Theor Appl Genet (2013) 126:2835–2848
DOI 10.1007/s00122-013-2175-9

ORIGINAL PAPER

Optimizing the allocation of resources for genomic selection in one breeding cycle

Christian Riedelsheimer · Albrecht E. Melchinger

Received: 15 May 2013 / Accepted: 29 July 2013 / Published online: 27 August 2013

Abstract

Key message **We developed a universally applicable planning tool for optimizing the allocation of resources for one cycle of genomic selection in a biparental population. The framework combines selection theory with constraint numerical optimization and considers genotype × environment interactions.**

Abstract Genomic selection (GS) is increasingly implemented in plant breeding programs to increase selection gain but little is known how to optimally allocate the resources under a given budget. We investigated this problem with model calculations by combining quantitative genetic selection theory with constraint numerical optimization. We assumed one selection cycle where both the training and prediction sets comprised double haploid (DH) lines from the same biparental population. Grain yield for testcrosses of maize DH lines was used as a model trait but all parameters can be adjusted in a freely available software implementation. An extension of the expected selection accuracy given by Daetwyler et al. (2008) was developed to correctly balance between the number of environments for phenotyping the training set and its population size in the presence of genotype × environment interactions. Under small budget, genotyping costs mainly determine whether GS is superior over phenotypic selection. With increasing budget, flexibility in resource allocation increases greatly but selection gain leveled off quickly requiring balancing the number of populations with the budget spent for each population. The use of an index combining phenotypic and GS predicted values in the training set was especially beneficial under limited resources and large genotype × environment interactions. Once a sufficiently high selection accuracy is achieved in the prediction set, further selection gain can be achieved most efficiently by massively expanding its size. Thus, with increasing budget, reducing the costs for producing a DH line becomes increasingly crucial for successfully exploiting the benefits of GS.

Introduction

Plant breeding programs are currently transformed by the implementation of genomic selection (GS) approaches (e.g. Bernardo and Yu 2007; Heffner et al. 2009; Lorenz et al. 2011; Xu et al. 2012). After initially proposing GS in the field of animal breeding (Meuwissen et al. 2001), many statistical techniques suitable for GS have been intensively studied (de los Campos et al. 2013; Riedelsheimer et al. 2012b; Heslot et al. 2012). GS has rapidly evolved into its own research field and developed standards to render results comparable across studies (Daetwyler et al. 2013). For GS in maize, different application scenarios have been proposed. These range from predicting the breeding value of lines from diverse collections (Riedelsheimer et al. 2012a) to biparental populations (Lorenzana and Bernardo 2009; Riedelsheimer et al. 2013) and single-cross hybrids from their parental inbred lines (Technow et al. 2012). Despite this remarkable progress, little efforts have been made to explicitly optimize the allocation of resources to

---

Communicated by M. Frisch.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-013-2175-9) contains supplementary material, which is available to authorized users.

C. Riedelsheimer · A. E. Melchinger (✉)
Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany
e-mail: melchinger@uni-hohenheim.de

maximize the economic return of investment in terms of selection gain per unit time and money.

This is surprising as in the early beginning of genomics-assisted breeding, large efforts have been made to analyze the economic efficiency of marker-assisted selection (MAS), the conceptual predecessor of GS. Lande and Thompson (1990) showed how selection efficiency can be increased by using an index incorporating both phenotypic and marker information. Later, specific aspects of the relative efficiency of MAS were studied. Moreau et al. (1998) investigated the overestimation of marker effects in populations of finite size and later the relative costs for genotyping versus phenotyping (Moreau et al. 2000). Further, the long-term efficiency of MAS over several cycles of selection was investigated in both plant breeding (Moreau et al. 2004) and animal breeding scenarios (Villanueva et al. 2004; Dekkers and Hospital 2002).

Meanwhile, MAS was found to be less efficient than initially advocated, especially for highly polygenic traits. The main reasons for this include the small proportion of the genetic variance explained by the detected quantitative trait loci (QTL) and the low transferability across populations (Melchinger et al. 1998; Bernardo 2008; Xu and Crouch 2008), especially, if no major QTL are present. Although GS is often seen as conceptually different from MAS, it is similar to MAS to the extent that marker effects also have to be estimated with sufficient precision. In addition, the transferability across populations is not always guaranteed but heavily depends on the design of the training set (Riedelsheimer et al. 2013). Hence, modeling prediction accuracy as a function of the trait and population parameters is crucial for optimizing the allocation of resources under GS.

