

# Origin Specific Genomic Selection: A Simple Process To Optimize the Favorable Contribution of Parents to Progeny

Chin Jian Yang,\* Rajiv Sharma,\* Gregor Gorjanc,<sup>†</sup> Sarah Hearne,<sup>‡</sup> Wayne Powell,\* and Ian Mackay\*<sup>§,1</sup>

\*Scotland's Rural College (SRUC), Kings Buildings, West Mains Road, Edinburgh, EH9 3JG, UK, <sup>†</sup>The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh, EH25 9RG, UK, <sup>‡</sup>CIMMYT, KM 45 Carretera Mexico-Veracruz, El Batan, 56237, Texcoco, Edo. De Mexico, Mexico, and <sup>§</sup>Implant Consultancy Ltd., Chelmsford, UK

ORCID IDs: 0000-0002-4953-5589 (C.J.Y.); 0000-0003-3488-8688 (R.S.); 0000-0001-8008-2787 (G.G.); 0000-0003-2015-2450 (S.H.); 0000-0002-5612-3398 (W.P.); 0000-0002-1109-4730 (I.M.)

**ABSTRACT** Modern crop breeding is in constant demand for new genetic diversity as part of the arms race with genetic gain. The elite gene pool has limited genetic variation and breeders are trying to introduce novelty from unadapted germplasm, landraces and wild relatives. For polygenic traits, currently available approaches to introgression are not ideal, as there is a demonstrable bias against exotic alleles during selection. Here, we propose a partitioned form of genomic selection, called Origin Specific Genomic Selection (OSGS), where we identify and target selection on favorable exotic alleles. Briefly, within a population derived from a bi-parental cross, we isolate alleles originating from the elite and exotic parents, which then allows us to separate out the predicted marker effects based on the allele origins. We validated the usefulness of OSGS using two nested association mapping (NAM) datasets: barley NAM (elite-exotic) and maize NAM (elite-elite), as well as by computer simulation. Our results suggest that OSGS works well in its goal to increase the contribution of favorable exotic alleles in bi-parental crosses, and it is possible to extend the approach to broader multi-parental populations.

## KEYWORDS

introgression  
barley  
maize  
NAM  
genomic  
selection  
plant breeding  
Genomic  
Prediction  
GenPred  
Shared Data  
Resources

There is a general concern that the genetic base of elite varieties of many crops has become very narrow, diminishing the ability of the farming landscape to respond positively and quickly to new challenges. To continue to introduce novel, high value genetic variation into the elite gene pool, breeding programs can select from crosses between their germplasm and materials from plant genetic resources; including wild species, landraces, and improved germplasm that are

unadapted to the target environment. In these exotic crosses, marker assisted selection and backcrossing can effectively track a limited number of QTL accounting for a large proportion of the genetic variation for traits such as disease resistances. For highly polygenic traits, the introgression of novel variation from exotic sources is more complex for a number of reasons. First, QTL mapping for polygenic traits is ineffective or may capture only a small proportion of the genetic variation. Second, the breeding scheme and population size needed to effectively pyramid many QTL are unmanageable. Third, loci at which the exotic lines carry a favorable allele are often linked in repulsion with loci at which the elite lines carry a favorable allele. Consequently, selection on segregating populations from elite-exotic crosses tends to select for the elite background and favorable contribution from the exotic may be lost through linkage drag, or equivalently through hitchhiking. Further, it can be difficult to phenotype adaptive traits, including yield, in populations with a high proportion of unadapted/exotic germplasm. For this reason, selection is often restricted to populations derived from the first or second

Copyright © 2020 Yang *et al.*

doi: <https://doi.org/10.1534/g3.120.401132>

Manuscript received December 22, 2019; accepted for publication May 15, 2020; published Early Online May 19, 2020.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Supplemental material available at figshare: <https://doi.org/10.25387/g3.12320093>.

<sup>1</sup>Corresponding author: Scotland's Rural College (SRUC), Kings Buildings, West Mains Road, Edinburgh, EH9 3JG, UK. E-mail: [ij.mackay@gmail.com](mailto:ij.mackay@gmail.com)

backcross to the elite parent, in which the average exotic contribution is one quarter or one eighth. However, this practice increases further the risk of loss of favorable alleles introduced from the exotic parent.

To overcome these problems, additional generations of crossing among progeny prior to selection can be made to reduce repulsion linkage. Recurrent selection programs have also been proposed to increase the frequency of favorable alleles from both elite and exotic donors over several generations (Hallauer and Carena 2012). Bernardo (2009) proposed genomic recurrent selection starting in the  $F_2$  before deriving recombinant lines and found this to be more effective than the conventional practice of selecting among lines derived from the backcross to the elite parent. More recently, Gorjanc *et al.* (2016) proposed genomic selection on a population established among exotic accessions to increase the frequency of favorable alleles prior to making crosses between the elite and (improved) exotic population. In simulation, this reduced the loss of favorable alleles from exotic sources compared to direct crossing. However, there is a risk that selective effort is wasted in increasing favorable allele frequencies in the improved exotic pool that are already at high frequency among elite lines. Ru and Bernardo (2019 and 2020) proposed backcrossing favorable linkage groups instead of QTL from exotic parent into elite parent using soybean nested association mapping (NAM) data as an example. Recently, Allier *et al.* (2019) proposed treating parental genome contribution as a trait in its own right, and suggested index or truncation selection on this and agronomic traits as a means of reducing the loss of favorable exotic alleles. In addition, Allier *et al.* (2020) proposed a method to identify exotic candidates that can provide the most benefit in elite-exotic crosses through maximizing favorable contributions from exotic parents.

The problems associated with introgression programs for quantitative traits also manifest in mainstream breeding programs. In a cross between two elite inbred lines, the favorable alleles at loci determining a polygenic trait are unlikely to be distributed equally between the two parents. For genetic progress, descendant lines must be selected in which both parents contribute favorable alleles, since only then can the performance of descendants exceed that of the best parent. Assuming for simplicity that all gene effects are equal, the selected line must be fixed for more favorable alleles than the best parent. However, selection among progeny may still result in a disproportionate contribution from the genome of the best parent. For example, Fradgley *et al.* (2019) found it common for an elite wheat line to share over 80% of its genetic material with one of its two elite parents.

In this paper we propose a simple process to quantify and therefore control the favorable contribution of parents to progeny with a technique called Origin Specific Genomic Selection (OSGS). We achieve this by partitioning a genomic prediction equation into two components: the first component is the contribution from markers where the favorable allele is carried by the primary (often elite) parent and the second component is the contribution from markers where the favorable allele is carried by the secondary (often exotic) parent. We test this method by within-cross prediction in two NAM datasets. The first is the HEB-25 barley NAM of backcross derived lines from an elite variety (Barke) and 25 wild barleys (Maurer *et al.* 2015). The second is the maize NAM of  $F_2$  derived lines from crosses between the inbred B73 and 25 lines selected to sample diversity among elite maize germplasm (Yu *et al.* 2008). We validate our results by computer simulations and discuss the implications of our results for introgression and pre-breeding together with

broader applications in plant breeding, including the use of OSGS in multi-parental populations.

## MATERIALS AND METHODS

### Genomic Selection (GS) and Origin Specific Genomic Selection (OSGS)

The mixed linear model commonly used in the training population of genomic selection (GS) can be generalized as:

$$y = Xb + Wu + e, \quad [1]$$

where  $y$  is a vector of observed trait values for each individual,

$X$  is a design matrix associating fixed effects with trait observations,

$b$  is a vector of fixed effects,

$W$  is a design matrix associating marker effects with trait observations,

$u$  is a vector of marker effects with an assumed distribution of  $N(0, I\sigma_u^2)$ ,

$e$  is a vector of residuals with an assumed distribution of  $N(0, I\sigma_e^2)$ .

