The preliminary yield trial (PYT) is a ubiquitous feature of grain breeding programs. In the absence of molecular markers, the key function of the PYT is to identify superior entries, which will then be evaluated the following year in more extensive yield trials and/or used as parents to begin another breeding cycle. For a given budget, there is a well-recognized tradeoff between the number of entries in the PYT and the number of plots allocated to each entry (Bos, 1983; Gauch and Zobel, 1996). Testing more entries enlarges the pool of selection candidates, but accuracy (the correlation between estimated and true genetic merit) is improved with additional resource allocation per entry, e.g., testing each entry in multiple locations and/or multiple replicates within location. Naturally, the plot-based heritability is a key parameter influencing the optimal strategy; as it decreases, more plots should be allocated to fewer entries to maximize genetic gain (Bos, 1983).

The preliminary yield trial (PYT) is a ubiquitous feature of grain breeding programs. In the absence of molecular markers, the key function of the PYT is to identify superior entries, which will then be evaluated the following year in more extensive yield trials and/or used as parents to begin another breeding cycle. For a given budget, there is a well-recognized tradeoff between the number of entries in the PYT and the number of plots allocated to each entry (Bos, 1983; Gauch and Zobel, 1996). Testing more entries enlarges the pool of selection candidates, but accuracy (the correlation between estimated and true genetic merit) is improved with additional resource allocation per entry, e.g., testing each entry in multiple locations and/or multiple replicates within location. Naturally, the plot-based heritability is a key parameter influencing the optimal strategy; as it decreases, more plots should be allocated to fewer entries to maximize genetic gain (Bos, 1983). The most common measure

Optimal Design of Preliminary Yield Trials with Genome-Wide Markers

Jeffrey B. Endelman,* Gary N. Atlin, Yoseph Beyene, Kassa Semagn, Xuecai Zhang, Mark E. Sorrells, and Jean-Luc Jannink

ABSTRACT

Previous research on genomic selection (GS) has focused on predicting unphenotyped lines. Genomic selection can also improve the accuracy of phenotyped lines at low heritability, e.g., in a preliminary yield trial (PYT). Our first objective was to estimate this effect within a biparental family, using multilocation yield data for barley (Hordeum vulgare L.) and maize (Zea mays L.). We found that accuracy increased with training population size and was higher with an unbalanced design spread across multiple locations than when testing all entries in one location. The latter phenomenon illustrates that when seed is limited, genome-wide markers enable broader sampling from the target population of environments. Our second objective was to explore the optimum allocation of resources at a fixed budget. When PYT selections are advanced for further testing, we propose a new metric for optimizing genetic gain: \( R_{max} \), the expected maximum genotypic value of the selections. For budgets up to 250 yield plot equivalents per family, the optimal design did not involve genotyping more progeny than were phenotyped, even when the cost of creating and genotyping each line was only 0.25 the cost of one yield plot unit (YPU). At a genotyping cost of 0.25 YPU, GS offered up to a 5% increase in genetic gain compared with phenotypic selection. To increase genetic gains further, the training population must be expanded beyond the full-sib family under selection, using close relatives of the parents as a source of prediction accuracy.


Abbreviations: BLUP, best linear unbiased prediction; cdf, cumulative distribution function; EGV, estimated genotypic value; GS, genomic selection; pdf, probability density function; PS, phenotypic selection; PYT, preliminary yield trial; SD, standard deviation; TP, training population; YPU, yield plot unit.
of genetic gain is the expected mean genotypic value of the selected entries, denoted by $R_{\text{mean}}$, which for truncation selection of normally distributed values follows the well-known “breeder’s equation” (Eq. [5] in Methods). When the selected entries are intermated, their expected mean is the mean of the base population for the next stage of selection, and thus $R_{\text{mean}}$ is an appropriate measure of genetic gain. However, when the PYT is only the first of several stages of testing, and ultimately it is the best of the PYT selections that will be potentially commercialized or used as parents, the expected mean for the first stage may not be the most appropriate metric for optimization. One difficulty with $R_{\text{mean}}$ is that, at a fixed population size, it decreases with the number of selected entries (Fig. 1). In the context of multistage selection, this leads to the erroneous conclusion that selecting fewer entries in the PYT increases genetic gain, when in fact, the more entries that are selected in the PYT, the greater the probability of finding a superior entry in advanced testing.

To address this shortcoming, several groups have instead optimized the probability of selecting an entry with genotypic value greater than some threshold (Robson et al., 1967; Johnson, 1989; Knapp, 1998; Longin et al., 2006), but this metric has the drawback of requiring the specification of a threshold. We propose instead to optimize the expected maximum genotypic value of the selected entries, denoted by $R_{\text{max}}$, which is readily interpreted as a measure of genetic gain. Furthermore, as shown in Fig. 1, $R_{\text{max}}$ increases with the number of selected entries. Although there is no simple expression for $R_{\text{max}}$ analogous to the breeder’s equation, exact results can be achieved with numerical integration (see Methods).

