinant inbred lines. Theor. Appl. Genet. 99: 1143–1148. https://doi.org/10.1007/s001220051318

Dudley, J. W., 1984 A Method of Identifying Lines for Use in Improving Parents of a Single Cross. Crop Sci. 24: 355–357. https://doi.org/10.2135/cropsci1984.0011183X002400020034x

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics. Ed. 4th.* Pearson. Harlow, England.

Frisch, M., M. Bohn, and A. E. Melchinger, 1999 Comparison of Selection Strategies for Marker-Assisted Backcrossing of a Gene. Crop Sci. 39: 1295–1301. https://doi.org/10.2135/cropsci1999.3951295x

Frisch, M., and A. E. Melchinger, 2007 Variance of the Parental Genome Contribution to Inbred Lines Derived From Biparental Crosses. Genetics 176: 477–488. https://doi.org/10.1534/genetics.106.065433

Gallais, A., 1990 *Théorie de la sélection en amélioration des plantes*, Masson, Paris.

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

Giraud, H., C. Lehermeier, E. Bauer, M. Falque, V. Segura *et al.*, 2014 Linkage Disequilibrium with Linkage Analysis of Multiline Crosses Reveals Different Multiallelic QTL for Hybrid Performance in the Flint and Dent Heterotic Groups of Maize. Genetics 198: 1717–1734. https://doi.org/10.1534/genetics.114.169367

Goodman M. M., 2000 Incorporation of exotic germplasm into elite maize lines: Maximizing favorable effects of the exotic source. Theor. Pop. Biol.

Gorjanc, G., J. Jenko, S. J. Hearne, and J. M. Hickey, 2016 Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. BMC Genomics 17: 30. https://doi.org/10.1186/s12864-015-2345-z

Gorjanc, G., R. C. Gaynor, and J. M. Hickey, 2018 Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. Theor. Appl. Genet. 131: 1953–1966. https://doi.org/10.1007/s00122-018-3125-3

Haldane, J., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. J. Genet. 8: 299–309.

Han, Y., J. N. Cameron, L. Wang, and W. D. Beavis, 2017 The Predicted Cross Value for Genetic Introgression of Multiple Alleles. Genetics 205: 1409–1423. https://doi.org/10.1534/genetics.116.197095

Hill, W. G., 1993 Variation in Genetic Composition in Backcrossing Programs. J. Hered. 84: 212–213. https://doi.org/10.1093/oxfordjournals.jhered.a111319

Horrace, W. C., 2005 Some results on the multivariate truncated normal distribution. J. Multivariate Anal. 94: 209–221. https://doi.org/10.1016/j.jmva.2004.10.007

Hospital, F., and A. Charcosset, 1997 Marker-Assisted Introgression of Quantitative Trait Loci. Genetics 147: 1469–1485.

Iwata, H., T. Hayashi, S. Terakami, N. Takada, T. Saito *et al.*, 2013 Genomic prediction of trait segregation in a progeny population: a case study of Japanese pear (Pyrus pyrifolia). BMC Genet. 14: 81. https://doi.org/10.1186/1471-2156-14-81

Jiao, Y., P. Peluso, J. Shi, T. Liang, M. C. Stitzer *et al.*, 2017 Improved maize reference genome with single-molecule technologies. Nature 546: 524–527.

Lehermeier, C., C.-C. Schön, and G. de Los Campos, 2015 Assessment of Genetic Heterogeneity in Structured Plant Populations Using Multivariate Whole-Genome Regression Models. Genetics 201: 323–337. https://doi.org/10.1534/genetics.115.177394

Lehermeier, C., G. de los Campos, V. Wimmer, and C.-C. Schön, 2017a Genomic variance estimates: With or without disequilibrium covariances? J. Anim. Breed. Genet. 134: 232–241. https://doi.org/10.1111/jbg.12268

Lehermeier, C., S. Teyssèdre, and C.-C. Schön, 2017b Genetic Gain Increases by Applying the Usefulness Criterion with Improved Variance Prediction in Selection of Crosses. Genetics 207: 1651–1661.

Lian, L., A. Jacobson, S. Zhong, and R. Bernardo, 2015 Prediction of genetic variance in biparental maize populations: Genomewide marker effects *vs.* mean genetic variance in prior populations. Crop Sci. 55: 1181–1188. https://doi.org/10.2135/cropsci2014.10.0729

Mohammadi, M., T. Tiede, and K. Smith, 2015 PopVar: A Genome-Wide Procedure for Predicting Genetic Variance and Correlated Response in Biparental Breeding Populations. Crop Sci. 55: 2068–2077. https://doi.org/10.2135/cropsci2015.01.0030

Müller, D., P. Schopp, and A. E. Melchinger, 2018 Selection on Expected Maximum Haploid Breeding Values Can Increase Genetic Gain in Recurrent Genomic Selection. G3 (Bethesda)g3.200091.2018.

Osthushenrich, T., M. Frisch, and E. Herzog, 2017 Genomic selection of crossing partners on basis of the expected mean and variance of their derived lines. PLoS One 12: e0188839. https://doi.org/10.1371/journal.pone.0188839

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

Rio, S., T. Mary-Huard, L. Moreau, and A. Charcosset, 2019 Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. Theor. Appl. Genet. 132: 81–96. https://doi.org/10.1007/s00122-018-3196-1

Schnell, F., and H. Utz, 1975 F1-Leistung und Elternwahl in der Züchtung von Selbstbefruchtern. pp. 243–248 in *Bericht über die Arbeitstagung der Vereinigung österreichischer Pflanzenzüchter*, BAL Gumpenstein, Gumpenstein, Austria.

Schopp, P., D. Müller, Y. C. J. Wientjes, and A. E. Melchinger, 2017 Genomic Prediction Within and Across Biparental Families: Means and Variances of Prediction Accuracy and Usefulness of Deterministic Equations. G3 (Bethesda)7: 3571–3586.

Servin, B., O. C. Martin, M. Mézard, and F. Hospital, 2004 Toward a Theory of Marker-Assisted Gene Pyramiding. Genetics 168: 513–523. https://doi.org/10.1534/genetics.103.023358

Shepherd, R. K., and B. P. Kinghorn, 1998 A tactical approach to the design of crossbreeding programs. In Proceedings of the sixth world congress on genetics applied to livestock production, Armidale, 11–16: 431–438.

Troyer, A. F., 1999 Background of U.S. Hybrid Corn. Crop Sci. 39: 601–626. https://doi.org/10.2135/cropsci1999.0011183X003900020001x

Van Inghelandt, D., A. E. Melchinger, J.-P. Martinant, and B. Stich, 2012 Genome-wide association mapping of flowering time and northern corn leaf blight (Setosphaeria turcica) resistance in a vast commercial maize germplasm set. BMC Plant Biol. 12: 56. https://doi.org/10.1186/1471-2229-12-56

Visscher, P. M., C. S. Haley, and R. Thompson, 1996 Marker-Assisted Introgression in Backcross Breeding Programs. Genetics 144: 1923–1932.

Visscher, P. M., S. E. Medland, M. A. R. Ferreira, K. I. Morley, G. Zhu *et al.*, 2006 Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2: e41. https://doi.org/10.1371/journal.pgen.0020041

Visscher, P. M., 2009 Whole genome approaches to quantitative genetics. Genetica 136: 351–358. https://doi.org/10.1007/s10709-008-9301-7

Wang, J., and R. Bernardo, 2000 Variance of Marker Estimates of Parental Contribution to F 2 and BC 1 -Derived Inbreds. Crop Sci. 40: 659–665. https://doi.org/10.2135/cropsci2000.403659x

Woolliams, J. A., P. Berg, B. S. Dagnachew, and T. H. E. Meuwissen, 2015 Genetic contributions and their optimization. J. Anim. Breed. Genet. 132: 89–99. https://doi.org/10.1111/jbg.12148

Yu, X., X. Li, T. Guo, C. Zhu, Y. Wu *et al.*, 2016 Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat. Plants 2: 16150. https://doi.org/10.1038/nplants.2016.150

Zhong, S., and J.-L. Jannink, 2007 Using Quantitative Trait Loci Results to Discriminate Among Crosses on the Basis of Their Progeny Mean and Variance. Genetics 177: 567–576. https://doi.org/10.1534/genetics.107.075358

Communicating editor: D. J. de Koning

Chapter 4 Improving short- and long-term genetic gain by accounting for within family variance in optimal cross selection

This chapter has been published in the peer-reviewed journal Frontiers in Genetics in 2019. The electronic version of this article is open-access on the publisher website:

https://www.frontiersin.org/articles/10.3389/fgene.2019.01006

frontiers in Genetics

ORIGINAL RESEARCH
published: 29 October 2019
doi: 10.3389/fgene.2019.01006

Improving Short- and Long-Term Genetic Gain by Accounting for Within-Family Variance in Optimal Cross-Selection

Antoine Allier[1,2], Christina Lehermeier[2], Alain Charcosset[1], Laurence Moreau[1] and Simon Teyssèdre[2]*

[1] GQE-Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, Gif-sur-Yvette, France,
[2] Genetics and Analytics Unit, RAGT2n, Druelle, France

The implementation of genomic selection in recurrent breeding programs raises the concern that a higher inbreeding rate could compromise the long-term genetic gain. An optimized mating strategy that maximizes the performance in progeny and maintains diversity for long-term genetic gain is therefore essential. The optimal cross-selection approach aims at identifying the optimal set of crosses that maximizes the expected genetic value in the progeny under a constraint on genetic diversity in the progeny. Optimal cross-selection usually does not account for within-family selection, i.e., the fact that only a selected fraction of each family is used as parents of the next generation. In this study, we consider within-family variance accounting for linkage disequilibrium between quantitative trait loci to predict the expected mean performance and the expected genetic diversity in the selected progeny of a set of crosses. These predictions rely on the usefulness criterion parental contribution (UCPC) method. We compared UCPC-based optimal cross-selection and the optimal cross-selection approach in a long-term simulated recurrent genomic selection breeding program considering overlapping generations. UCPC-based optimal cross-selection proved to be more efficient to convert the genetic diversity into short- and long-term genetic gains than optimal cross-selection. We also showed that, using the UCPC-based optimal cross-selection, the long-term genetic gain can be increased with only a limited reduction of the short-term commercial genetic gain.

Keywords: genomic prediction, optimal cross-selection, usefulness criterion, parental contributions, genetic diversity, Bulmer effect

OPEN ACCESS

Edited by:
Charles Chen,
Oklahoma State University,
United States

Reviewed by:
Changwei Shao,
Yellow Sea Fisheries Research Institute (CAFS), China
Zibei Lin,
La Trobe University, Australia

****Correspondence:***
Antoine Allier
antoine.allier@inra.fr

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *10 May 2019*
Accepted: *20 September 2019*
Published: *29 October 2019*

Citation:
Allier A, Lehermeier C, Charcosset A, Moreau L and Teyssèdre S (2019) Improving Short- and Long-Term Genetic Gain by Accounting for Within-Family Variance in Optimal Cross-Selection. Front. Genet. 10:1006. doi: 10.3389/fgene.2019.01006

INTRODUCTION

Successful breeding requires strategies that balance immediate genetic gain with the maintenance of population diversity to sustain long-term progress (Jannink, 2010). At each selection cycle, plant breeders are facing the choice of new parental lines and the way in which these are mated, to improve the mean population performance and generate the genetic variation on which selection will act. As breeding programs from different companies compete for short-term gain, breeders tend to use intensively the most performant individuals sometimes at the expense of genetic diversity (Rauf et al., 2010; Gerke et al., 2015; Allier et al., 2019a). The identification of the crossing plan that maximizes the performance in progeny and limits diversity reduction for long-term genetic gain is essential.

Historically, breeders used to select the best individuals based on phenotypic observations, considered as a proxy of their breeding value, i.e., the expected value of their progeny. In order to better estimate the breeding value of individuals, phenotypic selection has been complemented by pedigree-based prediction of breeding values (Henderson, 1984; Piepho et al., 2008) and more recently by genomic prediction of breeding values (Meuwissen et al., 2001), taking advantage of the availability of cheap high-density genotyping. In genomic selection (GS), a model calibrated on phenotype and genotype information of a training population is used to predict genomic estimated breeding values (GEBVs) from genome-wide marker information. A truncation selection is commonly applied on GEBVs, and the selected individuals are intercrossed to create the next generation. The interest of GS is due to the acceleration of selection progress by shortening generation interval, the increase in selection intensity, and the increase in accuracy (Hayes et al., 2010; Daetwyler et al., 2013; Heslot et al., 2015). As a consequence, compared to phenotypic selection, GS is expected to accelerate the loss of genetic diversity due to the rapid fixation of genomic regions with large effects, but also the higher probability to select individuals that are the closest to the training population and are therefore predicted more accurately (Clark et al., 2011; Pszczola et al., 2012). As a result, it has been shown in an experimental study (Rutkoski et al., 2015) and by stochastic simulations (Jannink, 2010; Lin et al., 2016) that GS increases the loss of diversity compared to phenotypic selection. Thus, the optimization of mating strategies in GS breeding programs is a critical area of theoretical and applied research.

Several approaches have been suggested to balance the short- and long-term genetic gain while selecting crosses in GS. In line with Kinghorn, (2011), Pryce et al. (2012), and Akdemir and Isidro-Sánchez (2016), the selection of a set of crosses requires two components: (i) a cross-selection index (CSI) that measures the interest of a set of crosses and (ii) an algorithm to find the set of crosses that maximizes the CSI.

The CSI may consider crosses individually; i.e., the interest of a cross does not depend on the other crosses in the selected set. In classical recurrent GS, candidates with the highest GEBVs are selected and intercrossed to maximize the expected progeny mean in the next generation. In this case, the CSI is simply the mean of parental GEBVs. However, such an approach maximizes neither the expected response to selection in the progeny, which involves genetic variance generated by Mendelian segregation within each family, nor the long-term genetic gain. Alternative measures of the interest of a cross have been proposed to account for parent complementarity, based on within cross variability and expected response to selection. Daetwyler et al. (2015) proposed the optimal haploid value (OHV) that accounts for the complementarity between parents of a cross for predefined haplotype segments. Using stochastic simulations, the authors observed that OHV selection yielded higher long-term genetic gain and preserved greater amount of genetic diversity than truncation GS. However, OHV accounts for neither the position of quantitative trait loci (QTLs) nor the linkage disequilibrium between QTLs (Lehermeier et al., 2017b; Müller et al., 2018). Schnell and Utz (1975) proposed the usefulness criterion (UC) of a cross to evaluate the expected response to selection in its progeny. The UC of a cross accounts for the progeny mean (μ) that is the mean of parental GEBVs and the progeny standard deviation (σ) the selection intensity (i) and the selection accuracy (h): $UC = \mu + ih\sigma$. Zhong and Jannink (2007) proposed to predict progeny variance using estimated QTL effects, accounting for linkage between loci. Genome-wide marker effects have also been considered to predict the progeny variance with computationally intensive stochastic simulations (e.g., Mohammadi et al., 2015). Recently, an unbiased predictor of progeny variance (σ^2) has been derived in Lehermeier et al. (2017b) for two-way crosses and extended in Allier et al. (2019b) for multiparental crosses implying up to four parents. Lehermeier et al. (2017b) observed that using UC as a CSI increased the short-term genetic gain compared to using OHV or mean parental GEBV. Similar results have been obtained by simulations by Müller et al. (2018), considering the expected maximum haploid breeding value (EMBV) that is akin to the UC for normally distributed and fully additive traits.

Alternatively, one can consider a more holistic CSI for which the interest of a cross depends on the other selected crosses. This is the case in optimal contribution selection (Wray and Goddard, 1994; Meuwissen, 1997; Woolliams et al., 2015), where a set of candidate parents is evaluated as a whole regarding the expected short-term gain and the associated risk on loosing long-term gain. Optimal contribution selection aims at identifying the optimal contributions (c) of candidate parents to the next generation obtained by random mating, in order to maximize the expected genetic value in the progeny (V) under a certain constraint on inbreeding (D). Optimal cross-selection, further referred as OCS, is an extension of the optimal contribution selection to deliver a crossing plan that maximizes V by considering additional constraints on the allocation of mates in crosses to limit D (Kinghorn et al., 2009; Kinghorn, 2011; Akdemir and Isidro-Sánchez, 2016; Gorjanc et al., 2018; Akdemir et al., 2018). In GS, the expected genetic value in progeny (V) to be maximized is the mean of parental GEBV ($\boldsymbol{a}$) weighted by parental contributions $\boldsymbol{c}$, i.e $\boldsymbol{c'a}$, and the constraint on inbreeding (D) to be minimized is $\boldsymbol{c'Kc}$ with $\boldsymbol{K}$ a genomic coancestry matrix. Differential evolutionary algorithms have been proposed to obtain optimal solutions for $\boldsymbol{c}$ and the crossing plan (Storn and Price, 1997; Kinghorn et al., 2009; Kinghorn, 2011). Optimal contribution selection is commonly used in animal breeding (Woolliams et al. 2015) and is increasingly adopted in plant breeding (Akdemir and Isidro-Sánchez, 2016; De Beukelaer et al., 2017; Lin et al., 2017; Gorjanc et al., 2018; Akdemir et al., 2018).

In plant breeding, one typically has larger biparental families than in animal breeding. Especially with GS, the selection intensity within-family can be largely increased so that plant breeders capitalize much more on the segregation variance within families than animal breeders. In previous works, the genetic gain (V) and constraint (D) have been defined at the level of the progeny before within-family selection. Exceptions are the work of Shepherd and Kinghorn (1998) and Akdemir and Isidro-Sánchez (2016); Akdemir et al. (2018), who added a term to V accounting for within cross variance assuming linkage equilibrium between QTLs. To our knowledge, no previous

study considered linkage disequilibrium (LD) between QTLs. Furthermore, as observed in historical wheat data (Fradgley et al., 2019) and using simulations in a maize context (Allier et al., 2019b), within-family selection also affects the effective contribution of parents to the next generation. This likely biases the prediction of inbreeding/diversity in the next generation, which to our knowledge has not been considered in previous studies.

In this study, we propose to adjust V and D terms so that within-family selection of the candidate parents for the next generation is accounted for. We propose to use the usefulness criterion parental contribution (UCPC) approach (Allier et al., 2019b) that enables to predict the expected mean performance of the selected fraction of progeny and to predict the contribution of parents to the selected fraction of progeny. We compared our OCS strategy based on UCPC with other cross-selection strategies, in a long-term simulated recurrent GS breeding program involving overlapping generations (**Figure 1A**). Our objectives were to demonstrate (1) the interest of UCPC to predict the genetic diversity in the selected fraction of progeny and (2) the interest of accounting for within-family selection in OCS for both short- and long-term genetic gains.

MATERIALS AND METHODS

Simulated Breeding Program

Breeding Program

We simulated a breeding program to compare the effect of different CSIs on short- and long-term genetic gain in a realistic breeding context considering overlapping and connected generations (i.e., cohorts) and the use of doubled haploid (DH) technology to derive progeny (**Figure 1A**). We considered that the process to derive DH progeny from a cross and to phenotype and genotype DH lines takes 3 years. Furthermore, we considered as candidate parents of a cohort T the selected fraction of DH progeny of the three last available cohorts, i.e., T-3, T-4 and T-5 (**Figures 1A**, **B**).

Each simulation replicate started from a population of 40 founders sampled among 57 Iodent maize genotypes from the Amaizing project (Rio et al., 2019; Allier et al., 2019b). We sampled 1,000 biallelic QTLs among the 40,478 high-quality single-nucleotide polymorphisms (SNPs) from the Illumina MaizeSNP50 BeadChip (Ganal et al. 2011), with consensus genetic positions from Giraud et al. (2014). The sampling process obeyed two constrains: a QTL minor allele frequency ≥ 0.2 and a distance between two consecutive QTLs ≥ 0.2 cM. Each QTL was assigned an additive effect sampled from a Gaussian distribution with a mean of zero and a variance of 0.05, and the favorable allele was attributed at random to one of the two SNP alleles.

We initiated a virtual breeding program starting from the founder genotypes with a burn-in period of 20 years that mimicked recurrent phenotypic selection. Burn-in started by randomly crossing the 40 founders into 20 biparental families, i.e., two-way crosses, during the first 3 years to initiate three overlapping cohorts. In each cohort, 80 DH progeny genotypes per cross were simulated. Phenotypes were simulated considering the genotype at QTLs, an error variance corresponding to a trait repeatability of 0.4 in the founder population and no genotype by environment interactions. For phenotyping, every individual was evaluated in four environments in 1 year. Since no secondary trait was considered and sufficient seed production for extensive progeny testing was assumed, we simulated a unique within-family selection of the 5% best progeny (i.e., 4 DHs) that is a common selection

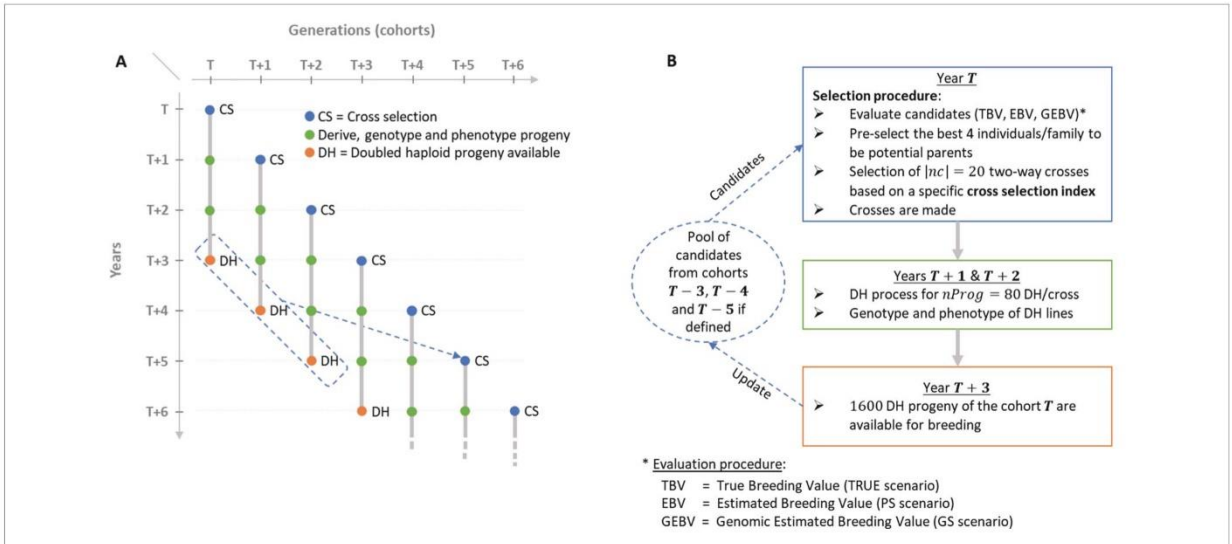


FIGURE 1 | Schematic view of the simulated breeding program. **(A)** Overall view of the breeding program and overlapping cohorts. **(B)** Life cycle of a given post burn-in cohort T depending on the scenario considered (TRUE with 1,000 known QTL effects, PS in absence of genomic information or GS with 2,000 noncausal SNPs estimated effects).

intensity in maize breeding. During burn-in, we first considered within-family phenotypic selection and then used the 50 DHs with the largest phenotypic mean as potential parents of the next cohort. These were randomly mated, i.e., without any constraint on parental contributions, to generate 20 biparental families of 80 DH lines. After 20 years of burn-in, this created extensive linkage disequilibrium as often observed in elite plant breeding programs (e.g., Van Inghelandt et al., 2011). We then compared different CSIs for 60 years of recurrent GS using DH technology (**Figure 1**). As in burn-in, each cohort T was generated by 20 two-way crosses ($|nc|$=20) of 80 DH progeny each ($nProg$ = 80). Candidate parents of cohort T were selected from the available DH of the three cohorts: T=3, T-4, and T-5 (**Figures 1A**, **B**). Per family, the 4 DH lines (i.e., 5%) with the largest breeding values, detailed in "Evaluation scenario" section, were considered as potential parents, yielding 4 DH lines/family × 20 families/cohort × 3 cohorts = 240 potential parents. Considering these N = 240 potential parents, $N(N\text{-}1)/2$ = 28,680 two-way crosses are possible. The set of $|nc|$ = 20 two-way crosses among these 28,680 candidate crosses was defined using different CSI detailed in the following sections. This simulated scheme yielded overlapping and connected cohorts as it is standard in practical plant breeding (**Figure 1A**). A detailed description of the simulated breeding program and the material is provided in **Supplementary Material (File S1)**.

Evaluation Scenarios

We considered different scenarios for genome-wide marker effects and progeny evaluation. In order to eliminate the uncertainty caused by the estimation of marker effects, we first compared several CSI assuming that we have access to the positions and effects of the 1,000 QTLs (referred to as TRUE scenario). For a representative subset of the CSI showing differentiated results in the TRUE scenario, we also considered a more realistic scenario where the effects of QTLs are unknown and selection was based on the effects of 2,000 noncausal SNPs randomly sampled over the genome. In this scenario, marker effects were obtained by back-solving (Wang et al., 2012) a G-BLUP model fitted using blupf-90 AI-REML solver (Misztal, 2008). This scenario was referred to as GS scenario, and marker effects used to predict the CSI were estimated every year with all candidate parents that were phenotyped and genotyped. The progeny were selected on their GEBV considering their phenotypes and their genotypes at noncausal SNPs. As a benchmark, we also considered a phenotypic selection scenario where progeny were selected based on their phenotypic mean (PS scenario). For details on the evaluation models, see **Supplementary Material (File S1)**. In the following, for sake of clarity, we present the different cross-selection strategies considering selection based on known QTL effects and positions (TRUE scenario). In GS scenario, QTL effects and positions were replaced by estimated marker effects and positions.

Cross-Selection Strategies

Optimal Cross-Selection Not Accounting for Within-Family Selection

Considering N homozygote candidate parents, $N(N\text{-}1)/2$ two-way crosses are possible. We define a crossing plan $\boldsymbol{nc}$ as a set of $|nc|$ crosses out of possible two-way crosses, giving the index of selected crosses, i.e., with the i^{th} element $nc(i)\in[1,N(N\text{-}1)/2]$. The $(N \times 1)$ dimensional vector of candidate parents contributions $\boldsymbol{c}$ is defined as

$$\boldsymbol{c} = \frac{1}{|nc|}\left(\boldsymbol{Z}_1\boldsymbol{c}_1 + \boldsymbol{Z}_2\boldsymbol{c}_2\right), \quad (1)$$

where $\boldsymbol{Z}_1$ (respectively $\boldsymbol{Z}_2$) is a $(N \times |nc|)$ dimensional design matrix that links each N candidate parent to the first (respectively second) parent in the set of crosses $\boldsymbol{nc}$, $\boldsymbol{c}_1$ (respectively, $\boldsymbol{c}_2$) is a $(|nc| \times 1)$ dimensional vector containing the contributions of the first (respectively, second) parent to progeny, i.e., a vector of 0.5 when assuming no selection within crosses.

The $(N\times 1)$ dimensional vector of candidate parents true breeding values is $\boldsymbol{a} = \boldsymbol{X}\beta_T$ where $\boldsymbol{X} = (x_1,\ldots,x_N)'$ is the $(N \times m)$ dimensional matrix of known parental genotypes at m biallelic QTLs, where $\boldsymbol{x}_p$ denotes the $(m \times 1)$ dimensional genotype vector of parent $p\in[1,N]$ with the j^{th} element coded as 1 or −1 for the genotypes AA or aa at QTL j. β_T is the $(m \times 1)$ dimensional vector of known additive QTL effects for the quantitative agronomic performance trait considered. The genetic gain $V(\boldsymbol{nc})$ for this set of two-way crosses is defined as the expected mean performance in the DH progeny:

$$V(\boldsymbol{nc}) = \boldsymbol{c}'\boldsymbol{a}. \quad (2)$$

We define the constraint on diversity (D) as the mean expected genetic diversity in DH progeny (He, Nei, 1973):

$$D(\boldsymbol{nc}) = 1 - \boldsymbol{c}'\boldsymbol{K}\boldsymbol{c}, \quad (3)$$

where $\boldsymbol{K} = \frac{1}{2}\left(\frac{1}{m}\boldsymbol{X}\boldsymbol{X}' + 1\right)$ is the $(N \times N)$ dimensional identity by state (IBS) coancestry matrix between the N candidates. **Supplementary Material (File S2)** details the relationship between the IBS coancestry among parents ($\boldsymbol{K}$), the parental contributions to progeny ($\boldsymbol{c}$) and the mean expected heterozygosity in progeny $He = \frac{1}{m}\sum_{j=1}^{m} 2p_j(1-p_j)$ where p_j the frequency of the genotypes AA at QTL j in the progeny.

Accounting for Within-Family Selection in OCS

In the OCS, as defined above, the progeny derived from the $\boldsymbol{nc}$ crosses are all expected to contribute to the next generation. We propose to consider $V(\boldsymbol{nc})$ and $D(\boldsymbol{nc})$ terms accounting for the fact that only a selected fraction of each family will be candidate for the next generation (e.g., 5% per family in our simulation study). For this, we apply the UCPC approach proposed by Allier et al. (2019b) for two-way crosses and extend its use to evaluate the interest of a set $\boldsymbol{nc}$ of two-way crosses after selection in progeny.

UCPC for Two-Way Crosses

Two inbred lines P_1 and P_2 are considered as parental lines for a candidate cross $P_1 \times P_2$ and $(\boldsymbol{x}_1, \boldsymbol{x}_2)'$ denotes their genotyping

matrix. Following Lehermeier et al. (2017b), the DH progeny mean and progeny variance of the performance in the progeny before selection can be computed as follows:

$$\mu_T = 0.5\left(\boldsymbol{x}'_1\boldsymbol{\beta}_T + \boldsymbol{x}'_2\boldsymbol{\beta}_T\right), \quad (4a)$$

$$\sigma_T^2 = \boldsymbol{\beta}_T' \, \Sigma \, \boldsymbol{\beta}_T, \quad (4b)$$

where $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and β_T were defined previously, and Σ is the $(m\times m)$ -dimensional variance covariance matrix of QTL genotypes in DH progeny defined in Lehermeier et al. (2017b).

To follow parental contributions, we consider P_1 parental contribution as a normally distributed trait (Allier et al., 2019b). As we only consider two-way crosses and biallelic QTLs, we can simplify for computational reasons the formulas by using IBS parental contributions computed for polymorphic QTLs between P_1 and P_2 instead of using identity-by-descent parental contributions (Allier et al., 2019b). We define the $(m\times 1)$ -dimensional vector β_{C1} to follow P_1 genome contribution at QTLs as $\boldsymbol{\beta}_{C1} = \frac{\boldsymbol{x}_1 - \boldsymbol{x}_2}{(\boldsymbol{x}_1 - \boldsymbol{x}_2)'(\boldsymbol{x}_1 - \boldsymbol{x}_2)}$. We compute the mean of P_1 contribution in the progeny before selection $\mu_{C1}=0.5(\boldsymbol{x}'_1\beta_{C1}+\boldsymbol{x}'_2\beta_{C1}+1)$. The progeny variance σ^2_{C1} for P_1 contribution in the progeny before selection is computed using Eq. 4b by replacing β_T by β_{C1} The progeny mean for P_2 contribution is then defined as $\mu_{C2} = 1-\mu_{C1}$.

Following Allier et al. (2019b), we compute the covariance between the performance and P_1 contribution in progeny as follows:

$$\sigma_{T,\,C1} = \boldsymbol{\beta}'_T \Sigma \, \boldsymbol{\beta}_{C1}. \quad (5)$$

The expected mean performance of the selected fraction of progeny, i.e., UC (Schnell and Utz, 1975), of the cross $P_1\times P_2$ is as follows:

$$UC^{(i)} = \mu_T + ih\sigma_T, \quad (6)$$

where i is the within-family selection intensity, and the exponent (i) in UC expresses the dependency of UC on the selection intensity i. We considered a selection accuracy h=1 as in Zhong and Jannink (2007), which holds when selecting on true breeding values in TRUE scenario. As discussed further, we also considered h = 1 when selecting crosses based on UCPC in GS scenario. The correlated responses to selection on P_1 and P_2 genome contributions in the selected fraction of progeny are as follows (Falconer and Mackay, 1996):

$$c_1^{(i)} = \mu_{C1} + i\frac{\sigma_{T,C1}}{\sigma_T} \text{ and } c_2^{(i)} = 1 - c_1^{(i)}. \quad (7)$$

Cross-Selection Based on UCPC

Accounting for within-family selection intensity i, the genetic gain term $V^{(i)}(\boldsymbol{nc})$ for a set of two-way crosses $\boldsymbol{nc}$ is defined as the expected performance in the selected fraction of progeny:

$$V^{(i)}(\boldsymbol{nc}) = \frac{1}{|\boldsymbol{nc}|}\sum_{j\in\boldsymbol{nc}} UC^{(i)}(j). \quad (8)$$

The constraint on diversity $D^{(i)}(\boldsymbol{nc})$ in the selected progeny is defined as follows:

$$D^{(i)}(\boldsymbol{nc}) = 1 - \boldsymbol{c}^{(i)\prime}\boldsymbol{K}\boldsymbol{c}^{(i)}, \quad (9)$$

where $\boldsymbol{c}^{(i)}$ is defined like $\boldsymbol{c}$ in Eq. 1 but accounting for within-family selection by replacing the ante-selection parental contributions $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ by the post-selection parental contributions $\boldsymbol{c}_1^{(i)}$ and $\boldsymbol{c}_2^{(i)}$ (Eq. 7), respectively. Note that considering the absence of selection in progeny, i.e., $i = 0$, yields $V^{(i=0)}(\boldsymbol{nc})$ being the mean of parent breeding values (Eq. 2) and $D^{(i=0)}(\boldsymbol{nc})$ being the expected diversity in progeny before selection (Eq. 3), which is equivalent to optimal cross-selection as proposed by Gorjanc et al. (2018). The R code (R Core Team, 2017) to evaluate a set of crosses as presented in the UCPC-based optimal cross-selection is provided in **Supplementary Material (File S3)**.

Multiobjective Optimization Framework

In practice, one does not evaluate only one set of crosses but several ones in order to find the optimal set of crosses to reach a specified target that is a function of $V^{(i)}(\boldsymbol{nc})$ and $D^{(i)}(\boldsymbol{nc})$. We use the ε-constraint method (Haimes et al., 1971; Gorjanc and Hickey, 2018) to solve the multiobjective optimization problem:

$$\begin{aligned} &\max_{\boldsymbol{nc}} V^{(i)}(\boldsymbol{nc}) \\ &\text{with } D^{(i)}(\boldsymbol{nc}) \geq He(t), \end{aligned} \quad (10)$$

where $He(t)$, $\forall t\in[0,t^*]$ is the minimal diversity constraint at time t. A differential evolutionary (DE) algorithm was implemented to find the set of $\boldsymbol{nc}$ crosses that is a Pareto-optimal solution of Eq. 10 (Storn and Price, 1997; Kinghorn et al., 2009; Kinghorn, 2011). DE is an optimization process inspired by natural selection. It started from an initial population of 7,170 random candidate solutions that are improved during 1,000 iterations through mutation (random changes in candidate solutions), recombination (exchanges between candidate solutions), and selection (every iteration a candidate solution was replaced by its mutated and recombined version if superior). The direct consideration of $He(t)$ in the optimization allows to control the decrease in genetic diversity similarly to what was suggested for controlling inbreeding rate in animal breeding (Woolliams et al., 1998, Woolliams et al., 2015). The loss of diversity along time is controlled by the targeted diversity trajectory, i.e., $He(t)$,

$\forall t \in [0,t^*]$, where $t^* \in \mathbb{N}^*$ is the time horizon when the genetic diversity $He(t^*) = He^*$ should be reached. In this study, $He(t)$ is defined as follows:

$$He(t) = \begin{cases} He^0 + \left(\frac{t}{t^*}\right)^s \left(He^* - He^0\right), \forall\ t \in [\![0,t^*]\!], \\ He^*, \forall\ t > t^* \end{cases} \quad (11)$$

where He^0 is the initial diversity at $t = 0$, and s is a shape parameter with $s = 1$ for a linear trajectory. **Figure 2** gives an illustration of alternative trajectories that can be defined using Eq. 11.

Cross-Selection Indices

We considered different cross-selection approaches varying in the within-family selection intensity (i) in $V^{(i)}(\boldsymbol{nc})$, $D^{(i)}(\boldsymbol{nc})$ (Eq. 10) and in the targeted diversity trajectory $He(t)$ (Eq. 11). We first considered as a benchmark the absence of constraint $D^{(i)}(\boldsymbol{nc})$, i.e., $He(t) = 0$, $\forall t$. We defined two alternative CSIs PM (parental mean) and UC, respectively considering $V^{(i=0)}(\boldsymbol{nc})$ and $V^{(i=2.06)}(\boldsymbol{nc})$, with $i = 2.06$ corresponding to the selection of the 5% most performant progeny per family. PM is equivalent to cross the best candidates together without accounting for within cross variance, while UC is defined as crossing candidates based on the expected mean performance of the 5% selected fraction of progeny. Note that the absence of constraint on diversity also means the absence of constraint on parental contributions. To compare optimal cross-selection accounting or not for within-family selection, we considered three linear diversity trajectories (Eq. 11) with $He^* = \{0.01, 0.10, 0.15\}$ that should be reached in $t^* = 60$ years. We defined the OCS methods, further referred to as OCS-He*, with $V^{(i=0)}(\boldsymbol{nc})$ and $D^{(i=0)}(\boldsymbol{nc})$. We defined the UCPC cross-selection methods, further referred to as UCPC-He*, with $V^{(i=2.06)}(\boldsymbol{nc})$ and $D^{(i=2.06)}(\boldsymbol{nc})$. The eight CSIs considered are summarized in **Table 1**.

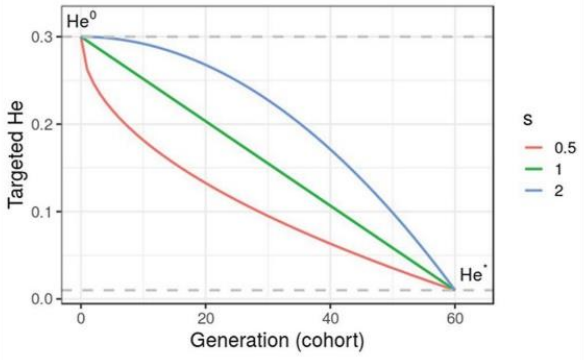


FIGURE 2 | Targeted diversity trajectories for three different shape parameters (s = 1, linear trajectory; s = 2, quadratic trajectory; and s = 0.5, inverse quadratic trajectory) for fixed initial diversity ($He^0 = 0.3$) at generation 0 and targeted diversity ($He^* = 0.01$) at generation 60 ($t^* = 60$). We considered in this study only linear trajectories (s = 1).

Simulation 1: Interest of UCPC to Predict the Diversity in the Selected Fraction of Progeny

Simulation 1 aimed at evaluating the interest to account for the effect of selection on parental contributions, i.e., post-selection parental contributions (using UCPC), compared to ignore selection, i.e., ante-selection parental contributions (similarly as in OCS), to predict the genetic diversity (He) in the selected fraction of progeny of a set of 20 crosses (using Eqs. 9 and 3, respectively). We considered a within-family selection intensity corresponding to selecting the 5% most performant progeny. We used the same genotypes, genetic map, and known QTL effects as for the first simulation replicate of the PM CSI in the TRUE scenario (**Table 1**). We extracted the simulated genotypes of 240 DH candidate parents of the first post burn-in cohort (further referred as E1) and of 240 DH candidate parents of the 20th post burn-in cohort (further referred as E2). Due to the selection process, E1 showed a higher diversity and lower performance compared to E2. We randomly generated 300 sets of 20 two-way crosses: 100 sets of intrageneration E1 crosses (E1 × E1), 100 sets of intrageneration E2 crosses (E2 × E2), and 100 sets of intergeneration and intrageneration crosses randomly sampled (E1 × E2, E1 × E1, E2 × E2). We derived 80 DH progeny per cross and predicted the ante- and post-selection parental contributions to evaluate the post-selection genetic diversity (He) for each set of crosses. We estimated the empirical post-selection diversity for each set of crosses and compared predicted and empirical values considering the mean prediction error as the mean of the difference between predicted He and empirical post-selection He, and the prediction accuracy as the squared correlation between predicted He and empirical post-selection He.

Simulation 2: Comparison of Different Csis

We ran 10 independent simulation replicates of all eight CSI summarized in **Table 1** for 60 years post burn-in considering known effects at the 1,000 QTLs (TRUE scenario). We also compared in 10 independent simulation replicates the CSI: PM,

TABLE 1 | Summary of tested cross-selection indices (CSI) in TRUE scenario defined for a set of crosses ***nc*** depending on the within-family selection intensity i.

Cross-selection index (CSI)	Gain term	Diversity term
PM	$V^{(i=0)}(\boldsymbol{nc})$	–
OCS-He* (3 different He*)	$V^{(i=0)}(\boldsymbol{nc})$	$D^{(i=0)}(\boldsymbol{nc})$
UC	$V^{(i=2.06)}(\boldsymbol{nc})$	–
UCPC-He* (3 different He*)	$V^{(i=2.06)}(\boldsymbol{nc})$	$D^{(i=2.06)}(\boldsymbol{nc})$

*$He^* = \{0.15; 0.10; 0.01\}$ to be reached linearly (s = 1) at the end of simulation ($t^* = 60$ years). $V^{(i=0)}(\boldsymbol{nc})$ is the averaged parental mean (PM) of crosses in **nc** and $V^{(i=2.06)}(\boldsymbol{nc})$ is the averaged usefulness criterion (UC) of crosses in **nc** considering a within-family selection intensity of 2.06. $D^{(i=0)}(\boldsymbol{nc})$ and $D^{(i=2.06)}(\boldsymbol{nc})$ are the expected genetic diversity in the progeny before and after within-family selection, respectively.*

UC, OCS-He* and UCPC-He* with He* = 0.01 considering estimated marker effect at the 2,000 SNPs (GS scenario) and PM based only on phenotypic evaluation (PS scenario). We followed several variables on the 80 DH progeny/family × 20 crosses realized every year. At each cohort $T \in [0,60]$ with $T = 0$ co rresponding to the last burn-in cohort, we computed the additive genetic variance as the variance of the 1,600 DH progeny true breeding values (TBVs): $\sigma_A^2(T) = var(TBV(T))$. We followed the mean genetic merit of all progeny $\mu(T)$ = mean($TBV(T)$) and of the 10 most performant progeny $\mu_{10}(T)$= mean(max($TBV(T)$)) as a proxy of realized performance that could be achieved at a commercial level by releasing these lines as varieties. Then, we centered and scaled the two genetic merits to obtain realized cumulative genetic gains in units of genetic standard deviation at the end of the burn-in ($T = 0$), at the whole progeny level $G(T) = (\mu(T) - \mu(0))/\sqrt{\sigma_A^2(0)}$ and at the commercial level $G_{10}(T) = (\mu_{10}(T) - \mu(0))/\sqrt{\sigma_A^2(0)}$.

The interest of long-term genetic gain relies on the ability to breed at long term, which depends on the short-term economic success of breeding. Following this rationale, we penalized strategies that compromised the short-term commercial genetic gain using the discounted cumulative gain following Dekkers et al. (1995) and Chakraborty et al. (2002). In practice, we computed the weighted sum of the commercial gain value in each generation $\sum_{T=1}^{60} w_T \, G_{10}(T)$, where the discounted weights $w_T = 1/(1+\rho)^T, \forall T \in [1,60]$ were scaled to have $\sum_{T=1}^{60} w_T = 1$ and ρ is the interest rate per generation. The discounted weights measure how much breeders will care about future genetic gain compared to today's genetic gain, also referred as the "net present value" of long-term gain in finance. For ρ = 0, the weights were $w_{T \in [1,60]} = 1/60$; i.e., the same importance was given to all cohorts. We compared different values of ρ and reported results for ρ = 0, ρ = 0.04 giving approximatively seven times more weight to short-term gain (after 10 years) compared to long-term gain (after 60 years) and ρ = 0.2 giving nearly no weight to gain after 30 years of breeding.

We also measured the additive genic variance at QTLs $\sigma_a^2(T) = \sum_{j=1}^{m} 4 \, p_j(T)(1 - p_j(T))\beta_j^2$, the mean expected heterozygosity at QTLs (He, Nei, 1973) $He(T) = m^{-1} \sum_{j=1}^{m} 2 \, p_j(T)(1 - p_j(T))$, and the number of QTLs where the favorable allele was fixed or lost in the progeny, with $p_j(T)$ the allele frequency at QTL $j \in [1,m]$ in the 1,600 DH progeny and β_j the additive effect of the QTL j. In addition, we considered the ratio of additive genetic over genic variance σ_A^2 / σ_a^2. hich provides an estimate of the amount of additive genic variance captured by negative covariances between QTLs, known as the Bulmer effect under directional selection (Bulmer, 1971, Bulmer, 1980; Lynch and Walsh, 1999). All these variables were further averaged on the 10 simulation replicates, and the standard error divided by the square root of the number of replicates is reported.

RESULTS

Simulation 1

Compared to the usual approach that ignores the effect of selection on parental contributions, accounting for the effect of within-family selection increased the squared correlation (R^2) between predicted genetic diversity and genetic diversity in the selected fraction of progeny (**Figures 3A, B**) for all three types of crosses. The squared correlation between predicted genetic diversity and post-selection genetic diversity for intrageneration crosses was only slightly increased (E1 × E1: from 0.811 to 0.822 and E2 × E2: from 0.880 to 0.888), while the squared correlation for sets of crosses involving also intergeneration crosses showed a larger increase (from 0.937 to 0.987) (**Figures 3A, B**). Using post-selection parental contributions instead of ante-selection parental contributions also reduced the mean prediction error of He (predicted – empirical He) (**Figures 4A, B**) for all three types of crosses. The mean prediction error for intrageneration crosses

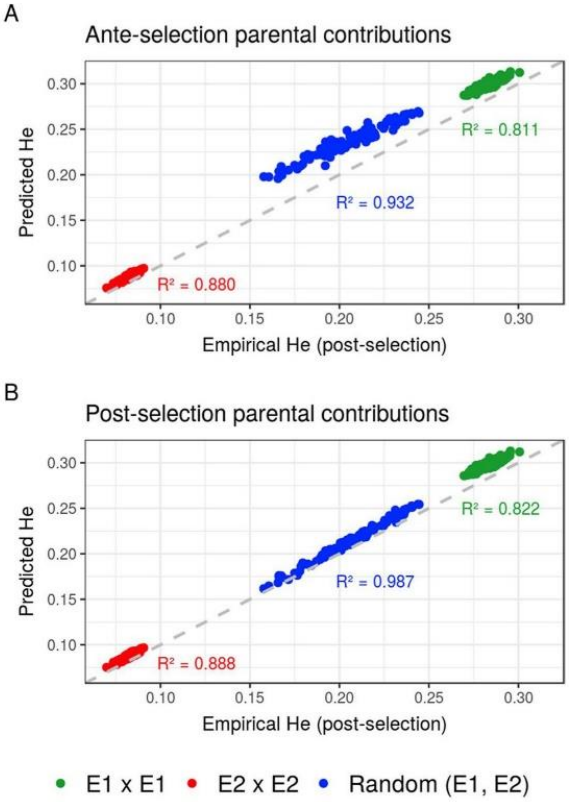


FIGURE 3 | Squared correlations (R^2) between predicted genetic diversity (He) and empirical He in the selected fraction of progeny of a set of 20 biparental crosses in the TRUE scenario considering **(A)** ante-selection parental contributions or **(B)** post-selection parental contributions to predict He. In total, 100 sets of each three types of crosses (intrageneration: E1xE1 and E2xE2 or randomly intragenerations and intergenerations): random (E1, E2) are shown, and the squared correlations between predicted and empirical post-selection He are given in the corresponding color.

was only slightly reduced (E1 × E1: from 0.006 to 0.005 and E2 × E2: from 0.016 to 0.015), while the mean prediction error for sets involving intergeneration crosses was more reduced (from 0.032 to 0.008) (**Figures 4A, B**). The mean prediction error of He was reduced but still positive when considering post-selection parental contributions, which means that the genetic diversity in the selected fraction of progeny remains overestimated. Note that the ante-selection contributions predicted well the empirical genetic diversity before selection for all three types of crosses (mean prediction error = 0.000 and $R^2 > 0.992$, results not shown).

