

## ORIGINAL RESEARCH ARTICLE

## Crop Breeding &amp; Genetics

## Outliers and their distribution in breeding populations

Rex Bernardo 

Dep. of Agronomy and Plant Genetics,  
Univ. of Minnesota, 411 Borlaug Hall, 1991  
Upper Buford Circle, Saint Paul, MN  
55108, USA

## Correspondence

Rex Bernardo, Dep. of Agronomy and Plant  
Genetics, Univ. of Minnesota, 411 Borlaug  
Hall, 1991 Upper Buford Circle, Saint Paul,  
MN 55108, USA.

Email: [bernardo@umn.edu](mailto:bernardo@umn.edu)

Assigned to Associate Editor Marcio  
Resende, Jr.

## Abstract

Outliers with highly superior performance are valuable in plant breeding, but their distribution in populations has not been well-studied. My objectives were to determine (a) if outliers behave in a predictable manner; (b) if they are distributed according to a normal distribution as is assumed for quantitative traits; and (c) which parental characteristics are indicative of the best chances of getting progeny with extreme performance. All possible biparental populations were made among 15 simulated barley (*Hordeum vulgare* L.) parental lines in *BreedingGames* software. Ten million lines were simulated within each cross for a total of 1.05 billion lines. Within each biparental population, recombinant inbreds in the top 1.0, 0.1, and 0.01% tails had a continuous distribution, indicating that outliers behave in a predictable manner but are rare in practice only because the population sizes used in breeding are small. Having a finite number of loci led to slight kurtosis, which caused a minor upward bias when the usefulness criterion was applied to the extreme upper tails. The midparent value was an excellent indicator of which biparental crosses had high upper-tail means, to the extent that modeling the genetic variance within each biparental population had little added benefit. In the simulations, selection for protein concentration and Fusarium (*F. graminearum*) head blight resistance decreased the gains for yield and changed the biparental population that led to the highest yield gains. Results indicated that having one very large breeding population in a select-only scheme is inferior to two or more select-and-recombine cycles with smaller populations.

## 1 | INTRODUCTION

Outliers are lines, clones, or hybrids with extreme performance for yield or other important traits. They are candidates that do not perform like the rest of the breeding population from which they were developed. Cultivars that are outliers in terms of their performance represent a step change rather than a gradual improvement. Historical examples of maize (*Zea mays* L.) cultivars with outlier-type performance include ‘LG11’ developed by Limagrain for northern Europe in the early 1970s (Barrière et al., 2006), the U.S. public hybrid

‘B73 × Mo17’ in the mid-1970s (Troyer, 2006), and ‘P3394’ developed by Pioneer Hi-Bred in the mid-1990s (Mikel, 2008). Outliers as defined herein exclude individuals with unusual performance because they resulted from accidental outcrossing or contamination.

The distribution of quantitative traits such as yield has classically been modeled according to a normal distribution, which can be fully described through the population mean and variance. In plant breeding textbooks, the response to selection or the mean of the upper tail of the population typically focuses on the best 20, 10, or 5% of individuals (Allard, 1960; Bernardo, 2020; Fehr, 1987). In contrast, the distribution of outliers in breeding populations has not been well-studied.

**Abbreviations:** QTL, quantitative trait loci.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Crop Science* published by Wiley Periodicals LLC on behalf of Crop Science Society of America.

Are outliers extreme, nonrepeatable occurrences, or are they empirically nonrepeatable because the population sizes used in breeding are too small to make them repeatable? Are individuals in the upper 1.0, 0.1, or 0.01% of a breeding population distributed according to a normal distribution, particularly when a trait is controlled by a finite number of loci (Chevalet, 1994; Fernando et al., 1994; Pong-Wong et al., 1999)?

Information on the distribution of outliers could be useful in choosing sizes of breeding populations and in maximizing the frequency of outliers through approaches such as the aggressive use of genomewide prediction (Bernardo, 2021). My objectives in this simulation study were to (a) determine if outliers behave in a predictable manner in large breeding populations; (b) determine if outliers are distributed according to a normal distribution as is assumed for quantitative traits; and (c) identify characteristics of parental lines that are indicative of the best chances of getting progeny with extreme performance.

## 2 | MATERIALS AND METHODS

### 2.1 | Genetic model and parental lines

The *BreedingGames* Fortran simulation software (Bernardo, 2017a; <https://bernardo-group.org/wp-content/uploads/2021/01/BreedingGames.zip>) was modified to allow studying outliers, which in terms of their distribution were defined as those in the 1.0, 0.1, or 0.01% upper tails. A founder population was first developed by crossing two simulated barley (*Hordeum vulgare* L.) homozygous lines that differed at 400 quantitative trait loci (QTL) for yield, 80 QTL for protein concentration, and 25 QTL for Fusarium (*F. graminearum*) head blight resistance. Subsets of the 400 QTL for yield controlled two or more traits, and such pleiotropy led to unfavorable genetic correlations of  $-0.47$  between yield and protein concentration,  $0.29$  between yield and Fusarium head blight incidence, and  $-0.44$  between protein concentration and Fusarium head blight incidence.