Then, once the marker effects are estimated ( $\hat{u}$ ), we can predict breeding values ( $\hat{a}$ ) for genotyped individuals (even non-phenotyped) as:

$$\hat{a} = W\hat{u}. \quad [2]$$

In a bi-parental cross, provided marker data are available on the parents, marker regression coefficients  $\hat{u}$  can be partitioned into those that pertain to the favored alleles of the primary parent  $\hat{u}_1$  and those that pertain to favored alleles of the secondary parent  $\hat{u}_2$  such that  $\hat{u} = \hat{u}_1 + \hat{u}_2$ . We define the primary parent as the better performing line, the elite parent in an introgression program. The prediction equation [2] can then be partitioned into:

$$\hat{a}_1 = W\hat{u}_1, \quad [3]$$

$$\hat{a}_2 = W\hat{u}_2, \quad [4]$$

and

$$\hat{a} = \hat{a}_1 + \hat{a}_2, \quad [5]$$

where  $\hat{a}_1$  is the contribution from the primary parent and analogously  $\hat{a}_2$  is the contribution from the secondary parent.

Among any set of individuals, we can then select based on  $\hat{a}$ , or on any index of  $\hat{a}_1$  and  $\hat{a}_2$ . Thus, we refer to the former method as GS and the latter method as OSGS. Table 1 provides a simple example of the computation of  $\hat{a}$ ,  $\hat{a}_1$  and  $\hat{a}_2$  in which three favorable alleles out of ten are contributed by the exotic parent.

### Data analysis

No modification of an existing method for genomic prediction is required for OSGS provided the method estimates  $u$ . OSGS requires only that (i) allele origins are identified and (ii) marker estimates are partitioned into two classes with favorable alleles carried by the primary or by secondary parent. If marker genotypes are coded -1, 0, 1 with 0 as the heterozygous class (Figure 1), this partition is simply on the basis of the sign of the regression coefficients. Also, identifying allele origin is trivial in plant breeding scenarios with inbred parents.

Performance differences among various genomic prediction methods are generally minimal, especially if predictions are among

**Table 1** An example of OSGS for ten unlinked markers segregating among inbred lines derived from the F<sub>2</sub> cross of an elite and exotic parent. At each marker, elite and exotic homozygotes are respectively coded -1 and +1. Negative regression coefficient indicates the increasing allele for the trait is carried by the elite parent and a positive coefficient that the increasing allele is carried by the exotic parent. Here, seven favorable alleles originate from the elite and three from the exotic parent. For each individual (ID1-5), the sum of the products of marker genotypes and regression coefficients gives an estimate of the total breeding value,  $\hat{a}$ . Totalling products over the first three and last seven markers partitions the breeding value into contribution respectively from the elite ( $\hat{a}_1$ ) and exotic ( $\hat{a}_2$ ) parent. For the coefficients given, the expected correlation between  $\hat{a}$  and  $\hat{a}_1$  is 0.89 and between  $\hat{a}$  and  $\hat{a}_2$  is 0.45. The expected correlation between  $\hat{a}_1$  and  $\hat{a}_2$  is zero, since the markers are not linked in this example

Marker	Favored allele	Regression coefficient	Marker genotypes				
			ID1	ID2	ID3	ID4	ID5
M1	Elite	-0.95	-1	1	1	-1	1
M2	Elite	-0.34	-1	-1	1	-1	-1
M3	Elite	-0.47	-1	1	-1	-1	1
M4	Elite	-0.11	-1	-1	-1	-1	1
M5	Elite	-0.49	1	1	-1	-1	1
M6	Elite	-0.63	1	-1	-1	1	1
M7	Elite	-1.24	-1	-1	1	1	-1
M8	Exotic	0.22	1	1	1	1	-1
M9	Exotic	0.38	-1	-1	1	1	-1
M10	Exotic	0.82	1	1	1	-1	1
Total breeding value ( $\hat{a}$ )			2.65	1.07	0.59	0.27	-0.85
Elite contribution ( $\hat{a}_1$ )			1.99	0.41	-0.83	0.49	-1.07
Exotic contribution ( $\hat{a}_2$ )			0.66	0.66	1.42	-0.22	0.22

closely related material over a limited number of generations (Daetwyler *et al.* 2013). In this paper therefore, we demonstrate only three standard methods: ridge regression (Hoerl and Kennard 1976) as implemented in rrBLUP (Endelman 2011), LASSO (Tibshirani 1996) as implemented in glmnet (Friedman *et al.* 2010), and BayesC $\pi$  (Habier *et al.* 2011) as implemented in BGLR (Pérez and de los Campos 2014) to test that OSGS is robust to the choice of method. All three methods are available as R packages and all analyses were performed with R 3.6.3 (R Core Team 2020).

In genomic selection, regression coefficients are typically estimated from one population of lines, the training or reference population, and the prediction equation is applied to a set of candidates with no trait information. Here, the emphasis is different since we are primarily interested in partitioning the observed phenotype of individuals into contributions from the two parents. As such, the question is not about training and testing, but what are parent contributions. We performed our analysis of the NAM datasets in two ways: (i) joint analysis of all 25 families, and (ii) independent analysis of all 25 families. The joint

analysis ignores the variations among the 25 families and thus allows us to test the robustness of OSGS to familial variation. The independent analysis is limited to variations within each bi-parental cross and thus provides a good platform for demonstrating the use of OSGS.

For each analysis, we estimated correlation coefficients between the observed ( $y - X\hat{b}$ ) and predicted trait values of  $\hat{a}$ ,  $\hat{a}_1$  and  $\hat{a}_2$ .  $\hat{a}$  is estimated from all (A) markers,  $\hat{a}_1$  is estimated from subset of markers with favorable primary (P) parent alleles, and  $\hat{a}_2$  is estimated from subset of markers with favorable secondary (S) parent alleles. The relative importance of the primary and secondary parents as contributors of favorable genetic variation was quantified by the correlations between the pairs of  $\hat{a}$ ,  $\hat{a}_1$  and  $\hat{a}_2$ , and by the number and distribution of favorable marker effects among the two parents.

In addition, we compared the distributions of favorable primary (P) and secondary (S) marker effects. We first extracted the P and S marker effects based on the signs of rrBLUP coefficients and the favorable direction for each trait. We then converted the marker effects into absolute values and compared the two distributions using the Kolmogorov-Smirnov test as implemented in the *ks.test* function in R (R Core Team 2020). Results were shown as  $-\log_{10}(p)$ .

To evaluate the potential of OSGS in optimizing favorable parental contributions to progeny, we compared the simulation outcomes from selection in the NAMs using OSGS and GS. To begin this simulation, we took the results from the independent analysis of the 25 NAM families. Within each family, we selected top four lines based on the estimated breeding values (EBVs) determined by OSGS and GS. In GS, the EBVs are essentially  $\hat{a}$ . In OSGS, we first ranked  $\hat{a}_1$  and  $\hat{a}_2$  such that the most favorable  $\hat{a}_1$  and  $\hat{a}_2$  have the highest rank value. Next, we calculated the EBVs as  $rank(\hat{a}_1) \cdot \omega + rank(\hat{a}_2) \cdot (1 - \omega)$ , in which  $\omega$  is the selection weight for P and ranges from 0 to 1. We made all  $\binom{4}{2} = 6$  possible crosses among these four selected lines and generated 10 double haploids (DHs) from each cross. This process was done using AlphaSim (Faux *et al.* 2016) to simulate the recombination events. We calculated the average fold change in proportions of favorable primary alleles (P) and favorable secondary alleles (S) alleles and average normalized change in EBVs before and

	Marker (SNP)						Marker (Phased)					
	1	2	3	4	5	6	1	2	3	4	5	6
Primary parent	A	T	G	C	A	T	-1	-1	-1	-1	-1	-1
Secondary parent	T	G	A	G	A	C	1	1	1	1	1	1
F <sub>1</sub>	W	K	R	S	A	Y	0	0	0	0	0	0
NAM Line 1	A	T	A	G	A	C	-1	-1	1	1	1	1
NAM Line 2	T	?	G	C	A	T	1	0	-1	-1	-1	-1
NAM Line 3	W	K	?	G	A	C	0	0	.5	1	1	1

**Figure 1** Example of recoding marker data for OSGS. SNP markers are originally coded according to standard IUPAC genetic codes (W = A/T, K = G/T, R = A/G, S = C/G, Y = C/T). In the phased markers, homozygous alleles from the primary parent are coded as -1, homozygous alleles from the secondary parent are coded as 1, and heterozygotes are coded as 0. Missing data and monomorphic SNPs are coded as the average between the two flanking, non-missing marker data.

after selection. Lastly, we compared these changes between OSGS and GS across all 25 families.