Simulation 2

Interest of UC Over PM

Considering known QTL effects (TRUE scenario), we observed that UC yielded significantly higher short- and long-term genetic gain at commercial level (G_{10}) than PM (on average, G_{10} = 9.316 [±0.208] compared to 8.338 [±0.195] 10 years post burn-in and G_{10} = 18.293 [±0.516] compared to 15.744 [±0.449] 60 years post burn-in; **Figures 5B, C**; **Supplementary Material [Table S1 File S4]**). When considering the whole progeny mean performance (G), PM nonsignificantly outperformed UC for the first 5 years (on average, G = 4.647 [±0.174] compared to 4.633 [±0.138] 5 years post burn-in), and after 5 years, UC significantly outperformed PM (on average, G=7.620 [±0.158] compared to 7.197 [±0.199] 10 years post burn-in) [**Figure 5A**, **Supplementary Material (Table S1 File S4)**]. UC showed higher genic (σ^2_a) and genetic (σ^2_A) additive variances than PM (**Figures 6A, B**), but both yielded a genic and genetic variance near zero after 60 years of breeding. The genetic over genic variance ratio (σ^2_A / σ^2_a) was also higher for UC compared to PM (**Figure 6C**). The evolution of genetic diversity (He) along years followed the same tendency as the genic variance (**Figure 7A**, **Figure 6A**). UC fixed more favorable alleles at QTLs after 60 years (**Figure 7B**) and lost less favorable alleles at QTLs than PM in all 10 simulation replicates, with an average of 243.1 (±4.547) QTLs where the favorable allele was lost compared to 274.9 (±4.283) QTLs for PM [**Figure 7C**; **Supplementary Material (Table S1 File S4)**].

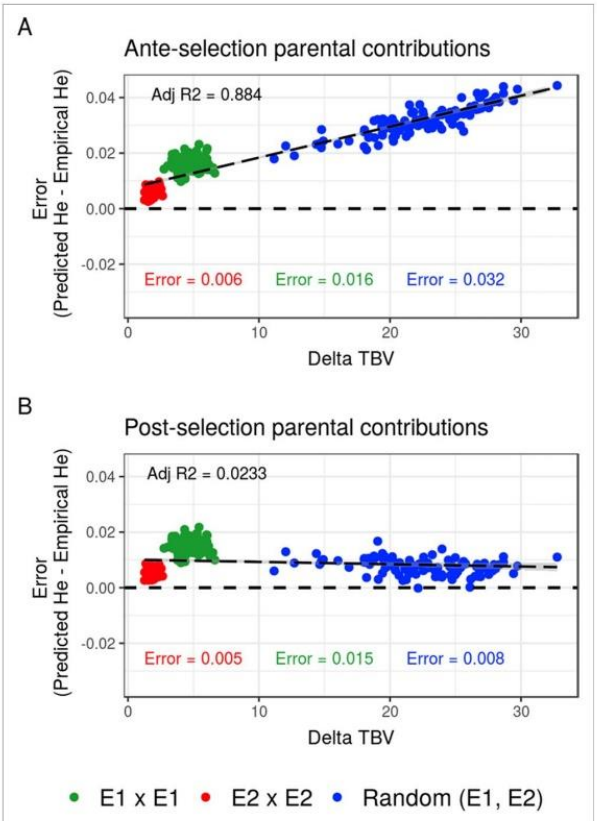


FIGURE 4 | Mean prediction error (predicted – empirical) of predicting the genetic diversity (He) in the selected fraction of progeny of a set of 20 biparental crosses in the TRUE scenario depending on the mean difference of performance between parents (Delta true breeding value TBV). Mean prediction error is measured as the predicted He – empirical post-selection He, considering **(A)** ante-selection parental contributions or **(B)** post-selection parental contributions to predict He. In total, 100 sets of each three types of crosses (intrageneration: E1 × E1 and E2 × E2 or randomly intra and inter-generations): random (E1, E2) are shown, and the averaged errors are given in the corresponding color.

Targeted Diversity Trajectory

Considering known QTL effects (TRUE scenario), the tested optimal cross-selection methods OCS-He* and UCPC-He* showed lower short-term genetic gain at the whole progeny level (G; **Figure 5A**) and at the commercial level (G_{10}; **Figures 5B, C**) but significantly higher long-term genetic gains than UC at 60 years **Supplementary Material (Table S1 File S4)**. The lower the targeted diversity He*, the higher the short-term and midterm genetic gain at both whole progeny (G; **Figure 5A**) and commercial (G_{10}; **Figures 5B, C**) levels. The higher the targeted diversity He*, the higher the long-term genetic gain except for OCS-He* = 0.10 and OCS-He* = 0.01 that performed similarly after 60 years (on average, G_{10} = 21.925 [±0.532] and 21.892 [±0.525]; **Figure 5B**, **Supplementary Material [Table S1 File S4]**). The highest targeted diversity (He* = 0.15) showed a strong penalty at the short term and midterm, while the intermediate targeted diversity (He* = 0.10) showed a lower penalty at the short term and midterm compared to the lowest targeted diversity (He* = 0.01) (**Figures 5A–C**).

For all targeted diversities and all simulation replicates, accounting for within-family selection (UCPC-He*) yielded a significantly higher short-term commercial genetic gain (G_{10}) after 5 and 10 years compared to OCS-He* [**Figures 5B, C**; **Supplementary Material (Table S1 File S4)**]. Long-term commercial genetic gain (G_{10}) after 60 years was also higher for UCPC-He* than for OCS-He* with He* = 0.01 in the 10 simulation replicates (on average, G_{10} = 22.869 [±0.641] compared to 21.892 [±0.525]) and less importantly with He* = 0.10 in nine out of 10 replicates (on average, G_{10}= 22.474 [±0.645] compared to 21.925 [±0.532]). However, for He* = 0.15, UCPC-He* outperformed OCS-He* at the long term in only three out of 10 replicates (on average, G_{10}= 20.665 [±0.573] compared to 20.938 [±0.553]) [**Figures 5B, C**; **Supplementary Material (Table S1 File S4)**]. The discounted cumulative gain giving more weight to short-term than to long-term gain (ρ = 0.04) was higher for UCPC-He* than

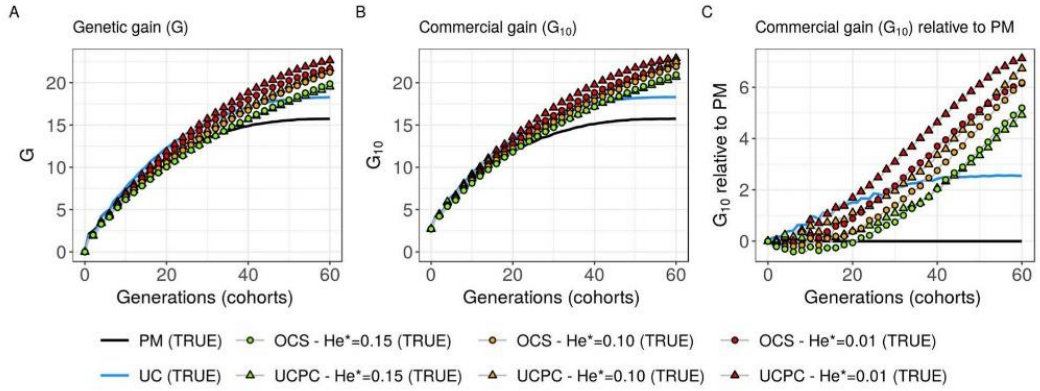


FIGURE 5 | Genetic gains for different cross-selection indices in the TRUE scenario (PM: parental mean, UC: usefulness criterion, OCS-He*: optimal cross-selection and UCPC-He*: UCPC-based optimal cross-selection) according to the generations. **(A)** Genetic gain (G) measured as the mean of the whole progeny, **(B)** commercial genetic gain (G_{10}) measured as the mean of the 10 best progeny, and **(C)** G_{10} relative to selection based on parental mean (PM).

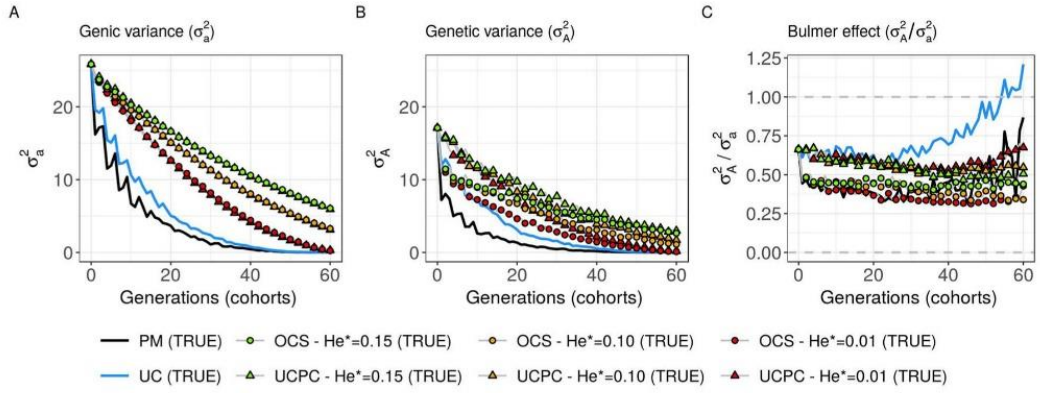


FIGURE 6 | Genetic and genic additive variances for different cross-selection indices in the TRUE scenario (PM: parental mean, UC: usefulness criterion, OCS-He*: optimal cross-selection, and UCPC-He*: UCPC-based optimal cross-selection) according to the generations. **(A)** Additive genic variance (σ^2_a) measured on the whole progeny, **(B)** additive genetic variance (σ^2_A) measured on the whole progeny, and **(C)** ratio of genetic over genic variance (σ^2_A / σ^2_a) reflecting the Bulmer effect.

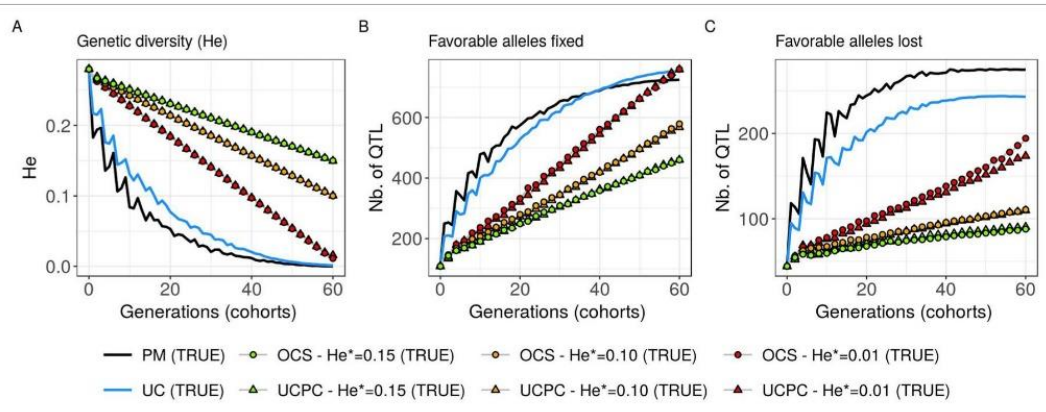


FIGURE 7 | Genetic diversity at QTLs for different cross-selection indices in the TRUE scenario (PM: parental mean, UC: usefulness criterion, OCS-He*: optimal cross-selection, and UCPC-He*: UCPC-based optimal cross-selection) according to the generations. **(A)** Genetic diversity at QTLs in the whole progeny (*He*), **(B)** number of QTLs where the favorable allele is fixed in the whole progeny, and **(C)** number of QTLs where the favorable allele is lost in the whole progeny.

OCS-He* in all simulation replicates for He* = 0.01 (on average, 12.321 [±0.284] compared to 11.675 [±0.262]), in all simulation replicates for He* = 0.10 (on average, 11.788 [±0.280] compared to 11.278 [±0.264]) and in nine out of 10 simulation replicates for He* = 0.15 (on average, 11.176 [±0.250] compared to 10.884 [±0.250]) (**Table 2**). Discounted cumulative gain giving the same weight to short- and long-term gain (ρ = 0) was also higher for UCPC-He* compared to OCS-He* (**Table 2**). When giving almost no weight to long-term gain after 30 years (ρ = 0.2), the best CSI appeared to be UC [on average, 6.822 (±0.145)] followed by the UCPC-He* with the lowest constraint on diversity (i.e., He* = 0.01) [on average, 6.682 (±0.143)].

For a given He*, the additive genic variance (σ^2_a; **Figure 6A**) and genetic diversity at QTLs (He; **Figure 7A**) were constrained by the targeted diversity trajectory for both UCPC-He* or OCS-He*. However, UCPC-He* and OCS-He* behaved differently for genetic variance (σ^2_A ; **Figure 6A**) resulting in differences for the ratio genetic over genic variances (σ^2_A / σ^2_a; **Figure 6C**). UCPC-He* yielded a higher ratio than OCS-He* (**Figure 6C**) independently of the targeted diversity He* at short term and midterm. For low targeted diversity (He* = 0.01), UCPC-He* showed in all 10 replicates a lower number of QTLs where the favorable allele was lost compared to OCS-He* (**Figure 7C**; **Supplementary Material** [**Table S1 File S4**], on average 173.6 [±4.031] QTLs-194.3 [±2.633] QTLs).

GS Scenario With Estimated Marker Effects

Considering estimated marker effects (GS scenario) yielded lower genetic gain than when considering known marker effects [**Figures 5–8** and **Supplementary Material (Tables S1 and S2 File S4)**]. However, the short- and long-term superiority of the UC over the CSI ignoring within cross variance (PM) was consistent with estimated effects (on average, G_{10} = 8.338 [±0.237] compared to 7.713 [±0.256] 10 years post burn-in and G_{10} = 15.367 [±0.358] compared to 13.287 [±0.436] 60 years post burn-in; **Figure 8**, **Supplementary Material** [**Table S2 File S4**]). Similarly, the long-term superiority of UCPC-He* = 0.01 over UC was conserved in all 10 replicates (on average, G_{10} = 16.398 [±0.426] compared to 14.438 [±0.320] 40 years post burn-in and G_{10} = 18.161 [±0.470] compared to 15.367 [±0.358] 60 years post burn-in; **Figure 8**, **Supplementary Material** [**Table S2 File S4**]). Before the 40th year, UC and UCPC-He* = 0.01 performed similarly **Supplementary Material (Table S2 File S4)**. In GS scenario, UCPC-He* = 0.01 outperformed OCS-He* = 0.01 during the first 20 years in all 10 replicates (on average, G_{10} = 8.162 [±0.208] compared to 7.734 [±0.237] 10 years post burn-in and G_{10} = 11.881 [±0.272] compared to 11.313 [±0.323] 20 years post burn-in; **Figure 8**, **Supplementary Material** [**Table S2 File S4**]). After 20 years, UCPC-He* = 0.01 outperformed OCS-He* = 0.01 in eight out of 10 replicates (on average, G_{10} = 16.398 [±0.426] compared to 15.850 [±0.384] 40 years post burn-in and G_{10} = 18.161 [±0.470] compared to 17.528 [±0.438] 60 years post burn-in; **Figure 8**, **Supplementary Material** [**Table S2 File S4**]). Observations on the genic variance (σ^2_a) and genetic variance (σ^2_A) were consistent as well. We also observed that UCPC-He* = 0.01 yielded a lower number of QTLs where the favorable allele was lost (on average, 218.8 [±3.852]) compared to OCS-He* = 0.01 (on average, 234.5 [±3.908]) (**Figure 8**). PM not considering the marker information, i.e., phenotypic selection (PS scenario), yielded lower short- and long-term genetic gains than PM considering marker information (GS scenario) (on average, G_{10} = 6.402 [±0.166] compared to 7.713 [±0.256] 10 years post burn-in and G_{10} = 10.810 [±0.329] compared to 13.287 [±0.436) 60 years post burn-in; **Figure 8**, **Supplementary Material** [**Table S2 File S4**]).

DISCUSSION

Predicting the Next-Generation Diversity

Accounting for within-family selection increased the squared correlation and reduced the mean error of post-selection genetic diversity prediction (**Figures 3**, **4**). The gain in squared correlation (**Figure 3**) and the reduction in mean error (**Figure 4**), were more important for parents showing differences in performance. This result is consistent with observations in Allier et al. (2019b), where crosses between two phenotypically distant parents yielded post-selection parental contributions that differ from their expectation before selection (i.e., 0.5). The mean prediction error was always positive, which can be explained by the use in Eq. 9 of genome-wide parental contributions to progeny in lieu of parental contributions at individual QTLs to predict allelic frequency changes due to selection **Supplementary Material (File S2)**. As a result, the predicted extreme frequencies at QTLs in the progeny are shrunk toward the mean frequency, leading to an overestimation of the

TABLE 2 | Discounted cumulative gain in TRUE scenario for three different parameters ρ giving more weight to short-term gain in different levels and assuming known QTL effects (TRUE scenario).

Cross-selection index (CSI)	Discounted cumulative gain		
	ρ = 0	ρ = 0.04	ρ = 0.2
UCPC - He* = 0.01	15.949 (±0.398)	12.321 (±0.284)	6.682 (±0.143)
UCPC - He* = 0.10	15.174 (±0.386)	11.788 (±0.280)	6.593 (±0.158)
UC	14.408 (±0.355)	11.689 (±0.266)	6.822 (±0.145)
OCS - He* = 0.01	15.148 (±0.346)	11.675 (±0.262)	6.360 (±0.149)
OCS - He* = 0.10	14.630 (±0.349)	11.278 (±0.264)	6.230 (±0.149)
UCPC - He* = 0.15	14.205 (±0.334)	11.176 (±0.250)	6.454 (±0.149)
OCS - He* = 0.15	14.056 (±0.337)	10.884 (±0.250)	6.103 (±0.155)
PM	12.609 (±0.280)	10.392 (±0.217)	6.345 (±0.155)

Mean discounted cumulative gain with ρ = 0 (constant weight along years), ρ = 0.04 (decreasing weight along years) and ρ = 0.2 (nearly null weights after 30 years) on the ten independent replicates. CSI are ordered in decreasing discounted cumulative gain with ρ = 0.04.

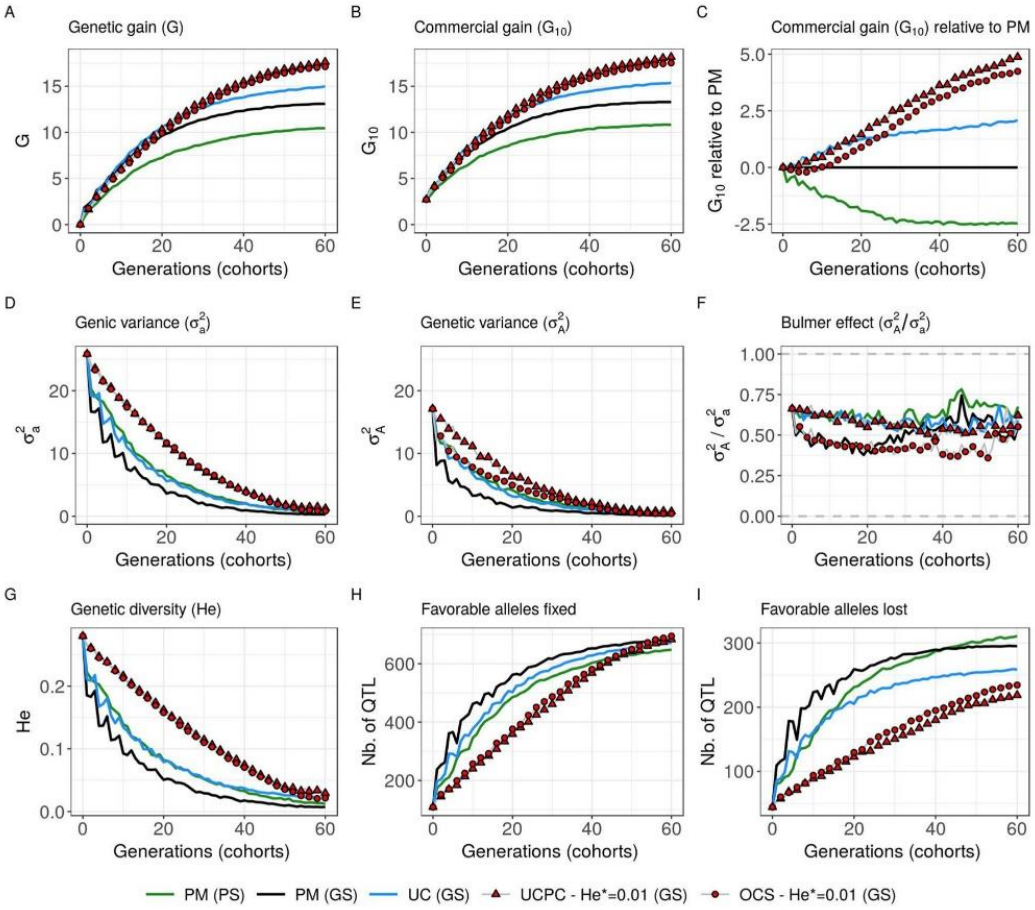


FIGURE 8 | Evolution of different variables for different cross-selection indices according to the generations in the GS scenario (PM, parental mean; UC, usefulness criterion; OCS-He*, optimal cross-selection; and UCPC-He*, UCPC-based optimal cross-selection for He* = 0.01) and in the PS scenario (PM, parental mean). **(A)** Genetic gain at whole progeny level (G), **(B)** genetic gain at commercial level (G_{10}), and **(C)** G_{10} relatively to PM (GS), genetic gain is measured on true breeding values. **(D)** Genic variance at QTLs (σ^2_a). **(E)** genetic variance of true breeding values (σ^2_A) and **(F)** ratio of genic over genetic variance (σ^2_A / σ^2_a). **(G)** genetic diversity at QTLs and number of QTLs where the favorable allele was fixed **(H)** and lost **(I)**.

expected heterozygosity (He) (results not shown). Local changes in allele frequency under artificial selection could be predicted following Falconer and Mackay (1996) and Gallais et al. (2007), but this approach would assume linkage equilibrium between QTLs, which is a strong assumption that does not correspond to the highly polygenic trait that we simulated.

Effect of UC on Short- and Long-Term Recurrent Selection

In a first approach, we considered no constraint on diversity during cross-selection and compared cross-selection maximizing the UC or maximizing the PM in the TRUE scenario, assuming known QTL effects and positions. The UC yielded higher short-term genetic gain at commercial level (G_{10}; **Figures 5B**, **C**). This was expected because UC predicts the mean performance of the best fraction of progeny. When considering the genetic gain at the mean progeny level (G; **Figure 5A**), UC needed 5 years to outperform PM. These results underline that UC maximizes the mean performance of the next generation issued from the intercross of selected progeny, sometimes at the expense of the current generation progeny mean performance. This observation is consistent with the fact that candidate parents of the sixth cohort came all from the three first cohorts generated considering UC and thus the sixth cohort took full advantage of the use of UC (**Figure 1A**). This tendency was also observed in simulations by Müller et al. (2018) considering the EMBV approach, akin to the UC for normally distributed additive traits. The UC also showed a higher long-term genetic gain at both commercial (G_{10}) and whole progeny level (G) compared to intercrossing the best candidate parents (PM). This long-term gain was driven by a higher additive genic variance at QTLs (σ^2_a ; **Figure 6A**) and a lower genomic covariance between QTLs (σ^2_A / σ^2_a ; **Figure 6C**) resulting in a higher additive genetic

variance in UC compared to PM (σ_A^2; **Figure 6B**). Note that with lower σ_a^2 the ratio σ_A^2/σ_a^2 becomes less interpretable in the long-term (**Figure 6C**). UC also better managed the fixation (**Figure 7B**) or the maintenance (**Figure 7C**) of the favorable allele at QTLs compared to PM. These results highlight the interest of considering within cross variance in cross-selection for improving long-term genetic gain as observed in Müller et al. (2018).

Accounting for Within-Family Variance in Optimal Cross-Selection

Assuming known marker effects, we observed that considering a constraint on diversity, i.e., optimal cross-selection, always maximized the long-term genetic gain, at the cost of a variable penalty for short-term gain, compared to no constraint on diversity (e.g., UC). We further compared the OCS (Gorjanc et al., 2018) with the UCPC-based optimal cross-selection that accounts for the fact that only a selected fraction of each family contributes to the next generation. In the optimization framework considered, we compared the ability of UCPC (referred to as UCPC-He*) and OCS (referred to as OCS-He*) to convert a determined loss of diversity into genetic gain. For a given diversity trajectory, UCPC-He* yielded higher short-term commercial gain than OCS-He*. Both, OCS-He* and UCPC-He* yielded similar additive genic variance (σ_a^2), but we observed differences in terms of the ratio σ_A^2/σ_a^2. As expected under directional selection, the ratio σ_A^2/σ_a^2 was positive and inferior to one, revealing a negative genomic covariance between QTLs (Bulmer, 1971). UCPC-He* yielded a higher ratio, i.e., lower repulsion, and thus a higher additive genetic variance (σ_A^2) than OCS-He* for a similar He*. This explains the higher long-term genetic gain at commercial and whole progeny levels observed for UCPC-He*. This result supports the idea, suggested in Allier et al. (2019a), that accounting for complementarity between parents when defining crossing plans is an efficient way to favor recombination events to reveal part of the additive genic variance hidden by repulsion between QTLs. For low targeted diversity (He* = 0.01), UCPC-He* also appeared to better manage the rare favorable alleles at QTLs than OCS-He*. These results highlighted the interest of UCPC-based optimal cross-selection to convert the genetic diversity into genetic gain by maintaining more rare favorable alleles and limiting repulsion between QTLs. In case of higher targeted diversity (He* = 0.15), the loss of diversity was likely not sufficient to fully express the additional interest of UCPC compared to OCS to convert diversity into genetic gain. In this case, UCPC-He* and OCS-He* performed similarly. Accounting for within cross variance to measure the expected gain of a cross in optimal cross-selection was already suggested in Shepherd and Kinghorn (1998). More recently, Akdemir and Isidro-Sánchez (2016) and Akdemir et al. (2018) accounted for within cross variance considering linkage equilibrium between QTLs. Akdemir and Isidro-Sánchez (2016) also observed that accounting for within cross variance during cross-selection yielded higher long-term mean performance with a penalty at short-term mean progeny performance.

Short-term economic returns of a breeding program condition the resources invested to maintain/increase response to selection and therefore long-term competitive capacity. Hence, to fully take advantage of their benefit at long term, it is necessary to make sure that tested breeding strategies do not compromise too much the short-term commercial genetic gain. For this reason, we considered the discounted cumulative commercial gain following Dekkers et al. (1995) and Chakraborty et al. (2002) as a summary variable to evaluate CSI while giving more weight to short-term gain in different levels. UCPC-He* outperformed OCS-He* for a given He* either considering uniform weights ($\rho = 0$) or giving approximately seven times more weight to short-term gain compared to long-term gain ($\rho = 0.04$). This was also true when focusing only on short-term gain ($\rho = 0.2$), but in this case the best model was UC without accounting for diversity (**Table 2**).

Practical Implementations in Breeding

UCPC With Estimated Marker Effects

In simulations, we first considered 1,000 QTLs with known additive effects sampled from a centered normal distribution. For a representative subset of CSIs (PM, UC, UCPC-He*, and OCS-He* with He* = 0.01; **Figure 8**), we considered estimated effects at 2,000 SNPs. The main conclusions obtained with known and estimated marker effects were consistent, supporting the practical interest of UCPC-based optimal cross-selection (**Figure 8**). The difference was that the superiority of UCPC-based optimal cross-selection over optimal cross-selection not accounting for within-family selection in GS scenario was not significant after 60 years **Supplementary Material (Table S2 File S4)**. With estimated marker effects instead of known QTL effects, the predicted progeny variance (σ^2) corresponded to the variance of the predicted breeding values, which are shrunk compared to TBVs, depending on the model accuracy (referred to as variance of posterior mean [VPM] in Lehermeier et al.). An alternative would be to consider the marker effects estimated at each sample of a Monte Carlo Markov Chain process, e.g., using a Bayesian ridge regression, to obtain an improved estimate of the additive genetic variance (referred to as posterior mean variance [PMV] in Lehermeier et al., 2017a; Lehermeier et al., 2017b).

In practice, QTL effects are unknown, so the selection of progeny cannot be based on TBVs, and thus the selection accuracy (h) is smaller than one. In our simulation study assuming unknown QTLs (GS scenario), progeny were selected based on estimated breeding values taking into account genotypic information as well as replicated phenotypic information, which led to a high selection accuracy, as it can be encountered in breeding. Thus, the assumption $h = 1$ used in Eq. 6 for GS scenario is reasonable. In order to shorten the cycle length of the breeding scheme, selection of progeny can be based on predicted GEBVs of genotyped but not phenotyped progeny. In such a case, the selection accuracy (h) will be considerably reduced. In such a situation, one can advocate to use PMV instead of VPM in the computation of UCPC and to take into account the proper selection accuracy (h) within crosses adapted to the selection scheme. When selection is based on predicted values, i.e.,

genotyped but not phenotyped progeny, the shrunk predictor VPM should be a good approximation of $(h\sigma)^2$.

UCPC-Based Optimal Cross-Selection

In this study, we assumed fully homozygous parents and two-way crosses. However, neither the optimal cross-selection nor UCPC-based optimal cross-selection is restricted to homozygote parents. Considering heterozygote parents in optimal cross-selection is straightforward. Following the extension of UCPC to four-way crosses (Allier et al., 2019b), UCPC optimal cross-selection can be used for phased heterozygous individuals, as it is commonly the case in perennial plants or animal breeding. Animal breeders are interested in Mendelian sampling variance for individual and cross-selection (Segelke et al., 2014; Bonk et al., 2016; Bijma et al., 2018) and might be interested to incorporate it into OCS strategies. We considered an inbred line breeding program, but the extension to hybrid breeding is of interest for species such as maize. The use of testcross effects, i.e., estimated on hybrids obtained by crossing candidate lines with lines from the opposite heterotic pool, in UCPC-based optimal cross-selection is straightforward, and so the UCPC-based optimal cross-selection can be used to improve each heterotic pool individually. In order to jointly improve two pools, further investigations are required to include dominance effects in UCPC-based optimal cross-selection. In addition, this would imply that crossing plans in both pools are jointly optimized to manage genetic diversity within pools and complementarity between pools.

We considered a within-family selection intensity corresponding to the selection of the 5% most performant progeny as candidates for the next generation. Equal selection intensities were assumed for all families, but in practice due to experimental constraints or optimized resource allocation (e.g., generate more progeny for crosses showing high progeny variance but low progeny mean), within-family selection intensity can be variable. Different within-family selection intensities (see Eqs. 8 and 9) can be considered in UCPC-based optimal cross-selection, but an optimization regarding resource allocation of the number of crosses and the selection intensities within crosses calls for further investigations. However, in marker-assisted selection schemes based on QTL detection results (Bernardo et al., 2006), an optimization of selection intensities per family was observed to be only of moderate interest.

Proposed UCPC-based optimal cross-selection was compared to OCS in a targeted diversity trajectory context. We considered a linear trajectory, but any genetic diversity trajectory can be considered (e.g., **Figure 2**). The optimal diversity trajectory cannot be easily determined and depends on breeding objectives and data considered. Optimal contribution selection in animal breeding considers a similar ϵ-constraint optimization with a targeted inbreeding trajectory determined by a fixed annual rate of inbreeding (e.g., 1% advocated by the Food and Agriculture Organization (FAO), Woolliams et al., 1998). Woolliams et al. (2015) argued that the optimal inbreeding rate is also not straightforward to define. An alternative formulation of the optimization problem to avoid the use of a fixed constraint is to consider a weighted index $(1-\alpha)$ $V(\boldsymbol{nc})+\alpha D(\boldsymbol{nc})$, where α is the weight balancing the expected gain $V(\boldsymbol{nc})$ and constraint $D(\boldsymbol{nc})$ (De Beukelaer et al., 2017). However, the appropriate choice of α is difficult and is not explicit either in terms of expected diversity or expected gain.

Introgression of Diversity and Anticipation of a Changing Breeding Context

We considered candidate parents coming from the three last overlapping cohorts (**Figure 1**) in order to reduce the number of candidate crosses during the progeny covariances prediction (UCPC) and the optimization process. This yielded elite candidate parents that were not directly related (no parent–progeny) and that did not show strong differences in performances, which is standard in a commercial plant breeding program focusing on yield improvement. However, when the genetic diversity in a program is so low that long-term genetic gain is compromised, external genetic resources need to be introgressed by crosses with internal elite parents. As suggested by results of simulation 1, we conjecture that the advantage of UCPC-based optimal cross-selection over OCS increases in such a context where heterogeneous, i.e., phenotypically distant, genetic materials are crossed. This requires investigations that we hope to address in subsequent research.

Our simulations also assumed fixed environments and a single targeted trait over 60 years. However, in a climate change context and with rapidly evolving societal demands for sustainable agricultural practices, environments and breeders objectives will likely change over time. In a multitrait context, the multiobjective optimization framework proposed in Akdemir et al. (2018) can be adapted to UCPC-based optimal cross-selection. The upcoming but yet unknown breeding objectives make the necessity to manage genetic diversity even more important than highlighted in this study.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://doi.org/10.25387/g3.7405892.

AUTHOR CONTRIBUTIONS

ST, CL, AC, and LM supervised the study. AA performed the simulations and wrote the manuscript. ST worked on the implementation in the simulator. All authors reviewed and approved the manuscript.

FUNDING

This research was funded by RAGT2n and the ANRT CIFRE grant no. 2016/1281 for AA.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01006/full#supplementary-material

REFERENCES

Akdemir, D., and Isidro-Sánchez, J. (2016). Efficient breeding by genomic mating. *Front. Genet.* 7, 210. doi: 10.3389/fgene.2016.00210

Akdemir, D., Beavis, W., Fritsche-Neto, R., Singh, A. K., and Isidro-Sánchez, J. (2018). Multi-objective optimized genomic breeding strategies for sustainable food improvement. *Heredity* 122, 672. doi: 10.1101/209080

Allier, A., Teyssèdre, S., Lehermeier, C., Claustres, B., Maltese, S., Moreau, L., Charcosset, A. (2019a). Assessment of breeding programs sustainability: application of phenotypic and genomic indicators to a North European grain maize program. *Theor. Appl. Genet.* 132, 1321–1334. doi: 10.1007/s00122-019-03280-w

Allier, A., Moreau, L., Charcosset, A., Teyssèdre, S., and Lehermeier, C. (2019b). Usefulness criterion and post-selection parental contributions in multi-parental crosses: application to polygenic trait introgression. *G3 Genes Genomes Genet.* 9, 1469–1479. doi: 10.1534/g3.119.400129

Bernardo, R., Moreau, L., and Charcosset, A. (2006). Number and fitness of selected individuals in marker-assisted and phenotypic recurrent selection. *Crop Sci.* 46, 1972–1980. doi: 10.2135/cropsci2006.01-0057

Bijma, P., Wientjes, Y. C. J., and Calus, M. P. L. (2018). Increasing genetic gain by selecting for higher Mendelian sampling variance. *Proc. World Congr. Genet. Appl. Livest. Prod. Genet. Gain-Breed. Strategies* 2, 47.

Bonk, S., Reichelt, M., Teuscher, F., Segelke, D., and Reinsch, N. (2016). Mendelian sampling covariability of marker effects and genetic values. *Genet. Sel. Evol.* 48, 36. doi: 10.1186/s12711-016-0214-0

Bulmer, M. (1971). The stability of equilibria under selection. *Heredity* 27, 157–162. doi: 10.1038/hdy.1971.81

Bulmer, M. (1980). *The mathematical theory of quantitative genetics.* Newyork Oxford University Press.

Chakraborty, R., Moreau, L., and Dekkers, J. C. (2002). A method to optimize selection on multiple identified quantitative trait loci. *Genet. Sel. Evol.* 34, 145. doi: 10.1186/1297-9686-34-2-145

Clark, S. A., Hickey, J. M., and van der Werf, J. H. (2011). Different models of genetic variation and their effect on genomic evaluation. *Genet. Sel. Evol. GSE* 43, 18. doi: 10.1186/1297-9686-43-18

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., and Hickey, J.M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193, 347–365. doi: 10.1534/genetics.112.147983

Daetwyler, H. D., Hayden, M. J., Spangenberg, G. C., and Hayes, B. J. (2015). Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics* 200, 1341–1348. doi: 10.1534/genetics.115.178038

De Beukelaer, H. D., Badke, Y., Fack, V., and Meyer, G. D. (2017). Moving beyond managing realized genomic relationship in long-term genomic selection. *Genet.* 206: 1127–1138. doi: 10.1534/genetics.116.194449

Dekkers, J. C. M., Birke, P. V., and Gibson, J. P. (1995). Optimum linear selection indexes for multiple generation objectives with non-linear profit functions. *Anim. Sci.* 61, 165–175. doi: 10.1017/S1357729800013667

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to quantitative genetics. 4th ed.* Harlow, England: Pearson.

Fradgley, N., Gardner, K. A., Cockram, J., Elderfield, J., Hickey, J. M., Howell, P., et al. (2019). A large-scale pedigree resource of wheat reveals evidence for adaptation and selection by breeders. *PLoS Biol.* 17, e3000071. doi: 10.1371/journal.pbio.3000071

Gallais, A., Moreau, L., and Charcosset, A. (2007). Detection of marker–QTL associations by studying change in marker frequencies with selection. *Theor. Appl. Genet.* 114, 669–681. doi: 10.1007/s00122-006-0467-z

Ganal, M. W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E. S., Charcosset, A., 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. doi: 10.1371/journal.pone.0028334

Gerke, J. P., Edwards, J. W., Guill, K. E., Ross-Ibarra, J., and McMullen, M. D. (2015). The genomic impacts of drift and selection for hybrid performance in maize. *Genetics* 201, 1201–1211. doi: 10.1534/genetics.115.182410

Giraud, H., Lehermeier, C., Bauer, E., Falque, M., Segura, V., Bauland, C., et al. (2014). Linkage disequilibrium with linkage analysis of multiline crosses reveals different multiallelic qtl for hybrid performance in the flint and dent heterotic groups of maize. *Genetics* 198, 1717–1734. doi: 10.1534/genetics.114.169367

Gorjanc, G., Gaynor, R. C., and Hickey, J. M. (2018). Optimal cross-selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theor. Appl. Genet.* 131, 1953–1966. doi: 10.1007/s00122-018-3125-3

Gorjanc G., and Hickey J. M. (2018) AlphaMate: a program for optimizing selection, maintenance of diversity and mate allocation in breeding programs. *Bioinformatics* 34, 3408–3411. doi: 10.1093/bioinformatics/bty375

Haimes, Y., Lasdon, L. S., and Wimer, D. (1971). On a bicriterion formation of the problems of integrated system identification and system optimization. *IEEE Trans. Syst. Man Cybern.* SMC-1, 296–297. doi: 10.1109/TSMC.1971.4308298

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., and Goddard, M. E. (2010). genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in holstein cattle as contrasting model traits. *PLoS Genet.* 6, e1001139. doi: 10.1371/journal.pgen.1001139

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

Heslot, N., Jannink, J.-L., and Sorrells, M. E. (2015). Perspectives for genomic selection applications and research in plants. *Crop Sci.* 55, 1–12. doi: 10.2135/cropsci2014.03.0249

Jannink, J.-L. (2010). Dynamics of long-term genomic selection. *Genet. Sel. Evol.* 42, 35. doi: 10.1186/1297-9686-42-35

Kinghorn, B. P., Banks, R., Gondro, C., Kremer, V. D., Meszaros, S. A., Newman, S., et al. (2009). "Strategies to exploit genetic variation while maintaining diversity," in *adaptation and fitness in animal populations* (Dordrecht: Springer), 191–200. doi: 10.1007/978-1-4020-9005-9_13

Kinghorn, B. P. (2011). An algorithm for efficient constrained mate selection. *Genet. Sel. Evol.* 43, 4. doi: 10.1186/1297-9686-43-4

Lehermeier, C., de los Campos, G., Wimmer, V., and Schön, C.-C. (2017a). Genomic variance estimates: with or without disequilibrium covariances? *J. Anim. Breed. Genet.* 134, 232–241. doi: 10.1111/jbg.12268

Lehermeier, C., Teyssèdre, S., and Schön, C.-C. (2017b). Genetic gain increases by applying the usefulness criterion with improved variance prediction in selection of crosses. *Genetics* 207, 1651–1661. doi: 10.1534/genetics.117.300403

Lin, Z., Cogan, N. O. I., Pembleton, L. W., Spangenberg, G. C., Forster, J. W., Hayes, B. J. et al., (2016). Genetic gain and inbreeding from genomic selection in a simulated commercial breeding program for perennial ryegrass. *Plant Genome* 9. doi: 10.3835/plantgenome2015.06.0046

Lin, Z., Shi, F., Hayes, B. J., and Daetwyler, H. D. (2017). Mitigation of inbreeding while preserving genetic gain in genomic breeding programs for outbred plants. *Theor. Appl. Genet.* 130, 969–980. doi: 10.1007/s00122-017-2863-y

Lynch, M., and Walsh, B. (1999). *Evolution and selection of quantitative traits.* Sunderland, MA., Sinauer Associates.

Meuwissen, T. H. (1997). Maximizing the response of selection with a predefined rate of inbreeding. *J. Anim. Sci.* 75, 934–940. doi: 10.2527/1997.754934x

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

Misztal, I. (2008). Reliable computing in estimation of variance components. *J. Anim. Breed. Genet.* 125, 363–370. doi: 10.1111/j.1439-0388.2008.00774.x

Mohammadi, M., Tiede, T., and Smith, K. (2015). PopVar: a genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. *Crop Sci.* 55, 2068–2077. doi: 10.2135/cropsci2015.01.0030

Müller, D., Schopp, P., and Melchinger, A. E. (2018). Selection on expected maximum haploid breeding values can increase genetic gain in recurrent genomic selection. *G3 Genes Genomes Genet.* 8, 1173–1181. doi: 10.1534/g3.118.200091

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U. S. A.* 70, 3321–3323. doi: 10.1073/pnas.70.12.3321

Piepho, H. P., Möhring, J., Melchinger, A. E., and Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161, 209–228. doi: 10.1007/s10681-007-9449-8

Pryce, J. E., Hayes, B. J., and Goddard, M. E. (2012). Novel strategies to minimize progeny inbreeding while maximizing genetic gain using genomic information. *J. Dairy Sci.* 95, 377–388. doi: 10.3168/jds.2011-4254

Pszczola, M., Strabel, T., Mulder, H. A., and Calus, M. P. L. (2012). Reliability of direct genomic values for animals with different relationships within

and to the reference population. *J. Dairy Sci.* 95, 389–400. doi: 10.3168/jds.2011-4338

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

Rauf, S., Teixeira da Silva, J. A., Khan, A. A., and Naveed, A. (2010). Consequences of plant breeding on genetic diversity. *Int. J. Plant Breed.* 4, 1–21.

Rio, S., Mary-Huard, T., Moreau, L., and Charcosset, A. (2019). Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. *Theor. Appl. Genet.* 132, 81–96. doi: 10.1007/s00122-018-3196-1

Rutkoski, J., Singh, R. P., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J. L. et al., (2015). Genetic gain from phenotypic and genomic selection for quantitative resistance to stem rust of wheat. *Plant Genome* 8. doi: 10.3835/plantgenome2014.10.0074

Schnell, F., and Utz, H. (1975). "F1-Leistung und Elternwahl in der Züchtung von Selbstbefruchtern," in *Bericht* über *die Arbeitstagung der Vereinigung österreichischer Pflanzenzüchter* (Austria: BAL Gumpenstein), 243–248.

Segelke, D., Reinhardt, F., Liu, Z., and Thaller, G. (2014). Prediction of expected genetic variation within groups of offspring for innovative mating schemes. *Genet. Sel. Evol.* 46, 42. doi: 10.1186/1297-9686-46-42

Shepherd, R. K., and Kinghorn, B. P. (1998). A tactical approach to the design of crossbreeding programs, in *Proceedings of the sixth world congress on genetics applied to livestock production: 11-16 january*, (Armidale) 431–438.

Storn, R., and Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* 11, 341–359. doi: 10.1023/A:1008202821328

Van Inghelandt, D., Reif, J. C., Dhillon, B. S., Flament, P., and Melchinger, A. E. (2011). Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm. *Theor. Appl. Genet.* 123, 11–20. doi: 10.1007/s00122-011-1562-3

Wang, H., Misztal, I., Aguilar, I., Legarra, A., and Muir, W. M. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res.* 94, 73–83. doi: 10.1017/S0016672312000274

Woolliams, J. A., Gwaze, D. P., Meuwissen, T. H., Planchenault, D., Renard, J. P., Thibier, M., et al. (1998). Secondary guidelines for the development of national farm animal genetic resources management plans. *Manage. Small Popul. Risk.*

Woolliams, J. A., Berg, P., Dagnachew, B. S., and Meuwissen, T. H. E. (2015). Genetic contributions and their optimization. *J. Anim. Breed. Genet.* 132, 89–99. doi: 10.1111/jbg.12148

Wray, N., and Goddard, M. (1994). Increasing long-term response to selection. *Genet. Sel. Evol.* 26, 431. doi: 10.1186/1297-9686-26-5-431

Zhong, S., and Jannink, J.-L. (2007). Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. *Genetics* 177, 567–576. doi: 10.1534/genetics.107.075358

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Chapter 5 Genetic resources and optimal cross selection for broadening the genetic base of elite breeding programs

This chapter is a draft article and has not yet been peer-reviewed.

Antoine Allier[12*], Simon Teyssèdre[2], Christina Lehermeier[2], Laurence Moreau[1], Alain Charcosset[1*]

[1] GQE - Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France
[2] RAGT2n, Genetics and Analytics Unit, 12510 Druelle, France
[*] Corresponding authors: allierantoine@gmail.com and alain.chacrcosset@inra.fr

Author contribution statement

AC, ST, CL and LM supervised the study. AA and ST worked on the simulator. AA performed the simulations, analysis and wrote the early version of the manuscript. All authors reviewed and approved the manuscript.

Acknowledgments

This research was funded by RAGT2n and the ANRT CIFRE Grant n° 2016/1281 for AA.

Abstract

The narrow genetic base of elite germplasm compromises long-term genetic gain and increases the vulnerability to biotic and abiotic stresses in unpredictable environmental conditions. Therefore, an efficient strategy is required to broaden the genetic base of commercial breeding programs while not compromising short-term variety release. Optimal cross selection aims at identifying the optimal set of crosses that balances the expected genetic value and diversity. We propose to consider genomic selection and optimal cross selection to recurrently improve genetic resources (i.e. pre-breeding), to bridge the improved genetic resources with elites (i.e. bridging), and to manage introductions into the elite breeding population. Optimal cross selection is particularly adapted to jointly identify bridging, introduction and elite crosses to ensure an overall consistency of the genetic base broadening strategy. We compared simulated breeding programs introducing donors with different performance levels, directly or indirectly after bridging. We also evaluated the effect of the training set composition on the success of introductions. We observed that with recurrent introductions of improved donors, it is possible to maintain the genetic diversity and increase mid- and long-term performances with only limited penalty at short-term. Considering a bridging step yielded significantly higher mid- and long-term performance when introducing low performing donors. The results also suggested to consider marker effects estimated on a broad training population including donor by elite and elite by elite progeny to identify bridging, introduction and elite crosses.

Key message

With recurrent genetic base broadening after pre-breeding, commercial breeding programs can maintain genetic diversity and take advantage of introduced favorable alleles to reach significantly higher long-term performance.

Key words

genetic base broadening; pre-breeding; bridging; introduction; genomic prediction; optimal cross selection

Introduction

Modern breeding has been successful in exploiting crop diversity for genetic improvement. However, current yield increases may not be sufficient in view of rapid human population growth (Godfray *et al.* 2010). Moreover, modern intensive breeding practices have exploited a very limited fraction of the available crop diversity (Cooper *et al.* 2001; Reif *et al.* 2005). The narrow genetic base of elite germplasm compromises long-term genetic gain and increases the genetic vulnerability to unpredictable environmental conditions (McCouch *et al.* 2013). Efficient genetic diversity management is therefore required in breeding programs. This involves the efficient incorporation of new genetic variation and its conversion into short- and long-term genetic gain.