Under various simplifying assumptions (e.g., independent, biallelic and additive loci), Daetwyler et al. (2008) derived a functional relationship between prediction accuracy under GS and the training set size, heritability of the trait and the effective number of loci. Meanwhile, several empirical studies suggest that for a given set of environments, the proposed formula is fairly accurate for at least one selection cycle within biparental populations, although some variation among traits was observed (Combs and Bernardo 2013; Riedelsheimer et al. 2013); Thus, it could be used as a tool for guiding the allocation of resources in GS, at least within the space of inference for which the formula was derived.

Recently, Lorenz (2013) used the proposed formula in a simulation study to analyze the effect of different resource allocations on MAS and GS for biparental populations. By varying the number of genotypes and replications, the author found a greater flexibility for treating the trade-off between these two parameters in GS when compared with MAS. However, many aspects regarding the implementation of GS to gain maximum benefit from GS are still unclear. For example, the proposed formula of Daetwyler et al. (2008) does not incorporate the effect of genotype × environment interactions in a plant breeding context, which is important to reasonably balance between number of environments and total population size. In addition, when the variance components as well as the costs for phenotyping and genotyping are assumed to be known, several questions arise: (1) How should the budget be optimally allocated between the training set and the prediction set and to what extent is this influenced by the available budget? (2) How does the optimum allocation of resources depend on (a) the relative costs of genotyping versus phenotyping, (b) the ratio of the genotypic variances to the masking variances, and (c) the available budget? (3) Under which circumstances is phenotypic selection superior to GS?

Our objective was to derive a theoretical framework including an appropriate treatment of genotype × environment interactions under which these questions can be directly answered. This was done for one selection cycle where both training and prediction sets contain DH lines from the same biparental population. We combined classical selection theory for quantitative traits with constraint numerical optimization. As a point of reference, we focused on grain yield at the testcross level in maize with variance components drawn from the literature. In addition, we provide a user-friendly software for free use to encourage practitioners to analyze and optimize their resource allocation given their specific cost and trait parameters.

Theory

GS as a special case of the two subsample scenario

Selection accuracy (r_A) is generally defined as the correlation between the estimated breeding value of candidates and their true breeding value. If only phenotypic values are available, r_A corresponds to the square root of the heritability of the trait, $r_A = h$ (Falconer and Mackay 1996, p. 189). GS with selection performed in both the training set and prediction set can be seen as a special case of selecting candidates across two subsets of a population with different selection accuracies.

Suppose a base population of candidates with breeding values independently, identically distributed, $N(0,1)$, from which two sets of candidates, π_1 and π_2 with sample size $N_1 = \psi N$ and $N_2 = (1-\psi)N$ are chosen at random, where $\psi \in (0,1)$. The estimated breeding values of candidates from π_1 and π_2 are determined by different tests with selection accuracies r_{A_1} and r_{A_2}, respectively. We assumed that $N_{\mathrm{s}} = \alpha N$ candidates (for $\alpha \in (0, 1)$) are selected from $\pi_1 \cup \pi_2$ based on their estimated breeding values, with $N_{\mathrm{s}_1} = \theta N_{\mathrm{s}}$ candidates being selected from π_1 and $N_{\mathrm{s}_2} = (1-\theta) N_{\mathrm{s}}$ candidates being selected from π_2, where $\theta \in \left[\max\left(0, \frac{\alpha+\psi-1}{\alpha}\right), \min\left(1, \frac{\psi}{\alpha}\right)\right]$. Thus, the proportion of candidates selected in π_1 and π_2 is $\alpha_1 = \frac{N_{\mathrm{s}_1}}{N_1} = \frac{\theta}{\psi}\alpha$ and $\alpha_2 = \frac{N_{\mathrm{s}_2}}{N_2} = \frac{(1-\theta)}{(1-\psi)}\alpha$, respectively. Then, the expected genetic gain ΔG under this selection scheme is obtained as

