Usefulness criterion and post-selection parental contributions in multi-parental crosses: Application to polygenic trait Introgression

Antoine Allier, Laurence Moreau, Alain Charcosset, Simon Teyssèdre, Christina Lehermeier

G3 Genes | Genomes | Genetics

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

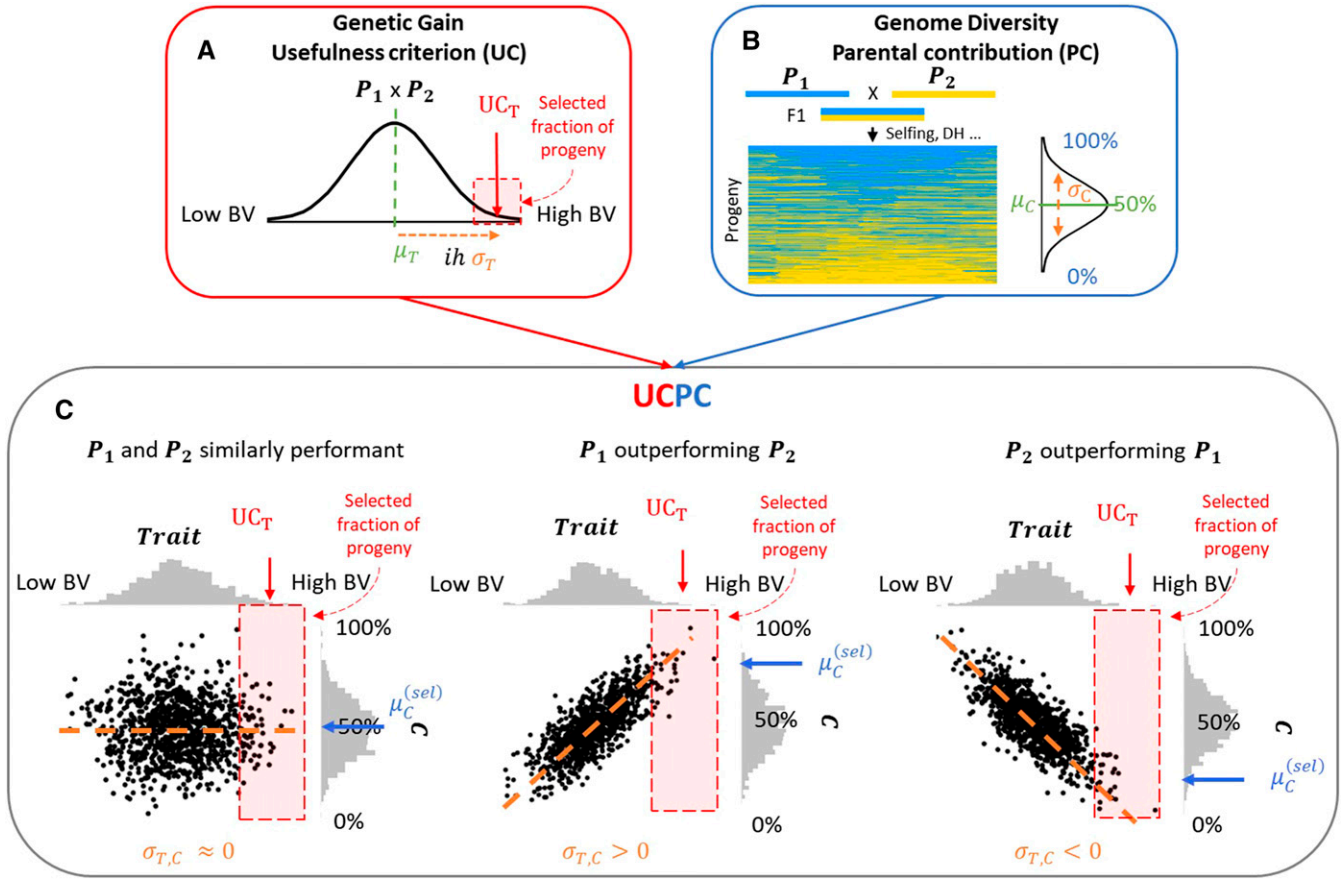

Figure 1 Illustration of Usefulness Criterion Parental Contribution (UCPC) for a two-way cross between P_1 and P_2. UCPC combines (A) the concept of usefulness criterion for an agronomic trait normally distributed ($N(\mu_T, \sigma_T)$) and (B) P_1 genome contribution considered as a normally distributed 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 $(4 \text{ x } p)$-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 $(N \text{ x } 4p)$-dimensional genotyping matrix of N progeny derived from a four-way cross (Figure 2) in a multi-allelic context as:

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

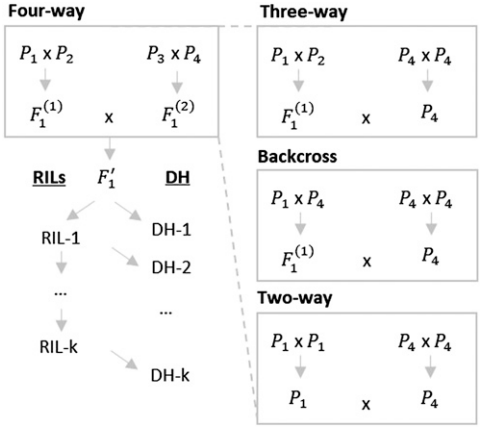

Figure 2 Illustration of four-way crosses (left) and derived crossing schemes (right). In the general case of four-way crosses, nomenclature is defined for recombinant inbred lines (RILs) after k generations of selfing (RIL-k) from pseudo F1 generation (F1′) and doubled haploid lines (DH) derived from 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 $(N \text{ x } p)$-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:

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

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 $(4p \text{ x } 4p)$-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 $(4p \text{ x } 3)$-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 $\Phi_{2jl} = D_{jl}^{14} + D_{jl}^{13} + D_{jl}^{24} + D_{jl}^{23}$
$c_{jl}^{(t)} = \frac{2c_{jl}^{(1)}}{1+2c_{jl}^{(1)}}\left(1-0.5^{t}(1-2c_{jl}^{(1)})^{t}\right)$