The objective of the present research was to explore the question of optimal resource allocation in the PYT when genome-wide markers are available and selection is based on genomic predictions rather than on phenotypic means. Earlier treatments of resource allocation with markers have focused on traditional marker-assisted selection, using an index that combines phenotype with a molecular prediction based on a small number of significant markers (Lande and Thompson, 1990; Knapp, 1998; Xie and Xu, 1998; Moreau et al., 2000). As originally conceived by Lande and Thompson (1990), the molecular score in such an index accounts for only a portion of the total genetic merit, and the marker effects are estimated from populations other than those under selection. By contrast, we will use the phenotypes from the PYT to predict total genetic merit for the PYT entries, using best linear unbiased prediction (BLUP) with a marker-derived relationship matrix (Habier et al., 2007; VanRaden, 2008). Because best linear prediction maximizes the correlation between the predicted and true random effects under normality (Searle et al., 1992), the use of an index combining phenotype and genomic prediction is superfluous in this case; the optimal index weight for the phenotype is zero (for a proof, see Supplemental File 1).

The availability of genome-wide markers introduces additional degrees of freedom for the design of the PYT. Within each biparental family, there can potentially be some progeny that are phenotyped in the PYT and some that are not. The prediction accuracy for the phenotyped progeny will naturally be higher, but this accuracy comes at a cost. Because of the different accuracies for the phenotyped vs. unphenotyped progeny, selection should be applied separately within each subset. Thus, for a given budget per family, the breeder must decide (i) How many progeny should be created (and genotyped)? (ii) What proportion of the progeny should be phenotyped, and how many plots should be allocated per entry? (iii) How many of the selected entries should come from the phenotyped vs. unphenotyped subsets?

To answer these questions, the prediction accuracy as a function of the training population size must be specified. In the present study, we empirically determined this relationship for two different families: (i) the Harrington x TR306 malting barley (Hordeum vulgare L.) population evaluated in diverse North America environments (Tinker et al., 1996), and (ii) a biparental maize (Zea mays L.) population from the CIMMYT East African breeding program. The empirical accuracy functions were then used to determine the post hoc optimal PYT design for each dataset, based on maximizing either $R_{\text{mean}}$ or $R_{\text{max}}$.

**MATERIALS AND METHODS**

**Genotypes and Phenotypes**

Genotype and grain yield data for 145 doubled haploid barley lines in the Harrington x TR306 population were downloaded from http://wheat.pw.usda.gov/ggpages/HxT/. Yield

---

**Figure 1.** The expected mean vs. the expected maximum response to selection (in units of genetic standard deviation [SD]), for population size 100 and heritability 0.1. As the number of selected entries increases, the mean response decreases while the maximum response increases.
was measured in single replicate trials in 25 environments; we retained 18 corresponding to the 9 locations for which 2 yr of data were available (ABb, SKa, SKb, SKc, MB, ONa, ONb, QC, PE). Four progeny (3, 76, 9, 33) were removed because of missing phenotype data. As downloaded, there were 127 markers coded A/B for the two parental haplotypes and a linkage map for the markers. Missing marker calls flanked by markers from the same parent (on a given chromosome) were imputed to agree with the flanking markers, and the remaining missing calls were imputed with the population mean.

The maize population consisted of 184 F2:3 lines from the cross CZL0719 × CZL0723. The lines were genotyped with a set of proprietary single nucleotide polymorphisms (SNPs) designed to cover the maize genome, resulting in 217 polymorphic markers. Three percent of the marker calls were missing and imputed with the population mean for each marker. Lines were crossed to a single tester, and hybrid yield was measured under optimal irrigation using an incomplete block, two-replicate trial in each of three locations in a single year. Using ASReml 3.0 (VSN International, 2009), the incomplete block effect was estimated for each trial assuming independent genotypes, and adjusted yields were used for subsequent analysis.

**Genomic Prediction Accuracy**

Average prediction accuracies were estimated by repeated sampling from the completely balanced datasets to simulate trials with different properties. In each simulated trial, the particular locations and their entry lists were chosen randomly subject to a specified number of locations, number of entries, and number of plots per entry. First, the number of locations was varied from one to three but using only a single plot per entry for training; thus, the entry lists for each location were disjoint. For the three-location scenario, the number of plots per entry was increased from one to three. As each entry was tested at most once in each location, a trial with three plots per entry was balanced (each entry in every location). With two plots per entry, the entries were evenly partitioned among the three possible pairs of locations (1 and 2, 1 and 3, or 2 and 3).

The phenotypes from each simulated trial were used to predict the average genotypic performance across all locations. Genomic predictions for all progeny (both phenotyped and unphenotyped) were made using a mixed model, with location as a fixed effect and genotypic value as a random effect with covariance proportional to the marker-derived relationship matrix, A_m. Variance components were estimated by REML using R package rrBLUP, version 4.1 (Endelman, 2011; R Development Core Team, 2012). Because of the low marker densities, the A_m matrix was estimated using the shrinkage estimation procedure of Endelman and Jannink (2012). For centered genotype matrix W (n lines × m markers), the shrinkage estimator is

\[ \hat{A}_m = \frac{\delta(S_S + \langle W_w \rangle^2)}{\langle W_w \rangle^2} \]

where \( S = m^{-1}WW' - \langle W_w \rangle^2 \) is the sample covariance matrix, \( p_j = 1 - 2q_j \) is the allele frequency at marker j, and the angular brackets denote an average with respect to the index. The shrinkage intensity \( \delta \) ranges from 0 to 1 and was chosen to minimize the expected mean-squared error of the genome-wide covariance matrix (for details, see Endelman and Jannink, 2012). When \( \delta = 0 \), there is no shrinkage, and Eq. [1] reduces to the first formula proposed by VanRaden (2008). The optimal shrinkage intensities were 0.25 and 0.20 for the barley and maize populations, respectively.