The QTL were located at random, according to a uniform distribution, across seven barley chromosomes. The sizes of each chromosome and of the genome (1,137 cM) corresponded to those in a barley consensus map (Muñoz-Amatriáin et al., 2011). The QTL effects for each trait had a geometric distribution, with few QTL having large effects and many QTL having small effects (Lande & Thompson, 1990), with the exception that one QTL for Fusarium head blight resistance had a major effect, reducing infection by about 5% (Bernardo, 2017b). The genotypic value of an individual was obtained by summing the effects of QTL alleles carried by the individual across the genome.

### Core Ideas

- Outliers in a breeding population behave in a predictable manner.
- Outliers are rare only because breeding populations are typically small.
- A finite number of loci causes slight kurtosis in the trait distribution.
- In choosing populations, midparent value is much more informative than genetic variance.
- Selection for two or more cycles is better than selection in one large population.

For each trait, the mean and variance among lines in the founder population were estimated by generating 10,000 random recombinant inbreds and calculating the mean and variance among the genotypic values of the lines. The genotypic values for each trait were standardized so that they had a mean of zero and a variance of 1, and they were subsequently scaled so that the trait means (genetic standard deviations in parentheses) in the founder population were  $5.12 \text{ Mg ha}^{-1}$  ( $0.42 \text{ Mg ha}^{-1}$ ) for yield,  $130 \text{ g kg}^{-1}$  ( $0.75 \text{ g kg}^{-1}$ ) for protein concentration, and  $11.0\%$  ( $3.0\%$ ) for Fusarium head blight incidence.

A total of 15 homozygous lines developed from the founder population were then selected to serve as parents in *BreedingGames* (Bernardo, 2017a) as well as in this study. These 15 simulated parental lines differed in yield, protein concentration, and Fusarium head blight incidence (Table 1). The parents were numbered in descending order according to their yield.

### 2.2 | Outliers in biparental populations

Outliers were identified in a single generation of selection among recombinant inbreds developed from multiple crosses, with the genotypic values being known without error. Biparental populations were simulated for all possible pairs of the 15 parents. Within each of the 105 resulting biparental populations, 10 million homozygous lines were developed. A total of 1.05 billion lines were therefore simulated in this study. Each homozygous line was developed through six generations of selfing followed by a final step of doubled haploidy to ensure complete homozygosity. The 10 million lines within each biparental population were generated in 10 sets of 1 million lines each to circumvent issues of stack overflow in the Fortran executable program. Genotypic values of the lines were obtained in the same manner as for the 15 parents.

**TABLE 1** Mean performance of 15 simulated barley parental lines for yield, protein concentration, and Fusarium head blight resistance

Parent	Yield Mg ha <sup>-1</sup>	Protein g kg <sup>-1</sup>	Fusarium head blight	
			Incidence %	Has major resistance allele
1	6.44	124	10.8	Yes
2	6.29	121	12.6	No
3	6.22	111	16.3	No
4	6.13	111	12.2	Yes
5	6.07	118	8.6	Yes
6	5.84	122	7.5	Yes
7	5.77	127	5.9	Yes
8	5.72	122	6.2	Yes
9	5.70	128	6.3	Yes
10	5.37	107	17.5	No
11	5.32	113	6.4	Yes
12	5.29	105	17.3	No
13	5.21	133	3.3	Yes
14	5.09	115	7.7	Yes
15	4.53	144	2.4	Yes

The analysis of outliers focused on yield, either with selection for yield alone or selection for yield after selecting for protein concentration and Fusarium head blight resistance (Section 2.4). The mean and genetic standard deviation among genotypic values were calculated among the 10 million lines within each biparental population. Outliers were characterized by calculating the mean of the top 1.0% (100,000 out of 10 million), top 0.1% (10,000 out of 10 million), and top 0.01% (1,000 out of 10 million) lines for yield. These means were denoted by  $\mu_{1\%}$ ,  $\mu_{0.1\%}$ , and  $\mu_{0.01\%}$ , respectively. Skewness and excess kurtosis coefficients were calculated among the 10 million lines within each biparental population.

The  $\mu_{1\%}$ ,  $\mu_{0.1\%}$ , and  $\mu_{0.01\%}$  values were compared with the usefulness criterion (Bernardo, 2020; Melchinger et al., 1988), which was the predicted mean of a given upper tail according to a normal distribution. Standardized selection differentials for 1.0, 0.1, and 0.01% selected were obtained via the NORMDIST (normal distribution) and NORMSINV (inverse of the standard normal distribution) functions in Microsoft Excel (Mackay, 2020). Furthermore, the  $\mu_{1\%}$ ,  $\mu_{0.1\%}$ , and  $\mu_{0.01\%}$  values were expressed as z-scores in two ways. First,  $\mu_{1\%}$ ,  $\mu_{0.1\%}$ , and  $\mu_{0.01\%}$  were converted into across-population z-scores relative to the mean (5.12 Mg ha<sup>-1</sup>) and genetic standard deviation (0.42 Mg ha<sup>-1</sup>) among lines in the founder population. Such z-scores allowed comparisons across the 105 biparental populations. Second,  $\mu_{1\%}$ ,  $\mu_{0.1\%}$ , and  $\mu_{0.01\%}$  values were converted into within-population

z-scores via the mean and genetic standard deviation among lines within the biparental population.

## 2.3 | Predictors of outliers

Across the 105 biparental populations, correlations were calculated between the  $\mu_{1\%}$ ,  $\mu_{0.1\%}$ , and  $\mu_{0.01\%}$  values and the mean, genetic standard deviation, and usefulness criterion (with the same proportion selected) in each biparental population. Correlations were also calculated between the  $\mu_{1\%}$ ,  $\mu_{0.1\%}$ , and  $\mu_{0.01\%}$  values and genomewide predictions of the performance of the lines in the corresponding upper tails.