### Barley NAM population

We analyzed two polygenic traits in the HEB-25 barley NAM population: days to heading (DTH) and yield (YLD), which were respectively taken from Herzig *et al.* (2018) and Sharma *et al.* (2018). Since only raw data on DTH and YLD were provided, we calculated the least squares means of DTH and YLD for 1,420 lines based on the fixed effects of location, nitrogen treatment and year.

We also obtained the accompanying marker genotype data from Maurer *et al.* (2015), which consisted of 1,427 lines and 5,709 polymorphic markers. We removed five markers that did not map to reference genome, resulting in 5,704 markers. The markers were initially coded as 0 for homozygous elite allele, 1 for heterozygous, 2 for homozygous wild allele and 5 for non-polymorphic within family. To maintain consistency between the barley and maize NAM data, we set all the markers coded as 5 to missing and imputed these missing markers using the same method as for the maize NAM (Buckler *et al.* 2009), where any missing data were imputed as an average of two non-missing flanking markers. Markers with missing data at the start and end of each chromosome were imputed to be the same as the nearest markers. Finally, we converted the marker from 0/1/2 to -1/0/1 format.

The trait and marker data combined resulted in 1,371 lines for analysis.

### Maize NAM population

We analyzed two polygenic traits in the maize NAM population that are comparable to DTH and YLD in the barley NAM population: days to silking (DTS) and cob length (CL), which were taken from Buckler *et al.* (2009) and Brown *et al.* (2011) respectively. Similar to the barley NAM trait data, we calculated the least squares means of DTS and CL for 4,910 and 4,884 lines respectively based on the fixed effects of location, year, replication within location and block.

We also obtained the accompanying marker genotype data from McMullen *et al.* (2009) for 4,699 lines and 1,106 polymorphic markers. The marker data are fully imputed and phased, so we only converted the marker format from 0/1/2 to -1/0/1 format.

The trait and marker data combined resulted in 4,697 lines for analysis.

### Computer simulations

For comparison with the barley and maize NAMs, we simulated three bi-parental populations: (1)  $F_2$ -derived, (2)  $BC_1$ -derived and (3) reverse  $BC_1$ -derived (secondary line as the recurrent parent), all of which were selfed for 4 generations prior to GS/OSGS. For each population, we simulated traits with varying ratios of favorable primary (P) and secondary (S) QTL (P:S = 50:50, 55:45, 60:40, 70:30, 80:20, 90:10) and with QTL densities of 2cM/QTL or 20cM/QTL. Within each simulation, same QTL were tested in  $F_2/BC_1/rBC_1$ -derived populations.

We used rrBLUP to calculate the marker effects in the  $F_6$ ,  $BC_1S_4$  and  $rBC_1S_4$  generations. These were used for predicting the breeding values of each line using GS and OSGS methods. Similar weighting schemes to the NAM simulation were used here to determine the breeding values in OSGS. We crossed the top 5 lines (identified by GS/OSGS) in a half diallel and generated 20 double haploids (DHs) from each cross. Similar to the previous simulation with the NAM datasets, we compared GS and OSGS impacts on P and S proportions and EBVs over a single generation of selection. In addition, we chose

the population with P:S = 60:40 and  $\omega = 0.5$  and performed recurrent selection for an additional four cycles. We used the previously calculated marker effects to predict EBVs for selection purposes in all subsequent generations. Details on the selection process can be found in Figure S1. All simulations were repeated 100 times.

All simulations were performed in R 3.6.3 (R Core Team 2020), in which marker data were generated using AlphaSim v0.11.1 (Faux *et al.* 2016) and trait data were generated using custom R scripts. For all populations, we simulated diploid individuals with 10 chromosomes and 7,750 markers distributed evenly across a total genetic distance of 1,550 cM. The markers were coded as -1 for the primary parent and 1 for the secondary parent. QTL positions were randomly sampled from a uniform distribution of all markers. QTL effects for the primary and secondary parent alleles were simulated from a half-normal distribution such that the QTL marker variances are equal between primary and secondary parent alleles, and the aggregated QTL marker variance is equal to  $p^{-1}$ , where  $p$  is the total number of QTL (Figure S2A). Markers selected as QTL markers were left in these analyses since their removal with such high marker density would have little effect, and our purpose is to compare the performance of OSGS and GS and not to test differences in prediction accuracy due to marker-QTL linkage. For any generation, the true breeding value of each line was calculated from its QTL marker genotypes and QTL effects, and the phenotypic trait value of each line was calculated by adding residual value drawn from a standard normal distribution with mean of 0 and variance of 1. Since we fixed the QTL marker variance and residual variance, the simulated mean trait heritabilities range from 0.40 to 0.95 depending on the proportion of favorable primary and secondary parent alleles and number of QTL markers (Figure S2B).

### Data availability

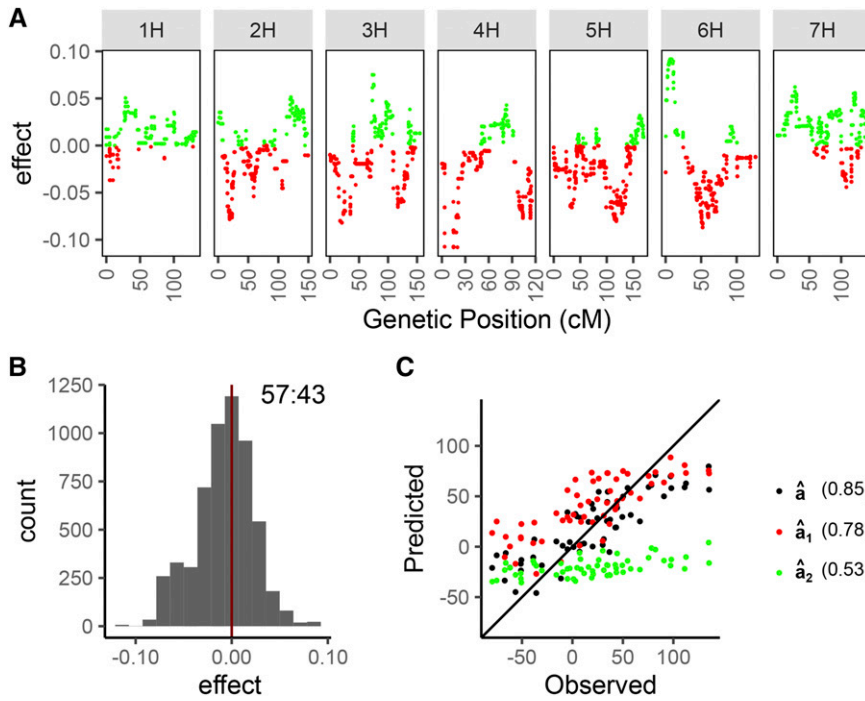
The barley NAM raw DTH, raw YLD and marker genotype datasets were downloaded from doi.org/10.5447/IPK/2017/6, doi.org/10.5447/IPK/2017/21 and Additional File 5 in doi.org/10.1186/s12864-015-1459-7 respectively. The maize NAM raw DTS, raw CL and marker genotype datasets were downloaded from the Cyverse Discovery Environment in the following folders respectively, (1) /iplant/home/shared/panzea/phenotypes/ Buckler\_etal\_2009\_Science\_flowering\_time\_data-090807.zip, (2) /iplant/home/shared/panzea/phenotypes/Brown\_etal\_2011\_PLoSGenet\_pheno\_data-120523.zip, (3) /iplant/home/shared/panzea/genotypes/SNPs/ NAM\_map\_and\_genos-120731.zip. All R scripts used in this manuscript can be accessed from <https://github.com/cjyang-sruc/OSGS>. File S1 contains all the supplemental materials. Supplemental material available at figshare: <https://doi.org/10.25387/g3.12320093>.