$$\begin{aligned}\Delta G &= \theta\,\Delta G_1 + (1-\theta)\,\Delta G_2 \\ &= \theta\, i(\alpha_1)\, r_{\mathrm{A}_1} + (1-\theta)\, i(\alpha_2)\, r_{\mathrm{A}_2},\end{aligned} \tag{1}$$

where $i(.)$ is the corresponding selection intensity, i.e., the standardized selection differential (Falconer and Mackay 1996, p. 189). In GS, π_1 corresponds to the training set π_{T} of size N_{T} with selection accuracy $r_{\mathrm{A_T}}$and $\pi_2$corresponds to the prediction set π_{P} of size N_{P} with selection accuracy$r_{\mathrm{A_P}}$ and our goal is to allocate the resources between π_{T} and π_{P} such that the selection gain across both sets is maximized.

Selection accuracy under GS in the prediction set

We assumed that the training set π_{T} is evaluated in a subset of E environments with one replication because this is expected to yield highest selection gain based on the results of phenotypic selection (Utz 1969) and simulations of GS (Lorenz 2013). In the prediction set π_{P}, selection aims to select the candidates with highest genotypic values on the basis of their genomic predicted values. Thus, $r_{\mathrm{A_P}}$ corresponds to $r(g, \hat{g}_{\mathrm{E}})$, the correlation between the true genotypic value (g) and its genomic predicted value for the set of E environments ($\hat{g}_{\mathrm{E}}$) at which π_{T} has been phenotyped and g_{E} denotes the expected genotypic value across the set of E environments. Thus, $\mathrm{cov}(g_{\mathrm{E}}, \hat{g}_{\mathrm{E}}) = \sigma^2_{g_{\mathrm{E}}} = \sigma^2_{\mathrm{g}} + \frac{\sigma^2_{\mathrm{ge}}}{E}$, where σ^2_{g} denotes the genotypic variance and σ^2_{ge} the genotype × environment variance component across the entire set of target environments. Then (Melchinger, unpublished results),

$$r(g, \hat{g}_{\mathrm{E}}) = r(g, g_{\mathrm{E}}) \times r(g_{\mathrm{E}}, \hat{g}_{\mathrm{E}}) \text{ with } r(g, g_{\mathrm{E}}) = \sqrt{\frac{\sigma^2_{\mathrm{g}}}{\sigma^2_{g_{\mathrm{E}}}}}. \tag{2}$$

For estimating $r(g_{\mathrm{E}}, \hat{g}_{\mathrm{E}})$, we used results from Daetwyler et al. (2008) yielding

$$r(g_{\mathrm{E}}, \hat{g}_{\mathrm{E}}) = \sqrt{\frac{N_{\mathrm{T}} h^2_{\mathrm{E}}}{N_{\mathrm{T}} h^2_{\mathrm{E}} + M_{\mathrm{e}}}} \tag{3}$$

where M_{e} is the number of independent loci and $h^2_{\mathrm{E}} = \frac{\sigma^2_{g_{\mathrm{E}}}}{\sigma^2_{g_{\mathrm{E}}} + \sigma^2_{\varepsilon}/E} = \frac{\sigma^2_{\mathrm{g}} + \sigma^2_{\mathrm{ge}}/E}{\sigma^2_{\mathrm{g}} + \sigma^2_{\mathrm{ge}}/E + \sigma^2_{\varepsilon}/E}$ refers to the heritability for phenotyping the training set in the E environments. Combining Eqs. (2) and (3) yields

$$r_{\mathrm{A_P}} = r(g, \hat{g}_{\mathrm{E}}) = \sqrt{\frac{\sigma^2_{\mathrm{g}}}{\sigma^2_{\mathrm{g}} + \sigma^2_{\mathrm{ge}}/E}}\sqrt{\frac{N_{\mathrm{T}} h^2_{\mathrm{E}}}{N_{\mathrm{T}} h^2_{\mathrm{E}} + M_{\mathrm{e}}}}. \tag{4}$$