For the barley dataset, the simulated trials were conducted using the phenotypes from Year 1, while the estimated genotypic values (EGVs) for validation were based on the data from Year 2. Estimated genotypic values were calculated by BLUP assuming independent genotypes (and including fixed effects for all 9 locations), and the accuracy of the EGVs (equal to the square-root of the repeatability) was determined from the prediction error variance (PEV) as (Clark et al., 2012)

\[ H_i = \sqrt{1 - \frac{\text{PEV}_{i}}{V_i}} \]

where \( V_i \) is the estimated genetic variance in the validation data. The mean \( H \) for the Year 2 EGVs in the barley dataset was 0.85. Genomic selection accuracy was calculated as the correlation coefficient with the EGVs, divided by their mean \( H \) (Dekkers, 2007). Separate accuracies were calculated for the phenotyped and unphenotyped entries in each simulated trial, denoted \( r_1 \) and \( r_2 \), respectively. (We will consistently use the subscript 1 for phenotyped entries and subscript 2 for unphenotyped entries.)

A slightly different procedure was required for the maize population as only 1 yr of data was available. Six yield plots were available for each entry, two per location, and the EGVs were estimated using all yield measurements not included in the genomic prediction model. For example, for predictions based on two plots per entry, four plots were available to estimate the EGV. As the number of plots per phenotyped entry increased, the number of plots available for estimating the EGV decreased and thus so did its accuracy (\( H \)). The average accuracy of the EGV was 0.70, 0.68, 0.66, and 0.63 for entries with 0, 1, 2, and 3 plots, respectively, in the training data. As with the barley case study, entries in the simulated maize trials were phenotyped with at most one plot per location. Thus, designs with multiple plots per entry imply testing each entry in multiple locations.

The size of the training population (TP) was varied from a minimum of 20 entries to a maximum of 140 for the barley family and 180 for the maize family, in increments of 20. For each combination of the three trial parameters (number of locations, number of phenotyped entries, number of plots per entry), 5000 samples were made for barley and 2000 for maize to estimate the mean GS accuracy with standard error (SE) less than 0.01.

Several nonlinear equations were compared for their fit to the empirically determined relationship between GS accuracy and TP size (not shown). Two parsimonious models were chosen for the economic optimization:

\[ r_2^2 = \frac{n_i}{b + n_i} \]

\[ r_1^2 = \frac{(1-c)n_i}{b + n_i} + c \]
where \( n_1 \) is the TP size. Equation [3] is the accuracy for the unphenotyped entries and depends on a single parameter \( h \). A second parameter \( i \) is required to model the intercept for the phenotyped entries in Eq. [4]. The unknown parameters were calculated by nonlinear least squares (nls) in R.

**Response to Selection**

For a single population with normally distributed genotypic values, the mean of the selected entries relative to the standard deviation (SD) of genotypic values was modeled by

\[
R_{\text{mean}} = ir
\]  

where \( r \) is selection accuracy and \( i \) is selection intensity. The selection intensity was calculated as \( i = p^{-1} f(\Phi^{-1}(1-p)) \), where \( f \) and \( \Phi \) are the pdf (probability density function) and cdf (cumulative distribution function) for the normal distribution, respectively, and for the proportion \( p \) we used the finite-population correction proposed by Bulmer (1980):

\[
p = \frac{k + \frac{1}{2}}{n + \frac{k}{2n}}
\]  

when selecting \( k \) entries from a population of size \( n \). As shown in the Appendix, the analog of Eq. [5] when selecting \( k \) entries from the phenotyped sample and \( k \) entries from the unphenotyped sample, with intensities \( i_1 \) and \( i_2 \), respectively, is

\[
R_{\text{mean}} = i_1 \gamma_1 + i_2 (1 - \gamma_1)
\]  

where \( \gamma_1 = k_1/(k_1 + k_2) \) is the proportion of the selected entries chosen from the phenotyped subset.

To formulate the expected maximum response (\( R_{\text{max}} \)), let \( g_{k,n} \) be the maximum genotypic value when selecting \( k < n \) entries from a population of size \( n \). The cdf \( F_{k,n}(y) \) of \( g_{k,n} \), which is the probability that \( g_{k,n} \leq y \), can be written as (Nagaraja and David, 1994)

\[
F_{k,n}(y) = \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} f_U(x) f_U(x) f_{X,n}(x) \right] dU
\]  

where

\[
f_{X,n}(x) = \frac{m! f(x)}{(j-1)! (m-j)!} [\Phi(x)]^{j-1} [1 - \Phi(x)]^{m-j}
\]  

is the pdf of the order statistic \( X_{j:n} \) and

\[
f_U(x) = f(U = u \mid X > x) = \frac{f(u) [1 - \Phi]\left(\frac{x-u}{\sqrt{\text{var}(x)}}\right)}{1 - \Phi(x)}
\]  

is the conditional probability of observing \( U = u \) given \( X > x \) for a bivariate \((X, U)\) standard normal distribution with correlation \( r \) (Rice, 2007). The expected maximum of \( g_{k,n} \) when selecting from one population is thus

\[
R_{\text{max}} = \int_{-\infty}^{\infty} dy \left[ \frac{dF_{k,n}(y)}{dy} \right] f_{X,n}(y)
\]  

where \( F'_{k,n} \) (the derivative of \( F_{k,n} \)) is the pdf for \( g_{k,n} \):

\[
F'_{k,n}(y) = \int_{-\infty}^{\infty} dx f_U(x) f_{X,n}(x)
\]  