## RESULTS

### Maize and barley NAM data analysis

OSGS is robust to the choice of GS methods as shown using three popular GS methods (rrBLUP, LASSO and BayesC $\pi$ ) (Figure S3 and Table S1). There were little differences in performances across these methods, especially in  $\hat{a}$  and  $\hat{a}_1$ . However, predictions on  $\hat{a}_2$  appeared slightly more variable when LASSO is used, which is likely due to a combination of small family size and fewer favorable exotic alleles in the barley NAM. In one example, LASSO failed to identify any favorable secondary parent alleles, resulting in zero prediction from these alleles (Table S1). In some barley NAM families, the prediction accuracies of  $\hat{a}$  from rrBLUP are perfect (Table S1), which suggest overfitting. However, these families also showed high accuracies with



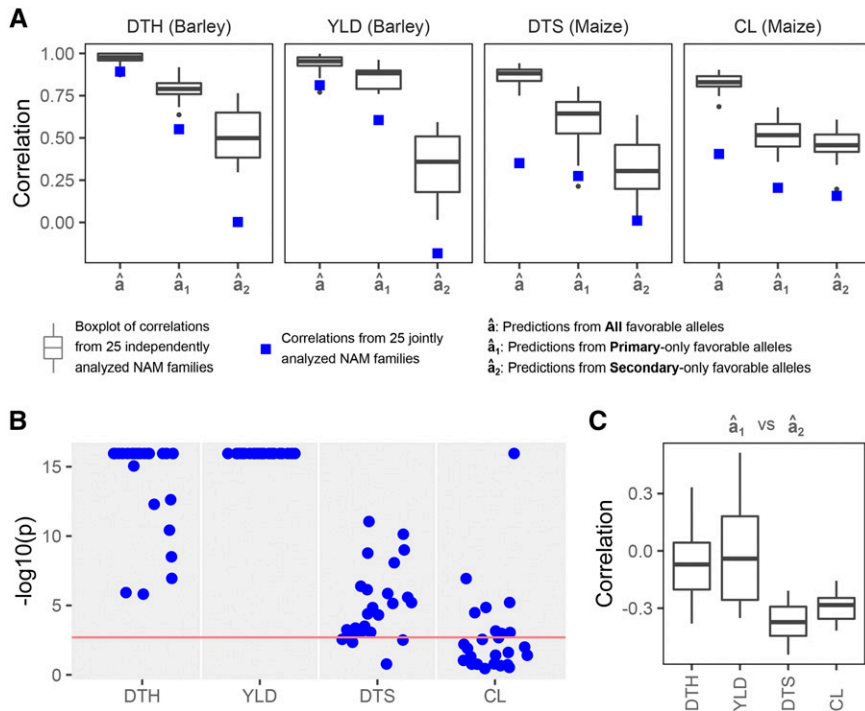


**Figure 2** Partitioning of favorable parental alleles in OSGS. Here, we took YLD in barley NAM family 1 as an example to illustrate how the markers can be partitioned into favorable primary (P) and secondary (S) parental alleles for breeding values prediction. [A] Marker effects are plotted along the chromosomes and genetic positions, with the P alleles colored red and S colored green. [B] Distribution of marker effects shows a bias for more P (57%) than S (43%), which suggests that the recurrent parent in barley NAM has more favorable YLD allele than the donor parent 1. [C] Predicted breeding values using all markers ( $\hat{a}$ ), P-only ( $\hat{a}_1$ ) and S-only ( $\hat{a}_2$ ) are plotted against the observed trait values, and the correlations are shown in parentheses.

the LASSO which selects for markers by cross-validation. Nonetheless, we are not overly concerned about these perfect predictions since our interest is to show how favorable alleles can be partitioned in OSGS. Overall, since there was little difference, we focus all of our analyses on rrBLUP.

Using YLD in barley NAM family 1 as an example (Figure 2), we showed the partitioning of all (A) markers into markers carrying

favorable primary (P) and secondary (S) parental alleles based on their effect signs. Given our marker coding and the favorable direction of YLD, P alleles are represented by markers with negative effects and S alleles are represented by markers with positive effects. We observed an uneven distribution of P and S alleles across the genome (Figure 2A). Overall counts of P and S alleles were unequal with a slight bias toward more P alleles, as shown in Figure 2B. In our



**Figure 3** Prediction accuracies and marker effect distributions across all 25 NAM families. [A] Prediction accuracies of  $\hat{a}$ ,  $\hat{a}_1$  and  $\hat{a}_2$  are shown as the correlations between the predicted and observed trait values from the joint and independent analyses of 25 NAM families. [B] P and S distributions estimated from the independent analyses were tested using Kolmogorov-Smirnov test and the results are shown as  $-\log_{10}(p)$ . The Bonferroni adjusted threshold of  $P = 0.05/25$  is shown as a red horizontal line. [C] Correlations between  $\hat{a}_1$  and  $\hat{a}_2$  represent the potential constraints when selecting for both P and S markers in OSGS, although the lack of strong negative correlations suggests these are small.

predictions using A, P or S alleles, *i.e.*,  $\hat{a}$ ,  $\hat{a}_1$  and  $\hat{a}_2$ , the accuracies decreased in the order of A, P and S (Figure 2C). Since P and S alleles are subsets of A, the prediction accuracies from either P or S can never exceed A's. Prediction accuracies for all families and traits can be found in Table S1.

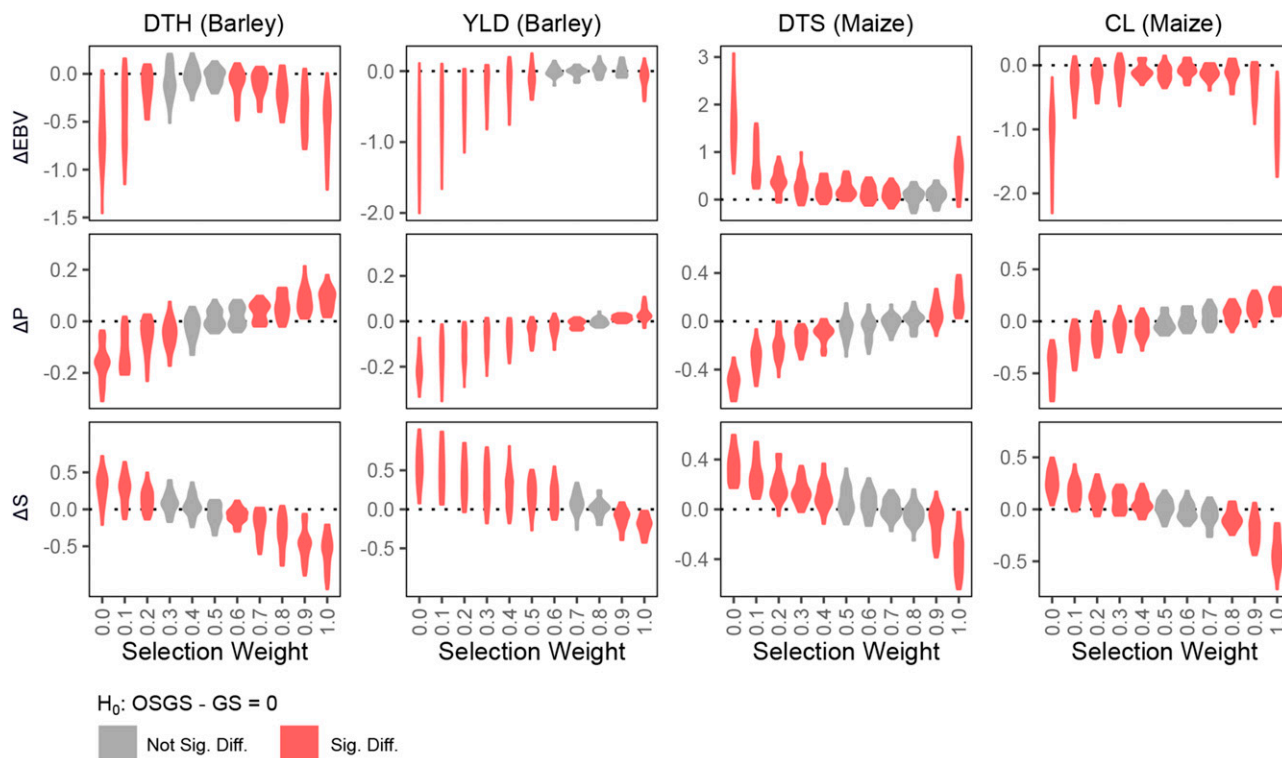
Between our joint and independent analyses of 25 NAM families, we found higher accuracies, but varying in degrees, in the independent analysis over joint analysis across all predictions (Figure 3A). In barley NAM, the discrepancies between the joint and independent analyses are less pronounced in  $\hat{a}$  and  $\hat{a}_1$  than  $\hat{a}_2$ . In maize NAM, the discrepancies are relatively similar across  $\hat{a}$ ,  $\hat{a}_1$  and  $\hat{a}_2$ . This observation is likely explained by how the NAMs were generated as the barley NAMs are BC<sub>1</sub>-derived and the maize NAMs are F<sub>2</sub>-derived. On average, the recurrent (common) NAM parent contributes approximately three-quarter in the BC<sub>1</sub> genome but only half in the F<sub>2</sub> genome. Unlike the recurrent parent, the donor parents are distinct and thus likely possess allelic variations, as shown in previous GWAS analyses (Buckler *et al.* 2009; Brown *et al.* 2011; Herzig *et al.* 2018; Sharma *et al.* 2018). Therefore, higher proportion of recurrent parent results in better predictions in the joint analysis.