In particular, the genomewide prediction model implemented within *BreedingGames* was used (Bernardo, 2017b). This prediction model was obtained by ridge regression-best linear unbiased prediction (Meuwissen et al., 2001) with marker and phenotypic data among 1,000 random lines developed in the founder population. The marker data were for 512 well-spaced biallelic markers that were polymorphic between the two parents of the founder population. The phenotypic data corresponded to heritabilities ( $h^2$ ) of 0.50 for yield, 0.75 for protein concentration, and 0.40 for Fusarium head blight resistance in the founder population (Bernardo, 2017b). The genomewide predictive abilities were 0.52 for yield, 0.78 for protein concentration, and 0.44 for Fusarium head blight incidence in the founder population.

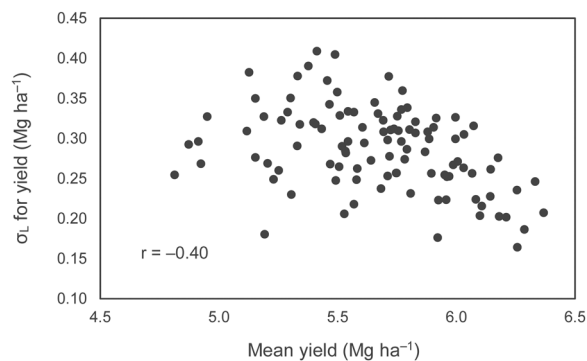
## 2.4 | Multiple-trait selection

The influence of multiple-trait selection on outliers was investigated by applying the same criteria used in *BreedingGames*, in which the objective was to develop a line that met standards for protein concentration and Fusarium head blight resistance (Bernardo, 2017a). Of the 10 million lines within each biparental population, only those lines that had  $\leq 5\%$  Fusarium head blight incidence and 110–130 g kg<sup>-1</sup> protein were retained (Bernardo, 2017b). The mean yields of the top 100,000 (1.0%), top 10,000 (0.1%), and top 1,000 (0.01%) of the remaining lines were then calculated if the number of remaining lines was equal to or exceeded the aforementioned numbers of lines. The z-scores were calculated as previously described.

# 3 | RESULTS

## 3.1 | Means and genetic standard deviation

The yields of the 15 simulated parents ranged from 4.53 to 6.44 Mg ha<sup>-1</sup> (Table 1). The mean of the 15 parents (5.67 Mg ha<sup>-1</sup>) was higher than the mean of the founder



**FIGURE 1** Mean versus genetic standard deviation among lines ( $\sigma_L$ ) for yield in each of 105 simulated biparental populations

population (5.12 Mg ha<sup>-1</sup>). The mean yields of the 105 biparental populations ranged from 4.81 (Parent 14 × Parent 15) to 6.37 Mg ha<sup>-1</sup> (Parent 1 × Parent 2; Figure 1). The mean yields of the biparental populations were perfectly correlated with the midparent values.

Within each biparental population, the genetic standard deviation among lines ( $\sigma_L$ ) ranged from 0.16 (Parent 1 × Parent 5) to 0.41 Mg ha<sup>-1</sup> (Parent 2 × Parent 15; Figure 1). The  $\sigma_L$  in the founder population was 0.42 Mg ha<sup>-1</sup>. Across the 105 biparental populations, the correlation between the mean and  $\sigma_L$  was  $-0.40$  (Figure 1), whereas the correlation between the mean and the genetic variance among lines was  $-0.39$ .

The Parent 1 × Parent 3 population had the highest values of  $\mu_{1\%}$  (6.95 Mg ha<sup>-1</sup>),  $\mu_{0.1\%}$  (7.09 Mg ha<sup>-1</sup>), and  $\mu_{0.01\%}$  (7.21 Mg ha<sup>-1</sup>). Relative to the mean and  $\sigma_L$  in the founder population, these extreme-tail means were equivalent to 4.37–4.98 units of  $\sigma_L$ . The Parent 14 × Parent 15 population had the lowest values of  $\mu_{1\%}$  (5.48 Mg ha<sup>-1</sup>),  $\mu_{0.1\%}$  (5.64 Mg ha<sup>-1</sup>), and  $\mu_{0.01\%}$  (5.77 Mg ha<sup>-1</sup>). Relative to the within-population means and  $\sigma_L$ , the  $\mu_{1\%}$  values across the 105 biparental populations corresponded to a mean gain (in units of  $\sigma_L$ , with range in parentheses) of 2.25 (2.02, 2.32), the  $\mu_{0.1\%}$  values corresponded to a mean gain of 2.86 (2.45, 3.06), and the  $\mu_{0.01\%}$  values corresponded to a mean gain of 3.32 (2.78, 3.67).

The mean of the biparental population was highly correlated with the means of the extreme upper tails. In particular, the mean of the biparental population had a correlation of 0.93 with  $\mu_{1\%}$ , 0.91 with  $\mu_{0.1\%}$ , and 0.89 with  $\mu_{0.01\%}$ . Furthermore, a slightly curvilinear relationship was found between the mean of the biparental population and the  $\mu_{1\%}$ ,  $\mu_{0.1\%}$ , and  $\mu_{0.01\%}$  values (Figure 2), with the curvilinear regression coefficient being statistically significant ( $p \leq .0002$ ) in all three cases.