When selecting from both phenotyped and unphenotyped sets, the cdf of the maximum genotypic value across all selected entries is the product of the cdfs for the two groups:

\[
P\{g_{k,n} \leq y\} = P\{g_{k,n} \leq y, g_{k,n} \leq y\} P\{g_{k,n} \leq y\} P\{g_{k,n} \leq y\}
\]  

The \( R_{\text{max}} \) for the two-population scenario (e.g., phenotyped and unphenotyped) is thus

\[
R_{\text{max}} = \int_{-\infty}^{\infty} dy \left[ \frac{dF_{k,n}(y)}{dy} \right] F_{k,n}(y) f_{X,n}(y) + F_{k,n}(y) F'_{k,n}(y)
\]  

where we have explicitly denoted the accuracies \( r_1 \) and \( r_2 \). Software to compute \( R_{\text{max}} \) is available as part of the R package GSdesign, which can be downloaded from http://potatobreeding.cals.wisc.edu/software. The \( R_{\text{max}} \) calculations are implemented in C and require the GNU Scientific Library (www.gnu.org/software/gsl).

**Optimal PYT Design**

The criterion for design was to maximize the response (either \( R_{\text{mean}} \) or \( R_{\text{max}} \)) for a fixed number of selected entries \( k \) and fixed budget \( B \), expressed in yield plot units (1 YPU = the cost of phenotyping one yield plot). If \( C \) denotes the cost of genotyping plus line creation (in YPU), then the cost of a trial with \( T \) plots per phenotyped entry is \( n_1 T + (n_1 + n_2) C \). When \( k, C, \) and \( B \) are fixed, there are two degrees of freedom for designing the trial, which we can take to be \( n_1 \) and \( k_2 \). The number of phenotyped entries selected is then \( k_1 = k - k_2 \), and equating the cost formula with the budget fixes the TP size at \( n_1 \) = floor\((B-n_1 C)/(C+T)\). The shape of the domain in the \((n_1, k_2)\) plane for the optimization problem is trapezoidal, bounded by the lines \( k_2 = 0, n_2 = n_2, k_2 = k-1, \) and \( n_1 = \text{floor}(B-n_1 C)/(C+T)\). Since \( n_1 \) is the minimum TP size and should be determined based on the range of validity for the model fits to \( r_1 \) and \( r_2 \) (Eq. [3] and [4]). For our case studies, \( n_1 \) = 20.

Finding the maximum response over this domain is a non-linear integer programming problem, which was solved by partial enumeration. The algorithm started at \( n_2 = 0, k_2 = 0 \), for which the response is given by the formulas for a single population. Complete enumeration for several situations revealed
that the maximum response was not necessarily unimodal with respect to \( n_2 \), but it was unimodal with respect to \( k_2 \) (Fig. S1). Thus, a grid of \( n_2 \) values was created (increments of 5), and for each value of \( n_2 \), the maximum response was found by hill-climbing, starting with \( k_2 = 1 \). The GSdesign software has been written to facilitate parallel computation of the maximum response across the \( n_2 \) grid, using R package multicore.

In the comparison between genomic and phenotypic selection (PS), both strategies were independently optimized at the same budget. For PS the cost of the trial was \((C+T)n\), where \( T \) is the number of plots per entry and \( C \) is the cost of line creation. At a fixed budget \( B \), the family size was thus \( n = \text{floor}[B/(C+T)] \).

Phenotypic selection accuracy was determined empirically by analyzing the same simulated trials used to determine GS accuracy. As with the GS strategy, predictions were made by BLUP (using fixed effects for location), but the covariance matrix for the random genotypic effects was an identity instead of the marker-derived relationship matrix. Phenotypic selection accuracy varied with \( T \) but was constant with respect to family size \( n \) (see Fig. 2), so no regression models were needed for the economic optimization.

**Two-stage Selection**

For comparison with the optimal \( R_{\text{max}} \) designs, a two-stage selection process was simulated in which all candidate progeny were genotyped and phenotyped. Genotypic values for the Stage 1 entries were drawn independently from a standard normal distribution. Selection values for Stage 1 were generated by adding a random normal deviate with zero mean and variance chosen such that the selection accuracy equaled that from the maize case (with markers) at that population size. Denoting the maize accuracy. As with the GS strategy, predictions were made by BLUP (using fixed effects for location), but the covariance matrix for the random genotypic effects was an identity instead of the marker-derived relationship matrix. Phenotypic selection accuracy varied with \( T \) but was constant with respect to family size \( n \) (see Fig. 2), so no regression models were needed for the economic optimization.