Among the possible sources of diversity, wild relatives, exotic germplasm accessions and landraces that predate modern breeding exhibit substantial genetic diversity. These *ex-situ* genetic resources are conserved worldwide in international gene banks and national collections. They provide a promising basis to improve crop productivity, crop resilience to biotic and abiotic stresses and crop nutritional quality (Salhuana and Pollak 2006; Wang *et al.* 2017). In case of traits determined by few genes of large effect, the favorable alleles can be identified and introgressed into elite germplasm following established marker-assisted backcross procedures (e.g. Charmet *et al.* 1999; Servin *et al.* 2004; Han *et al.* 2017). Such introgressions have been successful for mono- and oligogenic traits (e.g. earliness loci in maize, Simmonds 1979; Smith and Beavis 1996 and SUB1 gene in rice, Bailey-Serres *et al.* 2010). Introgressions also proved to be successful for more polygenic traits where few major causal regions have been identified. For instance, Ribaut and Ragot (2006) successfully introgressed five regions associated with maize flowering time and yield components under drought conditions. For complex traits controlled by numerous genes with small effect, e.g. grain yield in optimal conditions, the identification and introgression of favorable alleles into elite germplasm were mostly unsuccessful. This requires to go beyond the introgression of few identified favorable alleles toward the polygenic enrichment of elite germplasm (Simmonds 1962, 1993). Although plant breeders recognize the importance of genetic resources for elite genetic base broadening, only little use has been made of it (Glaszmann *et al.* 2010; Wang *et al.* 2017). The main reason is that breeding progress continues (Duvick 2005; Tadesse *et al.* 2019) and that breeders are reluctant to compromise elite germplasm with unadapted and unimproved genetic resources (Kannenberg and Falk 1995). Despite genetic resources carry novel favorable alleles that may counter balance their low genetic value by an increased genetic variance when crossed to elites (Longin and Reif 2014; Allier *et al.* 2019b), their progeny performance is mostly insufficient for breeders. Thus, breeding strategies are needed to bridge the performance gap between genetic resources and elites and to transfer beneficial genetic variations into elite germplasm while not compromising the performance of released varieties (Simmonds 1993; Gorjanc *et al.* 2016). Pre-breeding can be defined as the recurrent improvement of genetic resources to release donors that can be further introduced into the elite breeding population (Figure 1). According to Simmonds (1993), pre-breeding should start from a broad germplasm and should be carried out on several generations with low selection intensity to favor extensive recombination events and minimal inbreeding. The donor released from pre-breeding can be directly introduced into the elite breeding population. However, in cases where the performance gap between the donor released from pre-breeding and elites is too large, one may consider a buffer population between donor and elites before introduction in the elite breeding population, further referred to as bridging. The best progeny of bridging is then considered for introduction into the elite breeding population (Figure 1).

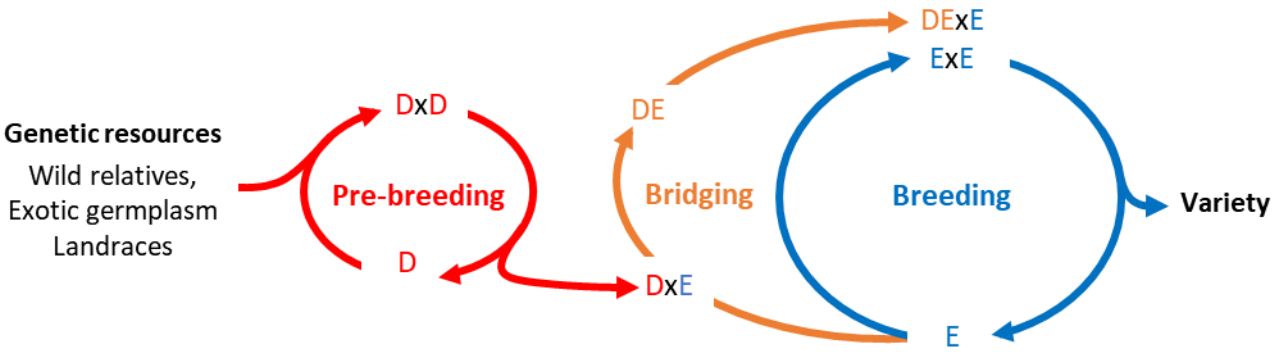


Figure 1 Diagram illustrating the respective positioning of pre-breeding, bridging and breeding from genetic resources to variety release.

Different sources of donors can be considered in autogamous and allogamous species for genetic base broadening. This includes landraces historically cultivated before modern breeding. For instance in maize, open pollinated varieties (OPVs) are landrace populations of heterozygous individuals cultivated before the hybrid maize breeding revolution in the 1950's (Anderson 1944; Troyer 1999). Inbred lines derived from OPVs present a large diversity and a potential interest for adaptation, but also a large performance gap with current varieties (Böhm *et al.* 2014; Melchinger *et al.* 2017; Böhm *et al.* 2017). These landraces can be further improved through pre-breeding that can be shared between the industry and public institutes in collaborative projects. In maize, the Latin American Maize Project (LAMP, Pollak 1990; Salhuana *et al.* 1997; Salhuana and Pollak 2006) provided breeders with useful characterization and evaluation of US and Latin American tropical germplasm accessions. Later, the Germplasm Enhancement of Maize project (GEM, Pollak and Salhuana 2001) improved the accessions identified in LAMP with elite lines furnished by private partners (Pollak 2003). Similarly, the Seed of Discovery project (SeeD, Gorjanc *et al.* 2016) aimed to harness favorable variations from landraces and to develop a bridging germplasm useful for genetic base broadening of commercial maize breeding programs. In this vein, Cramer and Kannenberg (1992) proposed the Hierarchical Open-ended Population Enrichment (HOPE) breeding system to release enriched maize inbreds for the industry. In its last version, the HOPE system is a breeding program with three hierarchical open ended gene pools permitting the transfer of favorable alleles from genetic resources to the elite pools (Popi 1997; Kannenberg 2001). Finally, breeders can consider the varieties released by breeding programs selecting on a different germplasm and in different environments as donors. In hybrid species, the ability to use one of the variety's inbred parent as a donor depends on the germplasm proprietary protection relative to species and countries (e.g. using reverse breeding, Smith *et al.* 2008). In the US, maize inbred parents of hybrid varieties become publically available after twenty years of plant variety protection act, these are referred to as ex-PVPA (Mikel and Dudley 2006). In inbred species such as wheat, using current varieties for breeding is straightforward if cultivated under the union for the protection of new varieties of plants convention (UPOV, Dutfield 2011). These donors are likely the most performing but also the less original that can be considered.

With the availability of cheap high density genotyping, Whittaker *et al.* (2000) and Meuwissen *et al.* (2001) have proposed to use genomewide prediction to fasten breeding progress by shortening generation intervals. In the most frequently used approaches of genomewide prediction, it is assumed that most genomic regions equally contribute with relatively small effects to polygenic traits. A large number of genomewide markers is employed, and their effects are estimated on a training set (TS) of

phenotyped and genotyped individuals. The genomic estimated breeding values (GEBVs) are further predicted considering the estimated marker effects and individuals' molecular marker information. Recurrent selection based on genomewide prediction, further referred to as genomic selection (GS), has been increasingly implemented in crop breeding programs (Heslot *et al.* 2015; Voss-Fels *et al.* 2019). GS efficiency depends on the relationship between individuals in the TS and the target population of individuals to predict (Habier *et al.* 2010; Pszczola *et al.* 2012). We assume that as a consequence, in commercial breeding programs, GS has been mostly implemented considering a narrow elite TS that optimizes the prediction accuracy on elite material. However, such a narrow TS limits the prediction accuracy on individuals carrying rare alleles, which is the case for the progeny of elite by donor crosses. Therefore, it is important to define the TS composition that maximizes the prediction accuracy in both elite and introduction families.

In the context of genetic base broadening, GS is also interesting to fasten and reduce the costs for the evaluation and identification of genetic resources in gene banks (Crossa *et al.* 2016; Yu *et al.* 2016). Furthermore, GS can fasten pre-breeding programs to reduce the performance gap between genetic resources and elite populations (Gorjanc *et al.* 2016). Instead of truncated selection (i.e. select and mate individuals with the largest estimated breeding values), Cowling *et al.* (2017) proposed to use the optimal contribution selection to improve genetic resources while maintaining a certain level of diversity in the pre-breeding population. Optimal contribution selection (Wray and Goddard 1994; Meuwissen 1997; Woolliams *et al.* 2015) aims at identifying the optimal parental contributions to the next generation in order to maximize the expected genetic value in the progeny under a certain constraint on diversity. Therefore, the optimal contribution selection is particularly adapted to pre-breeding and genetic diversity management. Cowling *et al.* (2017) considered the pedigree relationship information but considering the genomic relationship information can further improve the optimal cross selection (Clark *et al.* 2013). Considering optimal contribution selection on empirical cattle data, Eynard *et al.* (2018) observed that allowing for the introductions of old individuals in the breeding population supported long-term response to selection. The optimal cross selection (OCS) is the extension of optimal contribution selection to deliver a crossing plan (Kinghorn *et al.* 2009; Kinghorn 2011; Akdemir and Isidro-Sánchez 2016; Gorjanc *et al.* 2018; Akdemir *et al.* 2019). We propose to take advantage of OCS for selection of bridging, introduction and elite crosses (Figure 1). Using OCS, the donors and donor by elite crosses are selected complementarily to the elite by elite crosses in order to ensure an overall consistency of the genetic base broadening strategy. Allier *et al.* (2019c) proposed to account for within family variance and selection in a new version of OCS referred to as Usefulness Criterion Parental Contribution based OCS (UCPC based OCS). They observed both higher short- and long-term genetic gain compared to OCS in a simulated closed commercial breeding program.

We extend here the use of UCPC based OCS to pre-breeding, following Cowling *et al.*(2017), and to an open commercial breeding program with recurrent introductions of genetic resources, extending the work of Eynard *et al.* (2018). In this context, we aimed at evaluating the efficiency of genetic base broadening depending on the type of donors considered and the genetic base broadening scheme (Figure 1). We considered either donors corresponding to the generation of the founders of breeding pools or improved varieties released twenty years ago and five years ago. Our objectives were to evaluate (i) the interest of recurrent introductions of diversity in the breeding population, (ii) the interest to conduct or not bridging and (iii) the impact of the training set composition on within family genomewide prediction accuracies.

Material and methods

Simulated breeding programs

Material and simulations

We considered 338 Dent maize genotypes from the Amaizing project (Rio *et al.* 2019; Allier *et al.* 2020) as founders of genetic pools. This diversity was structured into three main groups: 82 Iowa Stiff Stalk Synthetics, 57 Iodents and 199 other dents. We sampled 1,000 biallelic quantitative trait loci (QTLs) with a minimal distance between two consecutive QTLs of 0.2 cM among the 40,478 single nucleotide polymorphisms (SNPs) from the Illumina MaizeSNP50 BeadChip (Ganal *et al.* 2011). Each QTL was assigned an additive effect sampled from a Gaussian distribution with a mean of zero and a variance of 0.05 and the favorable allele was attributed at random to one of the two SNP alleles. We sampled 2,000 SNPs as non-causal markers further used as genotyping information. The consensus genetic positions of sampled QTLs and SNPs was considered according to Giraud *et al.* (2014).

We simulated two different breeding programs: an external breeding program (Figure 2A) that released every year varieties that were later considered as potential donors for introduction in a commercial breeding program (Figure 2C-D). Both external and commercial programs used doubled haploid (DH) technology to derive progeny. We assumed a period of three years to derive, genotype and phenotype DH progeny. Every year T, progeny of the three last generations $T-3$, $T-4$ and $T-5$ were considered as potential parents of the next generation. It created overlapping and connected generations as it can be encountered in breeding. We first considered a burn-in period of twenty years with recurrent phenotypic selection from a population of founders. Burn-in created extensive linkage disequilibrium as often observed in elite breeding programs (Van Inghelandt *et al.* 2011). Every progeny was phenotyped and phenotypes were simulated considering the genotypes at QTLs, an error variance corresponding to a trait repeatability of 0.4 in the founder population, and no genotype by environment interactions (Appendix A). Every individual was evaluated in four environments in one year. After twenty years of burn-in, we simulated different breeding programs using GS. Every year, progeny phenotypes and genotypes of the three last available generations were used to fit a G-BLUP model (Appendix A). Progeny were selected based on GEBVs and marker effects were obtained by back-solving the G-BLUP model (Wang *et al.* 2012) and further used for optimal cross selection to generate the next generation (see optimal cross selection section and Appendix B).

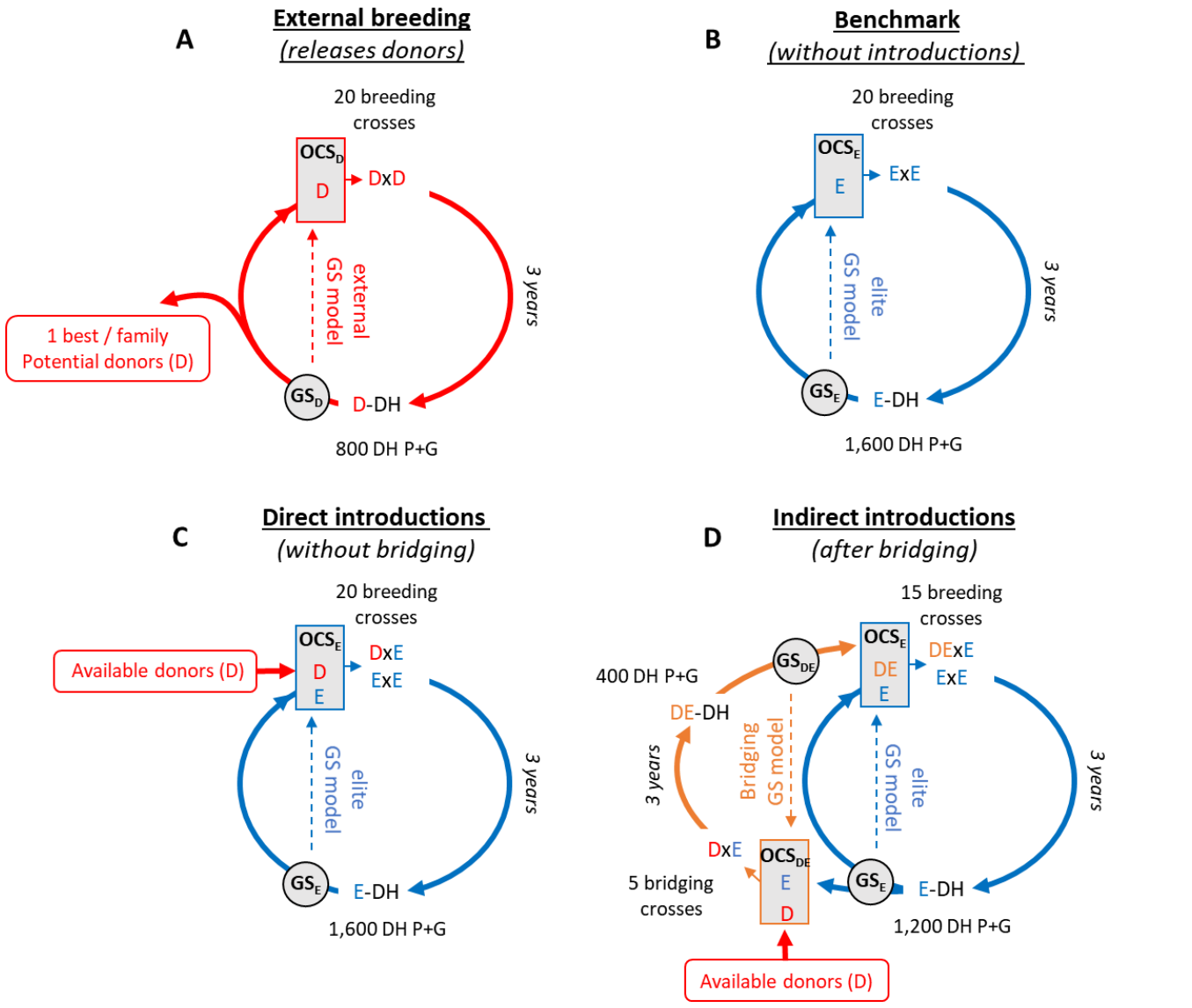


Figure 2 Diagram of simulated breeding programs. (**A**) External breeding program that generates potential donors, (**B**) commercial benchmark program without introductions, (**C**) commercial program with introductions without bridging or (**D**) commercial program with introductions after bridging.

External breeding program: Improvement of genetic resources

The external breeding program (Figure 2A), was simulated starting from a broad population of 40 founders sampled among the 338 maize genotypes. During the three first years, the founders were randomly crossed with replacement to generate each year 20 biparental families of 40 DH progeny to initiate the three overlapping generations. The genetic material in the external breeding is referred to as improved donors (D). During seventeen years, we first selected among the three last generations the 10% D progeny per family (i.e. 4 DH lines/family x 20 families x 3 years) with the largest phenotypic mean. We further randomly mated with replacement the 50 DH with the largest phenotypic mean to generate 20 biparental families of 40 DH lines. Note that we considered 20 biparental families to be consistent with the post burn-in simulations. After twenty years of burn-in, we considered GS trained on the D progeny of the three last generations (i.e. 2,400 D progeny, Figure 2A). Among these three last generations, we considered per family the 10% D progeny with the largest GEBVs as potential parents of the next generation, i.e. 4 DH lines/family x 20 families x 3 years = 240 potential parents. The 20 two-way crosses among the 240*239/2 = 28,680 candidate crosses were selected using optimal cross selection as detailed in the section: optimal cross selection.

Commercial breeding programs

The commercial breeding program (Figure 2B-D) started from a population of 10 founders sampled among the 57 Iodent genotypes. During the first three years, the founders were randomly crossed with replacement to generate each year 10 biparental families of 80 DH progeny to initiate the three overlapping generations. The elite genetic material in the internal breeding is referred to as elite progeny (E). During seventeen years, we considered as potential parents of the next generation the 50 E progeny with the largest phenotypic mean from the three last generations, i.e. without applying a preliminary within family selection. These were randomly mated to generate 20 biparental families of 80 DH lines. After twenty years of burn-in, we considered GS and differentiated three different scenarios: the benchmark commercial breeding program without introductions (Figure 2B), the commercial breeding program with direct introductions without bridging (Figure 2C) or the commercial breeding program with introductions after bridging (Figure 2D).

In absence of introductions (*benchmark*), the E progeny were selected based on the elite GS model trained on E progeny of the three last generations (i.e. 4,800 E progeny, Figure 2B). The 5% E progeny with the largest GEBVs within each family (i.e. 4 DH) in the three last breeding generations were considered as potential parents. The 20 two-way crosses among the 28,680 candidate crosses were defined using optimal cross selection as detailed in the next section: optimal cross selection.

For scenarios with introductions, we considered different sub-scenarios (i) for the genetic base broadening scheme including (*Bridging*) or not bridging (*Nobridging*) and (ii) for the potential donors considered, to cover different possibilities in both hybrid and inbred species. We considered as potential donors either the 338 genotypes from the Amaizing project or the D progeny with the largest GEBVs released by the external breeding program (i.e. 1 DH/family/year, 20 potential donors released every year). The scenario using the 338 genotypes from the Amaizing panel for genetic base broadening was identified with the suffix *Panel*. For the donors released by the external breeding program, we considered two time constraints for the access to diversity. To mimic a situation close to that of the US maize ex-PVPA system (Mikel and Dudley 2006), we first considered donors released 20

to 24 years before the current year (i.e. 5 years x 20 DH = 100 potential D) in scenarios with the suffix *20y*. To simulate a faster access to external diversity, as it would be the case in line breeding under UPOV convention (Dutfield 2011), we considered the donors released by the external breeding 5 to 9 years before the current year (i.e. 100 potential D) in scenarios with the suffix *5y*. For scenarios without bridging (Figure 2C), the E candidate parents were selected every year among the 5% E progeny showing the largest GEBVs per family in the three last breeding generations resulting in N_E = 4 DH x 20 families x 3 years = 240 potential E parents. The E progeny were selected based on the elite GS model trained on E progeny of the three last generations (i.e. 4,800 E progeny, Figure 2C). The 20 breeding crosses among the 28,680 candidate ExE elite crosses and DxE introduction crosses were selected using optimal cross selection without constraint on the type of crosses elite or introduction, using the elite GS model as described in section "Optimal cross selection". For scenarios with bridging (Figure 2D), the population was split into a bridging population of 5 families of 80 DH (i.e. 400 DE progeny) and a breeding population of 15 families of 80 DH (i.e. 1,200 E progeny). Every year, the E candidate parents for breeding were selected among the 5% E progeny per family showing the largest GEBVs from the three last breeding generations, resulting in N_E = 4 DH/family x 15 family x 3 year = 180 potential E parents. The E progeny were selected based on the elite GS model trained on all E progeny of the three last generations (i.e. 3,600 E progeny, Figure 2D). The DE candidate parents for introduction in the breeding population were similarly selected among the three last bridging generations, resulting in N_{DE} = 4 DH/family x 5 families x 3 years = 60 potential DE parents. The DE progeny were selected based on the bridging GS model trained on all DE progeny of the three last generations, i.e. 1,200 DE (Figure 2D). Among the $N_E(N_E - 1)/2$ = 16,110 ExE elite crosses and $N_{DE}N_E$= 10,800 DExE introduction crosses possible for breeding, the 15 breeding crosses were defined using optimal cross selection with the elite GS model and without constraint on the type of crosses ExE (elite) or DExE (introduction). The 5 DxE bridging crosses were selected among the possible crosses between the available D and potential E parents with the bridging GS model, conditionally to selected breeding crosses as described in the next section: optimal cross selection.

Optimal cross selection

The optimal cross selection selects the set of crosses ($\boldsymbol{nc}$) that maximizes the expected genetic value in the progeny (V) under a constraint on the genomewide genetic diversity in the progeny (D) (Kinghorn *et al.* 2009; Kinghorn 2011; Akdemir and Isidro-Sánchez 2016; Gorjanc *et al.* 2018; Akdemir *et al.* 2019). As proposed in Allier *et al.* (2019c), the effect of within family selection with intensity (i) and accuracy (h) on $V^{(i,h)}$ and $D^{(i,h)}$ can be accounted for in optimal cross selection by using UCPC based OCS (Appendix B). Similarly as in Allier *et al.* (2019c), we considered $h = 1$ for sake of simplicity.

For breeding crosses, the optimal set of $|\boldsymbol{nc}|$ = 20 crosses (in scenarios without bridging, Figure 2A-C) or $|\boldsymbol{nc}|$ = 15 crosses (in scenarios with bridging, Figure 2D) was selected to solve the multi-objective optimization problem:

$$\max_{\boldsymbol{nc}} V^{(i)}(\boldsymbol{nc})$$
$$\text{with } D^{(i)}(\boldsymbol{nc}) \geq He(t), \text{ (Eq. 1)}$$

where $He(t), \forall\, t \in [0, t^*]$ is the minimal genomewide diversity constraint at time t. The evolution of diversity along time was controlled by the targeted diversity trajectory, i.e. $He(t), \forall\, t \in [0, t^*]$ where $t^* \in \mathbb{N}^*$ is the time horizon when the diversity $He(t^*) = He^*$ should be reached. For the external and

the commercial benchmark without introductions breeding programs, we considered $He^* = 0.10$ and $He^* = 0.01$ reached after sixty years, respectively. As in Allier *et al.* (2019c), the constraint on $D^{(i)}$ followed a linear trajectory over time:

$$He(t) = \begin{cases} He^0 + \frac{t}{t^*}(He^* - He^0), \ \forall\, t \in [\![0, t^*]\!] \\ He^*, \ \forall\, t > t^* \end{cases}, \text{(Eq. 2)}$$

where He^0 is the initial diversity at $t = 0$, i.e. at the end of burn-in.

For the commercial breeding program with introductions, we maintained the genomewide diversity constant after the end of burn-in, i.e. $He(t) = He^0, \forall\, t \in [\![0, t^*]\!]$. Thus, the UCPC based OCS selected introduction crosses (i.e. DxE if no bridging and DExE if bridging) when necessary to maximize the performance while keeping genomewide diversity constant (Eq. 1). In case of bridging, we completed the 15 selected breeding crosses with 5 bridging crosses (DxE, Figure 2D) that maximized the following function on the full set of $|\boldsymbol{nc}| = 20$ crosses:

$$\max_{\boldsymbol{nc}} \alpha\, V^{(i)*}(\boldsymbol{nc}) + (1 - \alpha)\, D^{(i)*}(\boldsymbol{nc}), \text{(Eq. 3)}$$

where, $\alpha \in [0,1]$ is the relative weight given to performance compared to diversity, i is the within family selection intensity, $V^{(i)*}(\boldsymbol{nc}) = \frac{V^{(i)}(\boldsymbol{nc}) - V^{(i)}(\boldsymbol{nc}_D^*)}{V^{(i)}(\boldsymbol{nc}_V^*) - V^{(i)}(\boldsymbol{nc}_D^*)}$ and $D^{(i)*}(\boldsymbol{nc}) = \frac{D^{(i)}(\boldsymbol{nc}) - D^{(i)}(\boldsymbol{nc}_V^*)}{D^{(i)}(\boldsymbol{nc}_D^*) - D^{(i)}(\boldsymbol{nc}_V^*)}$ with $\boldsymbol{nc}_V^*$ and $\boldsymbol{nc}_D^*$ are the lists of crosses that maximize the performance (V) and the diversity (D), respectively. A differential evolution (DE) algorithm was used to find Pareto-optimal solutions of Eq. 1 and Eq. 3 (Storn and Price 1997; Kinghorn *et al.* 2009; Kinghorn 2011).

Interest of pre-breeding and bridging

We compared different commercial breeding programs with recurrent introductions considering or not bridging at constant cost (i.e. total of 1,600 DH/year) and considering three types of potential donors, resulting in the six genetic base broadening scenarios: *Bridging_Panel*, *Nobridging_Panel*, *Bridging_20y*, *Nobridging_20y*, *Bridging_5y*, *Nobridging_5y*. We ran ten independent simulation replicates of the external program that generated donors, the commercial benchmark without introductions, and the six genetic base broadening scenarios. Note that at a given simulation replicate the commercial breeding program accessed the potential donors released by the corresponding external breeding program simulation replicate.

We followed several indicators in the breeding families (i.e. E progeny, Figure 2). At each generation $T \in [0,60]$ with $T = 0$ corresponding to the last burn-in generation, we computed the mean genetic merit of E progeny $\mu(T) = mean(TBV(T))$ and of the ten most performing E progeny $\mu_{10}(T) = mean\left(\max_{10}(TBV(T))\right)$ as a proxy of the performance that could be achieved at the commercial level by releasing these lines as varieties. We also measured the frequency of the favorable allele in the E progeny $p_j(T)$ at each QTL j among the 1,000 QTLs. We further focused on the QTLs where the favorable allele was rare at the end of burn-in, i.e. $p_j(0) \leq 0.05$. The results were averaged and standard errors were computed over ten independent replicates.

Effect of a joint genomic selection model for bridging and breeding

For the three scenarios with bridging, we investigated the interest of a single TS grouping 3,600 DE and 1,200 E progeny to predict both breeding and bridging families. These three additional scenarios were referred to as *Bridging_Panel (Single TS)*, *Bridging_20y (Single TS)* and *Bridging_5y (Single TS)*. Every generation, we defined the prediction accuracies as the correlation between true breeding values and GEBVs ($cor(u, \hat{u})$) within breeding elite families (ExE), breeding introduction families (DExE) and bridging families (DxE). The prediction accuracies were averaged over the ten replicates and further averaged over the sixty generations. Note that considering a single GS model at constant cost yielded not only a broader but also a larger training set (4,800 DH progeny instead of 3,600 DH progeny for elite GS or 1,200 DH progeny for bridging GS, Figure 2).

We further investigated the effect of the proportion of DE and E progeny in the TS at constant size on within ExE and DExE family selection accuracy. We considered the 1,200 DE and 3,600 E progeny genotypes and phenotypes simulated at generations 18, 19, 20 in the first replicate of scenario *Bridging_20y*. We further selected the 5% DH per family with the highest GEBVs obtained using a GS model trained on all 4,800 progeny genotypes and phenotypes. These were randomly crossed to generate 50 elite (ExE) and 50 introduction (DExE) families of 80 DH progeny. These families were considered as the validation set (VS). We randomly sampled among the 4,800 DH progeny different TS of variable sizes and compositions (Table 1) and we evaluated the within elite (ExE) and introduction (DExE) family prediction accuracy ($cor(u, \hat{u})$). We also evaluated the within family variance prediction accuracy as the correlation between the variance of true breeding values and the estimated variance ($cor(\sigma, \hat{\sigma})$). We reported results for twenty independent samples.

Table 1 Description of the training sets compared: the full training sets considering all available progeny of the last three generations and training sets at constant size (1,200 progeny or 3,600 progeny) with variable proportion of DE progeny.

	TS name	Number of E	Number of DE
Full TS	Pure E (3,600)	3,600	0
	Pure DE (1,200)	0	1,200
	1/4 - DE (4,800)	3,600	1,200
Constant size (1,200)	Pure E (1,200)	1,200	0
	1/4 - DE (1,200)	900	300
Constant size (3,600)	1/3 - DE (3,600)	2,400	1,200
	1/4 - DE (3,600)	2,700	900
	1/6 - DE (3,600)	3,000	600
	1/12 – DE (3,600)	3,300	300
	1/24 - DE (3,600)	3,450	150
	1/36 - DE (3,600)	3,500	100

Results

Interest of pre-breeding and bridging

The interest of recurrent introductions in the commercial breeding program after or without bridging depended on the type of donor considered. Panel donors showed a large performance gap with the elites they were crossed to. This performance gap increased with advanced breeding generations (on average a true breeding value difference with elites increasing from -15 and -104 trait units). Improved donors showed a lower performance gap with elites. Twenty-year old donors showed an intermediate performance gap with elite (on average -22 trait units) and five-year old donors showed a reduced performance gap with elite (on average -8 trait units).

Direct introductions of panel donors without bridging (*Nobridging_Panel*) penalized the breeding population mean performance (μ) at short-term (at five years, μ = 8.168 +/- 0.282 compared to 9.239 +/- 0.237 without introductions, Figure 3A, Table S1) and long-term (at sixty years, μ = 9.651 +/- 0.958 compared to 38.837 +/- 1.563 without introductions, Figure 3A, Table S1). When considering the mean performance of the ten best progeny (μ_{10}), the short-term penalty was no more significant (at five years, μ_{10} = 15.802 +/- 0.341 compared to 15.746 +/- 0.391 without introductions, Figure 3B, Table S2) but the long-term penalty was still significant (at sixty years, μ_{10} = 29.767 +/- 1.108 compared to 39.567 +/- 1.571 without introductions, Figure 3B, Table S2). The introduction of panel donors after bridging (*Bridging_Panel*) did not significantly penalize the short-term mean performance of the breeding population (at five years, μ = 8.688 +/- 0.329 compared to 9.239 +/- 0.237 without introductions, Figure 3A, Table S1) and yielded significantly higher long-term performance (at sixty years, μ = 52.110 +/- 0.886 compared to 38.837 +/- 1.563 without introductions, Figure 3A, Table S1). When considering μ_{10}, the short-term penalty was reduced (at five years, μ_{10} = 15.605 +/- 0.477 compared to 15.746 +/- 0.391 without introductions, Figure 3B, Table S2) and the long-term gain increased (at sixty years, μ_{10} = 61.763 +/- 1.298 compared to 39.567 +/- 1.571 without introductions, Figure 3B, Table S2).

Direct introductions of twenty-year donors without bridging (*Nobridging_20y*) yielded a penalty in the mid-term compared to not introducing donors (at twenty years, μ = 16.818 +/- 2.397 compared to 23.182 +/- 1.446 without introductions, Figure 3A, Table S1). When considering μ_{10}, the mid-term penalty due to introductions was limited (Figure 3B, Table S2). After thirty years, this introduction scenario significantly outperformed the benchmark (μ = 33.546 +/- 1.519 compared to 30.006 +/- 1.319 without introductions, Figure 3A, Table S1) and this advantage increased until the end of the sixty years evaluated period (μ = 66.944 +/- 0.849 compared to 38.837 +/- 1.563 without introductions, Figure 3A, Table S1). The introduction of twenty-year old donors after bridging (*Bridging_20y*) penalized only the short-term performance (at five years, μ = 8.687 +/- 0.293 compared to 9.239 +/- 0.237 without introductions, Figure 3A, Table S1) and yielded significantly higher performance than the benchmark after twenty years (μ = 27.987 +/- 0.840 compared to 23.182 +/- 1.446 without introductions, Figure 3A, Table S1). Introductions after bridging significantly outperformed the direct introductions until the end of the sixty years evaluated period (μ = 69.154 +/- 0.868 with bridging compared to 66.944 +/- 0.849 without bridging and μ_{10} = 74.413 +/- 0.932 with bridging compared to 72.258 +/- 0.978 without bridging, Figure 3A-B, Table S1-S2).

Introducing five-year old donors after or without bridging yielded significantly higher mid- and long-term performances than all other tested scenarios, without any significant long-term advantage of introductions after bridging compared to direct introductions (at sixty years, μ = 74.074 +/- 0.869 with bridging compared to 74.662 +/- 0.938 without bridging, Figure 3, Table S1).

We observed that the recurrent introductions of donors impacted the genetic diversity of the commercial germplasm. The more the commercial program had access to recent germplasm of the external program, the more the varieties released by the commercial program where admixed with the external program elite germplasm (Figure 4B and Figure 4C). In the scenario where only panel donors were accessible for introductions, the internal program diversity did not converge toward the external program (Figure 4A).

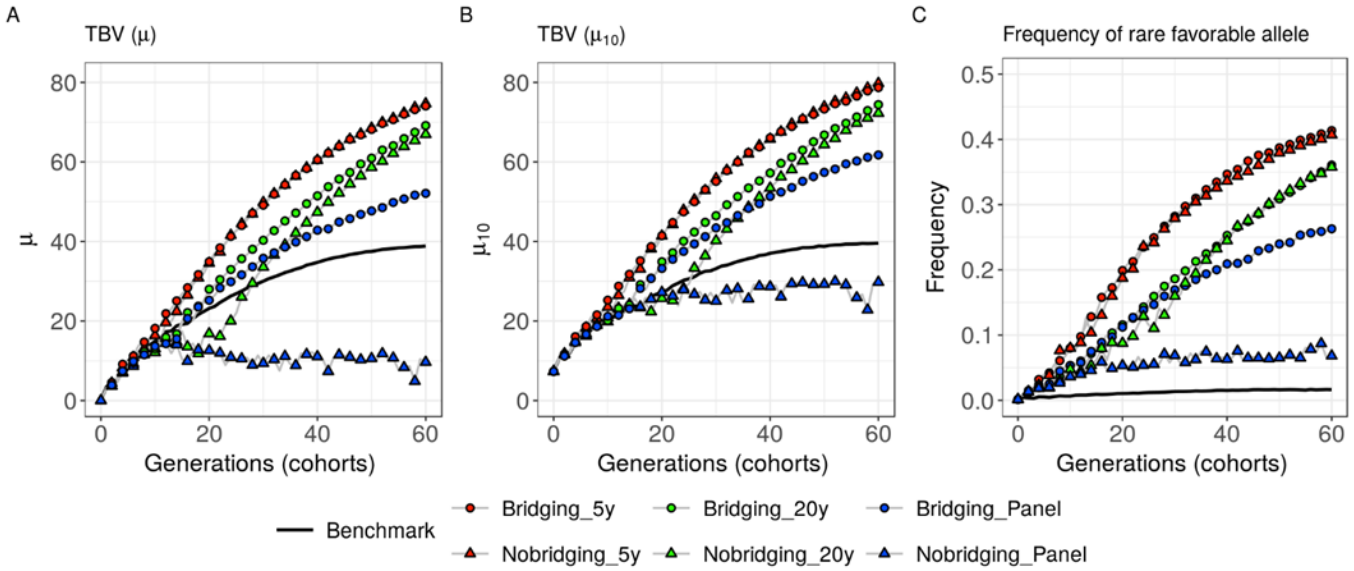


Figure 3 Evolution of the breeding population over generations. Scenarios considering presence or absence of bridging before introduction with different type of donors (panel, twenty-year old and five-year old donors). (**A**) Mean breeding population performance (μ), (**B**) mean performance of the ten best progeny (μ_{10}) and (**C**) frequency of the favorable alleles that were rare at the end of burn-in (i.e. $p(0) \leq 0.05$ corresponding on average to 269.9 +/- 23.6 QTLs).

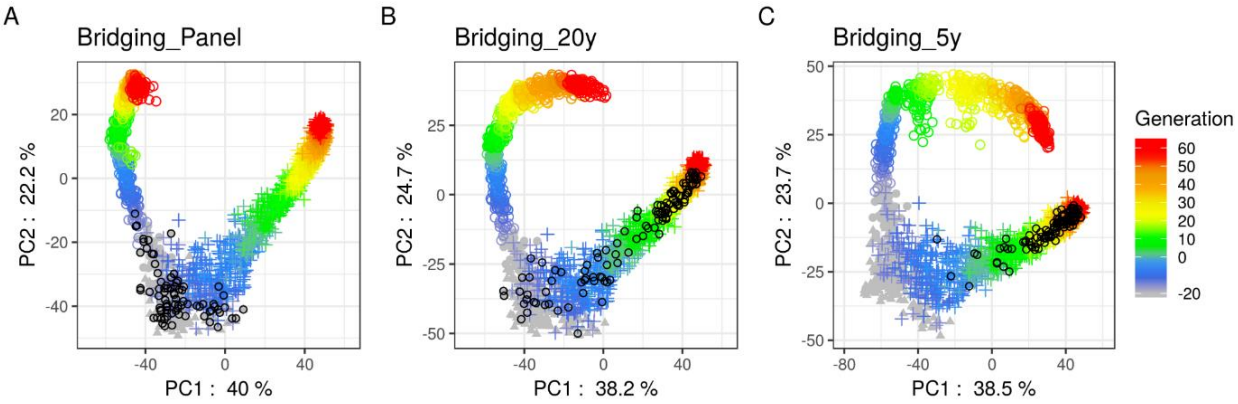


Figure 4 Principal component analysis of the modified Roger's genetic distance matrix (Wright 1978) of the 338 founders (gray: points for the 57 Iodent lines and triangles for the 281 remaining lines), the commercial ten best performing E progeny per generation (colored circles sign) and the twenty donors per generation released by the external program (colored plus sign). Both commercial and external lines are colored regarding their generation (note that negative generations correspond to burn-in). Black circles represent the donors that have been introduced into the commercial breeding program. Only three scenarios with bridging are represented for the first simulation replicate, (**A**) when only donors from panel were accessible, (**B**) when twenty-year old donors from the external breeding were accessible and (**C**) when five-year old donors from the external breeding were accessible.

The evolution of the mean frequency of initially rare favorable alleles (i.e. favorable allele that had a frequency at the end of burn-in ≤ 0.05 in the elite breeding population) also highlighted differences between strategies. The older the donors, the lower the increase in frequency of initially rare favorable alleles (at sixty years for scenario with bridging, the mean frequency was 0.414 +/- 0.012 for five-year old donors, 0.361 +/- 0.009 for twenty-year old donors, 0.263 +/- 0.008 for panel donors and 0.016 +/- 0.006 without introductions, Figure 3C, Table S3). For twenty-year old donors, omitting the bridging before introduction delayed the increase in frequency of initially rare favorable alleles (e.g. at twenty years, the mean frequency was 0.088 +/- 0.014 without bridging compared to 0.116 +/- 0.011 with bridging, Figure 3C, Table S3). More importantly, for panel donors the absence of bridging significantly penalized the increase in frequency of initially rare favorable alleles (at sixty years, 0.068 +/- 0.007 without bridging compared to 0.263 +/- 0.008 with bridging, Figure 3C, Table S3).

Effect of a joint genomic selection model for bridging and breeding

Scenarios considering a single TS of 3,600 E and 1,200 DE progeny yielded higher mid- and long-term μ and μ_{10} than scenarios considering two distinct TS for bridging and breeding (Figure 5A-B). After twenty years, single TS scenarios significantly outperformed scenarios with two distinct TS (μ = 40.111 +/- 1.149 compared to 34.900 +/- 0.905 for five-year old donors, μ = 30.497 +/- 1.135 compared to 27.987 +/- 0.840 for twenty-year old donors and μ = 29.292 +/- 0.802 compared to 25.212 +/- 1.314 for panel donors, Figure 5A, Table S1). After sixty years, the advantage of a single TS remained significant except for five-year old donors (μ = 75.749 +/- 1.093 compared to 74.074 +/- 0.869 for five-year old donors, μ = 71.130 +/- 1.028 compared to 69.154 +/- 0.868 for twenty-year old donors and μ = 57.067 +/- 1.444 compared to 52.110 +/- 0.886 for panel donors, Figure 5A, Table S1). When considering μ_{10}, a single TS was still more performing but its interest was less significant (e.g. for panel donors after sixty years, μ_{10} = 63.699 +/- 1.698 compared to 61.763 +/- 1.298, Figure 5 B, Table S1-S2). A single TS also favored the increase in frequency of initially rare favorable alleles introduced by five-year old donors and twenty-year old donors (e.g. for twenty-year old donors after sixty years, 0.380 +/- 0.010 compared to 0.361 +/- 0.009, Figure 5C, Table S3).

The observed within family prediction accuracies varied depending on the TS considered. For twenty-year old donors introduced after bridging, considering a single TS of 4,800 DE+E did not significantly improve the prediction accuracy within ExE families compared to using the pure elite TS of 3,600 E ($cor(u, \hat{u})$ = 0.73 +/- 0.06 compared to $cor(u, \hat{u})$ = 0.72 +/- 0.07, Table 2). However, it significantly improved the prediction accuracy within introduction DExE families compared to the pure elite TS of 3,600 E ($cor(u, \hat{u})$ = 0.77 +/- 0.07 compared to $cor(u, \hat{u})$ = 0.61 +/- 0.11, Table 2). A single TS also slightly but not significantly improved the prediction accuracy within bridging DxE families compared to the pure bridging TS of 1,200 DE ($cor(u, \hat{u})$ = 0.78 +/- 0.05 compared to $cor(u, \hat{u})$ = 0.73 +/- 0.06, Table 2). Similar observations were made on the other scenarios considering five-year old and panel donors. Prediction accuracies were larger in introduction DExE and bridging DxE families with older donors, i.e. phenotypically distant to elites, due to larger within family variances (e.g. for DExE families 14.43 +/- 4.40 for panel donors, 6.92 +/- 2.10 for twenty-year old donors and 5.00 +/- 1.41 for five-year old donors, Table 2).

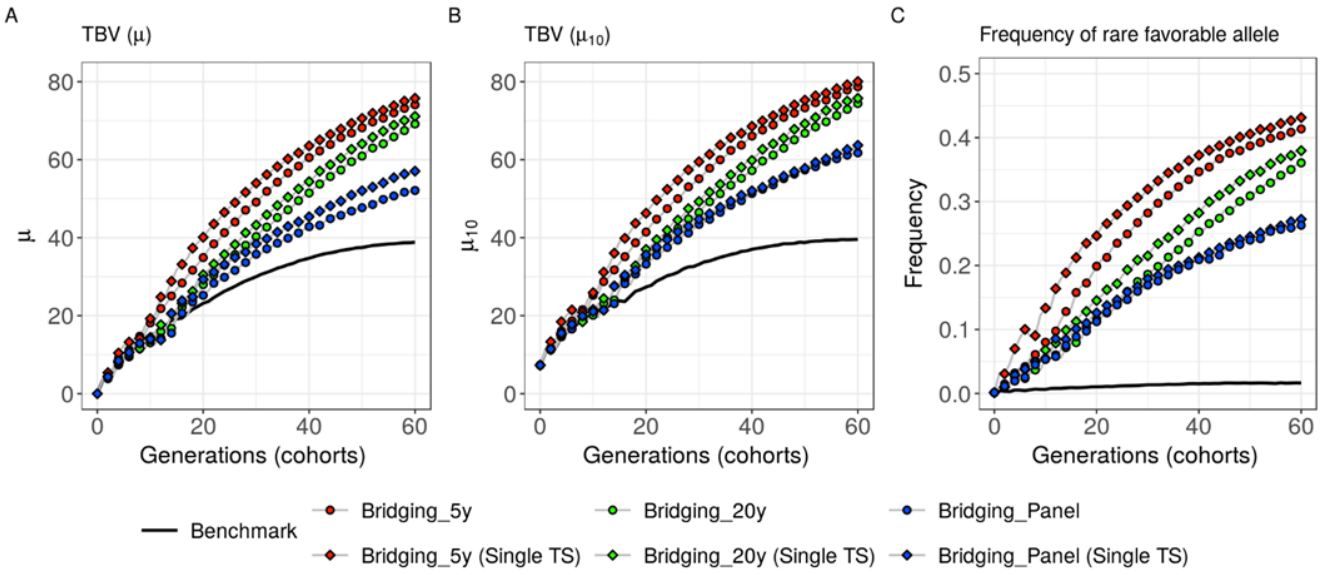


Figure 5 Evolution of the breeding population over generations. Scenarios considering bridging with different donors (panel, twenty-year old and five-year old donors) and either a single broad TS (*Single TS*) or two distinct training set for bridging and breeding (*default*). (**A**) Mean breeding population performance (μ), (**B**) mean performance of the ten best progeny (μ_{10}) and (**C**) frequency of the favorable alleles that were rare at the end of burn-in (i.e. $p(0) \leq 0.05$ corresponding on average to 269.9 +/- 23.6 QTLs).

Table 2 Within family prediction accuracies ($cor(u, \hat{u})$) depending on the validation set (VS): elite (ExE), introduction (DExE) and bridging (DxE) and the training set (TS) considered: pure elite (E), pure bridging (DE) and merged (E+DE). Results are given for scenarios with different donors, from the panel, twenty-year old and five-year old donors, considering a single TS and prediction accuracies are averaged over the ten replicates and all sixty generations. In brackets are given the standard errors averaged over sixty generations.

VS	Five-year old donor				Twenty-year old donor				Panel donor			
	Family variance	Prediction accuracy			Family variance	Prediction accuracy			Family variance	Prediction accuracy		
		TS = E (3,600)	TS = DE (1,200)	TS = E+DE (4,800)		TS = E (3,600)	TS = DE (1,200)	TS = E+DE (4,800)		TS = E (3,600)	TS = DE (1,200)	TS = E+DE (4,800)
ExE	3.76 (1.17)	0.69 [a] (0.07)	0.48 (0.1)	0.72 [b] (0.06)	3.93 (1.06)	0.72 [a] (0.07)	0.47 (0.10)	0.73 [b] (0.06)	4.02 (1.16)	0.72 [a] (0.05)	0.44 (0.10)	0.73 [b] (0.05)
DExE	5.00 (1.41)	0.60 [a] (0.1)	0.59 (0.1)	0.73 [b] (0.07)	6.92 (2.10)	0.61 [a] (0.11)	0.65 (0.10)	0.77 [b] (0.07)	14.43 (4.40)	0.65 [a] (0.12)	0.78 (0.07)	0.86 [b] (0.05)
DxE	9.69 (2.01)	0.61 (0.08)	0.66 [a] (0.08)	0.73 [b] (0.07)	18.31 (3.78)	0.65 (0.08)	0.73 [a] (0.06)	0.78 [b] (0.05)	64.15 (12.89)	0.74 (0.07)	0.82 [a] (0.04)	0.86 [b] (0.03)

[a] Prediction accuracies that would have been realized if the breeding (E) or bridging (DE) families had been each predicted only by the corresponding training set (to be compared with [b]).

[b] Realized prediction accuracies when considering a single training set (to be compared with [a]).

At constant TS size of 3,600 DH, the increase in proportion of DE progeny from 0 to 1/3 in the TS increased the prediction accuracy within introduction DExE families ($cor(u, \hat{u})$ = 0.58 +/- 0.02 to 0.73 +/- 0.01, Figure 6B) while it reduced the prediction accuracy within elite ExE families ($cor(u, \hat{u})$ = 0.70 +/- 0.01 to 0.65 +/- 0.02, Figure 6A). The TS with 3,000 E and 600 DE, appeared as a suitable compromise with within introduction DExE family $cor(u, \hat{u})$ = 0.70 +/- 0.02 and elite ExE families $cor(u, \hat{u})$ = 0.68 +/- 0.01. At constant TS size of 1,200 DH, the TS with 900 E and 300 DE progeny performed similarly as the pure bridging TS for prediction within DExE families ($cor(u, \hat{u})$ = 0.63 +/- 0.03 compared to 0.62 +/- 0.02, Figure 6B) but significantly outperformed the pure bridging TS for prediction within elite ExE families ($cor(u, \hat{u})$ = 0.52 +/- 0.04 compared to 0.34 +/- 0.02, Figure 6A). The within family variance prediction accuracy showed similar tendencies (Figure 7A-B). The increase in proportion of DE progeny from 0 to 1/3 in the TS increased the prediction accuracy within introduction DExE families ($cor(\sigma, \hat{\sigma})$ = 0.56 +/- 0.09 to 0.76 +/- 0.07, Figure 7B) while it reduced the prediction accuracy within elite ExE families ($cor(\sigma, \hat{\sigma})$ = 0.74 +/- 0.07 to 0.71 +/- 0.08, Figure 7A).