The order of  $\hat{a}$ ,  $\hat{a}_1$  and  $\hat{a}_2$  accuracies remained similar in all analyses, although the accuracy gaps among  $\hat{a}$ ,  $\hat{a}_1$  and  $\hat{a}_2$  differed when compared across traits (Figure 3A). Accuracy gaps between  $\hat{a}$  and  $\hat{a}_1$  are smallest in YLD than the others, while accuracy gaps between  $\hat{a}_1$  and  $\hat{a}_2$  are largest in YLD, intermediate in DTH and DTS, and smallest in CL. This observation can be partly attributed to the NAM population types as previously suggested, however, a more important factor is likely the difference in distributions of P and S

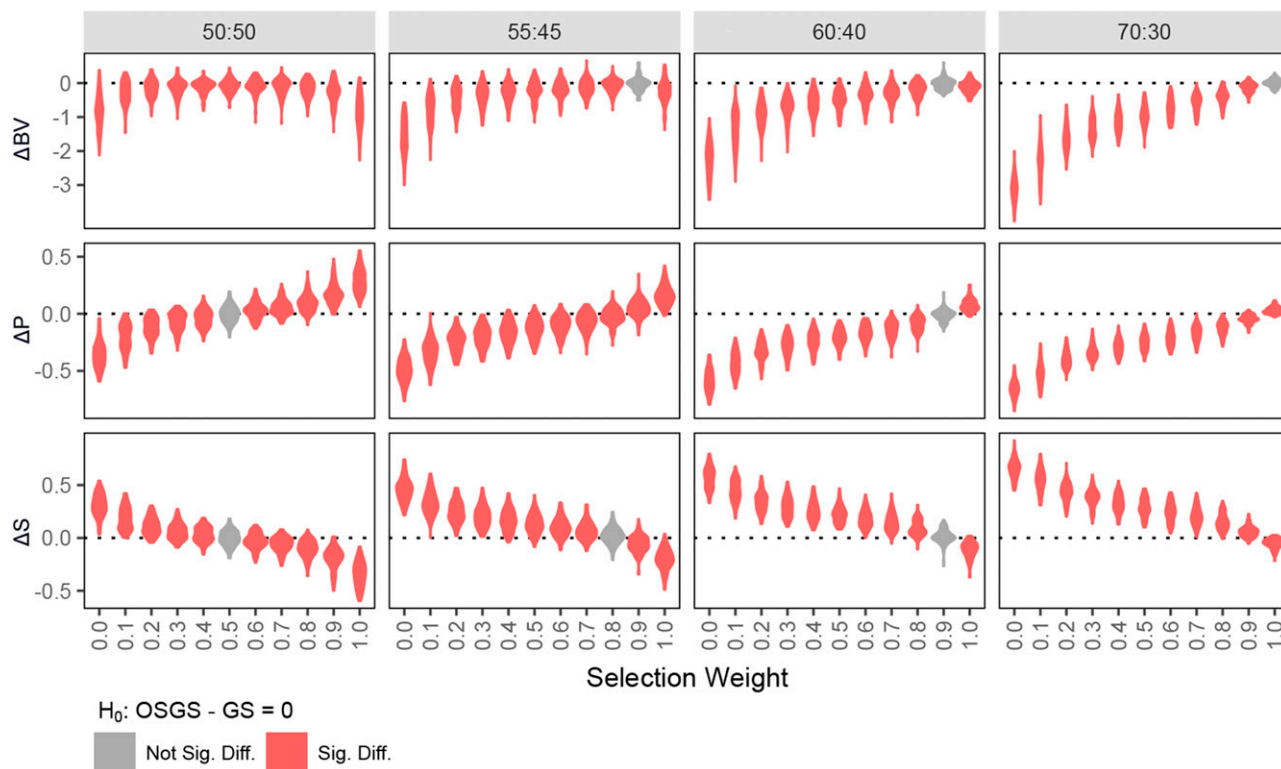
alleles across traits (Figure 3B). YLD showed the strongest difference between P and S distributions, followed by DTH, DTS and CL. Therefore, the greater the imbalance between P and S distributions, the smaller the gap between  $\hat{a}$  and  $\hat{a}_1$  as  $\hat{a}$  is largely predicted by P.

Distribution of marker effect estimates can inform about the proportion of favorable alleles contributed by each parent (Figure 2B, Figure S4-7). Late flowering in temperate environment (northern Europe) and high yield are favored in spring barley, while early flowering and large ear size are favored in maize, thus favorable DTH, YLD and CL are represented by positive marker effects and favorable DTS is represented by negative marker effects. Across all traits, we found variable proportions of favorable alleles (Table S1). The means and ranges of P proportions across all 25 families estimated from rrBLUP were 0.52 and 0.43 – 0.62 for barley DTH, 0.63 and 0.48 – 0.78 for barley YLD, 0.56 and 0.43 – 0.65 for maize DTS and 0.51 and 0.43 – 0.59 for maize CL. In barley, we found that the primary (elite) parent had slightly more favorable DTH alleles but many more favorable YLD alleles than the secondary (exotic) parents. In maize, we found that the primary parent had more favorable DTS alleles but about equal favorable CL alleles compared to the secondary parents. Provided that a trait is polygenic, results here suggested that the distribution of marker effects can be used as a reasonable approximation to the true proportions of favorable QTL.

In addition, most of the P and S distributions were significantly different, especially in the barley NAM population (Figure 3B). By comparing the P and S distributions for each trait and family using a Kolmogorov-Smirnov test, we found that all 25 barley NAM families but only about half of the maize NAM families had significant



**Figure 4** Performance comparison of OSGS under different  $\omega$  against GS in NAMs. We simulated a single generation of selection under OSGS/GS and evaluated the change in estimated breeding value ( $\Delta EBV$ ), P proportion ( $\Delta P$ ) and S proportion ( $\Delta S$ ).  $\Delta EBV$  is calculated as  $(\mu_{EBV,OSGS} - \mu_{EBV,GS})/\mu_{EBV,0}$ , where  $\mu_{EBV,OSGS}$  is the mean EBV of a family selected under OSGS,  $\mu_{EBV,GS}$  is the mean EBV of a family selected under GS, and  $\mu_{EBV,0}$  is the mean EBV of the initial (pre-selection) NAM family. Similarly,  $\Delta P$  is calculated as  $(\mu_{P,OSGS} - \mu_{P,GS})/\mu_{P,0}$  and  $\Delta S$  is calculated as  $(\mu_{S,OSGS} - \mu_{S,GS})/\mu_{S,0}$ . Significance is determined by t-test with Bonferroni correction ( $P = 0.05/25$ ).