<table>
<thead>
<tr>
<th>RESULTS</th>
<th>Prediction Accuracy</th>
</tr>
</thead>
</table>

To optimize the design of the PYT, a functional relationship between accuracy and training population (TP) size is needed, both for the entries with phenotypes (P and G; \( r_1 \)) and those without phenotypes (G; \( r_2 \)). Figure 2 shows this relationship for the case where each entry has been phenotyped with one yield plot. The blue symbols show the accuracy for the unphenotyped progeny, which tends toward zero as the TP size goes to zero. The black symbols show the accuracy for the TP without using the markers, which is not affected by TP size because the genotypic values of the entries are estimated independently. The red symbols show the accuracy for the TP when analyzed with the realized relationship matrix, and in both populations the markers led to a substantial increase in accuracy over the phenotypes. As with predicting unphenotyped progeny, the accuracy for the...
phenotyped progeny increased with TP size, approaching the phenotypic accuracy in the limit of zero TP size. The three symbols in Fig. 2 correspond to whether 1, 2, or 3 locations were used for training, with each entry appearing in only one location. For both the maize and barley data sets, spreading the training population across multiple locations increased accuracy only when markers were available.

Figure 3 shows how accuracy responded to greater allocation of resources per entry in the maize population, keeping the number of locations for the trial fixed at three. Because a single plot was used in each location, adding more plots per entry means testing each entry in more locations. The top panel shows that for a given training population size, testing entries in more locations improved accuracy for both the phenotyped (P and G) and unphenotyped (G only) progeny. For the unphenotyped progeny (bottom right), the fully replicated design (3 plots) had the same or slightly less accuracy than the unbalanced “sparse testing” design (each entry in only one location). These same trends were observed in the barley population (Fig. S2).

The solid lines in Fig. 3 are nonlinear regression models to the empirically determined accuracies. Because the accuracy for unphenotyped entries has zero intercept, while the intercept for phenotyped entries is nonzero, different functional forms are needed for these two cases. Good results were obtained for the unphenotyped entries with the one-parameter model $r^2 = n_i/(b+n_i)$, where the parameter $b$ was fit by least squares. Although no biological interpretation is needed to perform the economic optimization, theory suggests the $b$ parameter is approximately the ratio between the number of independent chromosome segments ($M_e$) and the heritability ($h^2$) (Daetwyler et al., 2010). Table 1 shows the calculated $M_e$ based on the least-squares fit for $b$ and observed $h^2$ for the 1, 2, and 3-plot scenarios. In theory, $M_e$ should not depend on the resource allocation strategy, and this expectation was met fairly well in both populations (Table 1). The mean calculated value of $M_e$ was 90 for the barley data and 120 for maize. For the accuracy of the phenotyped progeny, the above formula was

Table 1. Nonlinear regression estimates for the parameter $b$ and calculated values of $M_e$.

<table>
<thead>
<tr>
<th>Population</th>
<th>No. plots</th>
<th>$b^*$</th>
<th>$h^2$</th>
<th>$M_e = bh^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barley</td>
<td>1</td>
<td>637</td>
<td>0.14</td>
<td>87</td>
</tr>
<tr>
<td>Barley</td>
<td>2</td>
<td>375</td>
<td>0.24</td>
<td>90</td>
</tr>
<tr>
<td>Barley</td>
<td>3</td>
<td>286</td>
<td>0.32</td>
<td>93</td>
</tr>
<tr>
<td>Maize</td>
<td>1</td>
<td>755</td>
<td>0.14</td>
<td>103</td>
</tr>
<tr>
<td>Maize</td>
<td>2</td>
<td>447</td>
<td>0.27</td>
<td>121</td>
</tr>
<tr>
<td>Maize</td>
<td>3</td>
<td>327</td>
<td>0.40</td>
<td>130</td>
</tr>
</tbody>
</table>

*From the model $r^2 = n_i/(b+n_i)$, where $n_i$ is the training population size and $r^2$ is the prediction accuracy for unphenotyped lines.*
modified to allow for a nonzero intercept (Eq. [4] in Methods), and this two-parameter model (estimates in Table S1) provided a good fit to the maize and barley data (solid lines in upper left panel of Fig. 3 and Fig. S2).

**Optimal PYT Design**

Once the user has specified (i) the relationship between prediction accuracy and TP size, (ii) the budget, (iii) the cost of creating and genotyping a line (L+G) relative to one yield plot (YPU), and (iv) the number of entries that will be selected, the GSdesign software determines how many total progeny to create, how many to phenotype, and how many to select from each pool to maximize the response (either $R_{\text{max}}$ or $R_{\text{mean}}$). Figure 4 shows the results of this optimization for the maize case study, assuming 10 entries will be selected and for an L+G cost of 0.5 YPU. Since the relationship between prediction accuracy and TP size varies depending on the number of plots per entry, the optimization is performed separately for 1, 2, and 3 plots per entry, shown in black, blue, and red, respectively, in Fig. 4.

The optimal allocation is the one with the highest response. The upper left panel in Fig. 4 shows that, for $R_{\text{max}}$, the optimal allocation was one plot per entry regardless of budget in the range 100–250 YPU per family. For $R_{\text{mean}}$, there was a crossover in the optimal strategy between 125 and 150 YPU; below this budget the optimum was 2 plots per entry, and above it the optimum was 3 plots per entry. Thus when 10 entries were selected, the optimal strategy was different depending on whether it was the mean or the maximum of the selected entries that was most relevant.