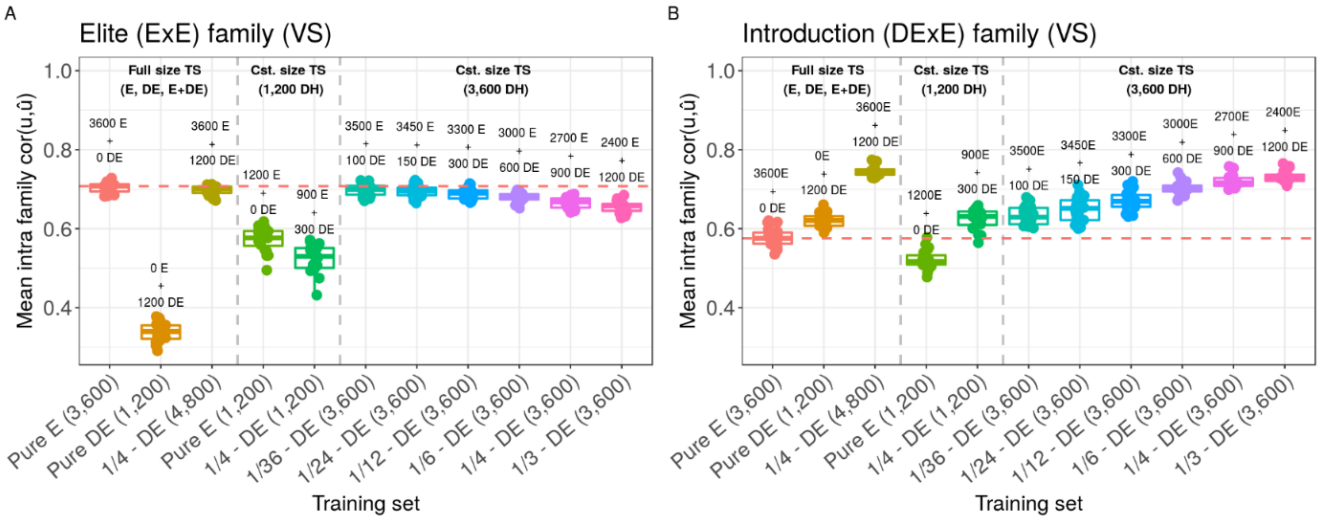


Figure 6 Effect of TS composition on intra family prediction accuracies ($cor(u, \hat{u})$) considering genotypes simulated at generations 18, 19, 20 in the scenario *Bridging_20y*. (**A**) Mean prediction accuracy within 50 elite (ExE) families and (**B**) mean prediction accuracy within 50 introduction (DExE) families. Boxplots represent the results for 20 independent replicates. One can distinguish three training set types (left to right): Full training set considering all 3,600 E progeny (Pure E), all 1,200 DE progeny (Pure DE) and all 3,600 E + 1,200 DE progeny; Training sets at constant size of 1,200 DH for comparison with Pure DE; Training sets at constant size of 3,600 DH and variable proportion of DE progeny for comparison with Pure E. The red dotted line represents the median value for Pure E TS.

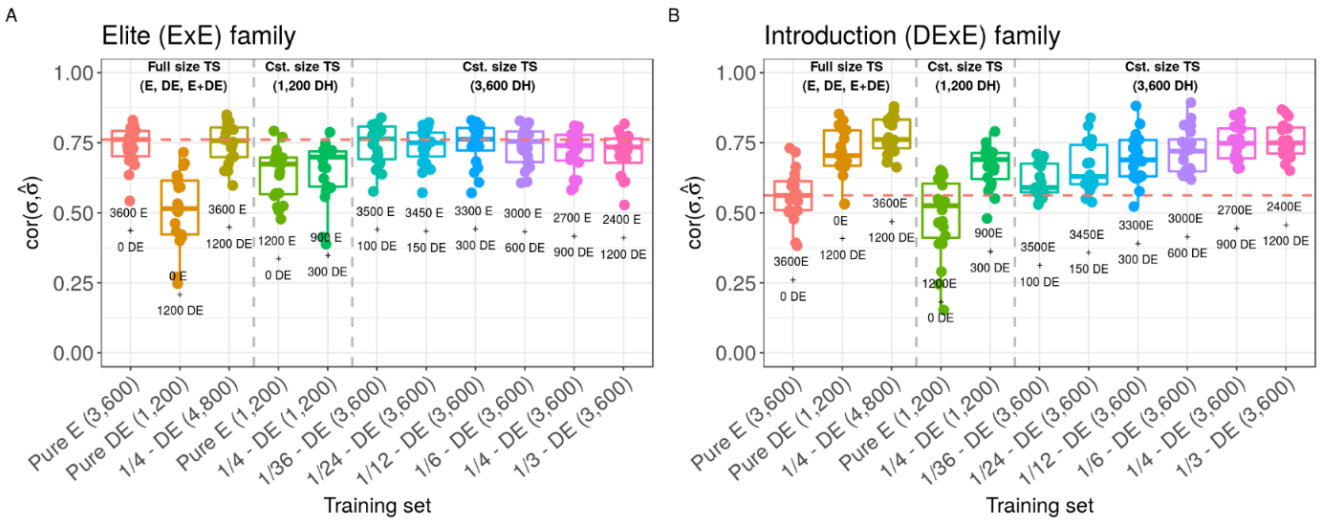


Figure 7 Effect of TS composition on family variance prediction accuracy ($cor(\sigma, \hat{\sigma})$) considering genotypes simulated at generations 18, 19, 20 in the scenario *Bridging_20y*. (**A**) Mean prediction accuracy in 50 elite (ExE) families and (**B**) mean prediction accuracy in 50 introduction after bridging (DExE) families. Boxplots represent the results for 20 independent replicates. One can distinguish three training set types (left to right): Full training set considering all 3,600 E progeny (Pure E), all 1,200 DE progeny (Pure DE) and all 3,600 E + 1,200 DE progeny; Training sets at constant size of 1,200 DH for comparison with Pure DE; Training sets at constant size of 3,600 DH and variable proportion of DE progeny for comparison with Pure E. The red dotted line represents the median value for Pure E TS.

Discussion

Genetic base broadening with optimal cross selection accounting for within family variance

Despite the recognition of the importance to broaden the elite genetic base in most crops, commercial breeders are reluctant to penalize the result of several generations of intensive selection by crossing these to unimproved genetic resources. Furthermore, among the large diversity available for genetic base broadening (e.g. landraces, public lines, varieties…), the identification of the useful genetic diversity to broaden the elite pool is difficult and might dishearten breeders. Consequently, there is a need for global breeding strategies that improve genetic resources to bridge the performance gap with elites, identify interesting sources of diversity that complement at best the elite germplasm and efficiently introduce them into elite germplasm.

The identification of genetic resources for polygenic enrichment of the elite pool should account for the complementarity between genetic resources and elites as reviewed in Allier *et al.* (2020). Allier *et al.* (2019b) proposed the Usefulness Criterion Parental Contribution (UCPC) approach to predict the interest of crosses between genetic resources and elite recipients based on the expected performance and diversity in the most performing fraction of the progeny. The interest of UCPC relies on the fact that it accounts for within family variance and selection when identifying crosses. For instance, when crossing phenotypically distant parents, e.g. genetic resource and elite recipient, we expect a higher cross variance that should be accounted for to properly evaluate the usefulness of the cross (Schnell and Utz 1975; Longin and Reif 2014; Allier *et al.* 2019b). Additionally, we expect the best performing fraction of the progeny to be genetically closer to the best parent. This deviation from the average parental value should be considered to evaluate properly the genetic diversity in the next generation (Allier *et al.* 2019b; d). Accounting for parental complementarity at marker linked to QTLs also favors effective recombination in progeny and breaks negative gametic linkage disequilibrium between QTLs (i.e. repulsion), which unleashes additive genetic variance and increases long-term genetic gain (Allier *et al.* 2019c). Therefore, the OCS is particularly adapted to genetic diversity management in pre-breeding and breeding programs (Akdemir and Isidro-Sánchez 2016; Cowling *et al.* 2017; Gorjanc *et al.* 2018; Allier *et al.* 2019c). The objective and the originality of this study were to consider UCPC based OCS to jointly select donors, introduction crosses and elite crosses to ensure an overall consistency of genetic base broadening accounting for the performance and diversity available in both bridging and breeding populations.

Genetic resources and simulated pre-breeding

Different sources of diversity can be considered by commercial breeders. The most original, but which show a large performance gap with elites, are landraces (e.g. DH libraries derived from landraces, Strigens *et al.* 2013; Melchinger *et al.* 2017; Böhm *et al.* 2017) and first varieties derived from landraces. Since breeding industry is highly competitive, breeders are likely reluctant to introduce unselected genetic resources directly into the breeding germplasm despite they might carry favorable adaptation alleles to face climatic changes (McCouch *et al.* 2013; Hellin *et al.* 2014; Böhm *et al.* 2017). Instead, commercial breeders will prefer to consider elite inbred lines from other than their own program (Kannenberg 2001).

In this study, the external breeding program was designed to release every generation several improved lines, later considered as donors for genetic base broadening of the commercial breeding program. The external program started from a broader genetic diversity than the commercial program (on average, He = 0.283 compared to He = 0.133 at the end of burn-in) and was designed to maintain higher genetic diversity during selection (on average, He = 0.101 compared to He = 0.014 after sixty years). This was done to mimic in a simple way the outcome of the activity of several companies conducting separate programs and therefore maintaining a global diversity. The external program can also be viewed as a pre-breeding program since it aimed at improving genetic resources to reduce their performance gap with elites while maintaining genomewide diversity among the pre-breeding population (Figure 1). The situation where the commercial breeding program can access donors released twenty years ago mimicked the situation of private lines with expired plant protection act in maize (Mikel and Dudley 2006) or old public lines. The situation where the commercial breeding program can access donors released five years ago mimicked either donors released by pre-breeding programs (e.g. in maize the SeeD project, Gorjanc *et al.* 2016) or donors released by programs working a different genetic basis and targeting different environments (e.g. commercial varieties in inbred species accessible for breeding under the UPOV convention, Dutfield 2011). The selection intensity was lower in the external breeding than in the commercial breeding programs (10% vs 5% of progeny selected, respectively). This was done to compensate the increased response to selection due to the higher genetic diversity and ensure that the donors released by the external program underperform the commercial breeding elites. It should be noted that donors outperforming elites might be encountered in practice when considering elite germplasm as source of diversity, but this situation was not considered in this study. In such a situation the direct introduction of donors would be clearly preferable.

Interest of introductions after bridging

When considering recent and performing donors (five-year old), scenarios with introductions after bridging or direct introductions performed similarly. Conversely, for panel and twenty-year old donors, introductions after bridging yielded significantly higher mid- and long-term performance compared to direct introductions. Note that introductions after bridging can be seen as a specific three-way cross with selection of the progeny of the first donor by elite recipient cross followed by crossing the selected progeny to a second more recent elite recipient. Assuming no selection between the first cross and the second cross, Allier *et al.* (2019b) predicted that three-way crosses were more prone to deliver performing progeny than back-crosses and F1 biparental crosses, when considering donors underperforming the elite germplasm. Since donors (D) were less performing than elites, the fraction of progeny selected in donor by elite bridging families (DE progeny) carried on expectation less than half of donor's genome (Allier *et al.* 2019b). Thus, progeny of introduction crosses after bridging (DExE) carried on expectation less than one fourth of the donor (D) genome. This D fraction includes favorable alleles but also unfavorable alleles brought by linkage drag, which number depends on the donor considered. Introductions penalized the mean breeding population performance in the first generations (Figure 3A-B). Next generations of recombination and selection partially broke the linkage between favorable and unfavorable alleles in introduced regions, resulting in a higher genetic gain than in the benchmark (Figure 3A-B) and an increase of the frequency of novel favorable alleles (Figure 3C). The more performing the donor, the less unfavorable alleles linked to favorable alleles and the more

rapidly novel favorable alleles were introduced and spread in the breeding population (Figure 3C). In absence of bridging, the introduction progeny (DxE) carried on expectation one half of the donor genome. Consequently, the penalty due to introductions was more important and the conversion into genetic gain required more recombination events, i.e. recycling generations (Figure 3A-B). For panel donors showing a large performance gap with elites, the direct introductions were not converted into genetic performance. The high inter-family additive variance in this scenario (Figure S1 A) reflected the structuration of the breeding population into badly performing introduction families and performing elite families with only limited gene flow between them. Such behavior might be corrected by adding a constraint to force the recycling of introduction progeny in Eq. 1 when donors are too badly performing, which requires further investigations.

Practical implementation in breeding programs

We considered a commercial breeding program with a genetic diversity at the end of the burn-in matching that of an experimental program reported by Allier *et al.* (2019a). Breeding programs ongoing for different species and breeders may present a diversity superior or inferior to the one that was simulated, which would make the importance of introductions lower or stronger than in the simulated scenarios, respectively. UCPC based OCS for genetic base broadening requires to genotype the candidate parents, including breeding material and potential donors, a genetic map and reliable marker effect estimates. This information is available in breeding programs that have already implemented genomic selection. In this study, we assumed fully homozygous inbred lines but considering heterozygote parents in UCPC based OCS is straightforward following the extension of UCPC to four-way crosses (Allier *et al.* 2019b). This is particularly interesting for perennial plants.

We proposed to implement bridging at constant cost by splitting the breeding population into a small bridging population and a large breeding population. This involves practical changes in the breeding organization that remain to be studied. We considered equal family sizes and within family selection intensities for bridging and breeding families. However, in practice different within family selection intensities can be considered in UCPC based OCS (Appendix B) and one may want to modulate the selection intensity among families, e.g. select less intensively in bridging and more intensively in breeding families. We could consider the selection intensities as fixed parameters regarding breeding objectives or as variable parameters to be optimized. The effect and the optimization of within family intensities in bridging and breeding requires further investigations. We considered a selection accuracy $h = 1$ for cross selection, for sake of facility. However, we observed that within family prediction accuracies were variable (Table 2, Figure 6). Note that *a priori* within family accuracy can be accounted for in UCPC based OCS (Appendix B). For instance it would give less importance to predicted variance for crosses with *a priori* low within family accuracy. The consequences on short- and long-term UCPC based OCS efficiency need to be investigated. In bridging, we gave more importance to performance than to diversity ($\alpha = 0.7$) when selecting bridging crosses in order to reduce the performance gap between donors derived materials and elites. When giving less weight to the performance than to the diversity, i.e. $\alpha = 0.3$, we observed non-significant changes on the short- or long-term performance for scenarios with five-year and twenty-year old donors and a significant increase of long-term performance and novel favorable allele frequency for the scenario with panel donors (Figure S2 A-C). This suggested that for unimproved donors, to select too strongly for performance in bridging favors

the first elite recipient genome contribution and limits the introduction of novel favorable alleles. Further investigations are required to better define this parameter for practical implementation.

In scenarios with bridging, we considered by default two distinct bridging and breeding GS models. The prediction of elite (ExE) and introduction (DExE) crosses usefulness and the prediction within crosses were based on a model trained on the breeding progeny of the three corresponding previous generations. Considering a unique genomic selection model trained on both bridging and breeding progeny increased the prediction accuracy within introduction families (DExE) (Table 2). This higher selection accuracy favored the spreading of the introduced favorable alleles in the breeding population and resulted in an increased mid- and long-term performance (Figure 5). Furthermore, compared to use two distinct TS, a single TS led to introduce more bridging progeny (DE) for scenarios considering good performing donors (five-years old) and less for scenarios considering bad performing donors (twenty-years old) (Figure S3 A). Also, as we likely selected more accurately the introduction crosses (DExE) with a single TS, there was an increase in the proportion of those that contributed to the ten best lines, especially for twenty-year old and panel donors (Figure S3 B).

It is well known that the prediction accuracy is increased for larger TS (Hickey *et al.* 2014). At constant TS size, increasing the proportion of bridging progeny (DE) up to one third in the TS significantly increased the family variance prediction accuracy ($cor(\sigma, \hat{\sigma})$) and within family prediction accuracy ($cor(u, \hat{u})$) in introduction families (DExE). Conversely, these higher proportions of bridging progeny (DE) in the TS significantly decreased $cor(\sigma, \hat{\sigma})$ and $cor(u, \hat{u})$ in elite families (ExE). The optimal balance between introduction and elite family prediction accuracies is likely data dependent as observed when considering genotypes and phenotypes simulated in different generations (Figure S4). For instance, considering later generations, a large proportion of DE in the TS penalized less the within elite prediction accuracy (Figure S4 C). The reason being that later breeding generations get closer to the external program germplasm (Figure 4). The optimal balance between bridging and breeding progeny in the training set might be defined using an optimization criterion such as the CDmean (Rincent *et al.* 2012) extended to account for linkage disequilibrium as suggested by Mangin *et al.* (2019).

Outlooks

We considered an inbred line breeding program corresponding to selecting lines on *per se* values for line variety development or on testcross values with fixed tester lines from the opposite heterotic pool for hybrid breeding. In this case, the use of testcross effects estimated on hybrids between candidate lines and tester lines is straightforward. The extension to hybrid reciprocal breeding is of interest for genetic broadening in several species such as maize and hybrid wheat (Longin and Reif 2014). In this context it is possible to account for the complementarity between heterotic groups in UCPC based OCS to complementarily enrich and improve both pools, ensuring a consistency of the hybrid program. This would require to include dominance effects in UCPC based OCS.

We considered a single trait selected in both the external and the commercial breeding programs in the same population of environments for a total of eighty years. These assumptions should be relaxed in further simulations. Firstly, it is well recognized that genetic resources suffer agronomic flaws (e.g. lodging, Tallury and Goodman 2001; Longin and Reif 2014) or miss adaptation (e.g. flowering time) that should be accounted for during pre-breeding and introduction in breeding. In such a multi-trait context, the multi-objective optimization framework proposed in Akdemir *et al.* (2019) can be

adapted to UCPC based OCS. Secondly, in practice several public pre-breeding programs or competitor programs can be considered as sources of candidate donors for genetic base broadening. These programs likely did not select for the same target environments and are themselves continuously enriched in new allelic variation. Thirdly, in a context of climate change and rapid evolving agricultural practices, breeding targets are expected to change (e.g. emerging biotic or abiotic stresses). Considering a more realistic context, where donors are released by different programs selecting in different environments and for different traits changing over time, likely makes the interest of maintaining genomewide genetic diversity through genetic base broadening even more important than highlighted in this study.

Appendix A

Simulation of progeny genotypes and phenotypes

Doubled haploid (DH) progeny genotypes were simulated considering meiosis events without crossover interference. The number of chiasmata was drawn from a Poisson distribution with λ equal to the chromosome length in Morgan, and crossover positions were determined using the recombination frequency obtained using the Haldane mapping function (Haldane 1919).
For phenotyping, we considered environmental effects sampled from a normal distribution of mean zero and variance 25 and did not consider genotype by environment interactions. Each generation was evaluated in $N_{loc} = 4$ locations in one year, i.e. four environments. Environmental errors were sampled from a normal distribution with mean zero and an error variance σ_ϵ^2 defined by the initial repeatability in the founder population $r = \frac{\sigma_G^2}{\sigma_G^2+\sigma_\epsilon^2} = 0.40$. This led to a heritability in the founder population of $h^2 = \frac{\sigma_G^2}{\sigma_G^2+\sigma_\epsilon^2/N_{loc}} = 0.73$ and $h^2 = 0.42$ at the end of burn-in in commercial breeding scenarios.

Genomewide prediction model

The genomic estimated breeding values of progeny (GEBV, $\hat{u}$) were estimated in Model 1 S1 fitted using mixed model software blup-f 90 (Misztal 2008) with AI-REML variance component estimates:

$$\boldsymbol{Y} = \mathbf{1}\mu + \boldsymbol{E\beta_{Env}} + \boldsymbol{Wu} + \boldsymbol{\epsilon}, \text{ (Model 1 S1)}$$

where $\boldsymbol{Y}$ is the vector of phenotypic values, μ is the intercept, $\boldsymbol{E}$ is the incidence matrix for environmental effects, $\boldsymbol{\beta_{Env}}$ is the vector of environmental fixed effects, $\boldsymbol{W}$ is the incidence matrix of individual breeding value random effects $\boldsymbol{u}$, $\boldsymbol{u} \sim N(\mathbf{0}, \sigma_G^2\boldsymbol{G})$ is the vector of breeding value random effects with $\boldsymbol{G}$ the genomic relationship matrix and $\boldsymbol{\epsilon}$ is the vector of independent residual random terms $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2\boldsymbol{I})$. $\boldsymbol{G}$ was estimated using the 2,000 non causal loci:

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ'}}{tr(\boldsymbol{ZZ'})/n}$$

where $\boldsymbol{Z}$ contains the centered allele counts, with elements computed as $x_{ij} + 1 - 2p_j$, where the element $x_{ij} \in \{-1,1\}$ is the genotype for individual i at non causal locus j and p_j is the frequency of the allele for which the homozygous genotype is coded 1 at non causal locus j. $tr(\boldsymbol{ZZ'})$ is the trace of $\boldsymbol{ZZ'}$ and $tr(\boldsymbol{ZZ'})/n$ forces the diagonal of $\boldsymbol{G}$ to be 1 on average (Legarra *et al.* 2009; Forni *et al.* 2011). Estimated marker effects $\widehat{\boldsymbol{\beta}}$ were obtained by back-solving: $\widehat{\boldsymbol{\beta}} = \boldsymbol{Z'}(\boldsymbol{ZZ'})^{-1}\widehat{\boldsymbol{u}}$ (Wang *et al.* 2012). The prediction accuracy was defined as $cor(\boldsymbol{u}, \widehat{\boldsymbol{u}})$ with $\boldsymbol{u}$ and $\widehat{\boldsymbol{u}}$ the vectors of true breeding values and genomic estimated breeding values, respectively.

Appendix B

We applied the Usefulness Criterion Parental Contributions approach (UCPC) proposed by Allier *et al.* (2019b) and further extended in Allier *et al.* (2019c) to evaluate the interest of a set of two-way crosses regarding the performance and the diversity in the best fraction of the progeny of each cross.

Prediction of the mean expected breeding value and parental contributions in the selected fraction of progeny

Considering two inbred lines P_1 and P_2 and the cross P_1 x P_2 and $(\boldsymbol{x}_1, \boldsymbol{x}_2)'$ denotes their $(2 \text{ x } m)$-dimensional genotyping matrix at the $m = 2{,}000$ SNP markers. $\boldsymbol{x_p}$ denotes the $(m \text{ x } 1)$-dimensional genotype vector of parent $P_{p\in\{1,2\}}$ with the j^{th} element coded as 1 or −1 for the genotypes AA or aa at QTL j. Following Lehermeier *et al.* (2017), the DH progeny mean and progeny variance of the breeding values in the progeny before selection can be computed as:

$$\hat{\mu}_T = 0.5\ (\boldsymbol{x'}_1\widehat{\boldsymbol{\beta}} + \boldsymbol{x'}_2\widehat{\boldsymbol{\beta}}), \text{ (Eq. 1a)}$$

$$\hat{\sigma}_T^2 = \widehat{\boldsymbol{\beta}}'\ \boldsymbol{\Sigma}\ \widehat{\boldsymbol{\beta}}, \text{ (Eq. 1b)}$$

where $\widehat{\boldsymbol{\beta}}$ is $(m \text{ x } 1)$-dimensional vector of estimated marker effects and $\boldsymbol{\Sigma}$ is the $(m \text{ x } m)$-dimensional variance covariance matrix of marker genotypes in DH progeny defined in Lehermeier *et al.* (2017). We define the $(m \text{ x } 1)$-dimensional vector $\boldsymbol{\beta}_{C1}$ to follow P_1 genome contribution to progeny as $\boldsymbol{\beta}_{C1} = \frac{\boldsymbol{x}_1 - \boldsymbol{x}_2}{(\boldsymbol{x}_1 - \boldsymbol{x}_2)'(\boldsymbol{x}_1 - \boldsymbol{x}_2)}$. The mean and variance of P_1 contribution in the progeny before selection are computed as:

$$\mu_{C1} = 0.5\ (\boldsymbol{x'}_1\boldsymbol{\beta}_{C1} + \boldsymbol{x'}_2\boldsymbol{\beta}_{C1} + 1), \text{ (Eq. 2a)}$$

$$\sigma_{C1}^2 = \boldsymbol{\beta}'_{C1}\ \boldsymbol{\Sigma}\ \boldsymbol{\beta}_{C1}. \text{ (Eq. 2b)}$$

The progeny mean for P_2 contribution is then $\mu_{C2} = 1 - \mu_{C1}$.
Following Allier *et al.* (2019b), the covariance between the breeding values and P_1 contribution in progeny is:

$$\hat{\sigma}_{T,C1} = \widehat{\boldsymbol{\beta}}'\ \boldsymbol{\Sigma}\ \boldsymbol{\beta}_{C1}. \text{ (Eq. 3)}$$

The expected mean breeding value of the selected fraction of progeny, i.e. usefulness criterion (Schnell and Utz 1975), of the cross P_1 x P_2 is:

$$\widehat{UC}^{(i,h)} = \hat{\mu}_T + ih\hat{\sigma}_T, \text{ (Eq. 4)}$$

where i is the within family selection intensity and h the within family selection accuracy. The correlated responses to selection on P_1 and P_2 contributions to the selected fraction of progeny are:

$$\hat{c}_1^{(i,h)} = \mu_{C1} + ih\frac{\hat{\sigma}_{T,C1}}{\hat{\sigma}_T} \text{ and } \hat{c}_2^{(i,h)} = 1 - \hat{c}_1^{(i,h)}. \text{ (Eq. 5)}$$

Optimal cross selection accounting for within family variance

Considering N homozygote candidate parents, $N(N-1)/2$ two-way crosses are possible. We define a crossing plan $\boldsymbol{nc}$ as a set of $|nc|$ crosses out of possible two-way crosses, giving the index of selected crosses, i.e. with the i^{th} element $nc\,(i) \in [1, N(N-1)/2]$. The ($N$ x 1)-dimensional vector of candidate parents estimated contributions in the selected fraction of progeny of each cross $\hat{\boldsymbol{c}}^{(i,h)}$ is:

$$\hat{\boldsymbol{c}}^{(i,h)} = \frac{1}{|nc|}\left(\boldsymbol{Z_1}\hat{\boldsymbol{c}}_{\boldsymbol{1}}^{(i,h)} + \boldsymbol{Z_2}\hat{\boldsymbol{c}}_{\boldsymbol{2}}^{(i,h)}\right), \text{(Eq. 6)}$$

where $\boldsymbol{Z_1}$ (respectively $\boldsymbol{Z_2}$) is a (N x $|nc|$)-dimensional design matrix that links each N candidate parent to the first (respectively second) parent in the set of crosses $\boldsymbol{nc}$, $\hat{\boldsymbol{c}}_{\boldsymbol{1}}^{(i,h)}$ (respectively $\hat{\boldsymbol{c}}_{\boldsymbol{2}}^{(i,h)}$) is a ($|nc|$ x 1)-dimensional vector containing the estimated contributions of the first (respectively second) parent to the selected fraction of the progeny of the crosses in $\boldsymbol{nc}$.
The expected performance $V(\boldsymbol{nc})$ for this set of two-way crosses is defined as the expected mean performance of the selected DH progeny, i.e. usefulness criterion:

$$\hat{V}^{(i,h)}(\boldsymbol{nc}) = \frac{1}{|\boldsymbol{nc}|}\sum_{j\in\boldsymbol{nc}} \widehat{\boldsymbol{UC}}^{(i,h)}(j). \text{(Eq. 7)}$$

The constraint on diversity $\widehat{D}^{(i,h)}(\boldsymbol{nc})$ in the selected progeny is:

$$\widehat{D}^{(i,h)}(\boldsymbol{nc}) = 1 - \hat{\boldsymbol{c}}^{(i,h)'}\boldsymbol{K}\,\hat{\boldsymbol{c}}^{(i,h)}, \text{(Eq. 8)}$$

where $\boldsymbol{K}$ is the (N x N)-dimensional identity by state (IBS) coancestry matrix at markers between the N candidates. Allier *et al.* (2019c) showed that $\widehat{D}^{(i,h)}(\boldsymbol{nc})$ is a good proxy of the genomewide diversity in the selected fraction of progeny $He^{(i,h)} = \frac{1}{m}\sum_{j=1}^{m} 2p_j^{(i,h)}(1-p_j^{(i,h)})$ where $p_j^{(i,h)}$ is the frequency of the genotypes AA at marker j in the selected fraction of progeny.

Cited literature

Akdemir D., and J. I. Isidro-Sánchez, 2016 Efficient Breeding by Genomic Mating. Front. Genet. 7: 210.

Akdemir D., W. Beavis, R. Fritsche-Neto, A. K. Singh, and J. Isidro-Sánchez, 2019 Multi-objective optimized genomic breeding strategies for sustainable food improvement. Heredity 122: 672.

Allier A., S. Teyssèdre, C. Lehermeier, B. Claustres, S. Maltese, et al., 2019a Assessment of breeding programs sustainability: application of phenotypic and genomic indicators to a North European grain maize program. Theor. Appl. Genet. 132: 1321–1334.

Allier A., L. Moreau, A. Charcosset, S. Teyssèdre, and C. Lehermeier, 2019b Usefulness Criterion and Post-selection Parental Contributions in Multi-parental Crosses: Application to Polygenic Trait Introgression. G3 Genes Genomes Genet. 9: 1469–1479.

Allier A., C. Lehermeier, A. Charcosset, L. Moreau, and S. Teyssèdre, 2019c Improving Short- and Long-Term Genetic Gain by Accounting for Within-Family Variance in Optimal Cross-Selection. Front. Genet. 10.

Allier A., S. Teyssèdre, C. Lehermeier, A. Charcosset, and L. Moreau, 2020 Genomic prediction with a maize collaborative panel: identification of genetic resources to enrich elite breeding programs. Theor. Appl. Genet. 133: 201–215.

Anderson E., 1944 The Sources of Effective Germ-Plasm in Hybrid Maize. Ann. Mo. Bot. Gard. 31: 355–361.

Bailey-Serres J., T. Fukao, P. Ronald, A. Ismail, S. Heuer, et al., 2010 Submergence Tolerant Rice: SUB1's Journey from Landrace to Modern Cultivar. Rice 3: 138–147.

Böhm J., W. Schipprack, V. Mirdita, H. F. Utz, and A. E. Melchinger, 2014 Breeding Potential of European Flint Maize Landraces Evaluated by their Testcross Performance. Crop Sci. 54: 1665–1672.

Böhm J., W. Schipprack, H. F. Utz, and A. E. Melchinger, 2017 Tapping the genetic diversity of landraces in allogamous crops with doubled haploid lines: a case study from European flint maize. TAG Theor. Appl. Genet. Theor. Angew. Genet. 130: 861–873.

Charmet G., N. Robert, M. R. Perretant, G. Gay, P. Sourdille, et al., 1999 Marker-assisted recurrent selection for cumulating additive and interactive QTLs in recombinant inbred lines. Theor. Appl. Genet. 99: 1143–1148.

Cooper H. D., C. Spillane, and T. Hodgkin, 2001 Broadening the Genetic Base of Crop Production. H.D. Cooper, C. Spillane and T. Hodgkin.

Cowling W. A., L. Li, K. H. M. Siddique, M. Henryon, P. Berg, et al., 2017 Evolving gene banks: improving diverse populations of crop and exotic germplasm with optimal contribution selection. J. Exp. Bot. 68: 1927–1939.

Cramer M. M., and L. W. Kannenberg, 1992 Five Years of HOPE: The Hierarchical Open-Ended Corn Breeding System. Crop Sci. 32: 1163–1171.

Crossa J., D. Jarquín, J. Franco, P. Pérez-Rodríguez, J. Burgueño, et al., 2016 Genomic Prediction of Gene Bank Wheat Landraces. G3 Genes Genomes Genet. 6: 1819–1834.

Dutfield G., 2011 The role of the international Union for the Protection of New Varieties of Plants (UPOV). Intellect. Prop. Issue Pap. 9

Duvick D. N., 2005 The Contribution of Breeding to Yield Advances in Maize (Zea mays L.). N Sparks Ed Adv Agron Acad. Press San Diego CA Vol. 86.: 83–145.

Eynard S. E., J. J. Windig, I. Hulsegge, S.-J. Hiemstra, and M. P. L. Calus, 2018 The impact of using old germplasm on genetic merit and diversity—A cattle breed case study. J. Anim. Breed. Genet. 135: 311–322.

Falconer D. S., and T. F. C. Mackay, 1996 Introduction to Quantitative Genetics. 4th ed. Pearson, Harlow, England.

Forni S., I. Aguilar, and I. Misztal, 2011 Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet. Sel. Evol. 43: 1.

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.

Giraud H., C. Lehermeier, E. Bauer, M. Falque, V. Segura, et al., 2014 Linkage Disequilibrium with Linkage Analysis of Multiline Crosses Reveals Different Multiallelic QTL for Hybrid Performance in the Flint and Dent Heterotic Groups of Maize. Genetics 198: 1717–1734.

Glaszmann J., B. Kilian, H. Upadhyaya, and R. Varshney, 2010 Accessing genetic diversity for crop improvement. Curr. Opin. Plant Biol. 13: 167–173.

Godfray H. C. J., J. R. Beddington, I. R. Crute, L. Haddad, D. Lawrence, et al., 2010 Food Security: The Challenge of Feeding 9 Billion People. Science 327: 812–818.

Gorjanc G., J. Jenko, S. J. Hearne, and J. M. Hickey, 2016 Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. BMC Genomics 17: 30.

Gorjanc G., R. C. Gaynor, and J. M. Hickey, 2018 Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. Theor. Appl. Genet. 131: 1953–1966.

Habier D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller, 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42: 5.

Haldane J., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. J Genet 8: 299–309.

Han Y., J. N. Cameron, L. Wang, and W. D. Beavis, 2017 The Predicted Cross Value for Genetic Introgression of Multiple Alleles. Genetics 205: 1409–1423.

Hellin J., M. R. Bellon, and S. J. Hearne, 2014 Maize Landraces and Adaptation to Climate Change in Mexico. J. Crop Improv. 28: 484–501.

Heslot N., J.-L. Jannink, and M. E. Sorrells, 2015 Perspectives for Genomic Selection Applications and Research in Plants. Crop Sci. 55: 1–12.

Hickey J., S. Dreisigacker, J. Crossa, S. Hearne, R. Babu, et al., 2014 Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci. 54.

Kannenberg L. W., and D. E. Falk, 1995 Models for activation of plant genetic resources for crop breeding programs. Can. J. Plant Sci. 75: 45–53.

Kannenberg L. W., 2001 HOPE, a Hierrarchical, Open-ended System for Broadening the Breeding Base of Maize, pp. 311–318 in Broadening the Genetic Base of Crop Production, H.D. Cooper, C. Spillane and T. Hodgkin.

Kinghorn B. P., R. Banks, C. Gondro, V. D. Kremer, S. A. Meszaros, et al., 2009 Strategies to Exploit Genetic Variation While Maintaining Diversity, pp. 191–200 in Adaptation and Fitness in Animal Populations, Springer, Dordrecht.

Kinghorn B. P., 2011 An algorithm for efficient constrained mate selection. Genet. Sel. Evol. 43: 4.

Legarra A., I. Aguilar, and I. Misztal, 2009 A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92: 4656–4663.

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.

Longin C. F. H., and J. C. Reif, 2014 Redesigning the exploitation of wheat genetic resources. Trends Plant Sci. 19: 631–636.

Mangin B., R. Rincent, C.-E. Rabier, L. Moreau, and E. Goudemand-Dugue, 2019 Training set optimization of genomic prediction by means of EthAcc. PLOS ONE 14: e0205629.

McCouch S., G. J. Baute, J. Bradeen, P. Bramel, P. K. Bretting, et al., 2013 Agriculture: Feeding the future. Nature 499: 23–24.

Melchinger A. E., P. Schopp, D. Müller, T. A. Schrag, E. Bauer, et al., 2017 Safeguarding Our Genetic Resources with Libraries of Doubled-Haploid Lines. Genetics 206: 1611–1619.

Meuwissen T. H., 1997 Maximizing the response of selection with a predefined rate of inbreeding. J. Anim. Sci. 75: 934–940.

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.

Mikel M. A., and J. W. Dudley, 2006 Evolution of North American Dent Corn from Public to Proprietary Germplasm. Crop Sci. 46: 1193–1205.

Misztal I., 2008 Reliable computing in estimation of variance components. J. Anim. Breed. Genet. 125: 363–370.

Pollak L. M. (Ed.), 1990 Evaluation of caribbean maize accessions in puerto rico, in Caribbean Food Crops Society 26th Annual Meeting, Mayaguez, Puerto Rico.

Pollak L. M., and W. Salhuana, 2001 The Germplasm Enhancement of Maize (GEM) Project: Private and Public Sector Collaboration, pp. 319–329 in Broadening the Genetic base of Crop Production, H.D. Cooper, C. Spillane and T. Hodgkin.

Pollak L. M., 2003 The history and success of the public-private project on germplasm enhancement of maize (GEM). Advances in agronomy 78: 46–89.

Popi J., 1997 A critical evaluation of the HOPE breeding system as a means for broadening the deployed germplasm base in maize.

Pszczola M., T. Strabel, H. A. Mulder, and M. P. L. Calus, 2012 Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95: 389–400.

Reif J. C., P. Zhang, S. Dreisigacker, M. L. Warburton, M. V. Ginkel, et al., 2005 Wheat genetic diversity trends during domestication and breeding. Theor. Appl. Genet. 110: 859–864.

Ribaut J.-M., and M. Ragot, 2006 Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations, and alternatives. J. Exp. Bot. 58: 351–360.

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.

Rio S., T. Mary-Huard, L. Moreau, and A. Charcosset, 2019 Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. Theor. Appl. Genet. 132: 81–96.

Salhuana W., R. Sevilla, and S. A. Eberhart, 1997 Latin american maize project (LAMP) final report. Pioneer Hi-Bred International Spec. Publ.

Salhuana W., and L. Pollak, 2006 Latin American Maize Project (LAMP) and Germplasm Enhancement of Maize (GEM) project: generating useful breeding germplasm. Maydica 51: 339–355.

Schnell F., and H. Utz, 1975 F1-Leistung und Elternwahl in der Züchtung von Selbstbefruchtern., pp. 243–248 in Bericht über die Arbeitstagung der Vereinigung österreichischer Pflanzenzüchter., BAL Gumpenstein, Austria.

Servin B., O. C. Martin, M. Mézard, and F. Hospital, 2004 Toward a Theory of Marker-Assisted Gene Pyramiding. Genetics 168: 513–523.

Simmonds N. W., 1962 Variability in Crop Plants, Its Use and Conservation. Biol. Rev. 37: 422–465.

Simmonds N. W., 1979 Principles of crop improvement. Longman, London.

Simmonds N. W., 1993 Introgression and Incorporation. Strategies for the Use of Crop Genetic Resources. Biol. Rev. 68: 539–562.

Smith S., and W. Beavis, 1996 Molecular Marker Assisted Breeding in a Company Environment, pp. 259–272 in The Impact of Plant Molecular Genetics, edited by Sobral B. W. S. Birkhäuser Boston, Boston, MA.

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.

Storn R., and K. Price, 1997 Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. J. Glob. Optim. 11: 341–359.

Strigens A., W. Schipprack, J. C. Reif, and A. E. Melchinger, 2013 Unlocking the Genetic Diversity of Maize Landraces with Doubled Haploids Opens New Avenues for Breeding. PLOS ONE 8: e57234.

Tadesse W., M. Sanchez-Garcia, S. G. Assefa, A. Amri, Z. Bishaw, et al., 2019 Genetic Gains in Wheat Breeding and Its Role in Feeding the World. Crop Breed. Genet. Genomics 1.

Tallury S. P., and M. M. Goodman, 2001 The State of the Use of Maize Genetic Diversity in the USA and Sub-Saharan Africa, pp. 159–179 in Broadening the Genetic Base of Crop Production, H.D. Cooper, C. Spillane and T. Hodgkin.

Troyer A. F., 1999 Background of U.S. Hybrid Corn. Crop Sci. 39: 601–626.

Van Inghelandt D., J. C. Reif, B. S. Dhillon, P. Flament, and A. E. Melchinger, 2011 Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm. Theor. Appl. Genet. 123: 11–20.

Voss-Fels K. P., M. Cooper, and B. J. Hayes, 2019 Accelerating crop genetic gains with genomic selection. Theor. Appl. Genet. 132: 669–686.

Wang H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir, 2012 Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. 94: 73–83.

Wang C., S. Hu, C. Gardner, and T. Lübberstedt, 2017 Emerging Avenues for Utilization of Exotic Germplasm. Trends Plant Sci. 22: 624–637.

Whittaker J. C., R. Thompson, and M. C. Denham, 2000 Marker-assisted selection using ridge regression. Genet. Res. 75: 249–252.

Woolliams J. A., P. Berg, B. S. Dagnachew, and T. H. E. Meuwissen, 2015 Genetic contributions and their optimization. J. Anim. Breed. Genet. 132: 89–99.

Wray N., and M. Goddard, 1994 Increasing long-term response to selection. Genet. Sel. Evol. 26: 431.

Wright S., 1978 Evolution and the genetics of populations. Volume 4: variability within and among natural populations. University of Chicago press.

Yu X., X. Li, T. Guo, C. Zhu, Y. Wu, et al., 2016 Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat. Plants 2: 16150.

General discussion and perspectives

There is an increasing awareness that plant breeding programs have to move from short-term to long-term perspectives in order to cope with future challenges. The advent of high density genotyping has opened new perspectives for breeding quantitative traits including genetic diversity assessment, genomic variance partitioning and genomic prediction of the genetic merit of individuals and parental crosses. The main objectives of this thesis were to develop indicators to assess plant breeding programs past efficiency and sustainability, and to develop strategies that balance the need for short-term genetic gain with that of maintaining and introducing diversity to enable long-term response to selection. In the next section the five chapters of this thesis are discussed and put into perspectives. For the sake of continuity of the general discussion, some chapters have been merged regardless of the chronology of publications. In the last section some perspectives for crops diversity management are discussed.

Contributions to diversity management

Diagnosis of breeding programs

Quantitative genetics theory provides breeders with the factors influencing short- and long-term breeding success. In **chapter 1** (Allier *et al.* 2019a), we proposed indicators based on quantitative genetics theory to quantify past breeding program efficiency and to forecast its near future evolution assuming past tendencies persist. These indicators are easily implemented and take advantage of the increasing amount of phenotyping and genotyping information available in most crop breeding programs that use genomic selection (Heslot *et al.* 2015; Voss-Fels *et al.* 2019). Phenotypic data can be used to estimate realized genetic gain and additive genetic variance evolution over breeding generations. The additive genetic variance trend enables to project the future response to selection on targeted traits based on response to selection theory models (Lush 1937; Robertson 1960). Complementarily, genotypic data inform about the genetic diversity without *a priori* on the trait(s) considered, i.e. the "neutral" diversity, which is of future importance to address yet unknown breeding targets raising in a context of societal and climatic changes (McCouch *et al.* 2013). In the illustrative hybrid maize breeding program considered, both breeding populations showed a significant positive genetic gain but contrasted evolutions of genetic variance and "neutral" genetic diversity, reflecting a complex open breeding system. In particular, we found in the Dent pool some large genomic regions with a very low diversity. As observed in Gerke *et al.* (2015), these regions were mainly located in low recombining pericentromeric regions. The different nonexclusive forces that can lead such hitchhiking were discussed in **chapter 1**, including founder effect, genetic drift and selection of favorable haplotypes. These regions raise several concerns: Do we really need to increase allelic diversity and/or recombination in these regions? As suggested in Gerke *et al.* (2015) the fixation of these regions may be important for group complementarity. This requires further investigations but the large size of low diversity and low recombining regions likely suggests that they may be composed of both favorable and unfavorable segments fixed by linkage drag. In order to test this assumption, one could think of haplotypic visualization approaches developed in **chapter 2** (Allier *et al.* 2020) using marker effects estimated on a broad panel where these regions are segregating.

Beyond estimating separately additive genetic variance and "neutral" genetic diversity, genomic regression models enable the estimation of the components of the additive genetic variance genomewide and per chromosome. The additive genetic variance can indeed be decomposed into the additive genic variance that corresponds to the sum of the additive variance at individual QTLs under the assumption of linkage equilibrium between QTLs and the covariance between QTLs (Bulmer 1971;

Lynch and Walsh 1998; Gianola *et al.* 2009; Lehermeier *et al.* 2017a). As expected under directional selection, negative covariances were observed in all chromosomes and captured between 17-63% of the additive genic variance. Based on the proportion of additive genic variance hidden by repulsion and the genic variance for each chromosome, we proposed to draw fine scale strategies to manage and increase the potential response to selection per chromosome. However, strategies to increase genic variance or unleash variance by recombination in specific chromosomes are far from evident. One solution would involve the selection of breeding crosses accounting for parental complementarity at markers linked to QTLs in these specific regions. As illustrated by simulations in **chapter 4** (Allier *et al.* 2019c), optimal cross selection (OCS) favors effective recombination events and unleashes parts of the hidden additive genic variance into additive genetic variance. Alternatively, modern plant breeding biotechnologies offer new opportunities to modify targeted loci and change recombination landscape and/or increase recombination, as it will be discussed in the last section.

We considered a private early maize breeding program as an application case but the advances in genotyping in most crops and animal species offer the opportunity to extend the use of global indicators to different breeding programs and species. Proposed indicators can be improved in different ways. For instance, we did not consider pedigree information in the analysis of genetic gain and additive genetic variance but, if of sufficient depth and quality, pedigree might be accounted for to better model the additive genetic component. We also considered a maize genotyping array of 50k SNPs (Ganal *et al.* 2011) as genotyping arrays are common routine genotyping technologies used in breeding companies (e.g. Van Inghelandt *et al.* 2010). However, such genotyping technology focuses on common variants only, which limits genetic diversity evaluation and management. This is referred to as the ascertainment bias caused by the SNP discovery process in which a small number of individuals are used in the discovery panel and by the selection of SNP with equilibrated frequencies (Albrechtsen *et al.* 2010). Alternatively, genotyping by sequencing (GBS, Elshire *et al.* 2011) that discovers and genotypes both common and rare variants, provides a robust diversity estimate with much reduced ascertainment bias (Heslot *et al.* 2013). For instance, Eynard *et al.* (2016) highlighted the interest of using common and rare SNP variants for genetic diversity quantification. The authors observed that whole-genome sequence revealed considerable losses of genetic diversity for rare variants that were unperceivable considering 50k SNP bead chip in cattle. GBS is also highly relevant for curating, identifying and harnessing variability in gene banks (Kilian and Graner 2012; Sehgal *et al.* 2015; Yu *et al.* 2016). Finally, sequencing technologies would enable to identify structural variations such as presence/absence and copy number variation (Springer *et al.* 2009; Alkan *et al.* 2011) that represent diversity untapped by SNP bead chips. In maize, GBS approaches are based on cost effective low depth sequencing of individuals (<1X genome coverage) and generate numerous missing data that need to be further imputed (e.g. up to 80% of missing data accurately imputed in Torkamaneh and Belzile 2015), which raises issues on the accuracy of imputation. One can expect that rapid progress in sequencing and the availability of large sequence database will alleviate this limitation for most crops.

We therefore believe that in practice such indicators of the genetic variances and diversity should be considered in routine in breeding programs to ensure the consistency between breeding long-term strategy and the breeding population. For instance, a joint reduction of the additive genetic variance and genetic diversity over time should indicate that a better management of the intrinsic diversity and introductions of extrinsic diversity is required. Waiting for a slowdown in genetic gain would be risky for genetic base broadening that usually takes several years or decades to be efficient. Alternatively,

sufficient additive genetic variance and genetic diversity stable over time would suggest that an optimization of the intrinsic diversity management is sufficient. In the following, we discuss the **chapters 2**, **3**, **4** and **5** considering first that intrinsic genetic diversity is sufficient and then that genetic base broadening is needed.

Optimization of mating design

First, let us assume the indicators proposed in **chapter 1** suggest that the genetic diversity is not limiting regarding the short- and long-term breeding objectives. In this context, the main breeder's objective is to efficiently convert the intrinsic diversity into long-term genetic gain while not compromising the variety performance at short-term. As suggested in Bernardo (2003) and Lado *et al.* (2017), cross selection is one of the most important decision in breeding. The ideal mating plan being the crosses that provide superior progeny performance and enough diversity to maintain genetic gain. Consequently, a shift should operate from the paradigm of recycling and crossing super elite lines together to the recognition of the interest of less performing but more complementary parents that will generate a longer term genetic variation.