**Figure 5** Performance comparison of OSGS under different  $\omega$  against GS in simulated  $F_2$ -derived populations. We applied a single generation of selection under OSGS/GS on 100 simulated  $F_2$ -derived populations (2cM/QTL) and evaluated the change in true breeding value ( $\Delta BV$ ), P proportion ( $\Delta P$ ) and S proportion ( $\Delta S$ ). Significance is determined by t-test with Bonferroni correction ( $P = 0.05/25$ ).

differences. The strongest difference in the distributions was observed in barley YLD, followed by barley DTH, maize DTS and maize CL. The distributions of P and S are more likely to be different in elite-exotic crosses (barley NAM) than elite-elite crosses (maize NAM). While rrBLUP assumes a single normal distribution of marker effects, the model is robust to the violation of the assumption given the good prediction accuracies from P and S.

There were weak negative correlations between  $\hat{a}_1$  and  $\hat{a}_2$  across all four traits (Figure 3C). While a strong positive correlation between the two would be ideal for selection, the lack of strong negative correlations implies that we can still select for both P and S without any severe constraints. To do so, we can apply index selection based on the ranks of  $\hat{a}_1$  and  $\hat{a}_2$  by treating the two predictions as two separate traits.

To evaluate the usefulness of OSGS in introgressing exotic alleles in a pre-breeding context, we simulated a single generation of selection on all four traits (Figure 4). In terms of estimated breeding values (EBVs), OSGS did not outperform GS in any of the tested selection weights ( $\omega$ ). However, OSGS can increase or decrease P and S in comparison to GS. As  $\omega$  decreased, P decreased and S increased, and vice versa. Based on these results, the ideal selection weights would be those that maximize the increase in S and minimize the EBV gap. Across all four traits,  $\omega$  of 0.4 to 0.6 appeared reasonable for efficient introgression of exotic alleles in pre-breeding programs.

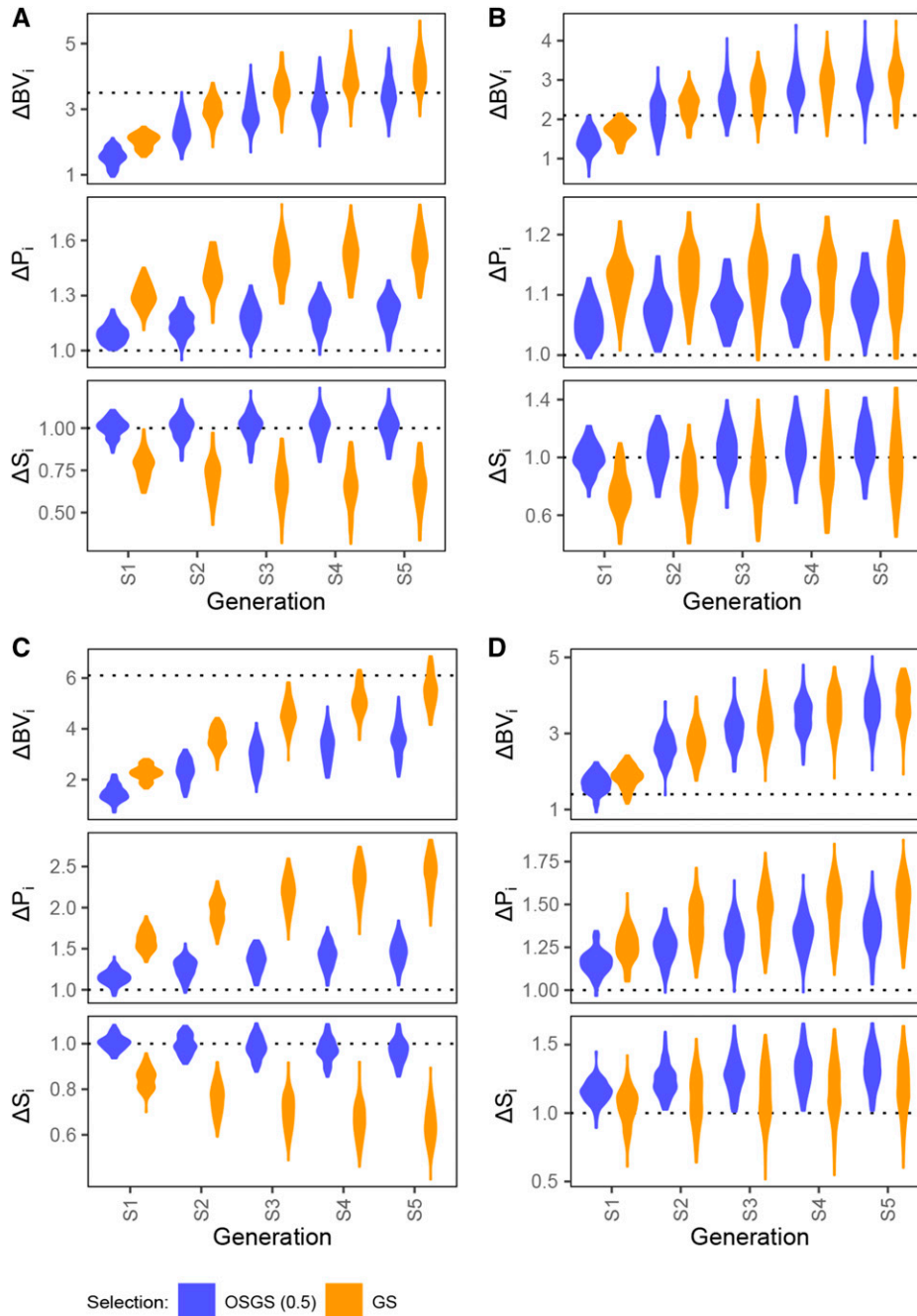
### Simulated data analysis

First, we evaluated the performance of OSGS under different proportion of P and S QTL and  $\omega$  (Figure 5, S8 and S9), and found that it can be optimized based on the proportion of P and S. In the case of P:S

= 50:50, OSGS with  $\omega$  of 0.5 resulted in similar true breeding values (BV) and P:S proportions to GS. As the proportion of P:S increases, a slight increase in  $\omega$  can minimize the BV gap between OSGS and GS, and still maintain a higher S proportion in OSGS than GS. Given that the proportions of estimated P and S marker effects reasonably approximated the true proportion of P and S QTL (Figure S10), we can adjust  $\omega$  in OSGS according to the estimated P and S proportions.

Comparing across  $F_2$ ,  $BC_1$  and  $rBC_1$ -derived populations, OSGS is best performed in an  $F_2$  population as it begins with an equal proportion of primary and secondary parent alleles (Figure 6 and S11, Table 2).  $F_2$  population provides a good starting ground for OSGS to elevate S proportion while keeping the BV gap with GS low. In a  $BC_1$  population, there is already a bias in the population toward primary parent alleles as the population has 75% primary parent alleles and 25% secondary parent alleles on average. While it is possible to minimize BV gap between OSGS and GS, there is little gain in S over multiple generations of recurrent selection. On the other hand, in a  $rBC_1$  population, the BV gap is too large to compensate for the gain in S. From a different perspective in the absence of OSGS, one is better off applying GS in a  $BC_1$  over an  $F_2$  population as it achieves higher breeding values faster (Figure 6 and S11, Table 2) without losing much S in the process.