In contrast, the correlations of  $\sigma_L$  with  $\mu_{1\%}$ ,  $\mu_{0.1\%}$ , and  $\mu_{0.01\%}$  were nonsignificant ( $p > .05$ ) and ranged from  $-0.05$  to 0.05. Correlations among the  $\mu_{1\%}$ ,  $\mu_{0.1\%}$ , and  $\mu_{0.01\%}$  values were  $> 0.99$ . Genomewide prediction was effective for identifying which biparental population had the highest upper-tail

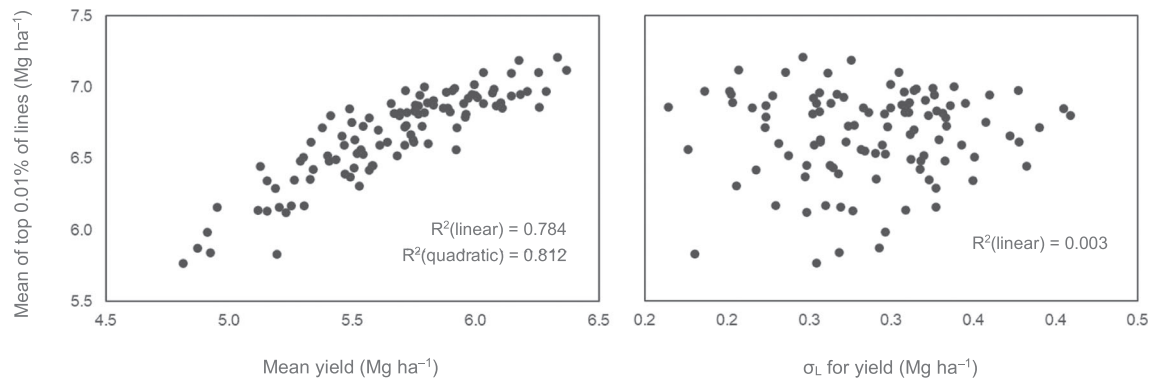
means. Specifically, the correlations between  $\mu_{1\%}$ ,  $\mu_{0.1\%}$ , or  $\mu_{0.01\%}$  and the corresponding upper-tail means obtained via genomewide prediction were all equal to 0.95.

### 3.2 | Upper-tail distributions and multiple-trait selection

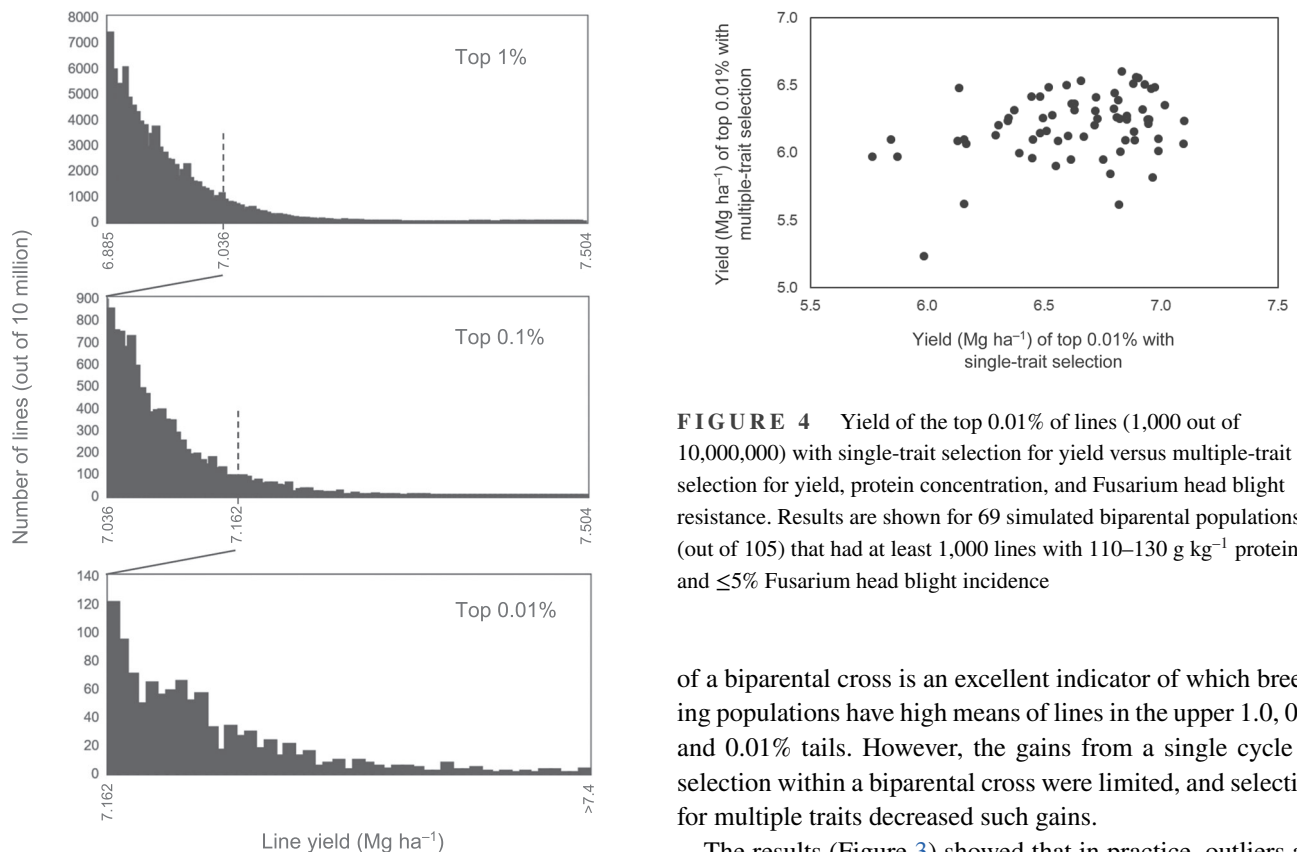
As exemplified by the upper-tail distribution in the Parent 1 × Parent 2 population, the best 1.0, 0.1, and 0.01% of lines had a continuous distribution (Figure 3). Skewness was absent for yield within each of the 105 biparental populations, with the observed skewness coefficients all being within 0.002 of the expected value of zero. Slight but consistent negative kurtosis was observed for yield within each biparental population, both when kurtosis was analyzed across the 10 million lines within each population and when kurtosis was analyzed within each of the 10 sets of 1 million lines within each population. The excess kurtosis, which has an expected value of zero when kurtosis is absent, ranged from  $-0.82$  in the Parent 1 × Parent 9 population to  $-0.04$  in the Parent 4 × Parent 5 population and had a mean of  $-0.27$ . Mean excess kurtosis for protein concentration ( $-0.29$ ) was close to that for yield, whereas mean excess kurtosis was strongest for Fusarium head blight incidence ( $-0.63$ ).