Figure 4. Determining the optimum allocation of resources (maize case study). Using the empirical accuracy models for 1, 2, and 3 plots per phenotyped entry, the optimum response was determined as a function of the budget (YPU = yield plots units), assuming 10 selected entries and an L+G cost of 0.5 YPU. Top: the maximum response; middle: the number of phenotyped (P and G) and unphenotyped (G) progeny; bottom: the number of selected entries. The optimal strategy for $R_{\text{max}}$ was 1 plot per phenotyped entry, while for $R_{\text{mean}}$, it was 2 to 3 plots per entry, depending on the budget.
The center panel in Fig. 4 shows the number of phenotyped (P and G) and unphenotyped (G) progeny corresponding to the responses in the top panel. For both types of response, the number of phenotyped progeny increased with budget and decreased with the number of plots per phenotyped entry. Recalling that for \( R_{\text{max}} \) the optimal allocation was one plot per entry, the center left panel shows that this corresponds to a strategy in which no lines were genotyped without phenotyping (dashed black line). With 2 or 3 plots per phenotyped entry, it was beneficial for \( R_{\text{max}} \) to genotype substantially more lines than were phenotyped, but not for \( R_{\text{mean}} \). The bottom panel shows the number of entries selected from the phenotyped and unphenotyped sets, which add up to 10.

The results in Fig. 4 were for the case of selecting 10 entries and for an L+G cost of 0.5 YPU. Figure 5 shows how the optimal solution for barley (see Fig. S3 for maize) changed when varying these parameters, assuming a budget of 200 YPU. Intuitively, one would expect that as the number of selected entries decreases, the optimal strategy would shift to predicting fewer entries with higher accuracy. This expectation was confirmed by the results; as the number of selected entries decreased, the optimal allocation shifted toward more plots per phenotyped entry (top panel) but fewer entries (center panel). This trend was observed for both types of response, but for \( R_{\text{max}} \) the shift to higher resource allocation per entry occurred at a lower

Figure 5. The effect of the line plus genotyping (L+G) cost and the number of selected entries on the optimal preliminary yield trial (PYT) design at a budget of 200 yield plot units (YPU) per family for barley (see Fig. S3 for maize). Left: \( R_{\text{max}} \); right: \( R_{\text{mean}} \); top: the number of plots per phenotyped entry; middle: the number of phenotyped progeny; bottom: the number of unphenotyped progeny. As the L+G cost decreased and the number selected increased, the optimal strategy shifted to more entries and fewer plots per entry. Even for L+G costs as low as 0.25 YPU, genetic gain was not improved by genotyping more progeny than were phenotyped (except when selecting at least eight entries for \( R_{\text{max}} \)).
number of selected entries than $R_{\text{mean}}$. For most of the explored parameter space, the optimal strategy did not involve genotyping more progeny than were phenotyped. Such a strategy was only favorable at an L+G cost of 0.25 and a high number of selected entries, and even then only for $R_{\text{max}}$ (bottom panel in Fig. 5 and S3).

Finally, we compared the expected genetic gain in the PYT based on genome-wide prediction (GS) against the expected gain for selection without markers based on phenotype means (PS). Figure 6 shows the ratio between the GS and PS gains when the two strategies are independently optimized at a budget of 200 YPU. As expected, the relative performance of GS improved as genotyping costs fell. Perhaps less intuitive is the result that the GS/PS ratio increased as fewer entries were selected. When selecting 5 entries, the breakeven point for incorporating genome-wide markers ($R_{\text{GS}}/R_{\text{PS}} = 1$) was at a genotyping cost between 0.5 and 0.75 YPU for $R_{\text{max}}$ and slightly higher (0.75–1.00 YPU) for $R_{\text{mean}}$. These results correspond to a budget of 200 YPU, and the relative performance of GS to PS increases with larger budgets (Fig. S5). Interestingly, Fig. 6 and S4 also show that markers were more helpful in improving $R_{\text{mean}}$ than $R_{\text{max}}$.

Two-Stage Selection
Because PYT selections enter Stage 2 trials, the number selected in the PYT depends on the number of Stage 2 entries, which in turn depends on the Stage 2 budget and resource allocation per entry. The design strategy we have described presumes a hierarchical approach to the two-stage design problem—first determine the Stage 2 design, then optimize the PYT based on $R_{\text{max}}$—but this is only an approximation. Because accuracy in the second stage is not perfect, the best entry may not be selected, and thus the genetic gain after two stages will be less than $R_{\text{max}}$ for the first stage (the PYT). As a consequence, the optimum allocation with respect to $R_{\text{max}}$ in Stage 1 may not be optimal with respect to the genetic gain after two stages.

To explore this issue we optimized a two-stage selection process using the parameters from the maize case study (see Methods). The resource allocation options for Stage 1 were 1, 2, or 3 plots per entry, while there was no limit to the number of plots per entry in Stage 2 (based on the earlier results, it was assumed that all selection candidates were phenotyped). The results are shown in Table 2, where the first two columns are the input parameters for the optimization, namely the L+G cost and the budget per family (for two stages). Columns 3 to 6 correspond to the resource allocation strategy with the highest genetic gain after two stages. Looking at Column 3, one sees the obvious trend that less of the total budget was spent on creating and genotyping lines as the L+G cost fell. For example, at an L+G cost of 1.5 YPU, 25% of the total budget was spent on creating and genotyping lines, compared with 13 to 15% of the budget at an L+G cost of 0.25 YPU. The ratio between the Stage 1 and Stage 2 phenotyping budgets was not related to the L+G cost and was nominally 2 but ranged from 1 to 3. The number of entries (per family) advanced from Stage 1 to Stage 2 was consistently 4 for L+G costs of 1.0 and 1.5 YPU but ranged from 5 to 8 at lower costs. Recapitulating the results from Fig. 5, at lower L+G costs the optimal Stage 1 design involved more entries and fewer plots per entry.