Different predictive tools have been proposed to support crop breeders with the implementation of their mating design. First, the optimal cross selection (OCS) has proven to be efficient to convert genetic diversity into long-term genetic gain (e.g. Akdemir and Isidro-Sánchez 2016; De Beukelaer *et al.* 2017; Gorjanc *et al.* 2018). When constraining on the genomic relationship matrix, OCS accounts indirectly for parental complementarity at neutral markers assuming independence of the loci and tends implicitly to favor crosses with higher Mendelian segregation variance. The usefulness criterion (UC) of a cross explicitly accounts for Mendelian segregation variance specific to the targeted trait(s). The concept of UC is quite ancient (Schnell and Utz 1975) but has long suffered the absence of accurate predictors of within cross variance. With recent advances in this domain the UC is more and more implemented in crops (Lehermeier *et al.* 2017b), and was also found as being of interest in animal breeding (Segelke *et al.* 2014; Bonk *et al.* 2016; Bijma *et al.* 2018). In **chapter 3** (Allier *et al.* 2019b), we proposed to consider a multivariate UC that predicts the expected performance in the best fraction of progeny and the parental contributions (PC) to the best fraction of progeny, namely the UCPC. We also extended the algebraic formulas for multi-parental crosses implying up to four parents, i.e. biparental crosses between heterozygote phased individuals which enables considering three-way or four-way crosses that are frequent in annual plants but also outbred animal or perennial plants. In **chapter 4** (Allier *et al.* 2019c), we then proposed the UCPC based OCS that differs from OCS in the sense that the parental complementarity for the traits considered is explicitly accounted for with consideration of linkage map and linkage disequilibrium. Furthermore, the next generation diversity at the whole genome level, which is derived from parental contributions, is optimized while anticipating the effect of within family selection. Simulations in **chapter 4** highlighted the importance to balance short-term performance and genetic diversity using OCS methods to more efficiently convert genetic diversity into genetic gain and maximize long-term performance. Constraining on diversity had a cost for short-term variety release compared to UC that might dishearten commercial breeders. Considering explicitly within family variance and selection in UCPC based OCS limited this penalty at short-term and yielded higher long-term performance. This involves crossing complementary parents to favor effective recombination events between complementary parental haplotypes. As a result, the recombination unleashes parts of the additive genic variance captured by the build-up of negative covariances observed in **chapter 1** (Bulmer 1971; Rasmusson and Phillips 1997; Bijma *et al.* 2018).

In practice, UCPC based OCS can be implemented in routine breeding programs to help breeders with cross selection regarding their short- and long-term objectives. UCPC based OCS requires parental genotype information, a genetic map, estimated marker effects and an optimization algorithm. We considered common SNP variants as markers but GBS data can be used to compute genomic relationship matrix between parents (Eynard *et al.* 2015, 2016). For instance, Eynard *et al.* (2016) observed that considering common and rare variants to estimate genomic relationship matrix in optimal contribution selection slightly reduced the loss of rare variants, while using 50k SNP bead chip data was sufficient to conserve common variants.

In **chapter 4**, since we aimed at comparing different crossing strategies, we considered a simplistic linear trajectory of diversity over generations and a fixed selection intensity within each family. More complex strategies can be applied but were not tested. The parametrization of the UCPC based OCS strategy regarding short- and long-term objectives (e.g. the constraint on diversity, within family selection intensity) is complex and requires quantified breeding objectives (e.g. targeted diversity, targeted annual genetic gain). The optimal parametrization of such an approach could be done using simulations based on breeding germplasm genotypes and assuming estimated marker effects as true QTL effects, i.e. assuming reliable estimates and neglecting the fact that estimated marker effects are allele frequency dependent.

We evaluated the interest of UCPC based OCS in an inbred plant breeding program and discussed its extension to crosses between heterozygote individuals. This is interesting for animal breeders and plant breeders working with heterozygous individuals (e.g. in perennial species). It also extends the use of UCPC based OCS to the two-part GS breeding program proposed by Gaynor *et al.* (2017) and Hickey *et al.* (2017). The authors proposed to distinguish the population improvement component to develop improved germplasm and the product development component to fix and identify new inbred parents for hybrids. In the population improvement component, the most performing progeny of parental crosses are selected and recycled before fixation to generate the next population improvement generation. In this context, Gorjanc *et al.* (2018) observed that OCS enabled optimal management and exploitation of population improvement germplasm and we can conjecture an additional gain to use UCPC based OCS.

Furthermore, as heterozygous individuals are conceptually crosses between two phased parental gametes, UCPC can be useful to select individuals accounting for the Mendelian segregation in their gametes (Segelke *et al.* 2014; Bonk *et al.* 2016; Bijma *et al.* 2018). For instance some individuals produce more variable progeny than others regardless of the second parent, i.e. more likely outstanding progenies of agricultural interest. In a breeding perspective, one may want to select individuals maximizing an index between their GEBV and expected gametic variances (Bijma *et al.* 2018). On the contrary, in a farmer perspective, one may select for high individual GEBV but low gametic variance to have more homogenous progeny (e.g. pig birth-weight) and simplify herd management (Cole and VanRaden 2011; Segelke *et al.* 2014). Such a balance between breeding and production objectives should also be considered for open pollinated plant species were the breeding population is also the production population (e.g. participatory breeding of maize landraces in developing countries, Bellon *et al.* 2003).

Genetic base broadening

Let us assume now, that the indicators proposed in **chapter 1** suggest that the genetic diversity is suboptimal regarding the long-term breeding objectives. A first step would consist in characterizing and identifying genetic resources for genetic base broadening using multi-environment trials. In **chapter 2** (Allier *et al.* 2020)**,** we reviewed, proposed and compared different criteria to identify genetic resources that can complement an elite population and that can compensate their low mean performance by an increased genetic variance when crossed to elites (Longin and Reif 2014). The different criteria account differently for parental complementarity at individual loci or haplotype segments. Criteria were parameterized to consider more or less recombination events and consequently evaluate the interest of genetic resources at more or less long-term when crossed to elites. Hence, the optimum parametrization cannot be provided and depends on the breeder's objectives. We observed that a genomewide prediction model trained on a collaborative panel including old material and elite material (Amaizing dent collaborative panel, Rio *et al.* 2019) had a relevant predictive ability on a large elite private material. This suggests that genomic predictions calibrated on such a collaborative panel can be used to identify interesting sources of diversity in the panel. This strategy might be extended to other collaborative diversity panels, libraries of DH lines derived from landraces (Strigens *et al.* 2013; Melchinger *et al.* 2017; Böhm *et al.* 2017; Hölker *et al.* 2019) or gene banks in other species to evaluate non phenotyped genetic resources as proposed in Yu *et al.* (2016) and Crossa *et al.* (2016). Methodological developments in **chapter 3** could enrich the proposal made in **chapter 2** in complementary ways. First, it would allow considering multi-parental crosses between genetic resources and elites which appeared to be of interest in case of low performing genetic resources (Allier *et al.* 2019b). UCPC would also make it possible to evaluate the genetic resources that balance performance and originality (as implemented in **chapter 5** in case of two way crosses). Finally, UCPC enables consideration of parental contributions in specific regions under the assumption that a sufficient number of loci are independently segregating in these regions to ensure the normality of the trait. Thus, UCPC could be used to identify donors that enrich specific regions in diversity, such as regions identified in **chapter 1**.

Finally in **chapter 5**, we evaluated strategies inspired from Simmonds (1993) for recurrent introductions in a simulated commercial breeding program. We considered different types of donors with variable performance gap with elites and compared two introduction strategies: direct introductions or indirect introductions in the breeding population. The latter involves a buffer population, namely bridging population, which bridges the most complementary genetic resources and elites before introduction in the breeding population. We considered the UCPC based OCS to manage recurrent genetic base broadening. In this context, an OCS holistic approach, where bridging crosses, introduction crosses and elite crosses are jointly optimized, ensures an overall consistency of the genetic base broadening strategy. We considered the UCPC based OCS to maximize genetic performance while maintaining genetic variation constant thanks to the intrinsic variability and introductions of extrinsic variability. Simulation demonstrated that recurrent introductions of pre-improved genetic resources (i.e. through pre-breeding or a minima bridging) can increase the genetic mid- and long-term genetic gain while maintaining genomewide genetic diversity constant. The less performant the introduced material, the more important was the short-term penalization of variety release. We also suggest to consider marker effects estimated on a large and broad TS that blends elites and progeny of elite by genetic resource crosses in order to balance the prediction accuracy in elite crosses and in introduction crosses.

Further investigations might be considered to complete this work. For instance, we discussed in **chapter 2** the practical interest of public-private collaborative pre-breeding projects. However, further investigations are required to identify key parameters of successful public-private diversity panels, including the origin of genetic resources, their improvement and the relative proportion of elite proprietary germplasm. Furthermore, simulations can also be performed to validate the interest of criteria proposed in **chapter 2** and evaluate the sensitivity of criteria accuracy to different training set compositions. In **chapter 5**, we simulated a breeding program with a reduced genetic diversity at the end of burn-in to be in the situation where genetic base broadening is required. In practice, the need of broadening genetic diversity might be variable depending on the adequacy of intrinsic diversity diagnosis and short- and long-term breeding objectives. We also assumed absence of mutations, epistasis and a single-trait breeding target that was constant during sixty years. Mutations and epistasis might reduce the importance of genetic base broadening by releasing additive genetic variance over generations as discussed in the next section. In a context of climatic and social expectation changes the breeding target is likely multi-trait and changing over time. Coupling different climatic scenarios with the simulation of a breeding program with a multi-trait target could be interesting to evaluate the interest of genetic base broadening in a more complex context. We believe that the need for genetic base broadening is likely more valuable than highlighted in **chapter 5** to be able to address yet unknown breeding objectives.

Altogether, this study supports breeders with tools to evaluate, manage and reveal intrinsic genetic variation, to identify and introduce extrinsic variation and efficiently convert genetic variation into genetic gain. Such quantitative genetics tools, among others, will support breeders toward integrated and sustainable breeding programs. In the next section we will discuss the importance of mutation and epistasis in open breeding populations. Then, we discuss the use of biotechnologies to fasten genetic base broadening in crops.

Perspectives

Is continued crop improvement sustainable?

In long-term simulations of **chapter 4**, nearly all the additive genetic variance was eroded and genetic merit plateaus were reached in most scenarios after sixty years. Other long-term simulation studies in plants also reached similar results (e.g. De Beukelaer *et al.* 2017; Gorjanc *et al.* 2018). Experimentally, Weber (2004) observed a selection plateau under directional selection in a large population of *Drosophila*. On the opposite, continued genetic gains are observed in most crops (e.g. in maize Duvick 2005, in wheat Tadesse *et al.* 2019) and the long-term Illinois divergent selection experiment for maize oil and protein content showed continuous genetic gains for hundred generations (Dudley and Lambert 2004). This raises questions about the sustainability of crop breeding but also the realism of the genetic model assumed in most long-term simulation studies. Several nonexclusive reasons may explain why continued improvement is possible in crops contrary to what simulations appear to claim.

Firstly, in simulations different approaches are compared for their efficiency to convert intrinsic variability into genetic gain for a clear breeding target trait (e.g. De Beukelaer *et al.* 2017; Gorjanc *et al.* 2018; Allier *et al.* 2019c). However, in practice commercial breeding programs are often more complex than simulated ones and extrinsic variation is used to maintain the response to selection (e.g.

Feng *et al.* 2006; Allier *et al.* 2019a; Bruce *et al.* 2019). Indeed, allowing for extrinsic variation introduction into the breeding population increased the selection limit and delayed the selection plateau in simulations (**chapter 5**). Furthermore, the breeding target is likely implying multiple traits showing different genetic (co)variances and changing over generations. Thus, selection efforts are spread over variable traits and it is less likely that breeders completely erode the additive genetic variation underlying the targeted traits in the breeding population.

Secondly a crop's genome is dynamic and new variations arise every generation while most simulated breeding programs assumed the absence of mutations. In maize the mutation rate is about 9-20 10^{-9} mutations per base pair per generation (Kremling *et al.* 2018). For a genome size of 2.4 Gb, this represents 20 to 50 mutations per generation of which most are neutral. Estimates of additional mutational variance per generation for a range of species and quantitative traits averaged on 0.1% of the environmental variance (Houle *et al.* 1996; Keightley 2004; Hill 2016). Despite mutation effects seem negligible, the multi-generation Illinois maize kernel content selection experiment (Dudley and Lambert 2004) and the long-term divergent selection experiment for flowering time in maize inbred lines (Durand *et al.* 2010, 2015) tend to nuance this *a priori*. In the Illinois experiment, lines have been selected for oil and protein content over hundred generations with variability still sufficient to achieve progress from selection, which can be explained only by mutations (Walsh 2004). In the divergent flowering time experiment, continued response to selection is observed after more than seventeen generations, which can be explained mainly by mutations following a phase of fixation of residual heterozygosity (Durand *et al.* 2010, 2015).

The third complementary explanation is that epistasis is neglected in most long-term simulation studies. Physiological epistasis arises from pleiotropies and interactions in metabolic pathways. Statistical epistasis is the statistical contribution of the interactions between loci to genetic variance and therefore depends on allelic frequencies (Lynch and Walsh 1998; Paixão and Barton 2016). As a consequence, in a finite population, additive by additive epistatic variance tends to be lost by genetic drift but is also partly converted into additive variance (Goodnight 1988), which maintains the response to selection (Barton and Turelli 2004; Carlborg *et al.* 2006; Paixão and Barton 2016; Barton 2017; Hill 2017). In the extreme regime where genetic drift drives allelic frequency changes, as it can be encountered in case of strong selection on a large number of loci in a finite population, Paixão and Barton (2016) observed that the total response to selection mostly depends on the initial standing variation. In the opposite regime, where directional selection drives the allelic changes in frequency, the authors observed that the total response to selection is greatly impacted by the conversion of epistatic to additive variance when initially neutral or deleterious alleles become favorable as the genetic background changes. Furthermore, mutations might present interactions with the genetic background (e.g. Durand *et al.* 2015) so that it is difficult to disentangle mutation from epistatic effects on long-term response to selection. In practice, which regime may correspond to breeding populations? The Iowa hybrid maize selection experiment (Gerke *et al.* 2015) showed that while most of allelic changes can be attributed to genetic drift, some regions showed signatures of selection, and seemed to indicate an intermediate regime. According to Hill (2017), the contribution of epistatic variance conversion to additive variance is likely more important than contribution of mutations in long-term experiments. However, as concluded by Hill (2017): "It is not obvious that we should be trying explicitly to exploit it by changing the focus of the selection to the epistasis itself. It seems better to concentrate on utilizing additive variance, and hope for a bonus from converting epistatic variance".

Simulations not accounting for mutations or epistasis might be pessimistic regarding the sustainability of crop improvement. Nevertheless, simulation results obtained in **chapter 4** and **chapter 5** provide important information on the optimal diversity management strategies. We can draw the following general recommendations. One should avoid too low genetic diversity in breeding populations to (i) maximize the conversion of additive genetic variance into gain and (ii) to hope for a bonus from converting epistatic variance in additive variance. One needs to evaluate frequently the additive genetic variance in the breeding population to assess if mutational and epistatic bonus are sufficient regarding long-term objectives. One needs to anticipate (e.g. participate to collaborative pre-breeding projects, routinely evaluate some available genetic resources) and introduce genetic resources to prevent a potential decrease of the additive genetic variance.

Biotechnologies for genetic base broadening

We highlighted in **chapter 1** the interest in increasing genetic diversity in specific chromosomic regions and in favoring recombination events to unleash genetic variation captured by repulsion between causal loci. Conventional solutions involve the selection of crosses between complementary parents in these regions using for instance UCPC based OCS (**chapter 3** and **chapter 4**). In **chapter 5**, we highlighted the interest of recurrent introductions of polygenic variation in breeding population on mid- and long-term genetic gain. We observed in our simulations, similarly as in **chapter 3**, that haplotypes introduced from the donor carried some original favorable alleles tightly linked with unfavorable alleles. However, recombination is often not sufficient to break this linkage and combine intrinsic and extrinsic favorable alleles in a single superior haplotype that will reach fixation. On the contrary, multiple sub-optimal haplotypes selectively interfere with one another so that none reach fixation, which is known as the Hill-Robertson interference (Felsenstein 1965; Hill and Robertson 1966). Advances in genome editing technics for adding, deleting or replacing a series of nucleotides in the genome are opening alternative perspectives to bypass the Hill-Robertson effect. In recent techniques this can be achieved using specific nuclease that cut DNA at specific predetermined places (e.g. zinc finger nuclease: ZNF, transcription activator-like effector nucleases: TALEN or clustered regulatory interspaced short palindromic repeats: CRISPR, Gaj *et al.* 2013; Belhaj *et al.* 2013).

Genome editing can be used to increase meiotic recombination rates (for a recent review, Blary and Jenczewski 2018) or induce mitotic recombination at precise locations (Sadhu *et al.* 2016). Battagin *et al.* (2016) performed simulations to explore the potential of manipulating recombination rates to increase response to selection in livestock breeding programs. The authors had to tremendously increase the genomewide recombination rate to 10-20 fold to significantly increase the response to selection. A disadvantage of increased recombination rate is the rapid decrease of linkage between QTLs and markers on which genomic selection predictive ability relies (Habier *et al.* 2013), requiring frequent updating of the genomic selection model (Battagin *et al.* 2016). Using simulations, Gonen *et al.* (2017) evaluated the interest to recombine in regions that did not recombine for several generations. The authors observed a release of additive genetic variance in the form of new allele combinations and thus an increased genetic gain. Tourrette et al. (2019) compared by simulations two different approaches to increase recombination in plants. The first approach increased the global recombination without affecting the recombination landscape and used a mutant of anti-crossover genes (developed in A. *thaliana*, Fernandes *et al.* 2018, pea, rice and tomato Mieulet *et al.* 2018). The second increased the recombination particularly in pericentromeric regions using differences of the

ploidy level between parents (developed in crosses between *Brassica rapa* and *Brassica napus*, Pelé *et al.* 2017). The authors found up to 30% of gain after twenty generations with an advantage to the recombination landscape modification.
However, all recombination events are not favorable. Recombination is advantageous if it uncouples favorable-unfavorable complexes that capture additive genetic variance but recombination is unfavorable if it breaks favorable-favorable complexes. Thus, the amount of variation that arises from induced recombination depends on the location of recombination points relative to the causal variants and gametic phase disequilibrium. Assuming that the estimated marker effects are accurate enough and that precise targeted meiotic recombination technology is available, Bernardo (2017) proposed an in silico simulation approach to identify one or two target recombination points in doubled haploid progeny of a cross between inbred maize lines to enhance the genetic gain. The author observed that one or two targeted recombination events per chromosome yielded 100 to 600% gain in response to selection compared to non-targeted recombination events. Still the feasibility and efficiency of such meiotic recombination in plants remains to be proved.

The CRISPR/Cas9 technology has also successfully been used to modify crop traits including drought tolerance in maize (Shi *et al.* 2017), sorghum, rice, wheat and soybean (Belhaj *et al.* 2013; Shalem *et al.* 2015). Beyond fastening few introgressions, genome editing is expected to fasten genetic base broadening while generating genetic diversity at multiple loci simultaneously (Ma *et al.* 2015; Sharon *et al.* 2018; Wolter *et al.* 2019). For instance, Jenko *et al.* (2015) evaluated the interest of multiple loci genome editing, referred to as promotion of alleles by genome editing (PAGE), in a simulated cattle breeding program where only sire where edited. The authors observed that PAGE had great potential in response to selection after 20 generations. However, the authors warned against the overuse of edited parents that would yield a rapid decrease in polygenic variation (Jenko *et al.* 2015).

These prospects assume reliable estimates of allelic effects to edit. It will require massive data from genotypes to phenotypes and at different integrated levels (gene expression, proteomic, etc.) to inverse the curse of dimensionality (i.e. from n <<p to n>p) (Jenko *et al.* 2015; Wallace *et al.* 2018; Ramstein *et al.* 2019). Despite promising preliminary simulation studies on the use of genome editing for quantitative trait breeding and improving accuracy of genome editing technics, there are still some unknown factors such as the approval by government agencies in Europe for food production and the acceptance by public opinion. Consequently, it is a necessity to manage and harness the “native” genetic diversity and continue to develop and optimize conventional introgression and genetic base broadening strategies.

Personal conclusion

I personally believe that sustainable and continued crop breeding for productivity and quality in a changing environment is possible and desirable. This involves an optimization of breeding strategies to ensure adequacy of the breeding germplasm with changing breeding targets. This involves a better environmental characterization and consideration of GxE in predictive breeding. This also involves the rapid and efficient conversion of intrinsic and extrinsic genetic variability into multivariate response to selection. Finally, this requires the management and usage of ex-situ genetic resources with strong public-private logistical and financial partnership to make it compelling for breeders.

Cited literature

Akdemir D., and J. I. Isidro-Sánchez, 2016 Efficient Breeding by Genomic Mating. Front. Genet. 7: 210.

Akdemir D., W. Beavis, R. Fritsche-Neto, A. K. Singh, and J. Isidro-Sánchez, 2019 Multi-objective optimized genomic breeding strategies for sustainable food improvement. Heredity 122: 672.

Albrechtsen A., F. C. Nielsen, and R. Nielsen, 2010 Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. Mol. Biol. Evol. 27: 2534–2547.

Alkan C., B. P. Coe, and E. E. Eichler, 2011 Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12: 363.

Allier A., S. Teyssèdre, C. Lehermeier, B. Claustres, S. Maltese, et al., 2019a Assessment of breeding programs sustainability: application of phenotypic and genomic indicators to a North European grain maize program. Theor. Appl. Genet. 132: 1321–1334.

Allier A., L. Moreau, A. Charcosset, S. Teyssèdre, and C. Lehermeier, 2019b Usefulness Criterion and Post-selection Parental Contributions in Multi-parental Crosses: Application to Polygenic Trait Introgression. G3 Genes Genomes Genet. 9: 1469–1479.

Allier A., C. Lehermeier, A. Charcosset, L. Moreau, and S. Teyssèdre, 2019c Improving Short- and Long-Term Genetic Gain by Accounting for Within-Family Variance in Optimal Cross-Selection. Front. Genet. 10.

Allier A., S. Teyssèdre, C. Lehermeier, A. Charcosset, and L. Moreau, 2020 Genomic prediction with a maize collaborative panel: identification of genetic resources to enrich elite breeding programs. Theor. Appl. Genet. 133: 201–215.

Anderson E., 1944 The Sources of Effective Germ-Plasm in Hybrid Maize. Ann. Mo. Bot. Gard. 31: 355–361.

Bailey-Serres J., T. Fukao, P. Ronald, A. Ismail, S. Heuer, et al., 2010 Submergence Tolerant Rice: SUB1's Journey from Landrace to Modern Cultivar. Rice 3: 138–147.

Barton N. H., and M. Turelli, 2004 Effects of Genetic Drift on Variance Components Under a General Model of Epistasis. Evolution 58: 2111–2132.

Barton N. H., 2017 How does epistasis influence the response to selection? Heredity 118: 96–109.

Battagin M., G. Gorjanc, A.-M. Faux, S. E. Johnston, and J. M. Hickey, 2016 Effect of manipulating recombination rates on response to selection in livestock breeding programs. Genet. Sel. Evol. 48: 44.

Beadle G., 1939 Teosinte and the origin of maize. J. Hered. 30: 245–247.

Beckett T. J., A. J. Morales, K. L. Koehler, and T. R. Rocheford, 2017 Genetic relatedness of previously Plant-Variety-Protected commercial maize inbreds. PloS One 12: e0189277.

Belhaj K., A. Chaparro-Garcia, S. Kamoun, and V. Nekrasov, 2013 Plant genome editing made easy: targeted mutagenesis in model and crop plants using the CRISPR/Cas system. Plant Methods 9: 39.

Bellon M. R., J. Berthaud, M. Smale, J. A. Aguirre, S. Taba, et al., 2003 Participatory landrace selection for on-farm conservation: An example from the Central Valleys of Oaxaca, Mexico. Genet. Resour. Crop Evol. 50: 401–416.

Bernardo R., 1994 Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. Crop Sci. 34: 20–25.

Bernardo R., 1996a Best Linear Unbiased Prediction of Maize Single-Cross Performance. Crop Sci. 36: 50–56.

Bernardo R., 1996b Best Linear Unbiased Prediction of the Performance of Crosses between Untested Maize Inbreds. Crop Sci. 36: 872–876.

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., 2014 Genomewide Selection of Parental Inbreds: Classes of Loci and Virtual Biparental Populations. Crop Sci. 54: 2586–2595.

Bernardo R., 2016 Genomewide Predictions for Backcrossing a Quantitative Trait from an Exotic to an Adapted Line. Crop Sci. 56: 1067–1075.

Bernardo R., 2017 Prospective Targeted Recombination and Genetic Gains for Quantitative Traits in Maize. Plant Genome 10.

Bijma P., Y. C. J. Wientjes, and M. P. L. Calus, 2018 Increasing genetic gain by selecting for higher Mendelian sampling variance. Proc. World Congr. Genet. Appl. Livest. Prod. Genetic Gain-Breeding Strategies 2: 47.

Blary A., and E. Jenczewski, 2018 Manipulation of crossover frequency and distribution for plant breeding. Theor. Appl. Genet. 132:575–592

Böhm J., W. Schipprack, H. F. Utz, and A. E. Melchinger, 2017 Tapping the genetic diversity of landraces in allogamous crops with doubled haploid lines: a case study from European flint maize. Theor. Appl. Genet. 130: 861–873.

Bonk S., M. Reichelt, F. Teuscher, D. Segelke, and N. Reinsch, 2016 Mendelian sampling covariability of marker effects and genetic values. Genet. Sel. Evol. 48: 36.

Brandenburg J.-T., T. Mary-Huard, G. Rigaill, S. J. Hearne, H. Corti, et al., 2017 Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. PLoS Genet. 13: e1006666.

Brandolini A., 1970 Maize, pp. 273–309 in Genetic resources in plants-their exploration and conservation, Frankel, O. H. Bennett, E., Oxford & Edinburgh.

Brauner P. C., D. Müller, P. Schopp, J. Böhm, E. Bauer, et al., 2018 Genomic Prediction Within and Among Doubled-Haploid Libraries from Maize Landraces. Genetics 210: 1185–1196.

Brauner P. C., W. Schipprack, H. F. Utz, E. Bauer, M. Mayer, et al., 2019 Testcross performance of doubled haploid lines from European flint maize landraces is promising for broadening the genetic base of elite germplasm. Theor. Appl. Genet. 132: 1897–1908.

Brown W. L., and E. Anderson, 1947 The Northern Flint Corns. Ann. Mo. Bot. Gard. 34: 1–29.

Brown W., 1979 Development and improvment of the germplasm base of modern maize, pp. 93–111 in Eucarpia Corn and Sorghum Section Proceedings.

Brown A. H. D., and M. T. Clegg, 1983 Isozyme assessment of plant genetic resources. Curr. Top. Biol. Med. Res. 11: 285–295.

Brown A. H. D., 1989 The case for core collections, pp. 136–156 in The use of plant genetic resources, Brown AHD, Frankel OH, Marshall DR, Williams JT, Cambridge, UK.

Bruce R. W., D. Torkamaneh, C. Grainger, F. Belzile, M. Eskandari, et al., 2019 Genome-wide genetic diversity is maintained through decades of soybean breeding in Canada. Theor. Appl. Genet. 132: 3089

Bruns H. A., 2017 Southern Corn Leaf Blight: A Story Worth Retelling. Agron. J. 109: 1218–1224.

Buckler E. S., B. S. Gaut, and M. D. McMullen, 2006 Molecular and functional diversity of maize. Curr. Opin. Plant Biol. 9: 172–176.

Bulmer M., 1971 The stability of equilibria under selection. Heredity 27: 157–162.

Camus-Kulandaivelu L., J.-B. Veyrieras, D. Madur, V. Combes, M. Fourmann, et al., 2006 Maize Adaptation to Temperate Climate: Relationship Between Population Structure and Polymorphism in the Dwarf8 Gene. Genetics 172: 2449–2463.

Carlborg Ö., L. Jacobsson, P. Åhgren, P. Siegel, and L. Andersson, 2006 Epistasis and the release of genetic variation during long-term selection. Nat. Genet. 38: 418–420.

Charmet G., N. Robert, M. R. Perretant, G. Gay, P. Sourdille, et al., 1999 Marker-assisted recurrent selection for cumulating additive and interactive QTLs in recombinant inbred lines. Theor. Appl. Genet. 99: 1143–1148.

Charmet G., 2011 Wheat domestication: Lessons for the future. C. R. Biol. 334: 212–220.

Clark S. A., B. P. Kinghorn, J. M. Hickey, and J. H. van der Werf, 2013 The effect of genomic information on optimal contribution selection in livestock breeding programs. Genet. Sel. Evol. 45: 44.

Clerc V. L., F. Bazante, C. Baril, J. Guiard, and D. Zhang, 2005 Assessing temporal changes in genetic diversity of maize varieties using microsatellite markers. Theor. Appl. Genet. 110: 294–302.

Coffman S.M., M.B. Hufford, C.M. Andorf, et al., 2020, Haplotype structure in commercial maize breeding programs in relation to key founder lines. Theor Appl Genet. early online

Cole J. B., and P. M. VanRaden, 2011 Use of haplotypes to estimate Mendelian sampling effects and selection limits. J. Anim. Breed. Genet. 128: 446–455.

Commission on Genetic Resources for Food, 2010 The second report on the state of the world's plant genetic resources for food and agriculture. Food and Agriculture Organization.

Cooper H. D., C. Spillane, and T. Hodgkin, 2001 Broadening the Genetic Base of Crop Production. H.D. Cooper, C. Spillane and T. Hodgkin.

Cramer M. M., and L. W. Kannenberg, 1992 Five Years of HOPE: The Hierarchical Open-Ended Corn Breeding System. Crop Sci. 32: 1163–1171.

Crossa J., D. Jarquín, J. Franco, P. Pérez-Rodríguez, J. Burgueño, et al., 2016 Genomic Prediction of Gene Bank Wheat Landraces. G3 Genes Genomes Genet. 6: 1819–1834.

Darwin C., 1876 The effects of cross and self fertilization in the vegetable kingdom. (Murry, London).

De Beukelaer H., Y. Badke, V. Fack, and G. De Meyer, 2017 Moving beyond managing realized genomic relationship in long-term genomic selection. Genetics 206: 1127–1138.

Doebley J. F., M. M. Goodman, and C. W. Stuber, 1986 Exceptional Genetic Divergence of Northern Flint Corn. Am. J. Bot. 73: 64–69.

Doebley J., J. D. Wendel, J. S. C. Smith, C. W. Stuber, and M. M. Goodman, 1988 The origin of cornbelt maize: The isozyme evidence. Econ. Bot. 42: 120–131.

Doebley J., 1990 Molecular Evidence and the Evolution of Maize. Econ. Bot. 44: 6–27.

Doebley J. F., B. S. Gaut, and B. D. Smith, 2006 The Molecular Genetics of Crop Domestication. Cell 127: 1309–1321.

Doyle J. J., 1988 5S ribosomal gene variation in the soybean and its progenitor. Theor. Appl. Genet. 75: 621–624.

Dubreuil P., and A. Charcosset, 1999 Relationships among maize inbred lines and populations from European and North-American origins as estimated using RFLP markers. Theor. Appl. Genet. 99: 473–480.

Dubreuil P., M. Warburton, M. Chastanet, D. Hoisington, and A. Charcosset, 2006 More on the introduction of temperate maize into Europe: large-scale bulk SSR genotyping and new historical elements [Zea mays L.; Simple Sequence Repeats]. Maydica Italy 51: 281-291

Dudley J., and R. Lambert, 2004 100 generations of selection for oil and protein in corn. Plant Breed. Rev. 24: 79–110.

Durand E., M. I. Tenaillon, C. Ridel, D. Coubriche, P. Jamin, et al., 2010 Standing variation and new mutations both contribute to a fast response to selection for flowering time in maize inbreds. BMC Evol. Biol. 10: 2.

Durand E., M. I. Tenaillon, X. Raffoux, S. Thépot, M. Falque, et al., 2015 Dearth of polymorphism associated with a sustained response to selection for flowering time in maize. BMC Evol. Biol. 15: 103.

Duvick D. N., 2001 Biotechnology in the 1930s: the development of hybrid maize. Nat. Rev. Genet. 2: 69–74.

Duvick D. N., J. S. C. Smith, and M. Cooper, 2004 Long-term selection in a commercial hybrid maize breeding program. Plant Breed. Rev. J Janick Ed Vol 24 Part 2 Long Term Sel. Crops Anim. Bact. Pp 109-151 John Wiley Sons N. Y.

Duvick D. N., 2005 The Contribution of Breeding to Yield Advances in Maize (Zea mays L.). N Sparks Ed Adv Agron Acad. Press San Diego CA Vol. 86.: 83–145.

East E., 1908 Inbreeding in corn, pp. 419–428 in Annual Report of the Connecticut Agricultural Experimental Station1907.

Elshire R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, et al., 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLOS ONE 6: e19379.

Eynard S. E., J. J. Windig, G. Leroy, R. van Binsbergen, and M. P. L. Calus, 2015 The effect of rare alleles on estimated genomic relationships from whole genome sequence data. BMC Genet. 16: 24.

Eynard S. E., J. J. Windig, S. J. Hiemstra, and M. P. L. Calus, 2016 Whole-genome sequence data uncover loss of genetic diversity due to selection. Genet. Sel. Evol. 48: 33.

Eynard S. E., P. Croiseau, D. Laloë, S. Fritz, M. P. L. Calus, et al., 2018 Which Individuals To Choose To Update the Reference Population? Minimizing the Loss of Genetic Diversity in Animal Genomic Selection Programs. G3 Genes Genomes Genet. 8: 113–121.

Falconer D. S., and T. F. C. Mackay, 1996 Introduction to Quantitative Genetics. 4th ed. Pearson, Harlow, England.

Felsenstein J., 1965 The effect of linkage on directional selection. Genetics 52(2): 349–63.

Feng L., S. Sebastian, S. Smith, and M. Cooper, 2006 Temporal trends in SSR allele frequencies associated with long-term selection for yield in maize. Maydica 293–300.

Fernandes J. B., M. Seguela-Arnaud, C. Larchevêque, A. H. Lloyd, and R. Mercier, 2018 Unleashing meiotic crossovers in hybrid plants. Proc. Natl. Acad. Sci. 115: 2431–2436.

Fischer R. A., 1930 The genetical theory of natural selection. Clarendon Press, Oxford, 2d ed. (1958), Dover, New York.

Fischer S., J. Möhring, C. C. Schön, H.-P. Piepho, D. Klein, et al., 2008 Trends in genetic variance components during 30 years of hybrid maize breeding at the University of Hohenheim. Plant Breed. 127: 446–451.

Fisher R. A., 1918 XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans. R. Soc. Edinb. 52: 399–433.

Fonceka D., H.-A. Tossim, R. Rivallan, H. Vignes, I. Faye, et al., 2012 Fostered and left behind alleles in peanut: interspecific QTL mapping reveals footprints of domestication and useful natural variation for breeding. BMC Plant Biol. 12: 26.

Food and Agriculture Organization, FAO, 2019 http://www.fao.org/faostat/en/#data/QC.

Frankel O., 1984 Genetic perspectives of germplasm conservation. Genet. Manip. Impact Man Soc. Camb. Univ. Press Camb. 61: 161–170.

Frascaroli E., T. A. Schrag, and A. E. Melchinger, 2013 Genetic diversity analysis of elite European maize (Zea mays L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs. Theor. Appl. Genet. 126: 133–141.

Fu Y. B., S. Kibite, and K. W. Richards, 2004 Amplified fragment length polymorphism analysis of 96 Canadian oat cultivars released between 1886 and 2001. Canadian Journal of Plant Science 23–30.

Fu Y.-B., 2006 Impact of plant breeding on genetic diversity of agricultural crops: searching for molecular evidence. Plant Genet. Resour. 4: 71–78.

Fu Y.-B., G. W. Peterson, J.-K. Yu, L. Gao, J. Jia, et al., 2006 Impact of plant breeding on genetic diversity of the Canadian hard red spring wheat germplasm as revealed by EST-derived SSR markers. Theor. Appl. Genet. 112: 1239–1247.

Fu Y.-B., 2015 Understanding crop genetic diversity under modern plant breeding. Theor. Appl. Genet. 128: 2131–2142.

Fukao T., T. Harris, and J. Bailey-Serres, 2008 Evolutionary analysis of the Sub1 gene cluster that confers submergence tolerance to domesticated rice. Ann. Bot. 103: 143–150.

Gaj T., C. A. Gersbach, and C. F. Barbas, 2013 ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. 31: 397–405.

Gallais A., A. Barrière, and J. P. Monod, 2001 A French Cooperative Programme for Management and Utilization of Maize Genetic Resources, pp. 331–340 in Broadening the Genetic Base of Crop Production, H.D. Cooper, C. Spillane and T. Hodgkin.

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.

Gaynor R. C., G. Gorjanc, A. R. Bentley, E. S. Ober, P. Howell, et al., 2017 A Two-Part Strategy for Using Genomic Selection to Develop Inbred Lines. Crop Sci. 57: 2372–2386.

Gerke J. P., J. W. Edwards, K. E. Guill, J. Ross-Ibarra, and M. D. McMullen, 2015 The Genomic Impacts of Drift and Selection for Hybrid Performance in Maize. Genetics 201: 1201–1211.

Gianola D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando, 2009 Additive Genetic Variability and the Bayesian Alphabet. Genetics 183: 347–363.

Glaszmann J., B. Kilian, H. Upadhyaya, and R. Varshney, 2010 Accessing genetic diversity for crop improvement. Curr. Opin. Plant Biol. 13: 167–173.

Goddard M., 2009 Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 245–257.

Gonen S., M. Battagin, S. E. Johnston, G. Gorjanc, and J. M. Hickey, 2017 The potential of shifting recombination hotspots to increase genetic gain in livestock breeding. Genet. Sel. Evol. 49: 55.

Goodman M. M., 1985 Exotic maize germplasm: Status, prospects, and remedies. Iowa State J. Res.

Goodman M. M., 1990 Genetic and germ plasm stocks worth conserving. J. Hered. 81: 11–16.

Goodman M. M., 1999 Broadening the genetic diversity in maize breeding by use of exotic germplasm, pp. 139–148 in The genetics and exploitation of heterosis in crops, J. G. Coors and S. Pandey.

Goodnight C. J., 1988 Epistasis and the effect of founder events on the additive genetic variance. Evolution 42: 441–454.

Gorjanc G., J. Jenko, S. J. Hearne, and J. M. Hickey, 2016 Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. BMC Genomics 17: 30.

Gorjanc G., R. C. Gaynor, and J. M. Hickey, 2018 Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. Theor. Appl. Genet. 131: 1953–1966.

Habier D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller, 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42: 5.

Habier D., R. L. Fernando, and D. J. Garrick, 2013 Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. Genetics 194: 597–607.

Hallander J., and P. Waldmann, 2009a Optimum contribution selection in large general tree breeding populations with an application to Scots pine. Theor. Appl. Genet. 118: 1133–1142.

Hallander J., and P. Waldmann, 2009b Optimization of selection contribution and mate allocations in monoecious tree breeding populations. BMC Genet. 10: 70.

Hallauer A. R., and J. B. Miranda, 1988 Quantitative genetics in maize breeding. Ames. Iowa State University Press.

Hammer K., N. Arrowsmith, and T. Gladis, 2003 Agrobiodiversity with emphasis on plant genetic resources. Naturwissenschaften 90: 241–250.

Han Y., X. Zhao, D. Liu, Y. Li, D. A. Lightfoot, et al., 2016 Domestication footprints anchor genomic regions of agronomic importance in soybeans. New Phytol. 209: 871–884.

Han Y., J. N. Cameron, L. Wang, and W. D. Beavis, 2017 The Predicted Cross Value for Genetic Introgression of Multiple Alleles. Genetics 205: 1409–1423.

Hayes B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard, 2009 Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92: 433–443.

Heerwaarden J. van, J. Doebley, W. H. Briggs, J. C. Glaubitz, M. M. Goodman, et al., 2011 Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc. Natl. Acad. Sci. 108: 1088–1092.

Heerwaarden J. van, M. B. Hufford, and J. Ross-Ibarra, 2012 Historical genomics of North American maize. Proc. Natl. Acad. Sci. 109: 12420–12425.

Hellin J., M. R. Bellon, and S. J. Hearne, 2014 Maize Landraces and Adaptation to Climate Change in Mexico. J. Crop Improv. 28: 484–501.

Henderson C. R., 1975 Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics 31: 423–447.

Heslot N., H.-P. Yang, M. E. Sorrells, and J.-L. Jannink, 2012 Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci. 52: 146–160.

Heslot N., J. Rutkoski, J. Poland, J.-L. Jannink, and M. E. Sorrells, 2013 Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PloS One 8: e74612.

Heslot N., J.-L. Jannink, and M. E. Sorrells, 2015 Perspectives for Genomic Selection Applications and Research in Plants. Crop Sci. 55: 1–12.

Hickey J. M., T. Chiurugwi, I. Mackay, W. Powell, Implementing Genomic Selection in CGIAR Breeding Programs Workshop Participants, et al., 2017 Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nat. Genet. 49: 1297–1303.

Hill W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. 8: 269–294.

Hill W. G., 1982a Rates of change in quantitative traits from fixation of new mutations. Proc. Natl. Acad. Sci. 79: 142–145.

Hill W. G., 1982b Predictions of response to artificial selection from new mutations. Genet. Res. 40: 255–278.

Hill W. G., 2016 Is Continued Genetic Improvement of Livestock Sustainable? Genetics 202: 877–881.

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.

Ho J. C., S. Kresovich, and K. R. Lamkey, 2005 Extent and Distribution of Genetic Variation in U.S. Maize. Crop Sci. 45: 1891–1900.

Hölker A. C., M. Mayer, T. Presterl, T. Bolduan, E. Bauer, et al., 2019 European maize landraces made accessible for plant breeding and genome-based studies. Theor. Appl. Genet. 132: 3333–3345

Horner E. S., W. H. Chapman, M. C. Lutrick, and H. W. Lundy, 1969 Comparison of Selection Based on Yield of Topcross Progenies and of S2 Progenies in Maize (Zea mays L.) 1. Crop Sci. 9: 539–543.

Houle D., B. Morikawa, and M. Lynch, 1996 Comparing Mutational Variabilities. Genetics 143: 1467–1483.

Hyten D. L., Q. Song, Y. Zhu, I.-Y. Choi, R. L. Nelson, et al., 2006 Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl. Acad. Sci. 103: 16666.

Isidro-Sanchez J., J.-L. Jannink, D. Akdemir, J. Poland, N. Heslot, et al., 2015 Training set optimization under population structure in genomic selection. Theor. Appl. Genet. 128: 145–158.

Jacobson A., L. Lian, S. Zhong, and R. Bernardo, 2015 Minimal Loss of Genetic Diversity after Genomewide Selection within Biparental Maize Populations. Crop Sci. 55: 783–789.

James J. W., and G. McBride, 1958 The spread of genes by natural and artificial selection in closed poultry flock. J. Genet. 56: 55.

Jannink J.-L., 2010 Dynamics of long-term genomic selection. Genet. Sel. Evol. 42: 35.

Jenko J., G. Gorjanc, M. A. Cleveland, R. K. Varshney, C. B. A. Whitelaw, et al., 2015 Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs. Genet. Sel. Evol. GSE 47: 55.

Jones D., 1918 The effects of inbreeding and cross-breeding upon development. Conn. Agric. Exp. Stn. Bull.

Kannenberg L. W., and D. E. Falk, 1995 Models for activation of plant genetic resources for crop breeding programs. Can. J. Plant Sci. 75: 45–53.

Kannenberg L. W., 2001 HOPE, a Hierrarchical, Open-ended System for Broadening the Breeding Base of Maize, pp. 311–318 in Broadening the Genetic Base of Crop Production, H.D. Cooper, C. Spillane and T. Hodgkin.

Keightley P. D., 2004 Mutational variation and long-term selection response. Plant Breed. Rev. 24: 227–248.

Kerr R. J., M. E. Goddard, and S. F. Jarvis, 1998 Maximising genetic response in tree breeding with constraints on group coancestry. Silvae Genet. Ger. 47: 2–3

Kilian B., and A. Graner, 2012 NGS technologies for analyzing germplasm diversity in genebanks*. Brief. Funct. Genomics 11: 38–50.

Kinghorn B. P., 2011 An algorithm for efficient constrained mate selection. Genet. Sel. Evol. 43: 4.

Koebner R., P. Donini, J. Reeves, R. Cooke, and J. Law, 2003 Temporal flux in the morphological and molecular diversity of UK barley. Theor. Appl. Genet. 106: 550–558.

Kremling K. A. G., S.-Y. Chen, M.-H. Su, N. K. Lepak, M. C. Romay, et al., 2018 Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature 555: 520–523.

Labate J. A., K. R. Lamkey, M. Lee, and W. L. Woodman, 1999 Temporal changes in allele frequencies in two reciprocally selected maize populations. Theor. Appl. Genet. 99: 1166–1178.

Ladizinsky G., 1985 Founder effect in crop-plant evolution. Econ. Bot. 39: 191–199.

Lado B., S. Battenfield, C. Guzmán, M. Quincke, R. P. Singh, et al., 2017 Strategies for Selecting Crosses Using Genomic Prediction in Two Wheat Breeding Programs. Plant Genome 10.

Laloë D., 1993 Precision and information in linear models of genetic evaluation. Genet. Sel. Evol. 25: 557.

Lande R., and R. Thompson, 1990 Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124: 743–756.

Lehermeier C., G. de los Campos, V. Wimmer, and C.-C. Schön, 2017a Genomic variance estimates: With or without disequilibrium covariances? J. Anim. Breed. Genet. 134: 232–241.

Lehermeier C., S. Teyssèdre, and C.-C. Schön, 2017b Genetic Gain Increases by Applying the Usefulness Criterion with Improved Variance Prediction in Selection of Crosses. Genetics 207: 1651–1661.

Lin Z., N. O. I. Cogan, L. W. Pembleton, G. C. Spangenberg, J. W. Forster, et al., 2016 Genetic Gain and Inbreeding from Genomic Selection in a Simulated Commercial Breeding Program for Perennial Ryegrass. Plant Genome 9.

Liu H., T. H. Meuwissen, A. C. Sørensen, and P. Berg, 2015 Upweighting rare favourable alleles increases long-term genetic gain in genomic selection programs. Genet. Sel. Evol. GSE 47.

Longin C. F. H., and J. C. Reif, 2014 Redesigning the exploitation of wheat genetic resources. Trends Plant Sci. 19: 631–636.

Lush J. L., 1937 Animal breeding plans. Iowa State College Press, Iowa.

Lynch M., and B. Walsh, 1998 Genetics and analysis of quantitative traits. Sinauer Sunderland, MA.

Ma X., Q. Zhang, Q. Zhu, W. Liu, Y. Chen, et al., 2015 A Robust CRISPR/Cas9 System for Convenient, High-Efficiency Multiplex Genome Editing in Monocot and Dicot Plants. Mol. Plant 8: 1274–1284.

Malézieux E., Y. Crozat, C. Dupraz, M. Laurans, D. Makowski, et al., 2009 Mixing Plant Species in Cropping Systems: Concepts, Tools and Models: A Review, pp. 329–353 in Sustainable Agriculture, edited by Lichtfouse E., Navarrete M., Debaeke P., Véronique S., Alberola C. Springer Netherlands, Dordrecht.

Mascher M., M. Schreiber, U. Scholz, A. Graner, J. C. Reif, et al., 2019 Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. Nat. Genet. 51: 1076–1081.

Matsuoka Y., Y. Vigouroux, M. M. Goodman, J. S. G, E. Buckler, et al., 2002 A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. 99: 6080–6084.

Maynard-Smith J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23–35.

McCouch S., G. J. Baute, J. Bradeen, P. Bramel, P. K. Bretting, et al., 2013 Agriculture: Feeding the future. Nature 499: 23–24.

Melchinger A. E., P. Schopp, D. Müller, T. A. Schrag, E. Bauer, et al., 2017 Safeguarding Our Genetic Resources with Libraries of Doubled-Haploid Lines. Genetics 206: 1611–1619.

Meuwissen T. H., 1997 Maximizing the response of selection with a predefined rate of inbreeding. J. Anim. Sci. 75: 934–940.

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.

Mieulet D., G. Aubert, C. Bres, A. Klein, G. Droc, et al., 2018 Unleashing meiotic crossovers in crops. Nat Plants 4: 480.

Mikel M. A., and J. W. Dudley, 2006 Evolution of North American Dent Corn from Public to Proprietary Germplasm. Crop Sci. 46: 1193–1205.

Mikel M. A., 2018 Progenitor lineage within proprietary dent corn germplasm, in Illinois Corn Breeders School Proceedings, University of Illinois, Urbana.

Nei M., 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. U. S. A. 70: 3321–3323.

Neimann-Sorensen A., and A. Robertson, 1961 The association between blood groups and several production characteristics in three Danish cattle breeds. Acta Agric. Scand. 11: 163–196.

Neyhart J. L., T. Tiede, A. J. Lorenz, and K. P. Smith, 2017 Evaluating Methods of Updating Training Data in Long-Term Genomewide Selection. G3 Genes Genomes Genet. 7: 1499–1510.

Paixão T., and N. H. Barton, 2016 The effect of gene interactions on the long-term response to selection. Proc. Natl. Acad. Sci. 113: 4422–4427.

Pelé A., M. Falque, G. Trotoux, F. Eber, S. Nègre, et al., 2017 Amplifying recombination genome-wide and reshaping crossover landscapes in Brassicas. PLOS Genet. 13: e1006794.