The usefulness criterion had an upward bias relative to the corresponding  $\mu_{1\%}$ ,  $\mu_{0.1\%}$ , and  $\mu_{0.01\%}$  values. Across the 105 biparental populations, the mean bias in the usefulness criterion (Mg ha<sup>-1</sup>, range in parentheses) was 0.043 (0.143, 0.002) relative to  $\mu_{1\%}$ , 0.091 (0.245, 0.006) relative to  $\mu_{0.1\%}$ , and 0.140 (0.347, 0.014) relative to  $\mu_{0.01\%}$ .

Compared with selection for yield alone, selection for multiple traits decreased the  $\mu_{1\%}$ ,  $\mu_{0.1\%}$ , and  $\mu_{0.01\%}$  values for yield. Many of the biparental populations, including the population with the highest mean yield (Parent 1 × Parent 2) and the population with the highest  $\mu_{1\%}$ ,  $\mu_{0.1\%}$ , and  $\mu_{0.01\%}$  values for yield alone (Parent 1 × Parent 3), did not have any lines that met the criteria of 110–130 g kg<sup>-1</sup> protein and  $\leq 5\%$  Fusarium head blight incidence. When the highest-yielding parent (Parent 1) led to lines that met the criteria for protein concentration and Fusarium head blight resistance, the frequencies (in parentheses) of such lines were very low and were found only in crosses with Parent 7 (0.19%), Parent 9 (0.03%), Parent 11 (0.02%), Parent 13 (1.67%), and Parent 15 (0.11%). Only 69 out of the 105 biparental populations had at least 1,000 lines with the desired values for protein concentration and Fusarium head blight incidence (Figure 4). With multiple-trait selection, the biparental population with the highest-yielding upper tail was Parent 2 × Parent 13 ( $\mu_{1\%} = 6.19$  Mg ha<sup>-1</sup>,  $\mu_{0.1\%} = 6.43$  Mg ha<sup>-1</sup>, and  $\mu_{0.01\%} = 6.60$  Mg ha<sup>-1</sup>). Relative to the mean and  $\sigma_L$  for yield in the founder population, these extreme-tail means with multiple-trait selection were equivalent to 2.56–3.53 units of  $\sigma_L$ .



**FIGURE 2** Mean yield of the top 0.01% of lines (1,000 out of 10 million) versus mean and genetic standard deviation among lines ( $\sigma_L$ ) in each of 105 simulated biparental populations



**FIGURE 3** Frequency distribution the top 1.0% (100,000 out of 10 million), top 0.01% (10,000), and top 0.01% (1,000) of lines in the simulated biparental population with the highest midparent value for yield (Parent 1  $\times$  Parent 2)

## 4 | DISCUSSION

### 4.1 | Distribution of outliers

The main findings in this study were that outliers in a breeding population behave in a predictable manner according to a slightly nonnormal distribution and that the midparent value

**FIGURE 4** Yield of the top 0.01% of lines (1,000 out of 10,000,000) with single-trait selection for yield versus multiple-trait selection for yield, protein concentration, and Fusarium head blight resistance. Results are shown for 69 simulated biparental populations (out of 105) that had at least 1,000 lines with 110–130 g kg<sup>-1</sup> protein and  $\leq 5\%$  Fusarium head blight incidence

of a biparental cross is an excellent indicator of which breeding populations have high means of lines in the upper 1.0, 0.1, and 0.01% tails. However, the gains from a single cycle of selection within a biparental cross were limited, and selection for multiple traits decreased such gains.