Lastly, comparing between QTL density of 2cM/QTL and 20cM/QTL, there is more merit to using OSGS when the number of QTL are large (Figure 6A, 6D and S11). In the case where the QTL density is low (20cM/QTL), there is little difference between GS and OSGS (Figure 6D and S11) aside from OSGS is slightly better in increasing the S proportion. However, as we increased the QTL density to



**Figure 6** Performance comparison of OSGS against GS under recurrent selection. We applied five generations of recurrent selection under OSGS/GS on 100 simulated  $F_2$ ,  $BC_1$  and  $rBC_1$ -derived populations with P:S of 60:40. We showed the change in true breeding value ( $\Delta BV_i$ ), P proportion ( $\Delta P_i$ ) and S proportion ( $\Delta S_i$ ), where  $i$  is either OSGS with  $\omega = 0.5$  or GS.  $\Delta BV_i$  is calculated as  $(\mu_{BV,i} - \sigma_{BV,0})/\mu_{BV,0}$ ,  $\Delta P_i$  as  $\mu_{P,i}/\mu_{P,0}$  and  $\Delta S_i$  as  $\mu_{S,i}/\mu_{S,0}$ . [A]  $F_2$  with 2cM/QTL. [B]  $BC_1$  with 2cM/QTL. [C]  $rBC_1$  with 2cM/QTL. [D]  $F_2$  with 20cM/QTL.

2cM/QTL, we found that OSGS was able to keep the balance between favorable primary and secondary parent alleles throughout selection, while GS resulted in a larger discrepancy (Figure 6A and S11). This highlights the issue with GS in an elite-exotic population as few exotic alleles manage to enter the final breeding population. OSGS can be used to address this issue.

## DISCUSSION

In the recent years, there has been a growing interest in exploring ways for efficient introduction of novel genetic variation from exotic germplasm like landraces and wild relatives into modern breeding populations (Mascher *et al.* 2019). Even in elite crosses, current

selection practices can be strongly biased in favor of one parent (Fradgley *et al.* 2019), and linkage drag may limit the potential for favorable alleles to be selected from the phenotypically weaker of the two genomes. To circumvent this problem, Gorjanc *et al.* (2016) suggested an approach to create improved lines from purely exotic materials prior to crossing with the elite materials. Samayoa *et al.* (2018) suggested a slightly different approach where the exotic improvement is only performed on adaptation-related traits. Alternatively, Han *et al.* (2017) formulated a method to identify candidate exotic lines for introgressing small numbers of favorable exotic alleles into elite populations. Allier *et al.* (2020) further extended this approach for introgressing a larger number of favorable exotic alleles



■ **Table 2** Consequences of OSGS/GS on the proportions of favorable alleles and breeding values. Using P:S ratio of 60:40 as an example, we compared the mean proportions of favorable primary (P) and favorable secondary (S), and the true BV after five generations of recurrent selection. First two rows are the parents, third row is the pre-selected population, and the remaining are selected populations. P and S in generation 0 are essentially weighted means of P and S in the parents where the weights are the mean proportion of parental markers. For example, in BC<sub>1</sub>,  $P = 0.75 \times 0.60 = 0.45$  and  $S = 0.25 \times 0.40 = 0.10$

Gen	$\omega$	F <sub>2</sub>			BC <sub>1</sub>			rBC <sub>1</sub>		
		P	S	BV	P	S	BV	P	S	BV
P	—	0.60	0.00	4.40	0.60	0.00	4.40	0.60	0.00	4.40
S	—	0.00	0.40	-4.40	0.00	0.40	-4.40	0.00	0.40	-4.40
0	—	0.30	0.20	0.00	0.45	0.10	2.19	0.15	0.30	-2.18
5	0.0	0.11	0.35	-1.41	0.25	0.26	1.42	0.03	0.39	-3.48
5	0.1	0.21	0.30	1.27	0.36	0.20	3.82	0.09	0.36	-1.76
5	0.2	0.28	0.26	3.02	0.42	0.16	4.85	0.14	0.34	-0.18
5	0.3	0.31	0.24	3.57	0.45	0.14	5.22	0.17	0.32	0.63
5	0.4	0.34	0.21	4.19	0.47	0.12	5.30	0.20	0.30	1.22
5	0.5	0.36	0.20	4.40	0.49	0.11	5.43	0.22	0.29	1.70
5	0.6	0.39	0.18	4.74	0.51	0.09	5.39	0.24	0.28	2.17
5	0.7	0.41	0.17	5.01	0.52	0.08	5.50	0.27	0.26	2.55
5	0.8	0.44	0.14	5.09	0.54	0.06	5.36	0.30	0.24	3.09
5	0.9	0.49	0.10	5.14	0.57	0.03	5.00	0.36	0.20	3.63
5	1.0	0.54	0.05	4.66	0.59	0.01	4.55	0.43	0.14	3.60
5	GS	0.46	0.13	5.29	0.50	0.10	5.48	0.36	0.19	3.83

by identifying exotic candidates with higher ratios of favorable over unfavorable alleles. In a slightly different approach, Allier *et al.* (2019) proposed the usefulness criterion parental contribution (UCPC) as a metric that combines both the usefulness criterion (Schnell and Utz 1975) and parental genomic contributions in identifying exotic materials for crossing with elite populations. Ru and Bernardo (2019 and 2020) proposed introgressing linkage groups over QTL via targeted recombination.

While these approaches seem promising, none of them directly addresses the issues of genomic selection in elite-exotic populations. These approaches focus on identifying the best possible exotic line for crossing, and none attempts to improve the exotic introgression potential after crossing exotic and elite lines. Improvement on solely exotic lines (Gorjanc *et al.* 2016) may risk selecting for favorable alleles that are already present in elite populations. Selecting for exotic lines with the best combination to the target elite lines (Han *et al.* 2017; Allier *et al.* 2019; Allier *et al.* 2020) likely requires accurate predictions on the crosses performances, which calls for large training population and/or close relationships among the selected lines that may not be available.

Here, we propose using OSGS as a generalized framework for partitioning favorable trait contributions among parents. When applied on a single elite-exotic cross population, high prediction accuracies will be possible without requiring a large sized population for phenotyping (Brandariz and Bernardo 2019). This subsequently allows us to partition these predictions into favorable primary and secondary parental contributions with high confidence. OSGS is flexible with respect to the choice of the exotic genome and is complementary to any of the previously described approaches to accommodate those selected exotic lines. In addition, we have demonstrated that OSGS works using the barley and maize NAMs, furthering the potential of community-generated genetic resources as potent breeding tools. Moreover, Bernardo (2009) and our results suggest that it is likely better to use F<sub>2</sub>-derived NAMs to backcross-derived NAMs for this purpose.

In general, OSGS is robust to the choice of a statistical method and should work for other untested methods provided marker effects can be estimated and partitioned into two or more classes. However, one

might consider models that are better suited for the presumed trait genetic architectures. For example, LASSO might be a better option for traits regulated by few QTL since LASSO reduces the effects of most markers to zero.

In this paper we have shown that OSGS can maintain the balance between favorable primary and secondary parent allele proportions over several generations of selection. Hence, OSGS may also play a similar role in optimal contribution selection initially suggested by Meuwissen (1997). Optimal contribution selection aims to maintain genetic diversity in a population under selection by penalizing the estimated breeding values with relationships among selected individuals (Woolliams *et al.* 2015). In genomic setting this penalty is based on genomic relationships identified from all markers, which does not distinguish between favorable primary and secondary parent alleles. Therefore, OSGS can be complementary to optimal contribution selection as we could partition the kinship matrix into two matrices based on markers carrying favorable primary or secondary parent allele effects. Similar approach has been advocated for optimal contribution selection in rare breeds of livestock in the presence of introgression from cosmopolitan breeds (Wang *et al.* 2017a, 2017b and 2019).

There are several applications of OSGS remaining to be explored. We found that the distributions of favorable primary and secondary parent effects are different, especially in elite-exotic crosses. This is expected because of the joint action of selection and drift during and after species domestication. OSGS may provide an approach to studying this effect by comparing distributions across populations and species. The application of OSGS could be extended to multi-parental crosses using predictions based on identity-by-descent relationships due to originating parents. Multi-parent populations based on two or more elite lines and a single exotic parent are already in use in pre-breeding (Hao *et al.* 2019; Singh *et al.* 2018). There is a strong risk that phenotypic or genomic selection in these populations will discriminate against favorable alleles carried by the exotic parent to an even greater extent than we have shown in bi-parental populations (see also simulations by Gorjanc *et al.* 2016).