Petersen L., H. Østergård, and H. Giese, 1994 Genetic diversity among wild and cultivated barley as revealed by RFLP. Theor. Appl. Genet. 89: 676–681.

Pollak L. M. (Ed.), 1990 Evaluation of Caribbean maize accessions in Puerto Rico, in Caribbean Food Crops Society 26th Annual Meeting, Mayaguez, Puerto Rico.

Pollak L. M., and W. Salhuana, 2001 The Germplasm Enhancement of Maize (GEM) Project: Private and Public Sector Collaboration, pp. 319–329 in Broadening the Genetic base of Crop Production, H.D. Cooper, C. Spillane and T. Hodgkin.

Pollak L. M., 2003 The history and success of the public-private project on germplasm enhancement of maize (GEM). Advances in agronomy 78: 46–89.

Popi J., 1997 A critical evaluation of the HOPE breeding system as a means for broadening the deployed germplasm base in maize. PhD Dissertation, University of Guelph, Ontario.

Pszczola M., T. Strabel, H. A. Mulder, and M. P. L. Calus, 2012 Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95: 389–400.

Ramstein G. P., S. E. Jensen, and E. S. Buckler, 2019 Breaking the curse of dimensionality to identify causal variants in Breeding 4. Theor. Appl. Genet. 132: 559–567.

Rasmusson D. C., and R. L. Phillips, 1997 Plant Breeding Progress and Genetic Diversity from De Novo Variation and Elevated Epistasis. Crop Sci. 37: 303–310.

Rebourg C., B. Gouesnard, and A. Charcosset, 2001 Large scale molecular analysis of traditional European maize populations. Relationships with morphological variation. Heredity 86: 574–587.

Rebourg C., M. Chastanet, B. Gouesnard, C. Welcker, P. Dubreuil, et al., 2003 Maize introduction into Europe: the history reviewed in the light of molecular data. Theor. Appl. Genet. 106: 895–903.

Reif J. C., P. Zhang, S. Dreisigacker, M. L. Warburton, M. V. Ginkel, et al., 2005a Wheat genetic diversity trends during domestication and breeding. Theor. Appl. Genet. 110: 859–864.

Reif J. C., S. Hamrit, M. Heckenberger, W. Schipprack, H. P. Maurer, et al., 2005b Trends in genetic diversity among European maize cultivars and their parental components during the past 50 years. Theor. Appl. Genet. 111: 838–845.

Reif J. C., S. Fischer, T. A. Schrag, K. R. Lamkey, D. Klein, et al., 2010 Broadening the genetic base of European maize heterotic pools with US Cornbelt germplasm using field and molecular marker data. TAG Theor. Appl. Genet. Theor. Angew. Genet. 120: 301–310.

Ribaut J.-M., and M. Ragot, 2006 Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations, and alternatives. J. Exp. Bot. 58: 351–360.

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.

Rincent R., S. Nicolas, S. Bouchet, T. Altmann, D. Brunel, et al., 2014 Dent and Flint maize diversity panels reveal important genetic potential for increasing biomass production. TAG Theor. Appl. Genet. Theor. Angew. Genet. 127: 2313–2331.

Rio S., T. Mary-Huard, L. Moreau, and A. Charcosset, 2019 Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. Theor. Appl. Genet. 132: 81–96.

Robertson A., 1960 A theory of limits in artificial selection. Proc. R. Soc. Lond. B Biol. Sci.

Robertson A., 1961 Inbreeding in artificial selection programmes. Genet. Res. 2: 189–194.

Roussel V., L. Leisova, F. Exbrayat, Z. Stehno, and F. Balfourier, 2005 SSR allelic diversity changes in 480 European bread wheat varieties released from 1840 to 2000. Theor. Appl. Genet. 111: 162–170.

Russell W. A., and S. A. Eberhart, 1975 Hybrid Performance of Selected Maize Lines from Reciprocal Recurrent and Testcross Selection Programs 1. Crop Sci. 15: 1–4.

Rutkoski J., R. P. Singh, J. Huerta-Espino, S. Bhavani, J. Poland, et al., 2015 Genetic Gain from Phenotypic and Genomic Selection for Quantitative Resistance to Stem Rust of Wheat. Plant Genome 8.

Sadhu M. J., J. S. Bloom, L. Day, and L. Kruglyak, 2016 CRISPR-directed mitotic recombination enables genetic mapping without crosses. Science 352: 1113–1116.

Salhuana W., R. Sevilla, and S. A. Eberhart, 1997 Latin american maize project (LAMP) final report. Pioneer Hi-Bred International Spec. Publ.

Salhuana W., and L. Pollak, 2006 Latin American Maize Project (LAMP) and Germplasm Enhancement of Maize (GEM) project: generating useful breeding germplasm. Maydica 51: 339–355.

Sanchez L., A. Caballero, and E. Santiago, 2006 Palliating the impact of fixation of a major gene on the genetic variation of artificially selected polygenes. 88: 105–118.

Schnell F., and H. Utz, 1975 F1-Leistung und Elternwahl in der Züchtung von Selbstbefruchtern., pp. 243–248 in Bericht über die Arbeitstagung der Vereinigung österreichischer Pflanzenzüchter., BAL Gumpenstein, Austria.

Segelke D., F. Reinhardt, Z. Liu, and G. Thaller, 2014 Prediction of expected genetic variation within groups of offspring for innovative mating schemes. Genet. Sel. Evol. 46: 42.

Sehgal D., P. Vikram, C. P. Sansaloni, C. Ortiz, C. S. Pierre, et al., 2015 Exploring and Mobilizing the Gene Bank Biodiversity for Wheat Improvement. PLOS ONE 10: e0132112.

Servin B., O. C. Martin, M. Mézard, and F. Hospital, 2004 Toward a Theory of Marker-Assisted Gene Pyramiding. Genetics 168: 513–523.

Shalem O., N. E. Sanjana, and F. Zhang, 2015 High-throughput functional genomics using CRISPR–Cas9. Nat. Rev. Genet. 16: 299.

Sharon E., S.-A. A. Chen, N. M. Khosla, J. D. Smith, J. K. Pritchard, et al., 2018 Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. Cell 175: 544–557.e16.

Shi J., H. Gao, H. Wang, H. R. Lafitte, R. L. Archibald, et al., 2017 ARGOS8 variants generated by CRISPR-Cas9 improve maize grain yield under field drought stress conditions. Plant Biotechnol. J. 15: 207–216.

Shull G. H., 1908 The composition of a field of maize. J. Hered. 296–301.

Shull G. H., 1909 A pure-line method in corn breeding. J. Hered. 1: 51–58.

Shull G. H., 1914 Duplicate genes for capsule-form in Bursa bursa-pastoris. Mol. Gen. Genet. MGG 12: 97–149.

Simmonds N. W., 1962 Variability in Crop Plants, Its Use and Conservation. Biol. Rev. 37: 422–465.

Simmonds N. W., 1979 Principles of crop improvement. Longman, London.

Simmonds N. W., 1993 Introgression and Incorporation. Strategies for the Use of Crop Genetic Resources. Biol. Rev. 68: 539–562.

Smith J. S. C., 1988 Diversity of United States Hybrid Maize Germplasm; Isozymic and Chromatographic Evidence. Crop Sci. 28: 63–69.

Smith S., and W. Beavis, 1996 Molecular Marker Assisted Breeding in a Company Environment, pp. 259–272 in The Impact of Plant Molecular Genetics, edited by Sobral B. W. S. Birkhäuser Boston, Boston, MA.

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.

Sorensen D., R. Fernando, and D. Gianola, 2001 Inferring the trajectory of genetic variance in the course of artificial selection. Genet. Res. 77: 83–94.

Spillane C., and P. Gepts, 2001 Evolutionary and genetics perspectives on the dynamics of crop genepools, pp. 25–53 in Broadening the Genetic Base of Crop Production, H.D. Cooper, C. Spillane and T. Hodgkin.

Springer N. M., K. Ying, Y. Fu, T. Ji, C.-T. Yeh, et al., 2009 Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. PLOS Genet. 5: e1000734.

Strigens A., W. Schipprack, J. C. Reif, and A. E. Melchinger, 2013 Unlocking the Genetic Diversity of Maize Landraces with Doubled Haploids Opens New Avenues for Breeding. PLOS ONE 8: e57234.

Sun C., and P. M. VanRaden, 2014 Increasing Long-Term Response by Selecting for Favorable Minor Alleles. PLOS ONE 9: e88510.

Tadesse W., M. Sanchez-Garcia, S. G. Assefa, A. Amri, Z. Bishaw, et al., 2019 Genetic Gains in Wheat Breeding and Its Role in Feeding the World. Crop Breed. Genet. Genomics 1.

Tallury S. P., and M. M. Goodman, 2001 The State of the Use of Maize Genetic Diversity in the USA and Sub-Saharan Africa, pp. 159–179 in Broadening the Genetic Base of Crop Production, H.D. Cooper, C. Spillane and T. Hodgkin.

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.

Tenaillon M. I., and A. Charcosset, 2011 A European perspective on maize history. C. R. Biol. 334: 221–228.

Torkamaneh D., and F. Belzile, 2015 Scanning and Filling: Ultra-Dense SNP Genotyping Combining Genotyping-By-Sequencing, SNP Array and Whole-Genome Resequencing Data. PLOS ONE 10: e0131533.

Tourrette E., R. Bernardo, M. Falque, and O. Martin, 2019 Assessing by modeling the consequences of increased recombination in genomic selection of Oryza sativa and Brassica rapa. G3 Genes Genomes Genet. 9: 4169-4181.

Troyer A. F., 1999 Background of U.S. Hybrid Corn. Crop Sci. 39: 601–626.

Ullstrup A. J., 1972 The Impacts of the Southern Corn Leaf Blight Epidemics of 1970-1971. Annual Review of Phytopathology 10: 37–50.

Unterseer S., E. Bauer, G. Haberer, M. Seidel, C. Knaak, et al., 2014 A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics 15: 823.

Van Inghelandt D. V., A. E. Melchinger, C. Lebreton, and B. Stich, 2010 Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor. Appl. Genet. 120: 1289–1299.

Vandermeer J., M. van Noordwijk, J. Anderson, C. Ong, and I. Perfecto, 1998 Global change and multi-species agroecosystems: Concepts and issues. Agric. Ecosyst. Environ. 67: 1–22.

VanRaden P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

Voss-Fels K. P., M. Cooper, and B. J. Hayes, 2019 Accelerating crop genetic gains with genomic selection. Theor. Appl. Genet. 132: 669–686.

Wallace J. G., E. Rodgers-Melnick, and E. S. Buckler, 2018 On the Road to Breeding 4.0: Unraveling the Good, the Bad, and the Boring of Crop Quantitative Genomics. Annu. Rev. Genet. 52: 421–444.

Walsh B., 2004 Population-and quantitative-genetic models of selection limits. Plant Breed. Rev. 24: 177–226.

Wang H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir, 2012 Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. 94: 73–83.

Wang C., S. Hu, C. Gardner, and T. Lübberstedt, 2017 Emerging Avenues for Utilization of Exotic Germplasm. Trends Plant Sci. 22: 624–637.

Weber K. E., and L. T. Diggins, 1990 Increased selection response in larger populations. II. Selection for ethanol vapor resistance in Drosophila melanogaster at two population sizes. Genetics 125: 585–597.

Weber K., 2004 Population size and long-term selection. Plant Breed. Rev. 24: 249–268.

Whittaker J. C., R. Thompson, and M. C. Denham, 2000 Marker-assisted selection using ridge regression. Genet. Res. 75: 249–252.

Wolter F., P. Schindele, and H. Puchta, 2019 Plant breeding at the speed of light: the power of CRISPR/Cas to generate directed genetic diversity at multiple sites. BMC Plant Biol. 19: 176.

Woolliams J. A., and R. Thompson, 1994 A theory of genetic contributions. Proc. 5th World Congr. Genet. Appl. Livest. Prod. Vol. 19.

Woolliams J. A., P. Bijma, and B. Villanueva, 1999 Expected genetic contributions and their impact on gene flow and genetic gain. Genetics 153: 1009–1020.

Woolliams J. A., P. Berg, B. S. Dagnachew, and T. H. E. Meuwissen, 2015 Genetic contributions and their optimization. J. Anim. Breed. Genet. 132: 89–99.

Wouw M. van de, C. Kik, T. van Hintum, R. van Treuren, and B. Visser, 2010 Genetic erosion in crops: concept, research results and challenges. Plant Genet. Resour. 8: 1–15.

Wray N. R., and R. Thompson, 1990 Prediction of rates of inbreeding in selected populations. Genet. Res. 55: 41–54.

Wray N., and M. Goddard, 1994 Increasing long-term response to selection. Genet. Sel. Evol. 26: 431.

Wright S., 1931 Evolution in Mendelian Populations. Genetics 16: 97–159.

Yamasaki M., S. I. Wright, and M. D. McMullen, 2007 Genomic Screening for Artificial Selection during Domestication and Improvement in Maize. Ann. Bot. 100: 967–973.

Yu X., X. Li, T. Guo, C. Zhu, Y. Wu, et al., 2016 Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat. Plants 2: 16150.

Supplementary Material Chapter 1

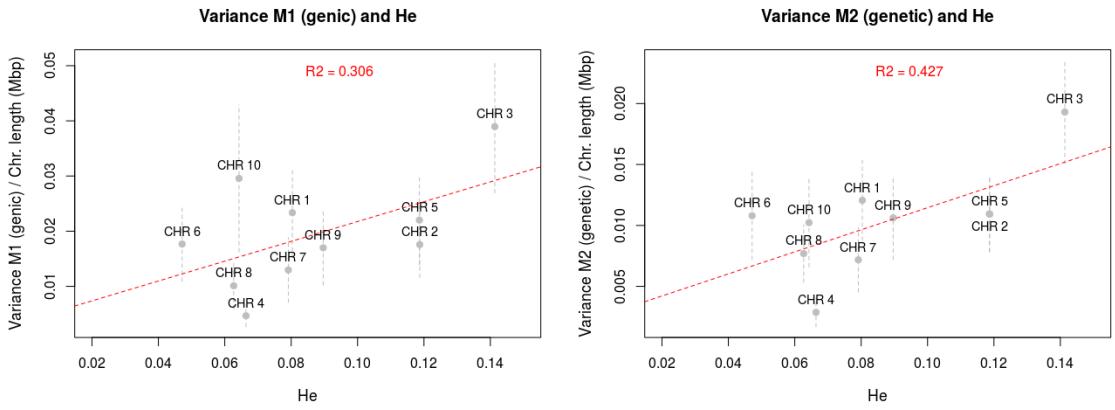


Figure S1 Relationship between genetic diversity (He) and genic $\hat{\sigma}^2_a$ (M1, **left**), genetic $\hat{\sigma}^2_A$ (M2, **right**) additive variances corrected by chromosome length (in Mbp). Vertical dashed bars represent posterior standard deviations of variance estimators.

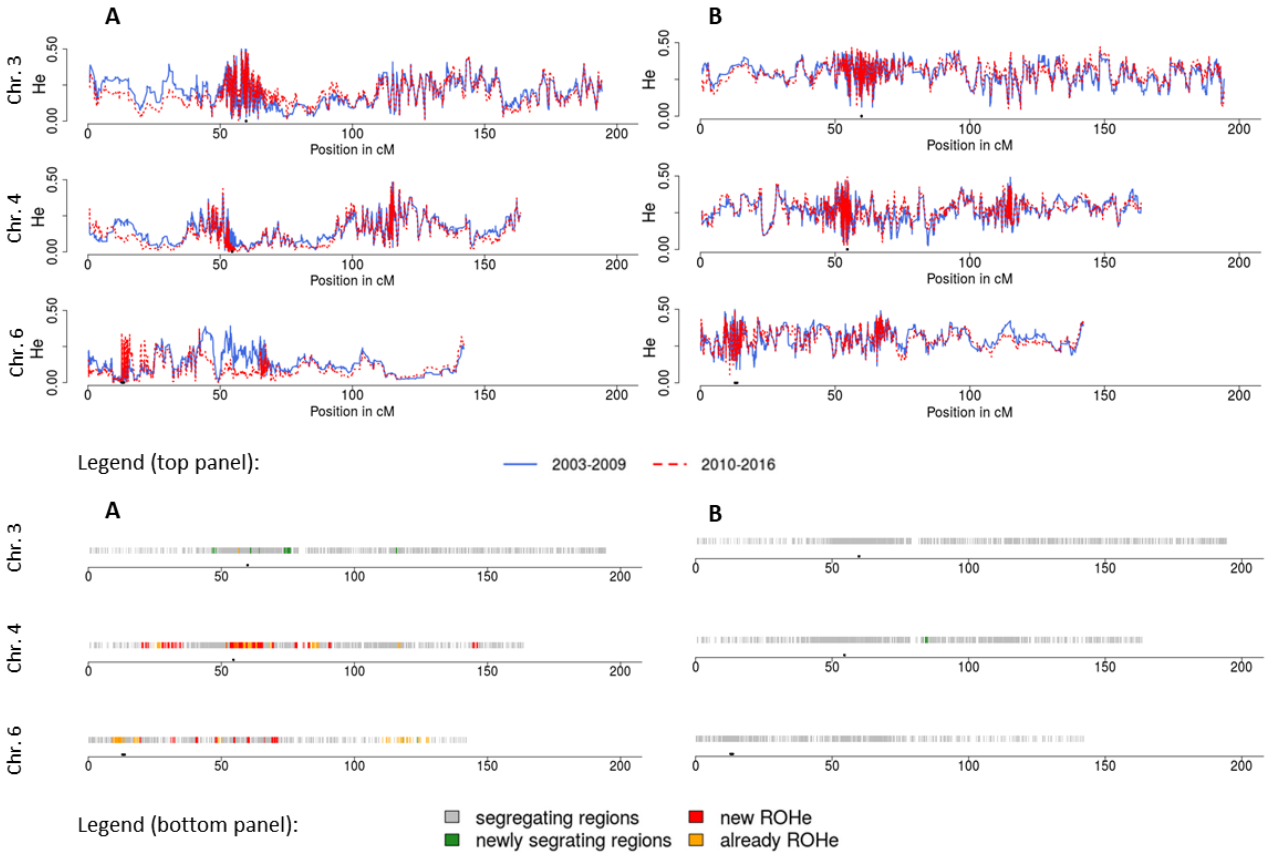


Figure S2 Genetic diversity (**Top panel**) and distribution of ROHe (**Bottom panel**) along genetic map. **Top panel**: Genetic diversity in Dent pool (**A**) and in Flint pool (**B**) for chromosomes 3, 4, 6 on genetic scale. Genetic diversity 2003-2009 in blue full line and 2010-2016 in red dotted line. Centromeres are marked in bold on the abscissa. **Bottom panel**: Evolution of ROHe in Dent pool (**A**) and in Flint pool (**B**) for chromosomes 3, 4, 6 on genetic scale. Regions are colored regarding their evolution between 2003-2009 and 2010-2016.

Table S1 Intra-cohort additive genetic variance (± standard error) in both pools estimated in Model 1b and sample size.

Cohort	Dent		Flint	
	GCA variance (±se)	Sample size	GCA variance (±se)	Sample Size
2003	21.35 (±5.87)	163	29.71 (±9.37)	83
2004	3.03 (±2.04)	208	28.27 (±7.97)	108
2005	13.79 (±3.39)	313	11.24 (±3.10)	256
2006	13.99 (±3.12)	353	11.00 (±2.82)	275
2007	10.67 (±3.11)	446	16.09 (±3.71)	275
2008	8.82 (±2.74)	403	12.79 (±3.16)	276
2009	14.21 (±4.07)	267	11.62 (±3.07)	349
2010	2.72 (±1.87)	315	11.50 (±2.69)	354
2011	6.17 (±1.99)	439	30.53 (±4.37)	356
2012	8.38 (±3.88)	340	5.80 (±2.93)	372
2013	3.40 (±3.66)	228	39.84 (±6.36)	290

Table S2 Intra-heterotic group genetic diversity and differentiation between heterotic groups using a five year sliding window with a one year increment.

Period	He		Fst
	Dent	Flint	
2003-2007	0.160	0.274	0.156
2004-2008	0.158	0.276	0.156
2005-2009	0.154	0.281	0.155
2006-2010	0.154	0.286	0.15
2007-2011	0.152	0.283	0.154
2008-2012	0.145	0.276	0.161
2009-2013	0.140	0.271	0.166
2010-2014	0.138	0.269	0.178
2011-2015	0.136	0.270	0.177
2012-2016	0.136	0.269	0.178

Table S3 Genetic diversity evolution between 2003-2009 and 2010-2016 in Dent and Flint pools and paired t-test significance on the difference between periods.

Chr.	Dent (He)			Flint (He)		
	2003-2009	2010-2016	ΔHe	2003-2009	2010-2016	ΔHe
1	0.148	0.113	-0.036***	0.260	0.269	0.009***
2	0.236	0.217	-0.019***	0.291	0.290	-0.001^{ns}
3	0.209	0.224	0.015***	0.290	0.296	0.006***
4	0.141	0.113	-0.027***	0.271	0.268	-0.003°
5	0.210	0.184	-0.026***	0.276	0.287	0.012***
6	0.115	0.105	-0.010***	0.311	0.312	0.001^{ns}
7	0.167	0.109	-0.058***	0.277	0.290	0.013***
8	0.115	0.075	-0.040***	0.271	0.287	0.016***
9	0.147	0.133	-0.014***	0.292	0.275	-0.017***
10	0.084	0.081	-0.004*	0.270	0.270	-0.001^{ns}

*p.value significance: <10^{-4} ***; <0.001 **; <0.01 *; <0.05 °°; <0.1 °; < 1 ^{ns}*

Table S4 Evolution of runs of expected homozygosity (ROHe) distribution between 2003-2009 and 2010-2016 in Dent pool in physical length. Column "% of chr" represents the percentage of the chromosome covered by ROHe.

	Period 2003-2009				Period 2010-2016			
Chr.	Nb. ROHe	Mean Length (Mb)	Max Length (Mb)	% of chr.	Nb. ROHe	Mean Length (Mb)	Max Length (Mb)	% of chr.
1	27	0.98	2.19	8.66	36	1.86	8.99	21.81
2	4	1.92	2.31	3.15	9	1.74	3.67	6.41
3	9	1.25	2.22	4.79	1	1.59	1.59	0.68
4	22	1.14	2.11	10.18	16	6.64	72.16	43.06
5	9	1.00	1.32	4.01	12	1.03	1.32	5.51
6	15	2.84	10.88	24.61	22	2.36	28.26	29.94
7	18	1.51	4.09	15.05	27	1.99	13.29	29.62
8	27	1.86	8.44	27.74	23	3.01	40.06	38.26
9	10	0.86	1.62	5.42	14	2.88	27.12	25.28
10	20	1.17	2.50	15.49	20	1.47	3.67	19.45
Mean	**16.10**	**1.45**	**3.77**	**11.91**	**18.00**	**2.46**	**20.01**	**22.00**

Table S5 Posterior means (± posterior standard deviation) of genomewide genomic variance accounting (M2, $\hat{\sigma}^2_A$) or not (M1, $\hat{\sigma}^2_a$) for covariance between QTLs in the 1,809 candidate RIL or DH Dent lines. Phenotypic variance (variance of BLUEs, Pheno) is also presented for comparison.

M1 (±sd)			**M2 (±sd)**			**Pheno**	**Ratio (±sd)**
$\hat{\sigma}^2_a$	Residual	Total	$\hat{\sigma}^2_A$	Residual	Total	Total	$\widehat{\sigma^2_A/\sigma^2_a}$
27.399 (±3.864)	34.606 (±1.424)	62.004 (±3.587)	20.599 (±1.459)	34.544 (±0.817)	55.143 (±1.230)	55.111	0.761 (±0.079)

Supplementary Material Chapter 2

File S1: Predictive ability on elite material

In the following, we evaluated the predictive ability of model in Eq. 1 trained across the Amazing dent panel (training population, TP) on elite private material (prediction population, PP). The PP lines consisted in lines produced in elite breeding from 2004 to 2016 and evaluated in hybrid combination on Flint testers for grain yield corrected at 15% of grain moisture (594 lines for GY; qx/ha), grain moisture (594 lines for GM; %) and male flowering time (539 lines for MF; days). These lines were genotyped with the MaizeSNP50 Illumina ® BeadChip (Ganal et *al.* 2011) and after quality control and imputation the same set of 40,478 SNPs as for the TP was kept, resulting in the genotyping matrix $\boldsymbol{X_{PP}}$.

The best linear unbiased estimators (BLUEs) of PP lines general combining ability (GCA) were estimated using the following model:

$$Y_{ijylr} = \mu + \beta_{yl} + \alpha_{1i} + \alpha_{2j} + \theta_{12\,ij} + \theta_{1Y\,iy} + \epsilon_{ijylr} \text{ (Model S1)}$$

where, Y_{ijylr} is observation r of the hybrid between line i and tester j evaluated in location l and year y. μ is the intercept, β_{yl} is the environment yl (Location x Year) fixed effect, α_{1i} is the tested PP line i GCA fixed effect, α_{2j} is the Flint tester j GCA fixed effect. $\boldsymbol{\theta_{12}} \sim N\left(\boldsymbol{O},\ \sigma^2_{\theta 12}\,\boldsymbol{I}\right)$ is the vector of hybrids between PP lines and Flint testers specific combining ability (SCA) random effects, $\boldsymbol{\theta_{1Y}} \sim N\left(\boldsymbol{O},\ \sigma^2_{1Y}\,\boldsymbol{I}\right)$ is the vector of tested PP line GCA by Year interaction random effects. Finally, $\boldsymbol{\epsilon} \sim N\left(\boldsymbol{O},\ \sigma^2_{\epsilon}\,\boldsymbol{I}\right)$ is the vector of independent random residual errors.

The heritability in the PP was estimated considering Model S1 where the tested PP line i GCA effect was modeled as random with $\boldsymbol{\alpha_1} \sim N\left(\boldsymbol{O}, \sigma^2_{\alpha 1}\,\boldsymbol{I}\right)$. The heritability in the PP was defined as: $h^2 = \hat{\sigma}^2_{\alpha 1} / \left(\hat{\sigma}^2_{\alpha 1} + \hat{\sigma}^2_{\theta 12}/n_{Hyb} + \hat{\sigma}^2_{1Y}/n_Y + \hat{\sigma}^2_{\epsilon}/n_{Obs}\right)$, where n_{Hyb} is the harmonic mean number of hybrids per line, n_Y is the harmonic mean number of years a given line was tested and n_{Obs} is the harmonic mean number of observations on a given line. The harmonic mean was considered instead of arithmetic mean as suggested in literature for unbalanced data set (Holland et *al.* 2010). The average coefficient of determination (referred as $\overline{\text{CD}}$, Laloë 1993) of $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{1}}$ best linear predictors (BLUP) was also considered as a proxy of trait heritability in the PP.

The BLUPs of genomic estimated breeding values of elite material were obtained as:

$$\boldsymbol{GEBV} = \boldsymbol{X_{PP}}\,\widehat{\boldsymbol{\beta}}$$

where, $\boldsymbol{X_{PP}}$ is the genotyping matrix of reference allele counts coded in 0 or 2 and $\widehat{\boldsymbol{\beta}}$ the vector of marker effects posterior mean obtained in Eq. 1. The predictive ability was evaluated as the correlation between the vector of GEBV and the vector of GCA in the PP: $r = cor(\boldsymbol{GEBV}, \widehat{\boldsymbol{\alpha}}_{\boldsymbol{1}})$.

Table S1 Square root of trait heritability in the prediction population PP and linear correlations between predictions and observations in the PP depending on the training population composition (TP, with or without elite private material) and the PP (all 13 years or a single year). For single-year predictions, the correlations were estimated on a subset of lines generated a given year and the minimum, maximum and mean correlations are reported.

Trait	Heritability in the prediction population (PP)		Predictive ability			
			Training population (TP)			
			338 public lines + 48 private lines		338 public lines	
			Predicted population (PP)			
	$\sqrt{h^2}$	$\sqrt{\overline{CD}}$	All 13 years	1 year min to max (mean)	All 13 years	1 year min to max (mean)
GY	0.347	0.371	0.404	-0.062 to 0.722 (0.305)	0.377	0.042 to 0.721 (0.282)
MF	0.519	0.548	0.495	0.222 to 0.715 (0.476)	0.509	0.260 to 0.728 (0.477)
GM	0.681	0.699	0.550	0.286 to 0.811 (0.560)	0.541	0.261 to 0.789 (0.542)

Cited literature:

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.

Holland J. B., W. E. Nyquist, and C. T. Cervantes-Martínez, 2010 Estimating and Interpreting Heritability for Plant Breeding: An Update, pp. 9–112 in Plant Breeding Reviews, John Wiley & Sons, Ltd.

Laloë D., 1993 Precision and information in linear models of genetic evaluation. Genet. Sel. Evol. 25: 557.

File S2: Supporting R code

Genomic prediction with a maize collaborative panel: identification of genetic resources to enrich elite breeding programs

File S2 - Supporting R code

Antoine Allier, Simon Teyssèdre, Christina Lehermeier, Alain Charcosset, Laurence Moreau

Contents

In this supporting file we provide the R script to compute the criterion H defined in the case of a donor and a population of elite lines $PopE$ (Eq. 5). We also implement the forward selection of donors in $PopD$ to enrich a population $PopE$ of elite inbred lines based on this H criterion. Finaly, we provide the code to compute the usefulness criterion for crosses between two inbred parents considering doubled haploid progeny derived from the $F1$ individual (Eq. 4) using the posterior mean variance (PMV) as described in Lehermeier et al. (2017) Genetics 207:1651-1661.

Examplary data and whole-genome regression model

For this documentation, we load a simulated maize data set from the `synbreedData` R package (Wimmer et al. 2012, Bioinformatics, 28:2086-2087), which includes genotypic data, phenotypic data, and a genetic map.

```
# Required packages
library(BGLR)
library(dplyr)
library(zoo)
library(synbreed)
library(synbreedData)
# Examplary data
data(maize)
maize <- codeGeno(maize)
Pheno <- maize$pheno[,1,1]
Geno <- maize$geno*2  # recode markers to 0,2

Map <- data.frame(CHROMOSOME = maize$map$chr,
                  POSITION = maize$map$pos,
                  MARKER = rownames(maize$map),
                  stringsAsFactors = FALSE)
```

First, a whole-genome regression model is fitted using phenotypic and genotypic data of a training population. For this, the function `BGLR` from R package BGLR is used (Perez and de los Campos 2014, Genetics, 198:483-495). With the option `saveEffects = T`, MCMC samples of marker effects are saved in a separate .bin file.

```
# Number of iterations and burn-in should be increased for proper inference
nIter <- 1000
burnIn <- 100
thin <- 2
# Run the Bayesian Ridge Regression
WGR <- BGLR(y = Pheno,
            ETA = list(list(X = Geno,model = "BRR",saveEffects = T)),
            nIter = nIter,burnIn = burnIn,thin = thin)
# Load .bin file including MCMC samples of marker effects
B <- readBinMat("ETA_1_b.bin")
# Posterior means of marker effects
Bhat <- colMeans(B)
# Genomic estimated breeding values
GEBV <- Geno%*%Bhat
```

Let us define the examplary $PopD$ of 50 candidate donors and $PopE$ of 10 elite lines:

```
# Assuming the Elites are the 10 best lines
PopE <- names(GEBV[order(GEBV,decreasing = TRUE),])[1:10]
# Donors are sampled among the remaining lines
PopD <- sample(setdiff(rownames(GEBV),PopE),50,replace = FALSE)
```

Definition of the criterion H

The following function computes the matrix $HEBV$ described in Eq. 2.

Function arguments:

- ObjectGeno: Genotype matrix with individuals (donors and elites) in line and markers in column
- ObjectBeta: Line vector of posterior mean marker effects
- ObjectMap: Data.frame of the genetic map with columns: CHROMOSOME, POSITION and MARKER (order by CHROMOSOME and POSITION)
- WindowSize: Integer for the size of the sliding haplotypes in nb. of markers
- StepSize: Integer for the increment of the sliding haplotypes in nb. of markers
- lambda: Scalling parameter to correct for overlapping segments (by default StepSize/WindowSize)

Value:

- Returns a list with the matrix of haplotypes estimated breeding values ($HEBV$) with individuals in line and haplotypes in column and a data.frame with the mean position (Mbp) and chromosome of each haplotype.

```
GetHEBVmat = function(ObjectGeno,ObjectBeta,ObjectMap,
                      WindowSize,StepSize,lambda = StepSize/WindowSize){
  # Matrix with loci effects for each individual in line and loci in column
  GEBVmat <- ObjectGeno*matrix(ObjectBeta,ncol = ncol(ObjectGeno),
                               nrow = nrow(ObjectGeno),byrow = TRUE)
  # Affect loci to haplotypes
  rownames(ObjectMap) <- ObjectMap$MARKER
  Haplo <- do.call(rbind,lapply(unique(ObjectMap$CHROMOSOME),function(chr_tmp){
    Lchr <- ObjectMap[intersect(ObjectMap$MARKER[ObjectMap$CHROMOSOME==chr_tmp],
                                colnames(GEBVmat)),"MARKER"]
    LHchr <- rollapply(Lchr,width = WindowSize,by = StepSize,align = "left",
                       FUN = function(d) {return(d)})
    rownames(LHchr) <- paste0("Hap_",seq(1,nrow(LHchr)),"_CHR_",chr_tmp)
    return(LHchr)
  }))
  # Design matrix Z affecting loci to haplotypes
  Z <- matrix(0,nrow = length(intersect(ObjectMap$MARKER,colnames(GEBVmat))),
              ncol = nrow(Haplo))
  rownames(Z) <- intersect(ObjectMap$MARKER,colnames(GEBVmat))
  colnames(Z) <- rownames(Haplo)
  invisible(lapply(1:ncol(Z),function(i){
    Z[Haplo[i,],i] <<- 1
  }))
  # Compute the HEBV matrix and mean position of each haplotypes
  HEBV <- (GEBVmat[,rownames(Z)]%*%Z)*lambda
  POS <- data.frame(HAP = colnames(Z),
                    CHR = sub(".*CHR_","",colnames(Z)),
                    POSITION = t(ObjectMap[rownames(Z),"POSITION"]%*%Z)/apply(Z,2,sum),
                    stringsAsFactors = FALSE)
  return(list(HEBV = HEBV,
              POSITION = POS))
}
```

The following function uses the HEBV matrix to compute the criterion H as described in the mansucript (Eq. 5).

Function arguments:

- ObjectHEBVmat: HEBV matrix with individuals in line and haplotypes in column
- ListPopE: Line vector of elite line names, i.e. *PopE*
- ListPopD: Line vector of candidate donor names, i.e. *PopD*

Value:

- Returns a data.frame with two columns: LINE (candidate donors) and corresponding H criterion value

```
ComputeH = function(ObjectHEBVmat,ListPopE,ListPopD){
  # Evaluate only the elite haplotypes
  if(is.null(ListPopD)) {
    tmp <-  apply(ObjectHEBVmat[ListPopE,],2,max)
    output <- data.frame(H = sum(tmp),
                         LINE = "PopE",
                         stringsAsFactors = FALSE)
  } else {
    # Evaluate the elite and donor haplotypes
    output <- do.call(rbind,lapply(ListPopD,function(i){
      tmp <-  apply(ObjectHEBVmat[c(ListPopE,i),],2,max)
      return(data.frame(H = sum(tmp),
                        LINE = i,
                        stringsAsFactors = FALSE))
    }))
  }
  return(output)
}
```

Forward selection of donors based on the criterion H

In the following we used previously defined functions in a forward iterative process.

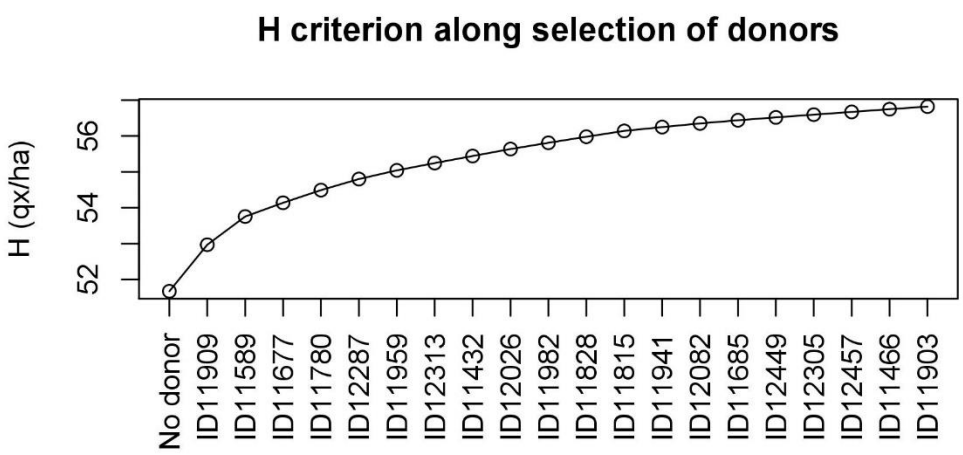
```
# User parameters
WindowSizeUI = 20
StepSizeUI = 5
nDonorUI = 20

# Compute the HEBV matrix
ObjectHEBV <- GetHEBVmat(ObjectGeno = Geno[c(PopE,PopD),],
                         ObjectBeta = Bhat,
                         ObjectMap = Map,
                         WindowSize = WindowSizeUI,
                         StepSize = StepSizeUI)
# Forward selection of donors
SelDonor <- c()
HSelDonor <- ComputeH(ObjectHEBVmat = ObjectHEBV$HEBV,# elites only
                      ListPopE = PopE,
                      ListPopD = NULL)
invisible(lapply(1:nDonorUI,function(iter){
  tmp <- ComputeH(ObjectHEBVmat = ObjectHEBV$HEBV,
                  ListPopE = c(SelDonor,PopE),# PopE + selected donors
```

```
                ListPopD = PopD[!PopD%in%c(SelDonor)] # yet unselected donors
  )
  # Increment with selected donor
  SelDonor <<- c(SelDonor,tmp[order(tmp$H,decreasing =TRUE),]$LINE[1])
  # Increment with H criterion value of elites + selected donors
  HSelDonor <<- rbind(HSelDonor,tmp[order(tmp$H,decreasing =TRUE),][1,])
}))
```

We can plot the increase in H during the forward selection of donors.

```
plot(HSelDonor$H,type = "o",xaxt = "n",
     main="H criterion along selection of donors",xlab="",ylab="H (qx/ha)")
axis(side = 1,at = 1:(length(SelDonor)+1),labels = c("No donor",SelDonor),las = 2)
```



Identification of donor by elite recipient crosses using the usefulness criterion

This function computes the covariance matrix Σ between marker genotypes in doubled haploid derived from $F1$ progeny of the cross P_1 x P_2 as presented in Lehermeier et al. (2017) Genetics 207:1651-1661.

Function arguments:

- ObjectGenoP1: Line vector of P_1 genotype
- ObjectGenoP2: Line vector of P_2 genotype
- ObjectMap: Data.frame of the genetic map with columns: CHROMOSOME, POSITION and MARKER (order by CHROMOSOME and POSITION)

Value:

- Returns the variance covariance matrix Σ of genotypes in DH progeny of the cross P_1 x P_2

```
GenCovProgeny   <- function(ObjectGenoP1,ObjectGenoP2,ObjectMap){
  # Computing expected frequency of recombinants c1
  myDist <- sapply(1:nrow(ObjectMap),
                   function(x) abs(ObjectMap$POSITION[x] - ObjectMap$POSITION))
  myCHR1 <- do.call(rbind,lapply(1:nrow(ObjectMap),function(x)
```

```
    rep(ObjectMap$CHROMOSOME[x],nrow(ObjectMap))))
  myCHR2 <- t(myCHR1)
  c1 <- 0.5*(1-exp(-2*(myDist/100)))
  c1[myCHR1!=myCHR2] <- 0.5
  Recomb <- (1-2*c1)
  # Computing diseqilibrium
  tmp <- rbind(ObjectGenoP1,ObjectGenoP2)/2
  D <- crossprod(scale(tmp,scale=F))/2
  # Sigma matrix
  MyVarCov <- 4*D*Recomb
  return(MyVarCov)
}
```

This function computes the posterior mean variance (PMV) estimate of DH progeny variance of the cross P_1 x P_2.

Function arguments:

- Bmat: Matrix including MCMC samples (in line) of marker effects (in column)
- ObjectGenoP1: Line vector of P_1 genotype
- ObjectGenoP2: Line vector of P_2 genotype
- ObjectMap: Data.frame of the genetic map with columns: CHROMOSOME,POSITION and MARKER (order by CHROMOSOME and POSITION)

Value:

- Returns the PMV of the trait in DH progeny of the cross P_1 x P_2

```
PMV <- function(Bmat,ObjectGenoP1,ObjectGenoP2,ObjectMap){
  # Variance-covariance matrix of genotypes in progeny
  Gcov <- GenCovProgeny(ObjectGenoP1,ObjectGenoP2,ObjectMap)
  # Posterior variance-covariance matrix of marker effects
  vB <- 1/nrow(Bmat)*crossprod(scale(Bmat,TRUE,FALSE))
  # Posterior mean of marker effects
  Bhat <- as.matrix(apply(Bmat,2,mean))
  # Posterior mean variance in progeny
  PMV <- sum(diag(Gcov %*% vB)) + t(Bhat) %*% Gcov %*% Bhat
  return(PMV)
}
```

This function computes the usefulness criterion (UC) considering DH progeny of the cross P_1 x P_2 using previously defined functions.

Function arguments:

- Bmat: Matrix including MCMC samples (in line) of marker effects (in column)
- ObjectGenoP1: Line vector of P_1 genotype
- ObjectGenoP2: Line vector of P_2 genotype
- ObjectMap: Data.frame of the genetic map with columns: CHROMOSOME, POSITION and MARKER (order by CHROMOSOME and POSITION)
- Psel: Selected fraction of progeny (by default 5%)
- h: Selection accuracy (by default 1)

Value:

- Returns the UC of the cross P_1 x P_2

```
UC <- function(Bmat,ObjectGenoP1,ObjectGenoP2,ObjectMap,Psel=0.05,h=1){
  Bhat <- colMeans(Bmat)
```

```
  i <- dnorm(qnorm(1-Psel))/Psel
  # Posterior mean variance
  VarG <- PMV(Bmat,ObjectGenoP1,ObjectGenoP2,ObjectMap)
  # Mean GEBVs of parents
  mu <- mean(crossprod(ObjectGenoP1,Bhat),crossprod(ObjectGenoP2,Bhat))
  return(mu + i*h*sqrt(VarG))
}
```

Finally, let us use the UC to identify the selected donor x elite recipient $\in PopE$ cross that maximizes the expected performance of selected DH progeny.

```
# Assuming we have selected the donors:
SelDonors4UC <-  HSelDonor$LINE[2:6]
# For each donor predict UC for all possible crosses
ResUC <-  do.call(rbind,lapply(SelDonors4UC,function(Donor){
  return(do.call(rbind,lapply(PopE,function(Recipient){
    UCtmp <- UC(Bmat = B,
                ObjectGenoP1 = Geno[Donor,],
                ObjectGenoP2 = Geno[Recipient,],
                ObjectMap = Map,
                Psel = 0.05,
                h = 1)
    return(data.frame(DONOR = Donor,
                      RECIPIENT = Recipient,
                      UC = UCtmp))
  })))
}))
# Selected crosses:
data.frame(ResUC %>%
             group_by(DONOR) %>%
             top_n(n = 1))
```

```
##     DONOR RECIPIENT       UC
## 1 ID11909   ID11857 23.56729
## 2 ID11589   ID12318 17.73277
## 3 ID11677   ID12318 19.78924
## 4 ID11780   ID11935 25.06862
## 5 ID12287   ID11666 30.04510
```

Supplementary Material Chapter 3

File S1: Derivation of linkage disequilibrium parameter in progeny for four-way cross and specific case of two-way cross, three-way cross and backcross

Here we derive the linkage disequilibrium parameter of doubled haploid progeny derived from the F_1' generation of a four-way cross (Figure 1 S1), while we give an extension for DH lines generated from higher selfing generations and for recombinant inbred lines in File S2. The crossing scheme for a four-way cross visualizing parental and potential progeny haplotypes is given in Figure 1 S1. Gametes from a four-way cross with four different parents (P1, P2, P3, and P4) correspond to gametes from six biparental crosses (P1xP2, P3xP4, P1xP3, P1xP4, P2xP3, P2xP4).

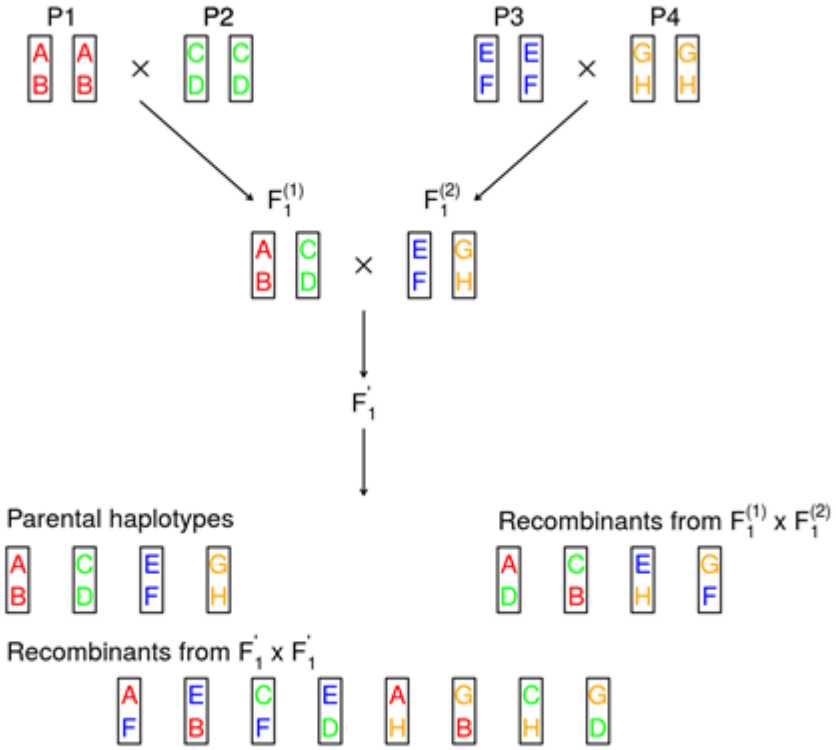


Figure 1 S1 Visualization of crossing scheme and two-locus parental as well as progeny haplotypes of a four-way cross from parents P1, P2, P3, and P4. Potential types of haplotypes are denoted with T1, T2, and T3.

To derive the entries of the Linkage Disequilibrium (LD) matrix $\boldsymbol{D}$ of the progeny of the four-way cross, we derive the frequencies of all different possible haplotypes. For this, three types of haplotypes can be differentiated (namely, T1, T2 and T3).