The results (Figure 3) showed that in practice, outliers are rare only because the sizes of biparental populations used in a breeding program are not large enough to allow outliers to appear consistently. The population size of 10 million lines per biparental cross simulated in this study was obviously unrealistic and was used simply to assess the distribution of lines in the extreme tails. The sizes of breeding populations vary across species and breeding programs (Bernardo, 2003; Yonezawa & Yamagata, 1978) but are typically in the order of 50 to 500 in maize, soybean [*Glycine max* (L.) Merrill], and apple (*Malus  $\times$  domestica* Borkh.) (Bernardo, 2020). Therefore, whereas steady genetic gains can be achieved by selecting the best 5 to 20% of individuals in a breeding population (Allard, 1960; Bernardo, 2020; Fehr, 1987), finding progeny

with extreme, outlier-type performance in a typical breeding program is largely a matter of chance.

The slight but consistent kurtosis in the distribution of trait values was due to a having a finite number of loci controlling the trait. The infinitesimal model, which assumes that a quantitative trait is controlled by a very large (conceptually infinite) number of loci each with very small effects, is expected to lead to the asymptotic property found in a normal distribution. But when a trait is controlled by a finite number of loci (Chevalet, 1994; Fernando et al., 1994; Pong-Wong et al., 1999), the expected distribution in the upper tail is not asymptotic but instead reaches a limit when all of the favorable alleles found in the two parents are accumulated in a single individual (Bailey & Comstock, 1976). Kurtosis was manifested by the distribution having thinner tails than expected from a normal distribution (i.e., platykurtic), and the tails were thinner for a trait controlled by fewer loci (25 QTL for Fusarium head blight) than for a trait controlled by more loci (400 QTL for yield).

It remains to be seen whether a statistical distribution known to be symmetric and platykurtic would be more useful than a normal distribution for modeling the extreme upper-tail behavior of a quantitative trait. For example, a raised cosine distribution (Rinne, 2010) is symmetric and has an excess kurtosis coefficient of  $-0.59$ , which was close to the mean excess kurtosis of  $-0.63$  observed for Fusarium head blight incidence. It seems more prudent, however, to instead simulate the trait distribution via a given linkage map and a hypothesized number of QTL or an estimated number of effective factors (Daetwyler et al., 2008; Li & Ji, 2005) controlling the trait.

## 4.2 | Breeding implications

A consequence of the observed kurtosis was a slight but consistent upward bias in the usefulness criterion, for which the standardized selection differential is calculated according to a normal distribution. Across the 105 biparental populations, the mean bias (range in parentheses) was 0.7% (0.1, 2.1%) with 1.0% selected, 1.4% (0.1, 3.5%) with 0.1% selected, and 2.1% (0.2, 5.0%) with 0.01% selected. The bias was therefore minor on average, and it suggested that a normal distribution could continue to be used to model or predict the means of the upper tails of the trait distribution via the usefulness criterion or, similarly, via the breeder's equation (Lush, 1937).

Breeding populations are typically created from good  $\times$  good crosses, and the results indicated that the mean of the two parents (i.e., midparent value) is an excellent indicator of which breeding populations would tend to have progeny with outlier-type performance. Given that both the population mean and variance determine the distribution of a quantitative trait, much research has been conducted on modeling the

genetic variance within a given cross. Such modeling or prediction of within-population genetic variance has been based on prior estimates of genetic variance in progenitor or related crosses (Bernardo & Nyquist, 1998; Lian et al., 2015) or on genomewide marker effects (Adeyemo & Bernardo, 2019; Bernardo, 2014; Lian et al., 2015; Neyhart & Smith, 2019; Osthusenrich et al., 2018; Tiede et al., 2016). The results herein, however, indicated only a limited value in considering both the mean and genetic variance (as opposed to the mean alone) in identifying breeding populations with high means of the extreme tails. This result was attributed to the variance among population means being much larger than the variance among  $\sigma_L$  values among the biparental crosses, as was previously found by Zhong and Jannink (2007). For yield, the variance among the means of the 105 biparental crosses was  $0.1217 \text{ Mg}^2 \text{ ha}^{-2}$ , whereas the variance among  $\sigma_L$  values was  $0.0026 \text{ Mg}^2 \text{ ha}^{-2}$ . Because the variance among means was 46 times as large as the variance among  $\sigma_L$  values, modeling  $\sigma_L$  in addition to the population mean had limited effectiveness.

In accordance with both theory (Hazel & Lush, 1942) and common experience among plant breeders, multiple-trait selection reduced the performance of the selected lines for the main trait of interest. When selection was conducted for protein concentration and Fusarium head blight resistance in addition to yield, the breeding population with the highest means of the extreme tails was no longer from a cross between two high-yielding parents (Parent 1  $\times$  Parent 3). Instead, the best cross was between Parent 2, which was a high-yielding ( $6.29 \text{ Mg ha}^{-1}$ ), Fusarium-susceptible (12.6%) line that did not carry the major QTL allele for Fusarium head blight resistance, and Parent 13, which was a low-yielding ( $5.21 \text{ Mg ha}^{-1}$ ), Fusarium-resistant (3.3%) line that carried the major QTL allele for Fusarium head blight resistance. Whereas selection for yield alone led to a maximum upper-tail mean of nearly  $5.0 \sigma_L$  relative to the founder population, multiple-trait selection reduced the maximum upper-tail mean to about  $3.5 \sigma_L$ . Multiple-trait selection therefore not only reduced genetic gains but also changed the breeding population that led to the maximum gains.