The next to last column in Table 2 is the $R_{\text{max}}$ for Stage 1 based on the number of Stage 1 entries and plots per entry in Column 4. As expected, the realized gains in Stage 2 (Column 6) were less than the Stage 1 $R_{\text{max}}$ values because selection accuracy was not perfect in Stage 2. Even so, the last column in Table 2 shows that the optimal number of plots per entry based on the Stage 1 $R_{\text{max}}$ (analogous to Fig. 5) was the same as the true optimum for Stage 1 when optimizing the two-stage gain. For example, at an L+G cost of 0.25 YPU and total budget of 300 YPU, the
optimal strategy for two-stage gain was to spend 0.75 × 300 = 225 YPU on developing, genotyping, and phenotyping the Stage 1 entries, use one plot per entry, and advance six entries to Stage 2. At a budget of 225 YPU and when selecting six entries, the one plot per entry strategy was also optimal with respect to $R_{\text{max}}$ while for $R_{\text{mean}}$ the optimum was three plots per entry. These results confirm that $R_{\text{max}}$ is a useful criterion for PYT design even though subsequent stages have imperfect accuracy.

**DISCUSSION**

Most of the previous research on genomic selection in plant breeding has focused on the determinants of prediction accuracy, such as the model, TP size, TP composition, and marker density (for reviews see Jannink et al., 2010; Lorenzana and Bernardo, 2011). Furthermore, the reported accuracies have been almost exclusively for unphenotyped lines. The present work extends our basic understanding of GS accuracy within a biparental population in several ways. One, we have reported empirical accuracies for grain yield for phenotyped lines. As with unphenotyped lines, the GS accuracy for phenotyped lines increases with TP size, and it equals the PS accuracy as the TP size goes to zero. Two, we have explored the effect of resource allocation on prediction accuracy. Allocating additional plots per phenotyped entry increases accuracy for both phenotyped and unphenotyped lines at a fixed TP size, but at a fixed number of plots only the phenotyped lines have higher accuracy. For unphenotyped lines, prediction accuracy slightly decreases when additional resources are devoted to existing lines instead of phenotyping new lines. Third, we have shown that even when limited to one plot per entry, spreading the PYT across multiple locations (each entry in only one location) is beneficial for predicting average genotypic performance across a range of environments. Genome–wide markers provide the connectivity in such unbalanced designs and allow multi–environment trials to be conducted earlier in the breeding cycle, which is advantageous for traits with high $G \times E$ (like yield).

Despite having very different contexts, the relationship between accuracy and training population size was remarkably similar between the maize and barley case studies. Further research is needed to understand the variability of this relationship across families within a breeding program and to assess its impact on PYT design. We expect that breeding programs will need to empirically determine the effect of TP size and resource allocation on accuracy by using balanced trials with multiple locations, years, and families. Such research is needed to build realistic regression models for the economic optimization, but after this initial investment, subsequent families can be evaluated more efficiently with optimized designs.

The design strategy we have presented is based on maximizing the expected genetic gain in the PYT. The traditional measure of genetic gain is the mean genotypic value of the selected entries, but this metric has drawbacks when lines are selected for further testing rather than for intermating. As an alternative to the mean, we have proposed using the expected maximum of the selected entries and have provided software for calculating this statistic. Because the expected mean and maximum converge as the number of selected entries decreases (and are identical when selecting one entry; see Fig. 1), the optimal design strategies for the two metrics are most similar when selecting few entries.

In Fig. 6 (barley) and S4 (maize), there was not much advantage to using GS for making selections in the PYT; even with genotyping costs at 0.25 YPU, there was at most a 5% gain over PS at a budget of 200 YPU per family. At higher budgets, the relative performance of GS to PS was higher (see Fig. S5). As the marker densities in this study were low (127 for barley, 217 for maize), higher marker densities may have improved the relative performance of GS, although within a biparental population several hundred markers can be adequate to maximize accuracy (Lorenzana and Bernardo, 2009; Zhao et al., 2013). Another caveat is that our analysis has been concerned solely with yield, while in many crops quality traits are also used to make PYT selections. As more traits are predicted from genome–wide markers, the cost of genotyping relative to phenotyping becomes more favorable. The comparative methodology we have developed can be used to explore this scenario, using empirical, within–family accuracies for a multitrait selection index.