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

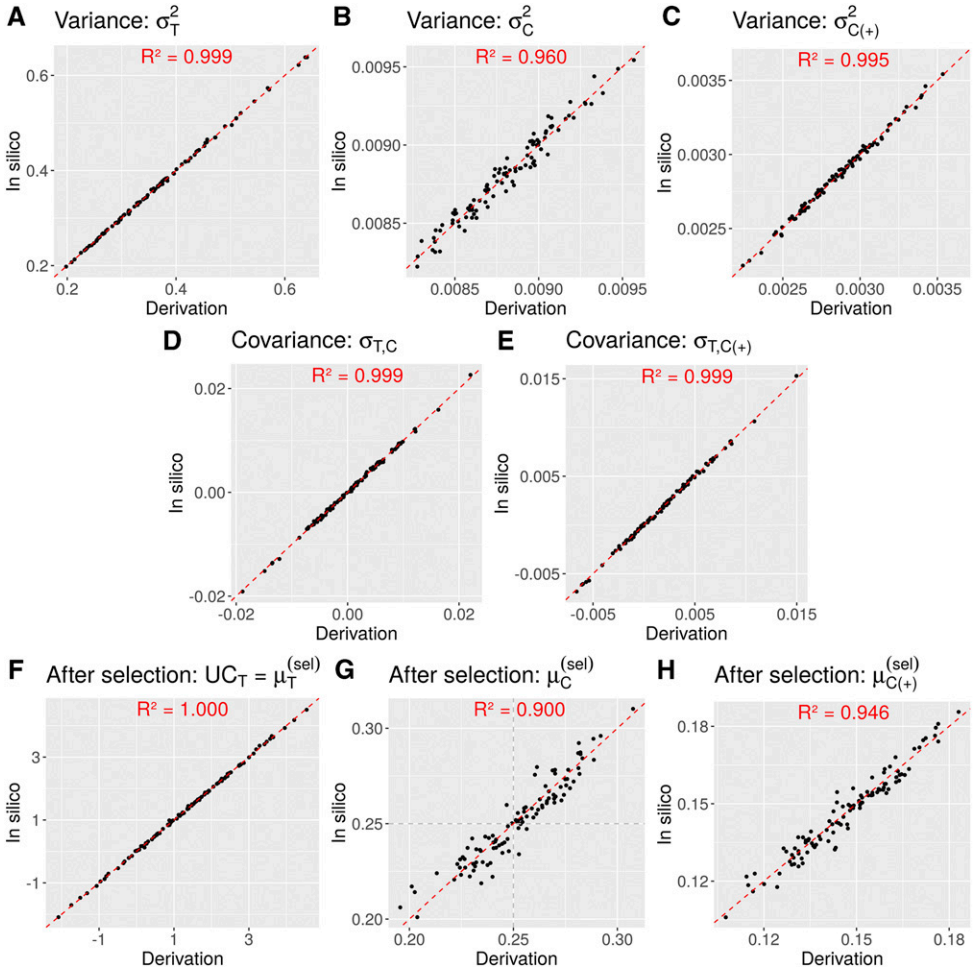

Figure 3 Comparison between predicted (derivation) and empirical (*in silico*) moments of the progeny distributions from 100 four-way crosses consisting of 50,000 DH-1 simulated progeny. Moments shown are (A) variance for the agronomic trait σ^2_T, (B) variance for the genome-wide contribution σ^2_C, (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

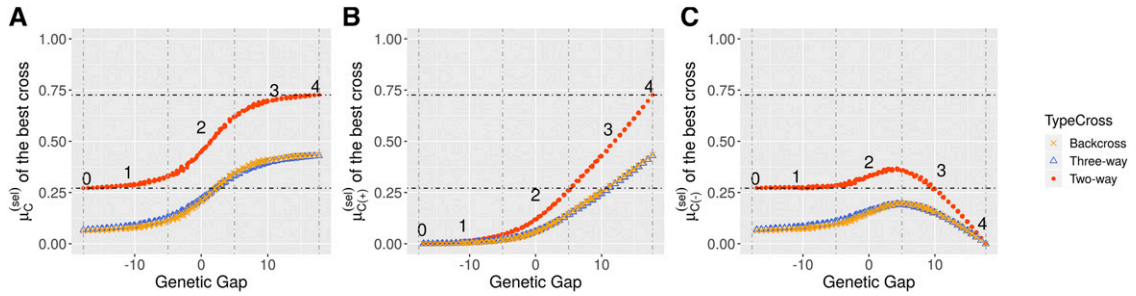

Figure 4 Donor contribution to the selected progeny of the best two-way cross (Donor*Elite), the best three-way cross ((Donor*Elite1)*Elite2) and the best backcross ((Donor*Elite1)*Elite1), depending on the genetic gap between donor line and the elite population. Each data point corresponds to the progeny 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.

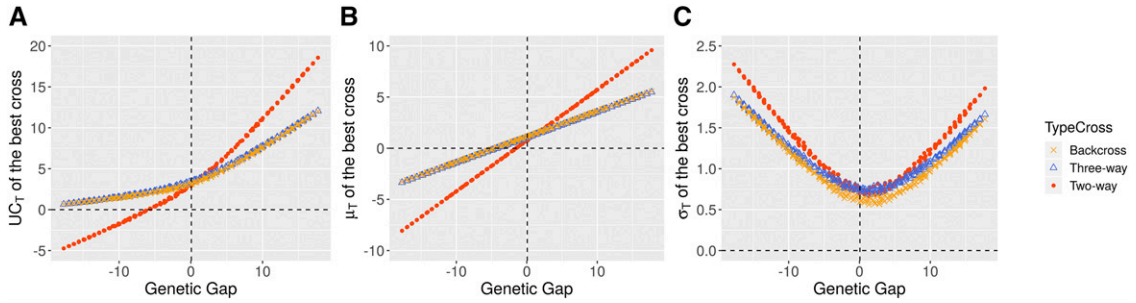

Figure 5 Comparison of the best two-way cross (Donor*Elite), the best three-way cross ((Donor*Elite1)*Elite2) and the best backcross ((Donor*Elite1)*Elite1), depending on the genetic gap (x-axis) with the elite population. Each data point corresponds to the progeny of the best cross and is colored depending 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