Relative to the mean and  $\sigma_L$  within each biparental cross, the 0.01% upper-tail means with selection for yield alone ranged from 2.78 to  $3.67 \sigma_L$  and had a mean of  $3.32 \sigma_L$  across the 105 biparental crosses. The potential gains from a single cycle of very stringent selection within an extremely large biparental population therefore exceeded  $3 \sigma_L$ . Realized gains are obtained by multiplying the maximum gains by the trait  $h^2$ . Assuming (for simplicity) that the yield  $h^2$  was 0.60 within each biparental cross, the mean potential gain of  $3.32 \sigma_L$  with 10 million lines per population corresponded to a mean realized gain of only  $2.00 \sigma_L$ . This mean realized gain is less than the gains of  $\geq 2.13 \sigma_L$  achieved through one cycle of genomewide selection followed by phenotypic selection, with  $h^2$  also being 0.60 (Bernardo, 2021).

In other words, breeders could try to obtain candidates with outlier-level performance via two general selection approaches. First, as investigated herein, breeders could use a select-only approach and try to maximize genetic gain by growing a very large breeding population. This select-only approach is appealing because it attempts to maximize short-term genetic gains (Sprague, 1984), and it has long been used as a standard breeding practice for cultivar development in many species (Allard, 1960). However, a select-only approach is limited because the standardized selection differential is not a linear function of the proportion selected; increasing the stringency of selection 100-fold from 1.0% to 0.01% increases the standardized selection differential (and the expected gain) by only about 1.5-fold. In practice, realized gains with few individuals selected would also be subject to genetic drift (Hanrahan et al., 1973; Smith, 1979; Weyhrich et al., 1998).

Second, breeders could use a select-and-recombine approach by identifying the best candidates in a biparental population via genomewide markers or phenotypic evaluation, recombining them to form the next cycle, and selecting lines from the resulting improved population. This approach, which relies on smaller but steady gains in the long term, is best exemplified by the Illinois long-term selection experiment wherein gains of 21 to 28 additive genetic standard deviations were achieved by analyzing fewer than 8,000 ears across 100 cycles of selection for oil and protein (Dudley & Lambert, 2004). A comparison of the gains in this study ( $2.0 \sigma_L$ , as described above) with 10 million lines versus the gains in an equal-time, equal-budget breeding scheme ( $\geq 2.13 \sigma_L$ ) that involved < 500 individuals in two cycles (Bernardo, 2021) indicates the superiority of a select-and-recombine approach. Overall, the results herein indicated that further studies that consider gains per unit time and cost should involve multiple cycles of selection rather than finding rare outliers in a single large population.

## AUTHOR CONTRIBUTIONS

Rex Bernardo: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

## CONFLICT OF INTEREST

The author declares that he has no conflict of interest.

## ORCID

Rex Bernardo  <https://orcid.org/0000-0003-3323-4690>

## REFERENCES

Adeyemo, E., & Bernardo, R. (2019). Predicting genetic variance from genomewide marker effects estimated from a diverse panel of maize inbreds. *Crop Science*, 59, 583–590. <https://doi.org/10.2135/cropsci2018.08.0525>

- Allard, R. W. (1960). *Principles of plant breeding*. John Wiley and Sons.
- Bailey, T. B., Jr., & Comstock, R. E. (1976). Linkage and the synthesis of better genotypes in self-fertilizing species. *Crop Science*, 16, 363–370. <https://doi.org/10.2135/cropsci1976.0011183X001600030012x>
- Barrière, Y., Alber, D., Dolstra, O., Lapiere, C., Motto, M., Ordás Pérez, A., Van Waes, J., Vlasminkel, L., Welcker, C., & Monod, J. P. (2006). Past and prospects of forage maize breeding in Europe. II. History, germplasm evolution and correlative agronomic changes. *Maydica*, 51, 435–449.
- Bernardo, R. (2003). Parental selection, number of breeding populations, and size of each population in inbred development. *Theoretical and Applied Genetics*, 107, 1252–1256.
- Bernardo, R. (2014). Genomewide selection of parental inbreds: Classes of loci and virtual biparental populations. *Crop Science*, 54, 2586–2595. <https://doi.org/10.2135/cropsci2014.01.0088>
- Bernardo, R. (2017a). *BreedingGames* software. *Crop Science*, 57, 2313–2313. <https://doi.org/10.2135/cropsci2017.07.04191e>
- Bernardo, R. (2017b). *BreedingGames* player guide. <https://bernardo-group.org/wp-content/uploads/2021/01/BreedingGames.zip>
- Bernardo, R. (2020). *Breeding for quantitative traits in plants* (3rd ed.). Stemma Press.
- Bernardo, R. (2021). Upgrading a maize breeding program via two-cycle genomewide selection: Same cost, same or less time, and larger gains. *Crop Science*, 61, 2444–2455. <https://doi.org/10.1002/csc2.20516>
- Bernardo, R., & Nyquist, W. E. (1998). Additive and testcross genetic variances in crosses among recombinant inbreds. *Theoretical and Applied Genetics*, 97, 116–121.
- Chevalet, C. (1994). An approximate theory of selection assuming a finite number of quantitative trait loci. *Genetics Selection Evolution*, 26, 379–400.
- Daetwyler, H. D., Villanueva, B., & Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genomewide approach. *PLOS One*, 3, e3395. <https://doi.org/10.1371/journal.pone.0003395>
- Dudley, J. W., & Lambert, R. J. (2004). 100 generations of selection for oil and protein in maize. *Plant Breeding Reviews*, 24(1), 79–110.
- Fehr, W. R. (1987). *Principles of cultivar development. Volume 1. Theory and technique*. Macmillan.
- Fernando, R. L., Stricker, C., & Elston, R. C. (1994). The finite polygenic mixed model: An alternative formulation for the mixed model of inheritance. *Theoretical and Applied Genetics*, 88, 573–580.
- Hanrahan, J. P., Eisen, E. J., & Lagates, J. E. (1973). Effects of population size and selection intensity of short-term response to selection for postweaning gain in mice. *Genetics*, 73, 513–530.
- Hazel, L. N., & Lush, J. L. (1942). The efficiency of three methods of selection. *Journal of Heredity*, 33, 393–399.
- Lande, R., & Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124, 743–756.
- Li, J., & Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95, 221–227.
- Lian, L., Jacobson, A., Zhong, S., & Bernardo, R. (2015). Prediction of genetic variance in biparental maize populations: Genomewide marker effects versus mean genetic variance in prior populations. *Crop Science*, 55, 1181–1188. <https://doi.org/10.2135/cropsci2014.10.0729>
- Lush, J. L. (1937). *Animal breeding plans*. Iowa State College Press.
- Mackay, I. (2020). *Selection intensity*. CGIAR Excellence in Breeding Platform. [https://excellenceinbreeding.org/sites/default/files/manual/EiB\\_M2\\_Selection%20Intensity\\_26-10-20.pdf](https://excellenceinbreeding.org/sites/default/files/manual/EiB_M2_Selection%20Intensity_26-10-20.pdf)

- Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*, 1819–1829.
- Melchinger, A. E., Schmidt, W., & Geiger, H. H. (1988). Comparison of testcrosses produced from  $F_2$  and first backcross populations of maize. *Crop Science*, *28*, 743–749. <https://doi.org/10.2135/cropsci1988.0011183X002800050004x>
- Mikel, M. A. (2008). Genetic diversity and improvement of contemporary proprietary North American dent corn. *Crop Science*, *48*, 1686–1695. <https://doi.org/10.2135/cropsci2008.01.0039>
- Muñoz-Amatriain, M., Moscou, M. J., Bhat, P. R., Svensson, J. T., Bartoš, J., Suchánková, P., Šimková, H., Endo, T. R., Fenton, R. D., Lonardi, S., Castillo, A. M., Chao, S., Cistué, L., Cuesta-Marcos, A., Forrest, K. L., Hayden, M. J., Hayes, P. M., Horsley, R. D., Makoto, K., ... Close, T. J. (2011). An improved consensus linkage map of barley based on flow-sorted chromosomes and single nucleotide polymorphism markers. *Plant Genome*, *4*, <https://doi.org/10.3835/plantgenome2011.08.0023>
- Neyhart, J. L., & Smith, K. P. (2019). Validating genomewide predictions of genetic variance in a contemporary breeding program. *Crop Science*, *59*, 1062–1072. <https://doi.org/10.2135/cropsci2018.11.0716>
- Osthushenrich, T., Frisch, M., Zenke-Philippi, C., Jaiser, H., Spiller, M., Cselényi, L., Krumnacker, K., Boxberger, S., Kopahnke, D., Habekuß, A., Ordon, F., & Herzog, E. (2018). Prediction of means and variances of crosses with genome-wide marker effects in barley. *Frontiers in Plant Science*, *9*, 1899, <https://doi.org/10.3389/fpls.2018.01899>
- Pong-Wong, R., Haley, C. S., & Woolliams, J. A. (1999). Behaviour of the additive finite locus model. *Genetics Selection Evolution*, *31*, 193–211.
- Rinne, H. (2010). *Location-scale distribution: Linear estimation and probability plotting using MATLAB*. Justus Leibig University.
- Smith, O. S. (1979). A model for evaluating progress from recurrent selection. *Crop Science*, *19*, 223–226. <https://doi.org/10.2135/cropsci1979.0011183X001900020013x>
- Sprague, G. F. (1984). Organization of breeding programs. In *20th annual Illinois corn breeders school* (pp. 20–31). University of Illinois at Urbana-Champaign.
- Tiede, T., Kumar, L., Mohamadi, M., & Smith, K. P. (2016). Predicting genetic variance in bi-parental breeding populations is more accurate when explicitly modeling the segregation of informative genomewide markers. *Molecular Breeding*, *35*, 199. <https://doi.org/10.1007/s11032-015-0390-6>
- Troyer, A. F. (2006). Adaptedness and heterosis in corn and mule hybrids. *Crop Science*, *46*, 528–543.
- Weyhrich, R. A., Lamkey, K. R., & Hallauer, A. R. (1998). Effective population size and response to  $S_1$ -progeny selection in the BS11 maize population. *Crop Science*, *38*, 1149–1158. <https://doi.org/10.2135/cropsci2005.0065>
- Yonezawa, K., & Yamagata, H. (1978). On the number and size of cross combinations in a breeding programme of self-fertilizing crops. *Euphytica*, *27*, 113–116.
- Zhong, S., & Jannink, J.-L. (2007). Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. *Genetics*, *177*, 567–576.

**How to cite this article:** Bernardo, R. (2022). Outliers and their distribution in breeding populations. *Crop Science*, *65*, 1107–1114. <https://doi.org/10.1002/csc2.20742>