The first type T1 corresponds to parental haplotypes, for example AB from Figure 1 S1. The frequency of the haplotype AB in the parents is:

$$p_{AB} = \frac{1}{4}$$

The frequency of AB in gametes from the cross $F_1^{(1)} \times F_1^{(2)}$ is:

$$p'_{AB} = \frac{1}{4}\left(1 - c^{(1)}\right),$$

with $c^{(1)}$ the recombination frequency and $(1 - c^{(1)})$ the frequency that no recombination takes place within the cross $F_1^{(1)} \times F_1^{(2)}$.
Similarly, the frequency of AB in gametes from the cross $F_1' \times F_1'$ is:

$$p''_{AB} = \frac{1}{4} * \left(1 - c^{(1)}\right)^2$$

As there are four different parental haplotypes, the frequency of the type T1 haplotypes is:

$$P(T_1) = p''_{AB} + p''_{CD} + p''_{EF} + p''_{GH} = \left(1 - c^{(1)}\right)^2 \qquad (1)$$

The second type T2 corresponds to haplotypes formed by recombination in the cross $F_1^{(1)} \times F_1^{(2)}$, for example AD. The frequency of this haplotype in the parents is

$$p_{AD} = 0$$

The frequency of AD in gametes from the cross $F_1^{(1)} \times F_1^{(2)}$ is:

$$p'_{AD} = \frac{1}{2} * \frac{c^{(1)}}{2} = \frac{1}{4} c^{(1)}$$

As $\frac{c^{(1)}}{2}$ is the frequency of recombinants within $F_1^{(1)}$, the frequency in the whole cross is reduced by a factor of 1/2. The frequency of AD in gametes from the cross $F_1' \times F_1'$ is:

$$p''_{AD} = \frac{1}{4} c^{(1)}\left(1 - c^{(1)}\right),$$

with $(1 - c^{(1)})$ the frequency that no recombination takes place within the cross $F_1' \times F_1'$.
Overall, the frequency of the type T2 haplotypes is:

$$P(T_2) = p''_{AD} + p''_{CB} + p''_{EH} + p''_{GF} = c^{(1)}\left(1 - c^{(1)}\right) \qquad (2)$$

The third type T3 corresponds to haplotypes formed by recombination in the cross $F_1' \times F_1'$, for example AF. The frequency of these haplotypes in the parents is:

$$p_{AF} = 0$$

The frequency of AF in gametes from the cross $F_1^{(1)} \times F_1^{(2)}$ is:

$$p'_{AF} = 0$$

The frequency of AF in gametes from the cross $F_1' \times F_1'$ can be calculated as:

$$p''_{AF} = \frac{1}{2}\left(1 - c^{(1)}\right) * \frac{1}{2}\left(1 - c^{(1)}\right) * \frac{c^{(1)}}{2} + \frac{c^{(1)}}{2} * \frac{1}{2}\left(1 - c^{(1)}\right) * \frac{c^{(1)}}{2}$$

$$+ \frac{1}{2}\left(1 - c^{(1)}\right) * \frac{c^{(1)}}{2} * \frac{c^{(1)}}{2} + \frac{c^{(1)}}{2} * \frac{c^{(1)}}{2} * \frac{c^{(1)}}{2} = \frac{1}{8} c^{(1)}$$

Overall, the frequency of the type T3 haplotypes is:

$$P(T_3) = p''_{AF} + p''_{EB} + p''_{CF} + p''_{ED} + p''_{AH} + p''_{GB} + p''_{CH} + p''_{GD} = c^{(1)} \qquad (3)$$

All the different haplotypes and frequencies are summarized in Table 1 S1.
We define $h_{jl} = (h_j, h_l)$ a haplotype including loci j and l, with h_j and h_l the alleles of the haplotype at loci j and l, $h_j, h_l \in \{0,1\}$. Using the frequencies of the three types of haplotypes, we derive the LD in the progeny between locus j and l as:

$$D_{jl}^{progeny} = p_{jl} - p_j p_l$$

$$= P\left(h_{jl} = (z_j, z_l)\right) - P(h_j = z_j)P(h_l = z_l)$$

$$= \sum_{k=1}^{3} P(h_{jl} = (z_j, z_l) \mid T_k)P(T_k) - P(h_j = z_j)P(h_l = z_l), \quad (4)$$

where z_j and z_l denotes realizations of h_j and h_l, respectively.
For the conditional haplotype probabilities it holds:

$$P(h_{jl} = (z_j, z_l) \mid T_k) = \frac{1}{|T_k|} \sum_{v_{jl} \in T_k} \mathbf{1}_{v_j == z_j} \times \mathbf{1}_{v_l == z_l},$$

with $|T_k|$ the number of haplotypes of type k, $v_{jl} = (v_j, v_l)$ a haplotype of type k, $\mathbf{1}_{v_j == z_l}$ ($\mathbf{1}_{v_j == z_l}$) an indicator equal to 1 if $v_j = z_j$ ($v_l = z_l$) and 0 otherwise.
For the allele frequencies it holds:

$$P(h_j = z_j) = \frac{1}{4}\left(\mathbf{1}_{A == z_j} + \mathbf{1}_{C == z_j} + \mathbf{1}_{E == z_j} + \mathbf{1}_{G == z_j}\right)$$

$$P(h_l = z_l) = \frac{1}{4}\left(\mathbf{1}_{B == z_l} + \mathbf{1}_{D == z_l} + \mathbf{1}_{F == z_l} + \mathbf{1}_{H == z_l}\right)$$

Table 1 S1 Different haplotype types, their frequency in the parents (G0), after the first cross (G1), after the second cross (G2) and the Linkage Disequilibrium (LD) in G2.

Type	G0	G1	G2	LD
T1[a]	$\frac{1}{4}$	$\frac{1}{4}(1 - c^{(1)})$	$\frac{1}{4} * (1 - c^{(1)})^2$	$\frac{1}{4} * (1 - c^{(1)})^2 - \frac{1}{16}$
T2[b]	0	$\frac{1}{4}c^{(1)}$	$\frac{1}{4}c^{(1)} * (1 - c^{(1)})$	$\frac{1}{4}c^{(1)} * (1 - c^{(1)}) - \frac{1}{16}$
T3[c]	0	0	$\frac{1}{8}c^{(1)}$	$\frac{1}{8}c^{(1)} - \frac{1}{16}$

[a] Haplotypes: AB, CD, EF, GH (parental haplotypes)
[b] Haplotypes: AD, BC, EH, FG (recombinant from $F_1^{(1)} \times F_1^{(2)}$)
[c] Haplotypes: AF, AH, CF, CH, EB, ED, GB, GD (recombinant from $F_1' \times F_1'$)

Further, we use the linkage disequilibrium among two parents between loci j and l, which is exemplified for parent 1 and 2:

$$D_{jl}^{12} = p_{jl}^{12} - p_j^{12} p_l^{12}$$

$$= \frac{1}{2}\left(\mathbf{1}_{A==z_j} \times \mathbf{1}_{B==z_l} + \mathbf{1}_{C==z_j} \times \mathbf{1}_{D==z_l}\right) - \frac{1}{4}\left(\mathbf{1}_{A==z_j} + \mathbf{1}_{C==z_j}\right)\left(\mathbf{1}_{B==z_l} + \mathbf{1}_{D==z_l}\right)$$

$$= \frac{1}{4}\left(\mathbf{1}_{A==z_j} \times \mathbf{1}_{B==z_l} + \mathbf{1}_{C==z_j} \times \mathbf{1}_{D==z_l} - \mathbf{1}_{A==z_j} \times \mathbf{1}_{D==z_l} - \mathbf{1}_{C==z_j} \times \mathbf{1}_{B==z_l}\right).$$

For sake of clarity, we abbreviate in the following $\mathbf{1}_{A==z_j}$ with $\mathbf{1}_A$, $\mathbf{1}_{B==z_l}$ with $\mathbf{1}_B$ and accordingly for the rest (C, D, E, F, G, H). Then we can reform the LD in the progeny as a function of the recombination frequency $c_{jl}^{(1)}$ and the LD among two parents between loci j and l:

$$D_{jl}^{progeny} = \sum_{k=1}^{3} P\left(h_{jl} = (z_j, z_l) | T_k\right) P(T_k) - P\left(h_j = z_j\right) P(h_l = z_l)$$

$$\begin{aligned}
&= \frac{1}{4}\left(\mathbf{1}_A\mathbf{1}_B + \mathbf{1}_C\mathbf{1}_D + \mathbf{1}_E\mathbf{1}_F + \mathbf{1}_G\mathbf{1}_H\right)\left(1 - c_{jl}^{(1)}\right)^2 \\
&\quad + \frac{1}{4}\left(\mathbf{1}_A\mathbf{1}_D + \mathbf{1}_C\mathbf{1}_B + \mathbf{1}_E\mathbf{1}_H + \mathbf{1}_G\mathbf{1}_F\right)c_{jl}^{(1)}\left(1 - c_{jl}^{(1)}\right) \\
&\quad + \frac{1}{8}\left(\mathbf{1}_A\mathbf{1}_F + \mathbf{1}_A\mathbf{1}_H + \mathbf{1}_C\mathbf{1}_F + \mathbf{1}_C\mathbf{1}_H + \mathbf{1}_E\mathbf{1}_B + \mathbf{1}_E\mathbf{1}_D + \mathbf{1}_G\mathbf{1}_B + \mathbf{1}_G\mathbf{1}_D\right)c_{jl}^{(1)} \\
&\quad - \frac{1}{16}\left(\mathbf{1}_A + \mathbf{1}_C + \mathbf{1}_E + \mathbf{1}_G\right)\left(\mathbf{1}_B + \mathbf{1}_D + \mathbf{1}_F + \mathbf{1}_H\right)
\end{aligned}$$

$$\begin{aligned}
= \frac{1}{4}\Bigg[&\left(\left(1 - c_{jl}^{(1)}\right)^2 - \frac{1}{4}\right)\left(\mathbf{1}_A\mathbf{1}_B + \mathbf{1}_C\mathbf{1}_D + \mathbf{1}_E\mathbf{1}_F + \mathbf{1}_G\mathbf{1}_H\right) \\
&+ \left(c_{jl}^{(1)}\left(1 - c_{jl}^{(1)}\right) - \frac{1}{4}\right)\left(\mathbf{1}_A\mathbf{1}_D + \mathbf{1}_C\mathbf{1}_B + \mathbf{1}_E\mathbf{1}_H + \mathbf{1}_G\mathbf{1}_F\right) \\
&+ \left(\frac{c_{jl}^{(1)}}{2} - \frac{1}{4}\right)\left(\mathbf{1}_A\mathbf{1}_F + \mathbf{1}_A\mathbf{1}_H + \mathbf{1}_C\mathbf{1}_F + \mathbf{1}_C\mathbf{1}_H + \mathbf{1}_E\mathbf{1}_B + \mathbf{1}_E\mathbf{1}_D + \mathbf{1}_G\mathbf{1}_B + \mathbf{1}_G\mathbf{1}_D\right)\Bigg]
\end{aligned}$$

$$\begin{aligned}
= \frac{1}{4}\Bigg[&\left(1 - c_{jl}^{(1)}\right)^2\left(\mathbf{1}_A\mathbf{1}_B + \mathbf{1}_C\mathbf{1}_D + \mathbf{1}_E\mathbf{1}_F + \mathbf{1}_G\mathbf{1}_H\right) \\
&+ c_{jl}^{(1)}\left(1 - c_{jl}^{(1)}\right)\left(\mathbf{1}_A\mathbf{1}_D + \mathbf{1}_C\mathbf{1}_B + \mathbf{1}_E\mathbf{1}_H + \mathbf{1}_G\mathbf{1}_F\right) \\
&- \frac{1}{4}\left(\mathbf{1}_A\mathbf{1}_B + \mathbf{1}_C\mathbf{1}_D + \mathbf{1}_E\mathbf{1}_F + \mathbf{1}_G\mathbf{1}_H\right) - \frac{1}{4}\left(\mathbf{1}_A\mathbf{1}_D + \mathbf{1}_C\mathbf{1}_B + \mathbf{1}_E\mathbf{1}_H + \mathbf{1}_G\mathbf{1}_F\right) \\
&- \frac{1}{4}\left(1 - 2c_{jl}^{(1)}\right)\left(\mathbf{1}_A\mathbf{1}_F + \mathbf{1}_A\mathbf{1}_H + \mathbf{1}_C\mathbf{1}_F + \mathbf{1}_C\mathbf{1}_H + \mathbf{1}_E\mathbf{1}_B + \mathbf{1}_E\mathbf{1}_D + \mathbf{1}_G\mathbf{1}_B\right. \\
&\left. + \mathbf{1}_G\mathbf{1}_D\right)\Bigg]
\end{aligned}$$

$$= \frac{1}{4}\Big[\Big(1-c_{jl}^{(1)}\Big)\big(\mathbf{1}_A\mathbf{1}_B+\mathbf{1}_C\mathbf{1}_D+\mathbf{1}_E\mathbf{1}_F+\mathbf{1}_G\mathbf{1}_H\big) - c_{jl}^{(1)}\Big(1-c_{jl}^{(1)}\Big)\big(\mathbf{1}_A\mathbf{1}_B+\mathbf{1}_C\mathbf{1}_D+\mathbf{1}_E\mathbf{1}_F+\mathbf{1}_G\mathbf{1}_H-\mathbf{1}_A\mathbf{1}_D-\mathbf{1}_C\mathbf{1}_B-\mathbf{1}_E\mathbf{1}_H - \mathbf{1}_G\mathbf{1}_F\big) - \frac{1}{4}\big(\mathbf{1}_A\mathbf{1}_B+\mathbf{1}_C\mathbf{1}_D+\mathbf{1}_E\mathbf{1}_F+\mathbf{1}_G\mathbf{1}_H+\ \mathbf{1}_A\mathbf{1}_D+\mathbf{1}_C\mathbf{1}_B+\mathbf{1}_E\mathbf{1}_H+\mathbf{1}_G\mathbf{1}_F\big) - \frac{1}{4}\Big(1-2c_{jl}^{(1)}\Big)\big(\mathbf{1}_A\mathbf{1}_F+\mathbf{1}_A\mathbf{1}_H+\mathbf{1}_C\mathbf{1}_F+\mathbf{1}_C\mathbf{1}_H+\mathbf{1}_E\mathbf{1}_B+\mathbf{1}_E\mathbf{1}_D+\mathbf{1}_G\mathbf{1}_B + \mathbf{1}_G\mathbf{1}_D\big)\Big]$$

$$= \frac{1}{4}\Big[\Big(1-c_{jl}^{(1)}\Big)\big(\mathbf{1}_A\mathbf{1}_B+\mathbf{1}_C\mathbf{1}_D+\mathbf{1}_E\mathbf{1}_F+\mathbf{1}_G\mathbf{1}_H\big) - 4c_{jl}^{(1)}\Big(1-c_{jl}^{(1)}\Big)\big(D_{jl}^{12}+\ D_{jl}^{34}\big) - \frac{1}{4}\big(\mathbf{1}_A\mathbf{1}_B+\mathbf{1}_C\mathbf{1}_D+\mathbf{1}_E\mathbf{1}_F+\mathbf{1}_G\mathbf{1}_H+\ \mathbf{1}_A\mathbf{1}_D+\mathbf{1}_C\mathbf{1}_B+\mathbf{1}_E\mathbf{1}_H+\mathbf{1}_G\mathbf{1}_F\big) + \frac{1}{4}\Big(1-2c_{jl}^{(1)}\Big)\big(4D_{jl}^{13}+4D_{jl}^{14}+4D_{jl}^{23}+\ 4D_{jl}^{24}-2\mathbf{1}_A\mathbf{1}_B-2\mathbf{1}_C\mathbf{1}_D-2\mathbf{1}_E\mathbf{1}_F - 2\mathbf{1}_G\mathbf{1}_H\big)\Big]$$

$$= \frac{1}{4}\Big[\Big(1-c_{jl}^{(1)}-\frac{1}{4}-\frac{2}{4}\Big(1-2c_{jl}^{(1)}\Big)\Big)\big(\mathbf{1}_A\mathbf{1}_B+\mathbf{1}_C\mathbf{1}_D+\mathbf{1}_E\mathbf{1}_F+\ \mathbf{1}_G\mathbf{1}_H\big) - 4c_{jl}^{(1)}\Big(1-c_{jl}^{(1)}\Big)\big(D_{jl}^{12}+\ D_{jl}^{34}\big) - \frac{1}{4}\big(\mathbf{1}_A\mathbf{1}_D+\mathbf{1}_C\mathbf{1}_B+\mathbf{1}_E\mathbf{1}_H+\mathbf{1}_G\mathbf{1}_F\big) + \frac{1}{4}\Big(1-2c_{jl}^{(1)}\Big)\big(4D_{jl}^{13}+4D_{jl}^{14}+4D_{jl}^{23}+4D_{jl}^{24}\big)\Big]$$

$$= \frac{1}{4}\Big[\frac{1}{4}\big(\mathbf{1}_A\mathbf{1}_B+\mathbf{1}_C\mathbf{1}_D+\mathbf{1}_E\mathbf{1}_F+\mathbf{1}_G\mathbf{1}_H-\mathbf{1}_A\mathbf{1}_D-\mathbf{1}_C\mathbf{1}_B-\mathbf{1}_E\mathbf{1}_H-\mathbf{1}_G\mathbf{1}_F\big) - 4c_{jl}^{(1)}\Big(1-c_{jl}^{(1)}\Big)\big(D_{jl}^{12}+\ D_{jl}^{34}\big)+\Big(1-2c_{jl}^{(1)}\Big)\big(D_{jl}^{13}+\ D_{jl}^{14}+D_{jl}^{23}+D_{jl}^{24}\big)\Big]$$

$$= \frac{1}{4}\Big[\frac{1}{4}\big(4D_{jl}^{12}+4D_{jl}^{34}\big)-\ 4c_{jl}^{(1)}\Big(1-c_{jl}^{(1)}\Big)\big(D_{jl}^{12}+\ D_{jl}^{34}\big) + \Big(1-2c_{jl}^{(1)}\Big)\big(D_{jl}^{13}+D_{jl}^{14}+D_{jl}^{23}+D_{jl}^{24}\big)\Big]$$

$$= \frac{1}{4}\Big[\big(D_{jl}^{12}+\ D_{jl}^{34}\big)\Big(\frac{1}{4}-4c_{jl}^{(1)}(1-c_{jl}^{(1)})\Big)+\Big(1-2c_{jl}^{(1)}\Big)\big(D_{jl}^{13}+D_{jl}^{14}+D_{jl}^{23}+D_{jl}^{24}\big)\Big]$$

$$= \frac{1}{4}\Big[\ \Big(1-2c_{jl}^{(1)}\Big)^2\big(D_{jl}^{12}+\ D_{jl}^{34}\big)+\Big(1-2c_{jl}^{(1)}\Big)\big(D_{jl}^{13}+D_{jl}^{14}+D_{jl}^{23}+D_{jl}^{24}\big)\Big]$$

$$= \frac{1}{4}\Big(1-2c_{jl}^{(1)}\Big)\Big[\ \big(D_{jl}^{13}+D_{jl}^{14}+D_{jl}^{23}+D_{jl}^{24}\big)+\Big(1-2c_{jl}^{(1)}\Big)\big(D_{jl}^{12}+\ D_{jl}^{34}\big)\Big]$$

$$= \tfrac{1}{4}\Big(1-2c_{jl}^{(1)}\Big)\Big[\Phi_{2\ jl}+\Big(1-2c_{jl}^{(1)}\Big)\Phi_{1\ jl}\Big] \qquad (5)$$

with $\Phi_{1\ jl} = D_{jl}^{12} + D_{jl}^{34}$ summing the LD values among parents that can be considered to be involved as biparental crosses in $F_1^{(1)} \times F_1^{(2)}$ and with $\Phi_{2\ jl} = D_{jl}^{13} + D_{jl}^{14} + D_{jl}^{23} + D_{jl}^{24}$ summing the LD values among parents that can be considered to be involved as biparental crosses in $F_1' \times F_1'$.

The linkage disequilibrium parameter $\Phi_1$and Φ_2 and equation (5) can be simplified in the case of two-way, three-way and backcrosses (Table 2 S1). For two-way crosses we arrive at the same variance covariance matrix elements Σ_{jl} as given by Lehermeier *et al.* (2017).

Table 2 S1 Linkage disequilibrium parameter between QTLs j and l in pairs of parental lines depending on the mating design.

	$\Phi_{1\ jl}$	$\Phi_{2\ jl}$	Σ_{jl}
Four-way	$D_{jl}^{12} + D_{jl}^{34}$	$D_{jl}^{13} + D_{jl}^{14} + D_{jl}^{23} + D_{jl}^{24}$	$\left(1 - 2c_{jl}^{(1)}\right)\left(\left(D_{jl}^{13} + D_{jl}^{14} + D_{jl}^{23} + D_{jl}^{24}\right) + \left(1 - 2c_{jl}^{(1)}\right)\left(D_{jl}^{12} + D_{jl}^{34}\right)\right)$
Three-way	D_{jl}^{12}	$2\ (D_{jl}^{14} + D_{jl}^{24})$	$\left(1 - 2c_{jl}^{(1)}\right)\left(2\ \left(D_{jl}^{14} + D_{jl}^{24}\right) + \left(1 - 2c_{jl}^{(1)}\right) D_{jl}^{12}\right)$
Backcross	D_{jl}^{14}	$2\ D_{jl}^{14}$	$\left(1 - 2c_{jl}^{(1)}\right)\left(3\ - 2c_{jl}^{(1)}\right)\ D_{jl}^{14}$
Two-way	0	$4\ D_{jl}^{14}$	$4\ \left(1 - 2c_{jl}^{(1)}\right)\ D_{jl}^{14}$

Cited literature:

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.

File S2: Validation of four-way cross formulas for DH-k and RIL-k and evolution of RIL variance depending on selfing generations

In File S1, we considered DH lines generated from F1' (DH-1), i.e., only two meioses took place. Progeny variance for DH-1 is expressed in terms of parental expected recombination frequency $c^{(1)}$ (Table 2 S1). For recombinant inbred lines (RILs) or when DH lines are generated from higher selfing generations, the expected frequency of recombinants increases depending on the number of selfing generations. In the following k denotes the generation from which progeny are derived (Figure 1). The expected frequency of recombinants in generation k can be derived from the genotype probabilities given in Broman (2012) as done in File S1 of Lehermeier *et al.* (2017). Hence, for DH lines after k generations, $c^{(1)}$ in Table 2 S1 should then be replaced by $c^{(k)}$, leading to the general four-way DH-k formula as shown in Table 1:

$$c^{(k)} = \frac{2c^{(1)}}{1+2c^{(1)}}\left(1-0.5^k\left(1-2c^{(1)}\right)^k\right), \forall\, k \in \mathbb{N}^*$$

In case of RILs, no doubling of gametes takes place and the covariance for RILs after generation k is obtained by updating $c^{(k)}$ by $c^{(k)} + 0.5\left[0.5\left(1-2c^{(1)}\right)\right]^k, \forall\, k \in \mathbb{N}^*$ (Table 1). Note that the variance-covariance of DH-k and RIL-k converge with increasing k.

Formulas for DH-k and RIL-k in the general case of four-way crosses have been validated by simulations for $k \in [\![1,6]\!]$ (Table 1 S2 and Table 2 S2). The observed high positive correlations (Table 1 S2) and low mean squared differences (Table 2 S2) between predicted (derivation) and empirical (*in silico*) values validate the presented formulas. Lower squared correlations between predicted and empirical values were observed for $\boldsymbol{\mu}_{\boldsymbol{C}}^{(\boldsymbol{sel})}$ and $\boldsymbol{\mu}_{\boldsymbol{C(+)}}^{(\boldsymbol{sel})}$ compared to the variances and covariances. This can be explained by sampling bias in *in silico* simulations (50,000 progenies) where the P_1 parental genome contribution before selection slightly differed from the expected value of 0.25 for four way crosses (ranging from 0.249 to 0.251).

Predicted RIL progeny variance for the simulated agronomic trait increased with the number of selfing generations considered (k) and converged toward DH progeny variance after five generations of selfing ($k = 5$) (Figure 1 S2). We observed that some crosses profited more from an increase in selfing generations by generating more variance compared to others. An example with two crosses is shown in Figure 2 S2. While the cross visualized in blue showed a higher variance in generation RIL-1 than the cross visualized in orange, it reached a plateau faster and showed a lower variance than the orange cross with $k \geq 3$. Differences in the speed to release variance between crosses is likely due to differences in the recombination frequency between segregating QTLs in parental lines. This underlines the interest of predicting RIL progeny variance using proposed algebraic formula.

Table 1 S2 Squared correlations (R^2) between empirical values (*in silico*) and predictions (derivation) per generation and type of progeny.

Generation	σ_T^2	σ_C^2	$\sigma_{C(+)}^2$	$\sigma_{T,C}$	$\sigma_{T,C(+)}$	UC_T	$\mu_C^{(sel)}$	$\mu_{C(+)}^{(sel)}$
DH1	0.999	0.960	0.995	0.999	0.999	1.000	0.900	0.946
DH2	0.999	0.964	0.995	0.998	0.998	1.000	0.909	0.952
DH3	0.999	0.966	0.995	0.999	0.999	1.000	0.914	0.955
DH4	0.999	0.969	0.995	0.999	0.999	1.000	0.912	0.955
DH5	0.999	0.961	0.994	0.998	0.998	1.000	0.914	0.955
DH6	0.999	0.963	0.994	0.998	0.998	1.000	0.913	0.955
RIL1	0.999	0.957	0.994	0.999	0.999	1.000	0.938	0.967
RIL2	0.999	0.957	0.994	0.999	0.999	1.000	0.917	0.957
RIL3	0.999	0.960	0.994	0.998	0.998	1.000	0.918	0.958
RIL4	0.999	0.962	0.994	0.998	0.998	1.000	0.915	0.956
RIL5	0.999	0.962	0.994	0.998	0.998	1.000	0.912	0.955
RIL6	0.999	0.962	0.994	0.999	0.998	1.000	0.911	0.954

Table 2 S2 Mean squared difference between empirical values (*in silico*) and predictions (derivation) per generation and type of progeny.

Generation	σ_T^2	σ_C^2	$\sigma_{C(+)}^2$	$\sigma_{T,C}$	$\sigma_{T,C(+)}$	UC_T	$\mu_C^{(sel)}$	$\mu_{C(+)}^{(sel)}$
DH1	5.20E-06	3.28E-09	3.52E-10	5.99E-08	2.07E-08	8.44E-04	4.92E-05	1.42E-05
DH2	5.09E-06	2.81E-09	3.16E-10	6.65E-08	2.24E-08	7.02E-04	3.83E-05	1.12E-05
DH3	5.36E-06	2.56E-09	2.97E-10	4.74E-08	1.51E-08	6.49E-04	3.50E-05	1.03E-05
DH4	4.56E-06	2.30E-09	2.87E-10	5.16E-08	1.66E-08	6.85E-04	3.55E-05	1.05E-05
DH5	4.83E-06	2.88E-09	3.32E-10	5.95E-08	1.99E-08	6.40E-04	3.47E-05	1.03E-05
DH6	4.76E-06	2.74E-09	3.14E-10	6.08E-08	1.96E-08	6.77E-04	3.47E-05	1.04E-05
RIL1	2.25E-06	1.56E-09	1.81E-10	2.96E-08	9.80E-09	4.30E-04	2.51E-05	7.54E-06
RIL2	3.26E-06	2.29E-09	2.69E-10	4.09E-08	1.37E-08	5.73E-04	3.40E-05	1.00E-05
RIL3	3.93E-06	2.58E-09	3.05E-10	5.28E-08	1.72E-08	6.22E-04	3.34E-05	9.84E-06
RIL4	4.49E-06	2.59E-09	3.02E-10	5.64E-08	1.81E-08	6.59E-04	3.43E-05	1.01E-05
RIL5	4.91E-06	2.69E-09	3.10E-10	5.59E-08	1.83E-08	6.65E-04	3.53E-05	1.04E-05
RIL6	4.91E-06	2.71E-09	3.13E-10	5.54E-08	1.83E-08	6.63E-04	3.59E-05	1.06E-05

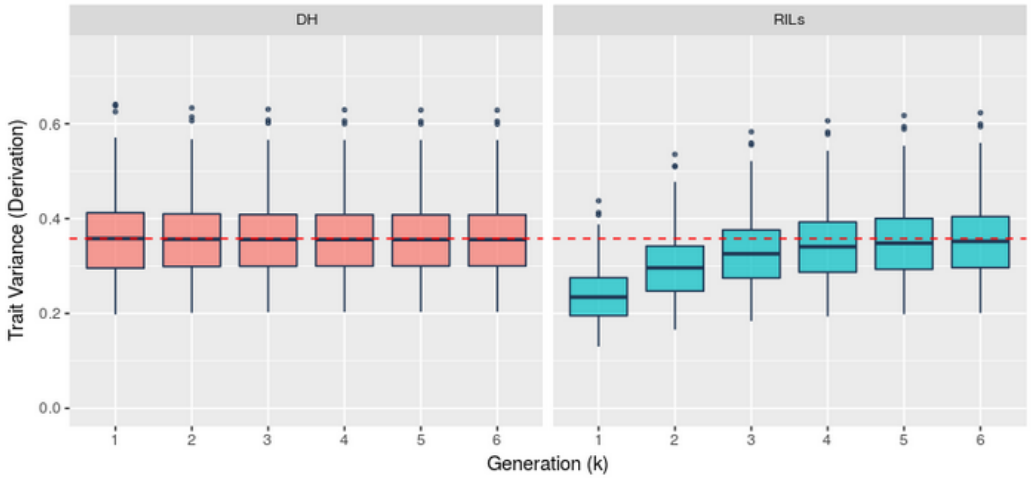


Figure 1 S2 Evolution of predicted progeny trait variance depending on progeny type (DH, left or RIL, right) and generation (k). The red dotted line presents the median DH progeny variance over 100 crosses.

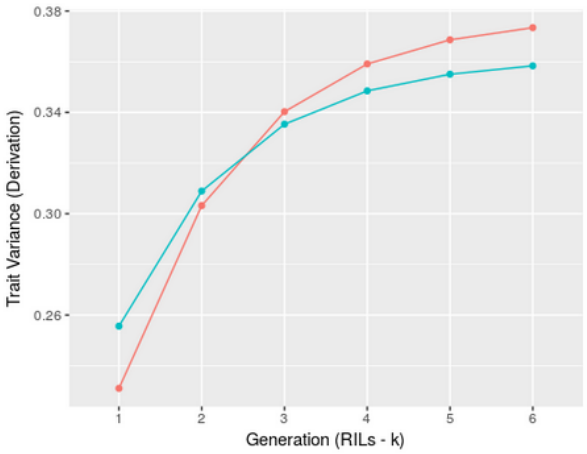


Figure 2 S2 Example of two crosses showing different evolutions of predicted RIL progeny variance depending on the selfing generation (k).

Cited literature:

Broman K. W., 2012 Genotype probabilities at intermediate generations in the construction of recombinant inbred lines. Genetics 190: 403–412.

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.

File S3: Comparison of IBD parental contribution variance with Frisch and Melchinger (2007) and simplification to IBS contribution

We used an algebraic formula to predict the variance of P_1 genome contribution in doubled haploid progeny derived from F1' plants. We considered two-way crosses DH-1 (called (F1)-DH) and backcrosses DH-1 (called (BC1)-DH) and compared our results with the results given by Frisch and Melchinger (2007). We considered one chromosome of 100cM for which Frisch and Melchinger (2007) derived a variance of parental contribution of 0.1419 for (F1)-DH and 0.0945 for (BC1)-DH. We varied the number of loci p used in our approach and for each, we ran ten independent samplings of loci. We observed that the results from our approach converged with increasing number of loci to the solution given by Frisch and Melchinger (2007) (Figure 1 S3).

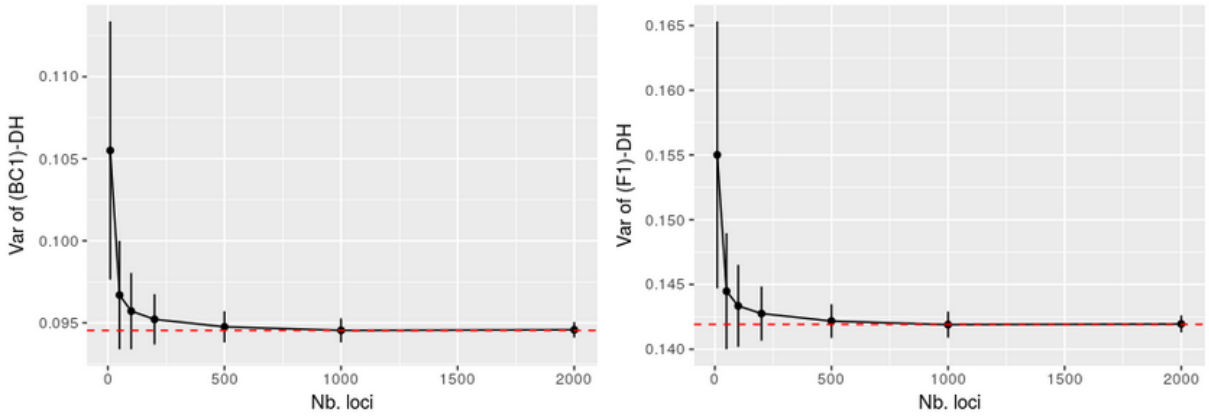


Figure 1 S3 Average parental genome contribution variance (black dots) for (BC1)-DH (left) and (F1)-DH (right) from ten simulation replications (+/- standard deviation represented by black vertical lines) with different number of considered loci. Red dotted line shows the results given by Frisch and Melchinger (2007).

In cases where the origin of the allele is not of interest and an identical by state (IBS) similarity between progeny and parental lines is sufficient, the multi-allelic coding can be simplified to a biallelic coding. This reduces the size of the covariance matrix from $(4p \text{ x } 4p)$ to $(p \text{ x } p)$, with p being the number of loci considered. For this, let us define the genotyping matrix of parental lines in biallelic coding:

$$\boldsymbol{X}_{IBS} = diag(\boldsymbol{X}_{Parental}) = (\boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3\ \boldsymbol{x}_4)'$$

where, $\boldsymbol{X}_{IBS}$ is a $(4 \text{ x } p)$-dimensional matrix of genotypes. The $(p \text{ x } 4)$-dimensional matrix of global parental contribution marker effects for each of the four parents can be defined as:

$$\boldsymbol{\beta}_{IBS} = \frac{1}{2p}\boldsymbol{X}_{IBS}{}'$$

where, $\forall\, i \in [1;4]$ $\boldsymbol{\beta}_{IBS}(.,i)$ is the p-dimensional vector of marker effect to follow the IBS contribution of parent i and p is the total number of loci considered.

We denote the $(N \text{ x } p)$-dimensional genotyping matrix of N doubled haploid (DH) progeny as $\boldsymbol{X}_{IBS-Progeny}$ with element $\boldsymbol{X}_{IBS-Progeny}\ (j,l), \forall\, j \in [1,N], l \in [1,p]$ the genotype of progeny j at

locus l coded as -1, 1 for the genotypes aa, AA, respectively. It results in the following $(N \text{ x } 4)$-dimensional matrix of parental IBS contribution to progeny:

$$\boldsymbol{C}_{IBS} = \boldsymbol{X}_{IBS-Progeny}\,\boldsymbol{\beta}_{IBS} + \frac{1}{2}\mathbf{1}_N\mathbf{1}_4'$$

where, $\forall\, j \in [1; N], \forall\, i \in [1; 4]$, $\boldsymbol{C}_{IBS}(j, i)$ is the parental line i contribution to progeny line j.

Cited literature:

Frisch M., and A. E. Melchinger, 2007 Variance of the Parental Genome Contribution to Inbred Lines Derived From Biparental Crosses. Genetics 176: 477–488.

Supplementary Material Chapter 4

File S1: Additional material

Material

We initiated simulations with the genome of 57 maize Iodent inbred lines (*Zea mays L.*) (Allier *et al.* 2019). These lines were genotyped with the Illumina MaizeSNP50 BeadChip (Ganal *et al.* 2011). After quality control and imputation, 40,478 high-quality SNPs were retained. The genetic map was obtained by predicting genetic positions from physical positions on the reference genome B73-v4 (Jiao *et al.* 2017) using a spline-smoothing interpolating procedure described in Bauer *et al.* (2013) and the dent genetic map in Giraud *et al.* (2014). At each simulation replicate we randomly sampled 40 lines to be the founder population. We randomly sampled 1,000 SNPs to be additive biallelic quantitative trait loci (QTL) of a polygenic trait. The sampling of QTL obeyed two constraints: QTL minor allele frequency $\geq$ 0.2 and distance between two consecutive QTL $\geq$ 0.2 cM. Each QTL was randomly assigned an additive effect from a Gaussian distribution with a mean of zero and a variance of 0.05. For the scenario where the 1,000 QTLs were unknown, we randomly sampled 2,000 non causal SNPs as genomewide markers used for evaluation (see "Evaluation model" section).

Simulation scheme

We aimed at comparing the effect of parent selection and allocation methods on short and long term genetic gains in a realistic breeding context using doubled haploid (DH) technology and considering overlapping and connected cohorts (i.e. generations) of three years as illustrated in Figure 1A. We considered that the process to derive DH lines from a cross and to phenotype and genotype DH lines took three years. Furthermore we considered as candidate parents of a new cohort only the DH progeny of the three last cohorts. For sake of clarity, the candidate parents of cohort T were selected from the available DH progeny of the three cohorts: $T-3$, $T-4$ and $T-5$ (Figure 1A-B). Within this breeding context, we defined a burn-in period of 20 years starting from founders that mimicked a phenotyping selection (PS) program using DH technology (more details in the "phenotyping" and "evaluation model" sections). Afterward, we compared different cross selection strategies during 60 years of breeding. We considered either that we had access to the 1,000 QTL effects (TRUE scenario) or that we estimated the effects of the 2,000 non causal SNPs (GS scenario). We also considered the absence of genomic information for selection, i.e. phenotypic selection (PS scenario).

We can distinguish the following simulation phases for the cohorts $\boldsymbol{T} \in [\![\mathbf{1}, \mathbf{80}]\!]$:

- **Burn-in Phase 1 ($\boldsymbol{T} \in [\![\mathbf{1}; \mathbf{3}]\!]$): Initialization**

Every year during the three first years, a cohort was initiated by randomly generating 20 biparental crosses from the 40 founders. We derived 80 DH lines per cross. Note that lines can contribute as parents to different crosses and cohorts, so that parental contributions are not controlled and different cohorts can share the same crosses at this stage.

- **Burn-in Phase 2 ($\boldsymbol{T} \in [\![\mathbf{4}; \mathbf{20}]\!]$)**

The second phase of burn-in mimicked 17 years of phenotypic selection to build up extensive linkage disequilibrium to compare scenarios in a realistic ongoing breeding context. In burn-in phase 2, phenotypic selection (PS) was used to estimate breeding value of candidate lines from the three last cohorts ($T-3$, $T-4$ and $T-5$, if available). After selecting the 4 best DH progeny per family (i.e.

5%), the overall 50 best progeny out of 3 cohorts x 20 families/cohort x 4 DH/family = 240 DH progeny were considered as potential parents of the cohort and were randomly mated to generate 20 biparental families of 80 DH lines. Note that lines can contribute as parents to different crosses and cohorts, so that parental contributions are not controlled and different cohorts can share the same crosses at this stage. Burn-in ended up with overlapping cohorts connected by the pedigree as it can be found in real breeding program.

- **Post burn-in ($\boldsymbol{T \in [\![21; 70]\!]}$)**

In post burn-in, the life cycle of a cohort was similar to burn-in phase 2 except changes in the way to evaluate, select and mate parents (Figure 1B).

Phenotyping

For phenotyping, we considered environmental effects sampled in a normal distribution of mean zero and variance 25 and did not consider genotype by environment interactions. Each cohort was evaluated in $N_{loc} = 4$ locations in one year, i.e. four environments. At each simulation replicate, five founder lines were randomly sampled to be check individuals phenotyped every year. Environmental errors were sampled from a normal distribution with mean zero and an error variance σ_ϵ^2 defined by the initial repeatability in the founder population $r = \frac{\sigma_G^2}{\sigma_G^2+\sigma_\epsilon^2} = 0.40$. This led to a heritability in the founder population of $h^2 = \frac{\sigma_G^2}{\sigma_G^2+\sigma_\epsilon^2/N_{loc}} = 0.73$. Note that the repeatability and heritability varied along selection cycles relatively to the evolution of additive genetic variance σ_G^2 (e.g. $h^2 = 0.73$ in founder population to $h^2 = 0.59$ at the end of burn-in and to $h^2 = 0.03$ after 60 years in the PS scenario).

Evaluation model

Different evaluation models were considered and should be distinguished at this stage. For phenotypic selection (PS scenario), the phenotypes of progeny were used to estimate their breeding values (EBV). We distinguished two scenarios using genomic information. On one hand, the 1,000 QTL positions and effects were known (TRUE scenario) and the evaluation consisted in summing the individual additive QTL effects to obtain the true breeding value (TBV) of progeny. On the other end, the 1,000 QTL positions and effects were unknown (GS scenario) and 2,000 SNP effects were estimated using the phenotypes and genotypes of the progeny from the three last cohorts. The progeny were selected on their genomic estimated breeding values (GEBV).

The breeding value of progeny (EBV in PS or GEBV in GS) were estimated in Model 1 S1 fitted using mixed model software blup-f 90 (Misztal 2008) with AI-REML variance component estimates:

$$\boldsymbol{Y} = \mathbf{1}\mu + \boldsymbol{E\beta_{Env}} + \boldsymbol{Wu} + \boldsymbol{\epsilon}, \text{ (Model 1 S1)}$$

where $\boldsymbol{Y}$ is the vector of phenotypic values, μ is the intercept, $\boldsymbol{E}$ is the incidence matrix for environmental effects, $\boldsymbol{\beta_{Env}}$ is the vector of environmental fixed effects, $\boldsymbol{W}$ is the incidence matrix of individual breeding value random effects $\boldsymbol{u}$, $\boldsymbol{u} \sim N(\mathbf{0}, \sigma_G^2\boldsymbol{U})$ is the vector of breeding value random effects with $\sigma_G^2\boldsymbol{U}$ its variance-covariance matrix and $\boldsymbol{\epsilon}$ is the vector of residual random terms $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2\boldsymbol{I})$ independent and identically distributed. For phenotypic selection (PS), the individuals were assumed independent, i.e. $\boldsymbol{u} \sim N(\mathbf{0}, \sigma_G^2\boldsymbol{I})$. For genomic selection (GS), the covariance between

individuals was modeled using the genomic relationship matrix $\boldsymbol{G}$, i.e. $\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma_G^2 \boldsymbol{G})$. Hereby, $\boldsymbol{G}$ was estimated using the 2,000 non causal loci as:

$$\boldsymbol{G} = \frac{\boldsymbol{Z}\boldsymbol{Z}'}{tr(\boldsymbol{Z}\boldsymbol{Z}')/n}$$

where, $\boldsymbol{Z}$ contains the centered allele counts, with elements computed as $x_{ij} + 1 - 2p_j$, where the element $x_{ij} \in \{-1,1\}$ is the genotype for individual i at non causal locus j and p_j is the frequency of the allele for which the homozygous genotype is coded 1 at non causal locus j. $tr(\boldsymbol{Z}\boldsymbol{Z}')$ is the trace of $\boldsymbol{Z}\boldsymbol{Z}'$ and $tr(\boldsymbol{Z}\boldsymbol{Z}')/n$ forces the diagonal of $\boldsymbol{G}$ to be 1 on average (Legarra *et al.* 2009; Forni *et al.* 2011). Note that for fully homozygous individuals $tr(\boldsymbol{Z}\boldsymbol{Z}')/n = 4\sum_j p_j(1-p_j)$. Estimated marker effects $\widehat{\boldsymbol{\beta}_T}$ were obtained by back-solving: $\widehat{\boldsymbol{\beta}_T} = \boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\hat{\boldsymbol{u}}$ (Wang *et al.* 2012) and used in lieu of known QTL effects $\boldsymbol{\beta_T}$.

Simulation of progeny genotypes

Doubled haploid progeny genotypes were simulated considering meiosis events without crossover interference. The number of chiasmata was drawn from a Poisson distribution with λ equal to the chromosome length in Morgan, and crossover positions were determined using the recombination frequency obtained using the Haldane mapping function (Haldane 1919).

Cited literature:

Allier A., L. Moreau, A. Charcosset, S. Teyssèdre, and C. Lehermeier, 2019 Usefulness Criterion and Post-selection Parental Contributions in Multi-parental Crosses: Application to Polygenic Trait Introgression. G3 Genes Genomes Genet. 9: 1469–1479.

Bauer E., M. Falque, H. Walter, C. Bauland, C. Camisan, et al., 2013 Intraspecific variation of recombination rate in maize. Genome Biol. 14: R103.

Forni S., I. Aguilar, and I. Misztal, 2011 Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet. Sel. Evol. 43: 1.

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.

Giraud H., C. Lehermeier, E. Bauer, M. Falque, V. Segura, et al., 2014 Linkage Disequilibrium with Linkage Analysis of Multiline Crosses Reveals Different Multiallelic QTL for Hybrid Performance in the Flint and Dent Heterotic Groups of Maize. Genetics 198: 1717–1734.

Haldane J., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. J Genet 8: 299–309.

Jiao Y., P. Peluso, J. Shi, T. Liang, M. C. Stitzer, et al., 2017 Improved maize reference genome with single-molecule technologies. Nature 546: 524–527.

Legarra A., I. Aguilar, and I. Misztal, 2009 A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92: 4656–4663.

Misztal I., 2008 Reliable computing in estimation of variance components. J. Anim. Breed. Genet. 125: 363–370.
Wang H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir, 2012 Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. 94: 73–83.

File S2: Relationship between IBS coancestry and genetic diversity in progeny

The identity by state (IBS) coancestry between N inbred parents is defined as:

$$\boldsymbol{K} = 0.5\left(\frac{1}{m}\,(\boldsymbol{X}\boldsymbol{X}') + \mathbf{1}_N\mathbf{1}_N'\right), \text{(Eq. 1)}$$

where, $\boldsymbol{X}$ is the genotyping matrix of the N parents in line and m loci in column, with elements coded -1 or 1 and $\mathbf{1}_N$ is a N-dimensional column vector of ones.

Considering the N-dimensional column vector of expected parental genomewide contributions $\boldsymbol{c}$, with $c(j),\ j \in [1, N]$ the contribution of the parent j to progeny, the mean expected IBS coancestry in progeny is:

$$IBS = \boldsymbol{c}'\boldsymbol{K}\,\boldsymbol{c} = 0.5\left[\frac{1}{m}(\boldsymbol{c}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{c}) + \boldsymbol{c}'\mathbf{1}_N\mathbf{1}_N'\boldsymbol{c}\right]. \text{(Eq. 2a)}$$

Note that $\boldsymbol{c}'\mathbf{1}_N\mathbf{1}_N'\boldsymbol{c} = 1$ since $\sum_{j=1}^{N}\boldsymbol{c}(j) = 1$. Then, Eq. 2a simplifies:

$$IBS = 0.5\left[\frac{1}{m}(\boldsymbol{c}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{c}) + 1\right] \text{(Eq. 2b)}$$

The mean expected genetic diversity (He) in progeny is:

$$He = \frac{1}{m}\,\mathbf{1}_m'\left(2\,\boldsymbol{p}\,\circ(\mathbf{1}_m - \boldsymbol{p})\right), \text{(Eq. 3a)}$$

where $\mathbf{1}_m$ is a m-dimensional column vector of ones, $\circ$ is the pairwise entry product and $\boldsymbol{p}$ is the m-dimensional column vector of expected allelic frequencies in progeny:

$$\boldsymbol{p} = 0.5\,((\boldsymbol{X} + \mathbf{1}_N\mathbf{1}_m')'\,\circ\mathbf{C})\,\mathbf{1}_N, \text{(Eq. 4a)}$$

where $\mathbf{C}$ is the $(m \text{ x } N)$-dimensional matrix of expected local parental contributions to progeny with $C(i,j),\ i \in [1, m],\ j \in [1, N]$ the contribution of parent j to progeny at the locus i. $C(i,j), \forall\, i \in [1, m]$ is further approximated by the genomewide parental contribution to progeny $c(j)$. Consequently, the m-dimensional column vector of expected allelic frequencies (Eq. 4a) is approximated as:

$$\widetilde{\boldsymbol{p}} = 0.5\,(\boldsymbol{X} + \mathbf{1}_N\mathbf{1}_m')'\boldsymbol{c}. \text{(Eq. 4b)}$$

We replace $\boldsymbol{p}$ by its approximation $\widetilde{\boldsymbol{p}}$ in Eq. 3a:

$$\widetilde{He} = \frac{1}{m}\,\mathbf{1}_m'\left((\boldsymbol{X}'\boldsymbol{c} + \mathbf{1}_m\mathbf{1}_N'\boldsymbol{c})\circ(\mathbf{1}_m - 0.5\,\boldsymbol{X}'\boldsymbol{c} - 0.5\,\mathbf{1}_m\mathbf{1}_N'\boldsymbol{c})\right). \text{(Eq. 5a)}$$

Note that $\mathbf{1}_m\mathbf{1}_N'\boldsymbol{c} = \mathbf{1}_m$ and Eq. 5a becomes:

$$\widetilde{He} = \frac{1}{m}\,\mathbf{1}_m'\left((\boldsymbol{X}'\boldsymbol{c} + \mathbf{1}_m)\circ(0.5\,\mathbf{1}_m - 0.5\,\boldsymbol{X}'\boldsymbol{c})\right)$$

$$= \frac{1}{m}\,\mathbf{1}_m'\left(0.5\,(\mathbf{1}_m - \boldsymbol{X}'\boldsymbol{c}\circ\boldsymbol{X}'\boldsymbol{c})\right)$$

$$= 0.5 \left(1 - \frac{1}{m} \mathbf{1}'_m \left(\boldsymbol{X}'\boldsymbol{c} \circ \boldsymbol{X}'\boldsymbol{c}\right)\right).$$ (Eq. 5b)

Let us note $\boldsymbol{v} = \boldsymbol{X}'\boldsymbol{c}$. It can be shown that $\mathbf{1}'_m \left(\boldsymbol{v} \circ \boldsymbol{v}\right) = \boldsymbol{v}'\boldsymbol{v}$, resulting in:

$$\widetilde{He} = 0.5 \left(1 - \frac{1}{m} \left(\boldsymbol{c}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{c}\right)\right) = 0.5(1 - 2\,IBS + 1) = 1 - IBS.$$ (Eq. 6)

Note that this equivalence is conserved whether we consider ante- or post-selection parental contributions ($\boldsymbol{c}$), respectively in OCS or in UCPC based OCS.