There might also be merits in combining OSGS with other approaches. For example, we can combine the parent selection

approaches of Allier *et al.* (2020) with OSGS. This may be particularly useful for breeding programs that attempt to use elite and exotic lines with high performance gaps in the traits of interest. In addition, OSGS can be extended to work with gametic variance-based selection (Bijma *et al.* 2020) by maintaining a balance in the parental contributions on gametic variance.

Lastly, the most promising application of OSGS may be its extension to multi-trait selection. This could be especially useful in elite-exotic crosses where the traits are not unanimously favorable in the elite lines. For example, the exotic parent may carry most favorable alleles for abiotic or biotic stress resistance, but the elite parent mostly for productivity traits.

## ACKNOWLEDGMENTS

G.G. acknowledges support from the BBSRC to The Roslin Institute (BBS/E/D/30002275) and The University of Edinburgh's Data-Driven Innovation Chancellor's Fellowship.

## REFERENCES

- 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 (Bethesda)* 9: 1469–1479. <https://doi.org/10.1534/g3.119.400129>
- 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. <https://doi.org/10.1007/s00122-019-03451-9>
- 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>
- Bijma, P., Y. C. J. Wientjes, and M. P. L. Calus, 2020 Breeding top genotypes and accelerating response to recurrent selection by selecting parents with greater gametic variance. *Genetics* 214: 91–107. <https://doi.org/10.1534/genetics.119.302643>
- Brandariz, S. P., and R. Bernardo, 2019 Small ad hoc vs. large general training populations for genomewide selection in maize biparental crosses. *Theor. Appl. Genet.* 132: 347–353. <https://doi.org/10.1007/s00122-018-3222-3>
- Brown, P. J., N. Upadyayula, G. S. Mahone, F. Tian, P. J. Bradbury *et al.*, 2011 Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS Genet.* 7: e1002383. <https://doi.org/10.1371/journal.pgen.1002383>
- Buckler, E. S., J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown *et al.*, 2009 The genetic architecture of maize flowering time. *Science* 325: 714–718. <https://doi.org/10.1126/science.1174276>
- Daetwyler, H. D., M. P. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey, 2013 Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193: 347–365. <https://doi.org/10.1534/genetics.112.147983>
- Endelman, J. B., 2011 Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4: 250–255. <https://doi.org/10.3835/plantgenome2011.08.0024>
- Faux, A. M., G. Gorjanc, R. C. Gaynor, M. Battagin, S. M. Edwards *et al.*, 2016 AlphaSim: software for breeding program simulation. *Plant Genome* 9. <https://doi.org/10.3835/plantgenome2016.02.0013>
- Fradgley, N., K. A. Gardner, J. Cockram, J. Elderfield, J. M. Hickey *et al.*, 2019 A large-scale pedigree resource of wheat reveals evidence for adaptation and selection by breeders. *PLoS Biol.* 17: e3000071. <https://doi.org/10.1371/journal.pbio.3000071>
- Friedman, J., T. Hastie, and R. Tibshirani, 2010 Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33: 1–22. <https://doi.org/10.18637/jss.v033.i01>
- 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>
- Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* 12: 186. <https://doi.org/10.1186/1471-2105-12-186>
- 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>
- Hao, M., L. Zhang, L. Zhao, S. Dai, A. Li *et al.*, 2019 A breeding strategy targeting the secondary gene pool of bread wheat: introgression from a synthetic hexaploid wheat. *Theor. Appl. Genet.* 132: 2285–2294. <https://doi.org/10.1007/s00122-019-03354-9>
- Hallauer, A. L., and M. J. Carena, 2012 Recurrent selection methods to improve germplasm in maize. *Maydica* 57: 266–283.
- Herzig, P., A. Maurer, V. Draba, R. Sharma, F. Draicchio *et al.*, 2018 Contrasting genetic regulation of plant development in wild barley grown in two European environments revealed by nested association mapping. *J. Exp. Bot.* 69: 1517–1531. <https://doi.org/10.1093/jxb/ery002>
- Hoerl, A. E., and R. W. Kennard, 1976 Ridge regression iterative estimation of the biasing parameter. *Commun. Stat-Theor. M.* 5: 77–88. <https://doi.org/10.1080/03610927608827333>
- 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. <https://doi.org/10.1038/s41588-019-0443-6>
- Maurer, A., V. Draba, Y. Jiang, F. Schnaithmann, R. Sharma *et al.*, 2015 Modelling the genetic architecture of flowering time control in barley through nested association mapping. *BMC Genomics* 16: 290. <https://doi.org/10.1186/s12864-015-1459-7>
- McMullen, M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. Li *et al.*, 2009 Genetic properties of the maize nested association mapping population. *Science* 325: 737–740. <https://doi.org/10.1126/science.1174320>
- Meuwissen, T. H. E., 1997 Maximizing the response of selection with a predefined rate of inbreeding. *J. Anim. Sci.* 75: 934–940. <https://doi.org/10.2527/1997.754934x>
- Pérez, P., and G. de los Campos, 2014 Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198: 483–495. <https://doi.org/10.1534/genetics.114.164442>
- R Core Team, 2020 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Ru, S., and R. Bernardo, 2019 Targeted recombination to increase genetic gain in self-pollinated species. *Theor. Appl. Genet.* 132: 289–300. <https://doi.org/10.1007/s00122-018-3216-1>
- Ru, S., and R. Bernardo, 2020 Predicted genetic gains from introgressing chromosome segments from exotic germplasm into an elite soybean cultivar. *Theor. Appl. Genet.* 133: 605–614. <https://doi.org/10.1007/s00122-019-03490-2>
- Samayoa, L. F., J. C. Dunne, R. J. Andres, and J. B. Holland, 2018 Harnessing maize biodiversity, pp. 335–366 in *The Maize Genome*, edited by Bennetzen, J., S. Flint-Garcia, C. Hirsch, and R. Tuberosa. Springer, New York. [https://doi.org/10.1007/978-3-319-97427-9\\_20](https://doi.org/10.1007/978-3-319-97427-9_20)
- Schnell, F. W., and H. F. Utz, 1975 F1-Leistung und Elternwahl in der Züchtung von Selbstbefruchtern, pp. 234–258 in *Bericht über die Arbeitstagung der Vereinigung Österreichischer Pflanzzüchter*. Gumpenstein, Österreich.
- Sharma, R., F. Draicchio, H. Bull, P. Herzig, A. Maurer *et al.*, 2018 Genome-wide association of yield traits in a nested association mapping population of barley reveals new gene diversity for future breeding. *J. Exp. Bot.* 69: 3811–3822. <https://doi.org/10.1093/jxb/ery178>
- Singh, S., P. Vikram, D. Sehgal, J. Burgueño, A. Sharma *et al.*, 2018 Harnessing genetic potential of wheat germplasm banks through impact-oriented-prebreeding for future food and nutritional security. *Sci. Rep.* 8: 12527. <https://doi.org/10.1038/s41598-018-30667-4>

- Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Series B (Methodological)* 58: 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Wang, Y., J. Bennewitz, and R. Wellmann, 2017a Novel optimum contribution selection methods accounting for conflicting objectives in breeding programs for livestock breeds with historical migration. *Genet. Sel. Evol.* 49: 45. <https://doi.org/10.1186/s12711-017-0320-7>
- Wang, Y., D. Segelke, R. Emmerling, J. Bennewitz, and R. Wellmann, 2017b Long-term impact of optimum contribution selection strategies on local livestock breeds with historical introgression using the example of German angler cattle. *G3 (Bethesda)* 7: 4009–4018 (Bethesda). <https://doi.org/10.1534/g3.117.300272>
- Wang, Y., J. Bennewitz, and R. Wellmann, 2019 Managing genomes of conserved livestock breeds with historical introgression to decrease genetic overlap with other breeds. *J. Anim. Breed. Genet.* 136: 505–517. <https://doi.org/10.1111/jbg.12405>
- 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, J., J. B. Holland, M. D. McMullen, and E. S. Buckler, 2008 Genetic design and statistical power of nested association mapping in maize. *Genetics* 178: 539–551. <https://doi.org/10.1534/genetics.107.074245>

*Communicating editor: P. Morrell*