### Table 2. Two-stage vs. one-stage optimization for the maize family.

<table>
<thead>
<tr>
<th>L+G cost</th>
<th>Total budget</th>
<th>Budget allocation L+G, Stage 1, Stage 2</th>
<th>No. entries, No. plots per entry, accuracy</th>
<th>Stage 2 realized maximum</th>
<th>Stage 1 $R_{\text{max}}$</th>
<th>Stage 1 $R_{\text{mean}}$</th>
<th>No. plots from one-stage optimization $R_{\text{max}}, R_{\text{mean}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>200</td>
<td>13, 52, 35</td>
<td>104, 1, 0.48</td>
<td>7.10, 0.78</td>
<td>1.82</td>
<td>2.13</td>
<td>1, 3</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>15, 60, 25</td>
<td>180, 1, 0.54</td>
<td>6.12, 0.81</td>
<td>2.03</td>
<td>2.19</td>
<td>1, 3</td>
</tr>
<tr>
<td>0.5</td>
<td>200</td>
<td>13, 52, 35</td>
<td>52, 2, 0.57</td>
<td>5.14, 0.83</td>
<td>1.77</td>
<td>1.98</td>
<td>2, 3</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>20, 40, 40</td>
<td>120, 1, 0.49</td>
<td>8.14, 0.83</td>
<td>1.96</td>
<td>2.21</td>
<td>1, 3</td>
</tr>
<tr>
<td>1.0</td>
<td>200</td>
<td>18, 52, 30</td>
<td>35, 3, 0.66</td>
<td>4.14, 0.83</td>
<td>1.72</td>
<td>1.91</td>
<td>3, 3</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>19, 56, 25</td>
<td>56, 3, 0.68</td>
<td>4.18, 0.86</td>
<td>1.90</td>
<td>2.07</td>
<td>3, 3</td>
</tr>
<tr>
<td>1.5</td>
<td>200</td>
<td>25, 50, 25</td>
<td>33, 3, 0.66</td>
<td>4.12, 0.81</td>
<td>1.68</td>
<td>1.89</td>
<td>3, 3</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>25, 50, 25</td>
<td>50, 3, 0.67</td>
<td>4.18, 0.86</td>
<td>1.87</td>
<td>2.02</td>
<td>3, 3</td>
</tr>
</tbody>
</table>
Regarding yield, our results indicate that GS offers little advantage when selections are made within the same full-sib family used for training. This is true regardless of whether the PYT entries are being selected as parents or for further testing. To increase genetic gain more than a few percentage points, the training population must be expanded beyond the full-sib family under selection. The magnitude of the within-family accuracy when training on close relatives is thus an important topic for research. From the few reports that exist, it appears that when the parents of the selection candidates have not been previously used as parents in the training population, the accuracy can be very low (e.g., < 0.1) (Massman et al., 2013; Riedelsheimer et al., 2013). In such cases, one should look to close relatives of the parents, such as full or half-sibs, as the source of prediction accuracy (Bernardo and Yu, 2007).

This analysis reinforces the idea that in the context of GS breeding, the PYT serves a dual purpose: the first is to identify superior entries for further testing, the second is to provide a source of accuracy for predicting the progeny of PYT entries. In the present study, we considered the possibility of genotyping more entries than are phenotyped, but this practice makes no contribution to accuracy. Since the optimal $R_{\text{max}}$ strategy at L+G costs above 0.25 YPU was to phenotype all progeny (Fig. 5 and S3), at those costs, there would not be a conflict between $R_{\text{max}}$ and the accuracy of progeny prediction. At lower L+G costs, a compromise between the two objectives may be needed.

Given the complexity of designing a GS-based breeding program, decision support software is needed to help breeders with this task (e.g., Gordillo and Geiger, 2008). The design framework and computational tools we have presented are a step in this direction.

#### APPENDIX

An expression is derived for the mean genotypic value of the selected entries from several populations, assuming selection is practiced within each population separately. The expected mean within population $j$ is given by the breeder’s equation:

$$R_j = \bar{x}_{s,j} - \bar{x}_j = i_j r_j \sigma_j$$

where $x$ denotes the genotypic value, $\bar{x}_j$ is the mean of population $j$, $\bar{x}_{s,j}$ is the mean of the selected entries in population $j$, $i_j$ is selection intensity, $r$ is selection accuracy, and $\sigma$ is the standard deviation of genotypic values. The mean of the selected entries across all populations, denoted $\bar{x}_s$, is given by the weighted average:

$$\bar{x}_s = \sum_j \gamma_j \bar{x}_{s,j} \quad \gamma_j = \frac{k_j}{\sum_i k_i}$$

where the weight $\gamma_j$ is the proportion of the total number of selections coming from population $j$ ($k_j$ denotes the number of selections from population $j$). Similarly, the overall mean is a weighted average of the population means, with weights given by the proportion of the total number of entries coming from each population ($n_j = \text{size of population } j$):

$$\bar{x}_j = \sum_j \beta_j \bar{x}_j \quad \beta_j = \frac{n_j}{\sum_j n_j}$$

Therefore, the overall response is

$$R = \bar{x}_i - \bar{x} = \sum_j [\gamma_j \bar{x}_{s,j} - \beta_j \bar{x}_j] = \sum_j [\gamma_j i_j r_j \sigma_j + (\gamma_j - \beta_j) \bar{x}_j]$$

For the specific case of two populations selected at random from the same biparental family, one phenotyped and one unphenotyped, the two populations have the same mean and same standard deviation of genotypic values: ($\bar{x}_j = \bar{x}, \sigma_j = \sigma$). In this case, Eq. [18] simplifies to Eq. [7] (with $R_{\text{mean}} = R/\sigma$).

#### Supplemental Data Available

Supplemental information is available at http://www.crops.org/publications/cs.

#### Acknowledgments

Financial support for this research was provided by the Bill and Melinda Gates Foundation and the USDA Agricultural Research Service.

#### References


Bos, I. 1983. The optimum number of replications when testing lines or families on a fixed number of plots. Euphytica 32:311–318. doi:10.1007/BF00021439


VSN International. 2009. ASReml 3. VSN Intl., Hemel Hempstead, UK.