File S3: Supporting R code

Improving short and long term genetic gain by accounting for within family variance in optimal cross selection

File S3 - Supporting R code

Allier Antoine, Christina Lehermeier, Alain Charcosset, Laurence Moreau and Simon Teyssèdre

Contents

This document illustrates how the UCPC (Usefulness Criterion Parental Contribution, described in Allier et al. (2019) in G3 5:1469-1479, doi: https://doi.org/10.1534/g3.119.400129) is used to predict the expected gain and parental contributions after progeny selection for a cross between two homozygous parental lines. We further extend it to the evaluation of a set of crosses (**nc**) underlying the UCPC based optimal cross selection.

1 Load examplary data

For this documentation, we load a simulated maize data set from the `synbreedData` R package, which includes genotypic data and a genetic map.

```
rm(list=ls())
library(synbreed)
library(synbreedData)
set.seed(1993)
# Use simulated maize data set from synbreedData package as example data
data(maize)
# Convert genotypes into -1, 1 coding
geno <- maize$geno*2-1
# Set a genetic map object
map <- data.frame(CHROMOSOME=maize$map$chr,
                  POSITION=maize$map$pos,
                  MARKER=rownames(maize$map))
```

2 Usefulness criterion and parental contribution (UCPC) for a single two-way cross

We sample two homozygous parents further referred to as P_1 and P_2.

```
# Sample illustrative parental lines
xP1 <- geno[10,]
xP2 <- geno[200,]
```

2.1 Genotypic covariance in doubled haploid progeny (DH-1) of P_1 x P_2

The following function computes the genotypic covariance matrix $\mathbf{\Sigma}$ in doubled haploid (DH) progeny derived from the cross between homozygous lines P_1 and P_2 as described in Lehermeier et al. (2017) Genetics 207:1651-1661.

Function arguments:

- ObjectGenoP1: Line vector of genotypes at loci of parent P_1 (coded -1/1)
- ObjectGenoP2: Line vector of genotypes at loci of parent P_2 (coded -1/1)
- ObjectMap: Data.frame of the genetic map with columns: CHROMOSOME, POSITION and MARKER

Value:

- Returns the covariance matrix of genotypes in DH progeny of the cross P_1 x P_2

```
GenCovProgeny   <- function(ObjectGenoP1, ObjectGenoP2, ObjectMap){
  # Compute expected frequency of recombinants c1
  myDist <- sapply(1:nrow(ObjectMap),
```

```
                    function(x) abs(ObjectMap$POSITION[x] - ObjectMap$POSITION))
  myCHR1 <- do.call(rbind, lapply(1:nrow(ObjectMap),
                                  function(x) rep(ObjectMap$CHROMOSOME[x],nrow(ObjectMap))))
  myCHR2 <- t(myCHR1)
  c1 <- 0.5*(1-exp(-2*(myDist/100)))
  c1[myCHR1!=myCHR2] <- 0.5
  Recomb <- (1-2*c1)
  # Compute disequilibrium
  tmp <- rbind(ObjectGenoP1,ObjectGenoP2)/2
  D <- crossprod(scale(tmp,scale=F))/2
  # Sigma matrix
  MyVarCov <- 4*D*Recomb
  return(MyVarCov)
}
```

The genotypic covariance matrix in the illustrative example is computed as:

```
Sigma <- GenCovProgeny(ObjectGenoP1 = xP1,
                       ObjectGenoP2 = xP2,
                       ObjectMap = map)
```

2.2 Ante-selection: progeny means and co-variances

2.2.1 Definition of marker effects

Let us simulate the column vector of effects β_T for the performance trait T as:

```
# Simulate marker effects (can be replaced by estimated marker effects)
BetaT <- matrix(rnorm(ncol(geno),0,sqrt(0.05)), ncol=1)
```

It results in the parent P_1 and P_2 breeding values:

```
# P1 performance is:
round(xP1%*%BetaT, digits=3)[1,1]
```

```
## [1] -8.438
```

```
# P2 performance is:
round(xP2%*%BetaT, digits=3)[1,1]
```

```
## [1] -8.836
```

The following function defines the column vector of marker effects β_{C1} to follow P_1 identity by state (IBS) contribution to progeny considering only polymorphic loci between parents P_1 and P_2.

Function arguments:

- ObjectGenoP1: Line vector of genotypes at loci of parent P_1 (coded -1/1)
- ObjectGenoP2: Line vector of genotypes at loci of parent P_2 (coded -1/1)

Value:

- Returns a column vector β_{C1} of effects to follow P_1 parental IBS contribution to progeny considering only polymorphic loci between P_1 and P_2.

```
GetBetaC1 = function(ObjectGenoP1, ObjectGenoP2){
  X1tmp <- matrix(ObjectGenoP1, ncol=1)
  X2tmp <- matrix(ObjectGenoP2, ncol=1)
```

```
  tmp <- X1tmp-X2tmp
  return(tmp/crossprod(tmp)[1])
}
BetaC1 <- GetBetaC1(ObjectGenoP1 = xP1,
                    ObjectGenoP2 = xP2)
```

It results in the parent P_1 contribution to parents:

```
# P1 contribution to itself is:
round(xP1%*%BetaC1 + 0.5, digits=3)[1,1]
```

```
## [1] 1
```

```
# P1 contribution to P2 is:
round(xP2%*%BetaC1 + 0.5, digits=3)[1,1]
```

```
## [1] 0
```

2.2.2 Progeny means

The effects β_T and β_{C1} are used to compute progeny means (μ_T and μ_{C1}) before selection:

```
# Progeny mean performance before selection is:
MuT <- 0.5*(xP1%*%BetaT+xP2%*%BetaT)
round(MuT, digits=3)[1,1]
```

```
## [1] -8.637
```

```
# Progeny mean P1 contribution before selection is:
MuC1 <- 0.5*(xP1%*%BetaC1 + xP2%*%BetaC1 + 1)
round(MuC1, digits=3)[1,1]
```

```
## [1] 0.5
```

2.2.3 Progeny co-variances

The following function computes the genetic co-variances of two traits in progeny based on marker effects and the genotypic covariance matrix in progeny $\mathbf{\Sigma}$.

Function arguments:

- ObjectBeta1: Column vector of marker effects for trait 1 (β_1)
- ObjectBeta2: Column vector of marker effects for trait 2 (β_2)
- ObjectSigma: Covariance matrix of genotypes in progeny ($\mathbf{\Sigma}$)

Value:

- Returns the genetic covariance $\beta_1'\mathbf{\Sigma}\beta_2 = \beta_2'\mathbf{\Sigma}\beta_1$.

```
VarCovProgeny = function(ObjectBeta1, ObjectBeta2, ObjectSigma){
  crossprod(ObjectBeta1,ObjectSigma%*%ObjectBeta2)
}
```

The genetic variance in progeny for the performance trait T (σ^2_T) is:

```
VarT <- VarCovProgeny(ObjectBeta1 = BetaT,
                      ObjectBeta2 = BetaT,
                      ObjectSigma = Sigma)
round(VarT, digits=3)[1,1]
```

```
## [1] 16.94
```

The genetic variance in progeny for P_1 contribution trait (σ^2_{C1}) is:

```
VarC1 <- VarCovProgeny(ObjectBeta1 = BetaC1,
                       ObjectBeta2 = BetaC1,
                       ObjectSigma = Sigma)
round(VarC1, digits=3)[1,1]
```

```
## [1] 0.014
```

The genetic covariance between performance trait and P_1 contribution trait ($\sigma_{T,C1}$) is:

```
CovTC1 <- VarCovProgeny(ObjectBeta1 = BetaT,
                        ObjectBeta2 = BetaC1,
                        ObjectSigma = Sigma)

round(CovTC1, digits=3)[1,1]
```

```
## [1] 0.048
```

2.3 Post-selection: usefulness criterion and parental contribution (UCPC)

The following function computes the usefulness criterion parental contribution (UCPC) for a single cross P_1 x P_2 based on progeny means and genetic co-variances previously predicted and a within cross selection of the $pSel$ most performant progeny.

Function arguments:

- ObjectMuT: Progeny mean for the performance trait T (μ_T)
- ObjectVarT: Progeny genetic variance for the performance trait T (σ^2_T)
- ObjectMuC1: Progeny mean for the P_1 contribution (i.e. $\mu_{C1} = 0.5$ for two-way cross)
- ObjectCovTC1: Progeny genetic covariance between the performance and the P_1 contribution ($\sigma_{T,C1}$)
- ObjectpSel: Percentage of selected progeny within the family
- h: Selection accuracy (by default h=1)

Value:

- Returns a data.frame giving the progeny mean performance before selection (μ_T), the usefulness criterion for the performance trait T ($UC^{(i)}_T$), the expected P_1 and P_2 contributions to progeny before selection (c_1, c_2) and the expected P_1 and P_2 contributions to the selected fraction of progeny ($c^{(i)}_1$, $c^{(i)}_2$).

```
GetUCPC = function(ObjectMuT, ObjectVarT, ObjectMuC1, ObjectCovTC1,
                   ObjectpSel, h=1){
  i <- dnorm(qnorm(1-ObjectpSel))/ObjectpSel
  # Usefulness criterion on T
  UCT <- ObjectMuT+i*h*sqrt(ObjectVarT)
  # Correlated response to selection for C1
  C1sel <- ObjectMuC1+i*h*ObjectCovTC1/sqrt(ObjectVarT)
  return(data.frame(MuT = ObjectMuT,
                    UCT = UCT,
                    MuC1 = ObjectMuC1,
                    MuC2 = 1-ObjectMuC1,
                    SelC1 = C1sel,
                    SelC2 = 1-C1sel))
  }

UCPC <- GetUCPC(ObjectMuT = MuT, ObjectVarT = VarT, ObjectMuC1 = MuC1,
```

```
                ObjectCovTC1 = CovTC1, ObjectpSel = 0.05)

UCPC
```

```
##         MuT        UCT MuC1 MuC2    SelC1    SelC2
## 1 -8.637359 -0.1475092  0.5  0.5 0.524217 0.475783
```

3 UCPC to evaluate the interest of a set of two-way crosses

In optimal cross selection, we evaluate a set of crosses as a whole instead of independent single crosses.

```
# Sample of a set of twenty two-way crosses
SampledP1 <- sample(1:nrow(geno),20, replace = TRUE)
SampledP2 <- sample(setdiff(1:nrow(geno),SampledP1),20, replace = TRUE)
# Create a cross object: with parents and within cross percentage of selected progeny
Crosses = data.frame(PARENT1 = SampledP1,
                     PARENT2 = SampledP2,
                     PSelect = 0.05,
                     stringsAsFactors = FALSE)
```

3.1 Compute the UCPC for several two-way crosses

The following function implements the UCPC previously defined for a single cross in a loop for several crosses.

Function arguments:

- ObjectCrosses: Data.frame of crosses with columns: PARENT1, PARENT2 and PSelect giving respectively the first and second parent of the cross and the within family selected fraction of progeny
- ObjectBetaT: Column vector of trait performance marker effects β_T
- ObjectGeno: Genotype of all candidate parents in lines and markers in column (coded in -1/1)
- ObjectMap: Data.frame of the genetic map with columns: CHROMOSOME, POSITION and MARKER

Value:

- Returns a data.frame giving for every cross: the progeny mean performance before selection (μ_T), the usefulness criterion for the performance trait T ($UC_T^{(i)}$), the expected P_1 and P_2 contributions to progeny before selection (c_1, c_2) and the expected P_1 and P_2 contributions to the selected fraction of progeny ($c_1^{(i)}$, $c_2^{(i)}$).

```
GetSetUCPC = function(ObjectCrosses, ObjectBetaT, ObjectGeno, ObjectMap){
  return(do.call(rbind,lapply(1:nrow(ObjectCrosses), function(nCross){
    # Parents and within cross selection parameters
    P1 <- ObjectCrosses$PARENT1[nCross]
    P2 <- ObjectCrosses$PARENT2[nCross]
    pSel <- ObjectCrosses$PSelect[nCross]
    # Genotype of parents
    xP1 <- ObjectGeno[P1,]
    xP2 <- ObjectGeno[P2,]
    # Get Sigma matrix
    Sigma <- GenCovProgeny(ObjectGenoP1 = xP1,
                           ObjectGenoP2 = xP2,
                           ObjectMap = ObjectMap)
    # Construct C1 effects
    BetaC1 <- GetBetaC1(ObjectGenoP1 = xP1,
```

```
                        ObjectGenoP2 = xP2)
    # Get ante-selection means and co-variances
    MuT <- 0.5*(xP1%*%ObjectBetaT+xP2%*%ObjectBetaT)
    VarT <- VarCovProgeny(ObjectBeta1 = ObjectBetaT,
                          ObjectBeta2 = ObjectBetaT,
                          ObjectSigma = Sigma)
    MuC1 <- 0.5*(xP1%*%BetaC1 + xP2%*%BetaC1 + 1)
    CovTC1 <- VarCovProgeny(ObjectBeta1 = ObjectBetaT,
                            ObjectBeta2 = BetaC1,
                            ObjectSigma = Sigma)
    # Get the UCPC considering within family selection
    UCPC <- GetUCPC(ObjectMuT = MuT, ObjectVarT = VarT,
                    ObjectMuC1 = MuC1, ObjectCovTC1 = CovTC1,
                    ObjectpSel = pSel)

    return(cbind(data.frame(Cross = paste0(P1,"x",P2),
                            PARENT1 = P1,
                            PARENT2 = P2,
                            PSel = pSel,
                            stringsAsFactors = FALSE),
                 UCPC))
  }))
}

SetUCPC <- GetSetUCPC(ObjectCrosses = Crosses, ObjectGeno = geno,
                      ObjectBetaT = BetaT, ObjectMap = map)
SetUCPC[1:5,-c(2,3)]
```

```
##      Cross PSel       MuT       UCT MuC1 MuC2     SelC1     SelC2
## 1   20x315 0.05 -9.797972 -1.656583  0.5  0.5 0.5473419 0.4526581
## 2 702x1227 0.05  1.580834 10.411533  0.5  0.5 0.4579492 0.5420508
## 3  822x367 0.05 -6.345915  2.739877  0.5  0.5 0.4696598 0.5303402
## 4  186x182 0.05 -4.039036  2.027207  0.5  0.5 0.5029892 0.4970108
## 5  626x913 0.05  2.735078 14.270303  0.5  0.5 0.4924262 0.5075738
```

3.2 Evaluate a set of crosses for expected genetic gain and genetic diversity

To evaluate this set of crosses we need to compute the gain term $V^{(i)}(\mathbf{nc})$ and diversity constraint term $D^{(i)}(\mathbf{nc})$ depending on the within family selection intensity i.

The IBS coancestry among candidate parents is defined using the following function.

Function arguments:

- ObjectGeno: Genotype of all candidate parents in lines and markers in column (coded in -1/1)

Value:

- Returns a matrix of IBS coancestry among candidate parents.

```
GetIBS = function(ObjectGeno){
  0.5*(tcrossprod(ObjectGeno)/ncol(ObjectGeno)+1)
}
```

Then, the following function is used to evaluate the set of crosses either accounting (UCPC) or not (OCS) for within family selection.

Function arguments:

- ObjectUCPC: UCPC data.frame obtained previously for the set of crosses
- ObjectGeno: Genotype of all candidate parents in lines and markers in column (coded in -1/1)

Value:

- Returns a data.frame with the number of crosses in the set ("nCROSS"), the number of unique parents ("nUniqParents"), $V^{(i=0)}(\mathbf{nc})$ and $D^{(i=0)}(\mathbf{nc})$ not accounting for within family selection ("OCS_Vnc" and "OCS_Dnc"), $V^{(i)}(\mathbf{nc})$ and $D^{(i)}(\mathbf{nc})$ accounting for within family selection ("UCPC_Vnc" and "UCPC_Dnc").

```
EvaluateSetUCPC = function(ObjectUCPC, ObjectGeno){
  # Incidence matrices Z1 and Z2
  UniqueParents <- unique(c(ObjectUCPC$PARENT1,ObjectUCPC$PARENT2))
  Z1 <- matrix(0,ncol=nrow(ObjectUCPC),nrow=length(UniqueParents))
  Z2 <- matrix(0,ncol=nrow(ObjectUCPC),nrow=length(UniqueParents))
  invisible(lapply(1:length(UniqueParents), function(x){
    Z1[x,ObjectUCPC$PARENT1==UniqueParents[x]] <<-1
    Z2[x,ObjectUCPC$PARENT2==UniqueParents[x]] <<-1
  }))
  # Compute IBS matrix
  K <- GetIBS(ObjectGeno = ObjectGeno[UniqueParents,])
  # UCPC based OCS: D(nc) and V(nc) terms accounting for within family selection
  c1 <- matrix(ObjectUCPC$SelC1, ncol=1)
  c2 <- matrix(ObjectUCPC$SelC2, ncol=1)
  c <- (Z1%*%c1+Z2%*%c2)/nrow(ObjectUCPC)
  UCPC_Dnc <- 1-crossprod(c,K%*%c)
  UCPC_Vnc <- mean(ObjectUCPC$UCT)
  # classical OCS: D(nc) and V(nc) terms not accounting for within family selection
  c1 <- matrix(ObjectUCPC$MuC1, ncol=1)
  c2 <- matrix(ObjectUCPC$MuC2, ncol=1)
  c <- (Z1%*%c1+Z2%*%c2)/nrow(ObjectUCPC)
  OCS_Dnc <- 1-crossprod(c,K%*%c)
  OCS_Vnc <- mean(ObjectUCPC$MuT)

  return(data.frame(nCROSS = nrow(ObjectUCPC),
                    nUniqParents = length(UniqueParents),
                    OCS_Vnc = OCS_Vnc,
                    OCS_Dnc = OCS_Dnc,
                    UCPC_Vnc = UCPC_Vnc,
                    UCPC_Dnc = UCPC_Dnc,
                    stringsAsFactors = FALSE))
}
```

```
SetEval <- EvaluateSetUCPC(ObjectUCPC = SetUCPC, ObjectGeno = geno)
SetEval
```

```
##   nCROSS nUniqParents    OCS_Vnc   OCS_Dnc UCPC_Vnc  UCPC_Dnc
## 1     20           40 -0.8399918 0.3373086 8.155683 0.3367695
```

File S4: Supplementary tables

Table S1 TRUE scenario: Mean commercial genetic gain (G_{10}) and genetic gain (G) at different generations (5, 10, 20, 40 and 60 years) and mean number of QTLs where the favorable allele has been lost after 60 years. In brackets is given the standard error (standard deviation divided by the square root of the number of independent replicates: $\sqrt{10}$).

CSI	Commercial genetic gain (G_{10})					Genetic gain (G)					# of QTL where the favorable allele is lost after 60 years
	5 years	10 years	20 years	40 years	60 years	5 years	10 years	20 years	40 years	60 years	
PM	6.184	8.338	11.861	15.118	15.744	4.647	7.197	11.085	14.869	15.735	274.9
(TRUE scenario)	(0.174)	(0.195)	(0.280)	(0.373)	(0.449)	(0.174)	(0.199)	(0.258)	(0.353)	(0.447)	(4.283)
UC	6.574	9.316	13.369	17.553	18.293	4.633	7.620	12.290	17.139	18.280	243.1
(TRUE scenario)	(0.170)	(0.208)	(0.316)	(0.460)	(0.516)	(0.138)	(0.158)	(0.286)	(0.441)	(0.513)	(4.547)
OCS - He*=0.01	5.924	8.563	12.743	18.821	21.892	3.918	6.810	11.326	18.017	21.656	194.3
(TRUE scenario)	(0.130)	(0.224)	(0.294)	(0.447)	(0.525)	(0.133)	(0.187)	(0.277)	(0.429)	(0.529)	(2.633)
UCPC - He*=0.01	6.317	9.164	13.550	19.752	22.869	4.024	7.018	11.859	18.832	22.626	173.6
(TRUE scenario)	(0.139)	(0.201)	(0.322)	(0.538)	(0.641)	(0.120)	(0.149)	(0.285)	(0.507)	(0.634)	(4.031)
OCS - He*=0.10	5.901	8.455	12.239	17.872	21.925	3.838	6.490	10.547	16.739	21.237	110.7
(TRUE scenario)	(0.136)	(0.193)	(0.310)	(0.445)	(0.532)	(0.119)	(0.183)	(0.276)	(0.411)	(0.507)	(3.768)
UCPC - He*=0.10	6.327	8.927	12.972	18.475	22.474	3.915	6.760	11.051	17.178	21.643	109.7
(TRUE scenario)	(0.175)	(0.198)	(0.326)	(0.510)	(0.645)	(0.129)	(0.187)	(0.294)	(0.471)	(0.621)	(3.876)
OCS - He*=0.15	5.785	8.118	11.800	17.148	20.938	3.708	6.185	10.042	15.747	19.867	87.9
(TRUE scenario)	(0.161)	(0.211)	(0.276)	(0.422)	(0.553)	(0.144)	(0.182)	(0.255)	(0.390)	(0.525)	(4.365)
UCPC - He*=0.15	6.215	8.643	12.248	17.187	20.665	3.803	6.402	10.246	15.670	19.528	90.3
(TRUE scenario)	(0.186)	(0.194)	(0.311)	(0.439)	(0.573)	(0.132)	(0.186)	(0.252)	(0.389)	(0.546)	(5.439)

Table S2 GS and PS scenarios: Mean commercial genetic gain (G_{10}) and genetic gain (G) at different generations (5, 10, 20, 40 and 60 years) and mean number of QTLs where the favorable allele has been lost after 60 years. In brackets is given the standard error (standard deviation divided by the square root of the number of independent replicates: $\sqrt{10}$).

CSI	Commercial genetic gain (G_{10})					Genetic gain (G)					# of QTL where the favorable allele is lost after 60 years
	5 years	10 years	20 years	40 years	60 years	5 years	10 years	20 years	40 years	60 years	
PM (PS scenario)	4.925 (0.165)	6.402 (0.166)	8.507 (0.270)	10.371 (0.343)	10.810 (0.329)	2.827 (0.156)	4.672 (0.154)	7.241 (0.222)	9.633 (0.339)	10.445 (0.318)	310.8 (4.250)
PM (GS scenario)	5.543 (0.198)	7.713 (0.256)	10.423 (0.331)	12.769 (0.414)	13.287 (0.436)	4.013 (0.175)	6.509 (0.170)	9.655 (0.326)	12.326 (0.403)	13.084 (0.427)	295.2 (3.708)
UC (GS scenario)	5.984 (0.211)	8.338 (0.237)	11.660 (0.314)	14.438 (0.320)	15.367 (0.358)	4.088 (0.174)	6.672 (0.226)	10.530 (0.285)	13.790 (0.311)	14.971 (0.336)	258.6 (4.571)
OCS - He*=0.01 (GS scenario)	5.546 (0.198)	7.734 (0.237)	11.313 (0.323)	15.850 (0.384)	17.528 (0.438)	3.418 (0.154)	5.894 (0.191)	9.896 (0.309)	15.114 (0.369)	17.128 (0.429)	234.5 (3.908)
UCPC - He*=0.01 (GS scenario)	5.930 (0.179)	8.162 (0.208)	11.881 (0.272)	16.398 (0.426)	18.161 (0.470)	3.544 (0.129)	6.049 (0.178)	10.290 (0.252)	15.486 (0.388)	17.633 (0.465)	218.8 (3.852)

Supplementary Material Chapter 5

Supplementary tables

Table S1 Mean breeding population performance (μ) at different generations (5, 10, 20, 30, 40, 50 and 60 years). In brackets is given the standard error (standard deviation divided by the square root of the number of independent replicates: $\sqrt{10}$).

Scenario	Population mean performance (μ)						
	5 years	**10 years**	**20 years**	**30 years**	**40 years**	**50 years**	**60 years**
Benchmark	9.239 (0.237)	14.913 (0.645)	23.182 (1.446)	30.006 (1.319)	34.741 (1.329)	37.486 (1.487)	38.837 (1.563)
Nobridging_Panel	8.168 (0.282)	13.022 (0.890)	12.589 (0.988)	9.355 (3.147)	11.059 (0.933)	10.581 (0.797)	9.651 (0.958)
Bridging_Panel	8.688 (0.329)	13.653 (0.867)	25.212 (1.314)	35.770 (1.077)	42.828 (1.091)	47.681 (1.256)	52.110 (0.886)
Bridging_Panel (Single TS)	9.635 (0.267)	13.988 (1.599)	29.292 (0.802)	38.398 (1.133)	45.374 (1.090)	52.036 (1.262)	57.067 (1.444)
Nobridging_20y	8.383 (0.271)	12.083 (2.563)	16.818 (2.397)	33.546 (1.519)	47.346 (1.096)	58.628 (1.087)	66.944 (0.849)
Bridging_20y	8.687 (0.293)	13.045 (1.328)	27.987 (0.840)	40.296 (1.010)	51.468 (0.957)	60.939 (1.010)	69.154 (0.868)
Bridging_20y (Single TS)	9.430 (0.229)	14.431 (1.419)	30.497 (1.135)	43.238 (1.384)	54.373 (1.220)	64.066 (1.121)	71.130 (1.028)
Nobridging_5y	9.820 (0.358)	16.356 (0.954)	34.541 (0.980)	49.837 (0.911)	60.424 (0.952)	68.524 (0.918)	74.662 (0.938)
Bridging_5y	10.152 (0.368)	18.146 (1.180)	34.900 (0.905)	49.131 (1.164)	60.542 (1.021)	68.184 (0.916)	74.074 (0.869)
Bridging_5y (Single TS)	12.348 (0.622)	19.246 (1.868)	40.111 (1.149)	53.981 (1.047)	63.540 (1.052)	70.603 (1.160)	75.749 (1.093)

Table S2 Mean performance of the ten best progeny (μ_{10}) at different generations (5, 10, 20, 30, 40, 50 and 60 years). In brackets is given the standard error (standard deviation divided by the square root of the number of independent replicates: $\sqrt{10}$).

Scenario	Ten best mean performance (μ_{10})						
	5 years	10 years	20 years	30 years	40 years	50 years	60 years
Benchmark	15.746	20.544	27.346	33.169	37.020	38.761	39.567
	(0.391)	(0.945)	(1.527)	(1.360)	(1.365)	(1.505)	(1.571)
Nobridging_Panel	15.802	20.776	27.215	25.015	28.625	29.255	29.767
	(0.341)	(0.499)	(0.740)	(2.638)	(0.775)	(0.855)	(1.108)
Bridging_Panel	15.605	21.148	33.215	43.408	51.348	57.376	61.763
	(0.477)	(0.773)	(1.257)	(1.379)	(1.361)	(1.447)	(1.298)
Bridging_Panel (Single TS)	16.727	21.063	35.619	44.730	52.026	57.796	63.699
	(0.527)	(1.530)	(0.843)	(1.399)	(1.352)	(1.524)	(1.698)
Nobridging_20y	15.991	19.827	25.694	40.211	53.449	64.336	72.258
	(0.476)	(1.855)	(1.882)	(1.487)	(1.087)	(1.138)	(0.978)
Bridging_20y	15.793	20.123	34.896	46.471	57.205	66.774	74.413
	(0.357)	(1.240)	(1.009)	(1.104)	(1.062)	(1.115)	(0.932)
Bridging_20y (Single TS)	16.364	21.588	36.989	49.263	59.814	69.221	75.747
	(0.520)	(1.342)	(1.179)	(1.460)	(1.252)	(1.233)	(1.052)
Nobridging_5y	17.077	23.499	41.389	55.828	65.767	73.923	79.776
	(0.476)	(0.730)	(0.978)	(0.910)	(1.036)	(1.005)	(1.058)
Bridging_5y	17.745	25.232	41.448	55.119	66.063	73.297	78.694
	(0.577)	(1.262)	(1.004)	(1.246)	(0.975)	(0.803)	(0.885)
Bridging_5y (Single TS)	20.110	25.900	46.272	59.490	68.559	75.284	80.058
	(0.725)	(2.208)	(1.124)	(1.074)	(1.080)	(1.289)	(1.125)

Table S3 Frequency of the rare favorable allele in the bridging population at different generations (5 , 10, 20, 40, 50 and 60 years). The rare favorable alleles were defined with a frequency ≤ 0.05 at the end of burn-in and concerned on average 269.9 (+/- 23.6) QTLs out of the 1,000 QTLs. In brackets is given the standard error (standard deviation divided by the square root of the number of independent replicates: $\sqrt{10}$).

Scenario	Mean frequency of rare favorable allele						
	5 years	10 years	20 years	30 years	40 years	50 years	60 years
Benchmark	0.005 (0.001)	0.006 (0.002)	0.010 (0.004)	0.013 (0.005)	0.015 (0.006)	0.016 (0.006)	0.016 (0.006)
Nobridging_Panel	0.019 (0.002)	0.036 (0.004)	0.053 (0.006)	0.069 (0.013)	0.063 (0.007)	0.065 (0.007)	0.068 (0.007)
Bridging_Panel	0.021 (0.003)	0.054 (0.010)	0.112 (0.010)	0.169 (0.011)	0.209 (0.010)	0.240 (0.009)	0.263 (0.008)
Bridging_Panel (Single TS)	0.030 (0.004)	0.055 (0.006)	0.126 (0.009)	0.179 (0.010)	0.213 (0.011)	0.245 (0.012)	0.272 (0.014)
Nobridging_20y	0.017 (0.002)	0.047 (0.010)	0.088 (0.014)	0.160 (0.010)	0.245 (0.012)	0.313 (0.012)	0.358 (0.010)
Bridging_20y	0.022 (0.002)	0.053 (0.007)	0.116 (0.011)	0.186 (0.014)	0.253 (0.009)	0.309 (0.009)	0.361 (0.009)
Bridging_20y (Single TS)	0.029 (0.004)	0.068 (0.006)	0.145 (0.013)	0.215 (0.013)	0.282 (0.012)	0.342 (0.011)	0.380 (0.010)
Nobridging_5y	0.038 (0.009)	0.080 (0.019)	0.187 (0.015)	0.279 (0.017)	0.336 (0.015)	0.379 (0.013)	0.407 (0.012)
Bridging_5y	0.040 (0.005)	0.080 (0.011)	0.199 (0.014)	0.282 (0.013)	0.347 (0.013)	0.387 (0.012)	0.414 (0.012)
Bridging_5y (Single TS)	0.087 (0.011)	0.133 (0.018)	0.246 (0.012)	0.319 (0.010)	0.372 (0.010)	0.406 (0.009)	0.431 (0.009)

Supplementary figures

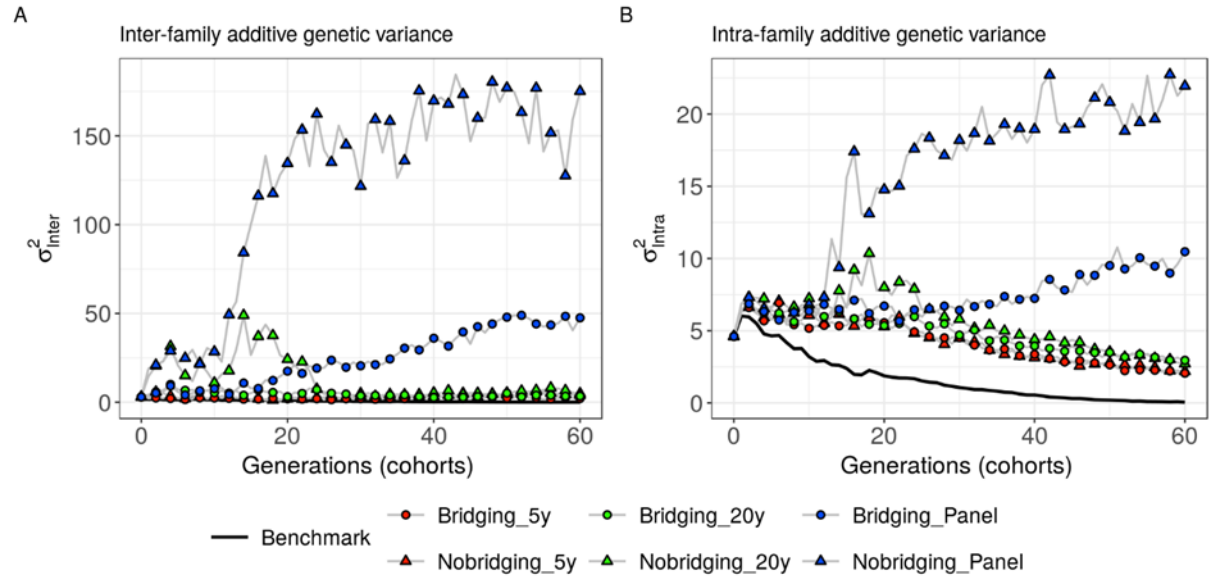


Figure S1 Evolution of the additive genetic variance intra- and inter-family components in the breeding population. Scenarios considering presence or absence of bridging before introduction and different type of donors (panel, twenty-year old and five-year old donors). (**A**) Inter-family additive variance and (**B**) intra-family additive variance.

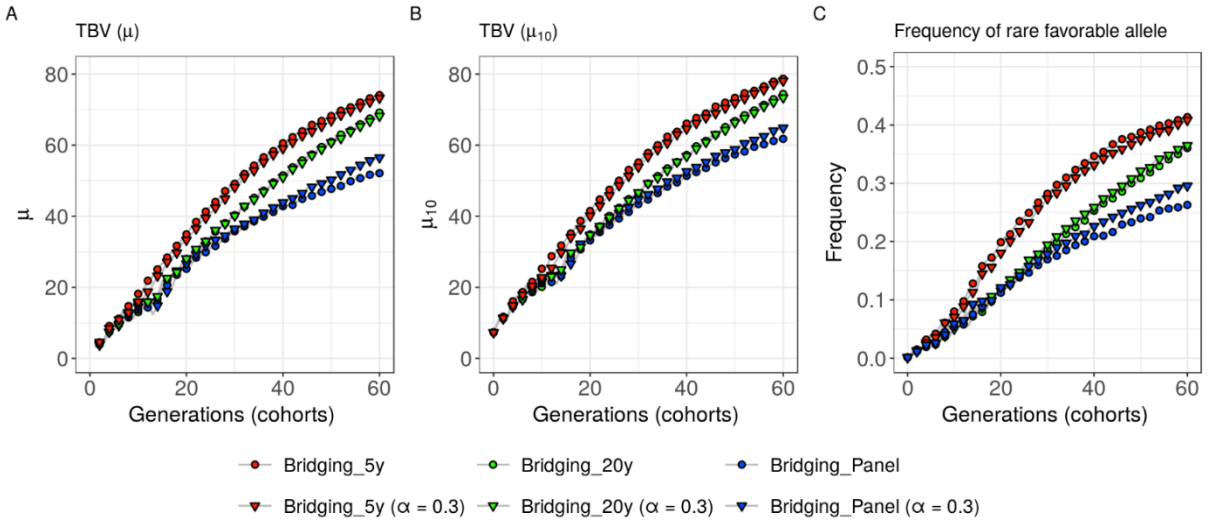


Figure S2 Evolution of the breeding population over generations. Scenarios considering presence of bridging with different type of donors (panel, twenty-year old and five-year old donors) and two weightings for the optimal cross selection in bridging (default is $\alpha = 0.7$). (**A**) Mean breeding population performance (μ), (**B**) mean performance of the ten best progeny (μ_{10}) and (**C**) frequency of the favorable alleles that were rare at the end of burn-in (i.e. $p(0) \leq 0.05$ corresponding on average to 269.9 +/- 23.6 QTLs).

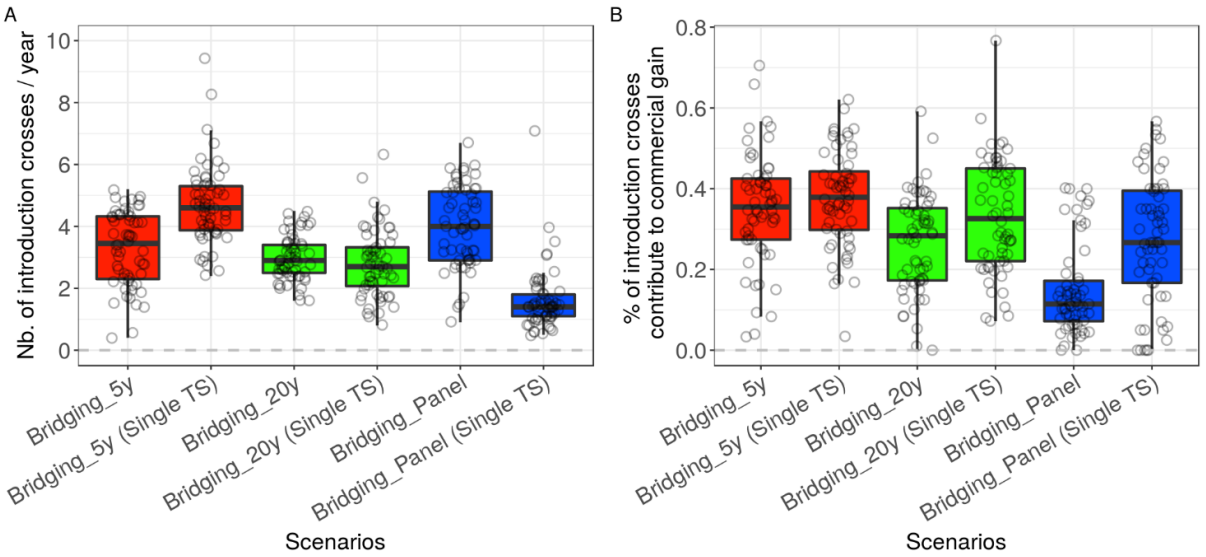


Figure S3 Summary statistics on the introduction crosses. Scenarios considering bridging, different donors (panel, twenty-year old and five-year old donors) and either a single training set (*Single TS*) or two distinct training sets for bridging and breeding (*default*). (**A**) Number of introduction crosses (DExE) per year and (**B**) the fraction of the introduction crosses (DExE) that contributed at least in one of the ten best progeny released by the internal breeding program. The distribution over the sixty generations is given after averaging over the ten replicates.

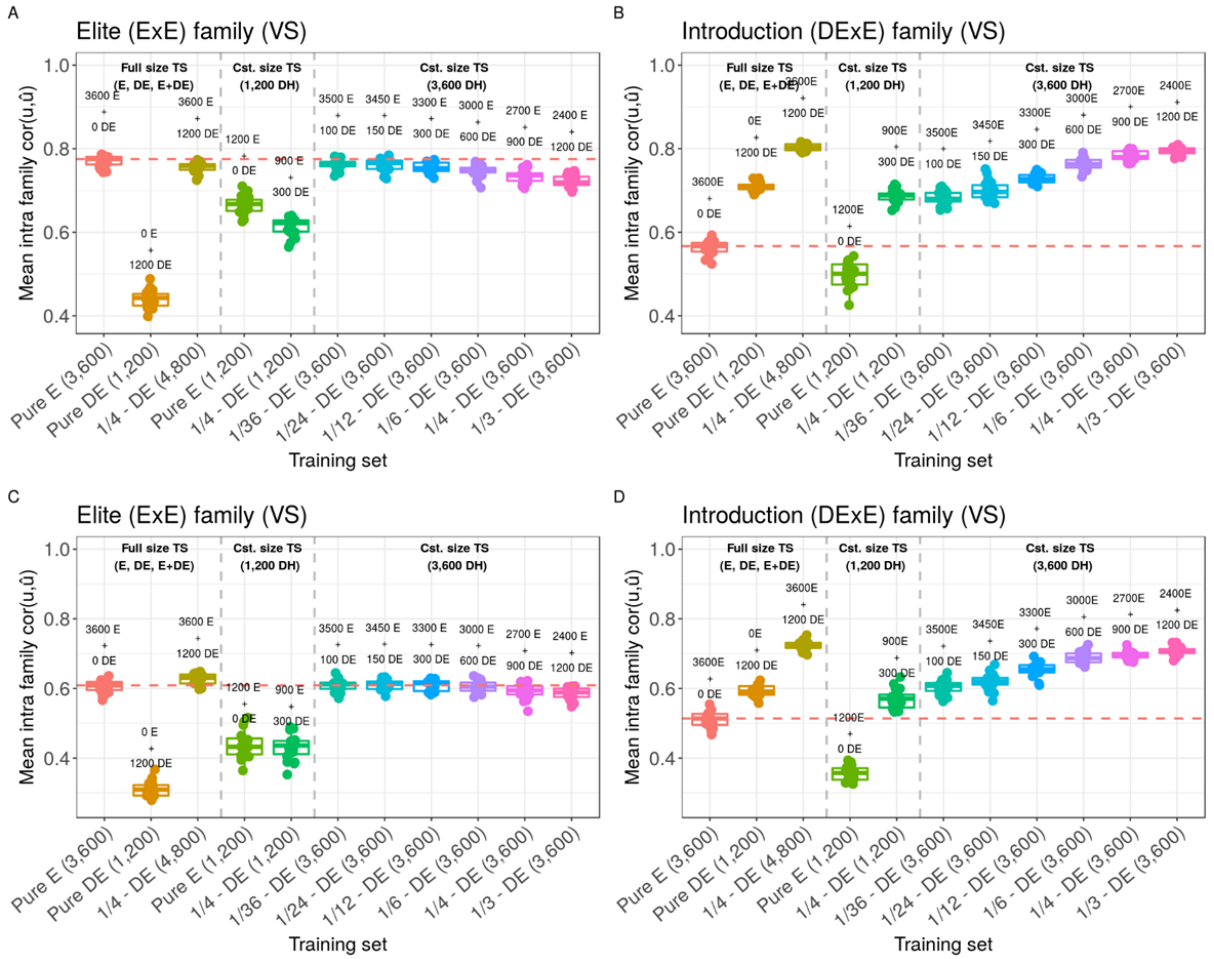


Figure S4 Effect of TS composition on intra-family prediction accuracies ($cor(u, \hat{u})$) considering genotypes simulated at generations 8, 9, 10 (**A**, **B**) or 38, 39, 40 (**C**, **D**) in the scenario *Bridging_20y*. (**A**, **C**) Mean prediction accuracy within 50 elite (ExE) families and (**B**, **D**) mean prediction accuracy within 50 introduction (DExE) families. Boxplots represent the results for 20 independent replicates. One can distinguish three training set types (left to right): Full training set considering all 3,600 E progeny (Pure E), all 1,200 DE progeny (Pure DE) and all 3,600 E + 1,200 DE progeny; Training sets at constant size of 1,200 DH for comparison with Pure DE; Training sets at constant size of 3,600 DH and variable proportion of DE progeny for comparison with Pure E. The red dotted line represents the median value for Pure E TS.

Titre: Contributions à la gestion de la diversité génétique dans les programmes de sélection génomique maïs
Mot clés: Maïs, Diversité génétique, Sélection génomique, Sélection optimale de plan de croisements, Critère d'utilité, Elargissement de la base génétique
Résumé :

Une sélection efficace et durable repose sur un compromis entre efforts à court terme afin de proposer aux agriculteurs des variétés compétitives, et le maintien d'une base génétique large garantissant des variétés futures qui répondront aux défis climatiques, biologiques et sociétaux de demain. Les avancées du génotypage haut débit ont ouvert de nouvelles perspectives de sélection pour les caractères quantitatifs telles que la prédiction génomique de performances individuelles, de l'intérêt de plans de croisements, ainsi que la gestion de la diversité. L'objectif de cette thèse est de contribuer au développement de méthodologies et schémas de sélection efficaces et durables. Cela inclue l'évaluation de la diversité génétique des populations élites, sa conversion efficace en gain génétique à court et long termes, ainsi que l'identification de sources de variabilité génétique d'intérêt et leurs introductions dans les populations de sélection.

Nous avons tout d'abord proposé d'exploiter des séries temporelles de phénotypes et génotypes afin d'évaluer l'effet de la sélection sur la diversité génétique des populations élites ainsi que leur réponse attendue à la sélection. Ces indicateurs ont été appliqués à un programme privé de sélection maïs grain et des stratégies de gestion et amélioration de la réponse à la sélection ont été discutées.

La sélection du plan de croisement qui génère des descendants performants et suffisamment de diversité est un facteur clef du succès à court et long termes des programmes de sélection. Le modèle prédictif de la distribution d'un caractère quantitatif dans une famille biparentale a été étendu au cas des familles multi-parentales. Une approche multi-caractères a été proposée, considérant les performances agronomiques et les contributions parentales comme des caractères corrélés et normalement distribués.

Cette approche dénommée critère d'utilité et contributions parentales (UCPC) permet de prédire la performance moyenne et la diversité attendues dans la fraction sélectionnée de la descendance d'un croisement. L'UCPC peut être utilisé afin d'étendre la sélection optimale de plan de croisements (OCS) qui a pour but de maximiser le gain génétique tout en limitant la perte de diversité. Nous avons montré par simulation que l'OCS basée sur l'UCPC converti plus efficacement la diversité génétique en gain à court et long termes que l'OCS.

La base génétique étroite des populations élites compromet le gain génétique à long terme. De ce fait, une stratégie d'élargissement de leur base génétique sans compromettre le gain à court terme est nécessaire. De nombreuses sources de diversité peuvent être considérées mais toutes ne peuvent être évaluées. Différents critères prédictifs ont été passés en revue et comparés afin d'évaluer l'utilité de ressources génétiques pour enrichir un pool élite. Ces critères s'appuient sur les effets aux marqueurs estimés dans un panel collaboratif constitué de lignées de diversité publiques et de lignées élites privées. L'UCPC permet de même l'identification du croisement multi-parental optimal entre ressources génétiques et lignées élites en fonction des caractéristiques d'originalité et de performance des ressources génétiques. Finalement, nous avons proposé d'utiliser l'approche OCS basée sur l'UCPC afin d'améliorer des ressources génétiques, puis de connecter les ressources génétiques améliorées au matériel élite avant de les introduire dans la population en sélection. Par simulations, nous avons montré l'intérêt de réaliser des introductions récurrentes de ressources génétiques préalablement améliorées afin de maximiser le gain génétique tout en maintenant la diversité constante.

Ces travaux ouvrent de nouvelles perspectives pour la gestion de la diversité génétique.

Tile: Contributions to Genetic Diversity Management in Maize Breeding Programs using Genomic Selection
Keywords: Maize, Genetic Diversity, Genomic Prediction, Optimal Cross Selection, Usefulness Criterion, Genetic Base Broadening
Abstract :

There is an increasing awareness that crop breeding programs should move from short- to long-term objectives by maintaining genetic diversity to cope with future challenges in a context of climatic changes. The advent of high density genotyping opened new avenues for breeding quantitative traits including genomic prediction of individual performances, of parental crosses usefulness, and genetic diversity management. This thesis aims at developing methodologies to further enhance the efficiency and sustainability of breeding programs. This involves the evaluation of genetic diversity in elite breeding pools, its efficient conversion into short- and long-term genetic gain and the efficient identification, improvement and introduction of extrinsic variability into breeding pools.

We first investigated how temporal phenotypic and genotypic data can be used to develop indicators of the genetic diversity and the potential response to selection of a breeding population. We applied these indicators on a commercial hybrid grain maize program and discussed strategies to manage and unlock potential response to selection in breeding populations.

Selection of parental crosses that generate superior progeny while maintaining sufficient diversity is a key success factor of short- and long-term breeding. We extended analytical solutions to predict the distribution of a quantitative trait in the progeny of biparental crosses to the case of multiparental crosses. We also proposed to consider a multitrait approach where agronomic trait and parental genome contributions are considered as correlated normally distributed traits.

This approach, called Usefulness Criterion Parental Contribution (UCPC), enables to predict the expected mean performance and diversity in the most performing fraction of progeny. We used UCPC to extend the Optimal Cross Selection (OCS) method, which aims at maximizing the performance in progeny while maintaining diversity for long-term genetic gain. In a long-term simulated recurrent genomic selection breeding program, UCPC based OCS proved to be more efficient than OCS to convert the genetic diversity into short- and long-term genetic gains.

The narrow genetic base of an elite population might compromise its long-term genetic gain in unpredictable environmental conditions. An efficient strategy to broaden the genetic base of commercial breeding programs is therefore required. Many genetic resources are accessible to breeders but cannot all be considered. We reviewed, proposed and compared different predictive criteria for selecting genetic resources that best complement elite recipients, based on genomewide marker effects estimated on a collaborative diversity panel. We also investigated which mating design should be implemented between a promising genetic resource and elite recipient(s) depending on its phenotypic and genetic distance to elites. Finally, we evaluated the interest of UCPC based OCS to improve genetic resources (pre-breeding), to bridge pre-breeding and breeding (bridging), and to manage recurrent introductions into the breeding population. In a long-term simulated commercial breeding program, we demonstrated that recurrent introductions from a pre-breeding population maximize long-term genetic gain while maintaining genetic diversity constant, with only limited short-term penalty.

The results of this thesis open new perspectives to manage genetic diversity in breeding.

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