Selection accuracy for index selection in the training set

In the training set π_{T}, we practice selection based on an optimal index combining the phenotypic value (y) and the genomic predicted genomic value ($\hat{g}_{\mathrm{E}}$) across the set of E environments. Applying results of Lande and Thompson (1990) initially developed for combined MAS, the optimum index is obtained as $I = b_{\mathrm{P}} y + b_{\mathrm{M}} \hat{g}_{\mathrm{E}}$, where the relative weights b_{P} and b_{M} are found to be

$$\frac{b_{\mathrm{M}}}{b_{\mathrm{P}}} = \frac{(1/h^2) - 1}{1 - r^2(g, \hat{g}_{\mathrm{E}})} \text{ with } h^2 = \frac{\sigma^2_{\mathrm{g}}}{\sigma^2_{\mathrm{g}} + \sigma^2_{\mathrm{ge}}/E + \sigma^2_{\varepsilon}/E}. \tag{5}$$

Selection accuracy in π_{T} is then obtained as the correlation of the index I with the true genotypic value (Lande and Thompson, 1990):

$$r_{\mathrm{A_T}} = r(g, I) = \sqrt{r^2(g, \hat{g}_{\mathrm{E}}) + \frac{h^2\left(1 - r^2(g, \hat{g}_{\mathrm{E}})\right)^2}{1 - h^2 r^2(g, \hat{g}_{\mathrm{E}})}}. \tag{6}$$

Selection gain in the training set and in the prediction set

Selection gain in π_{T} and π_{P} was calculated as $\Delta G_{\mathrm{T}} = i(\alpha_{\mathrm{T}})\, r_{\mathrm{A_T}}$ and $\Delta G_{\mathrm{P}} = i(\alpha_{\mathrm{P}}) r_{\mathrm{A_P}}$, respectively, where $i(.)$ is the selection intensity and α_T and α_P are the proportions of candidates selected in each set. The total selection gain in $\pi_{\mathrm{T}} \cup \pi_{\mathrm{P}}$ was then calculated as the weighted sum

$$\Delta G_{\mathrm{GS}} = \theta\, \Delta G_{\mathrm{T}} + (1-\theta)\, \Delta G_{\mathrm{P}}. \tag{7}$$

Note that we consider a single breeding cycle for a single DH population split into a training and prediction set. Thus, selection gain refers to the improvement of the genotypic values of the DH lines in this cycle only.

Assumptions of the parameters in the model calculations

Effective number of loci

Hitherto, at least three formulas for inferring M_{e} have been proposed. All three depend on the effective population size (N_{e}) and the length of the genome (L) or the average length of a chromosome (l):

1. $M_{\mathrm{e}} = 2 N_{\mathrm{e}} L$ (Meuwissen and Goddard 2010)
2. $M_{\mathrm{e}} = 2 N_{\mathrm{e}} L / \log(4 N_{\mathrm{e}} L)$ (Goddard 2009)
3. $M_{\mathrm{e}} = 2 N_{\mathrm{e}} L / \log(N_{\mathrm{e}} l)$ (Goddard et al. 2011).

Substituting $L = 17.96$ M taken from the Genetic 2008 Composite Map of Maize (ww.maizegdb.org) and $N_\mathrm{e} = 1$ for DH lines from a biparental population derived from a F_1 plant yields $M_\mathrm{e} = 35.9$, 8.4, and 61.3 for formulas 1–3, respectively. In this study, we assumed $M_\mathrm{e} = 30$, which lies in the middle of the proposed range. This value is also close to $M_\mathrm{e} = 28$ chosen by Lorenz (2013) for DH lines from biparental maize populations using another line of reasoning based on the number of chromosomes and the expected number of cross-over events.

Budget and costs of phenotyping and genotyping

We assumed a budget B for one population and breeding cycle expressed in plot equivalents (PE). From this budget, a fraction b is spent on the training set π_T and the remainder on the prediction set π_P. Thus, the budget allocated to the training set is $B_\mathrm{T} = bB$ and the budget for the prediction set is $B_\mathrm{P} = (1 - b)\,B$.

Costs for production of one candidate (C_C) in the training set and prediction set, such as production of one doubled haploid line or and recombinant inbred line is expressed in PE. Costs per phenotyping unit (C_Ph), corresponding to one PE, include the costs for producing the seeds of the test candidates, such as the proportional cost of producing testcross seed or seed increase for evaluation of per se performance, as well as the costs of the plot (land, sowing, harvesting, etc.) and all special assays (e.g., resistance tests, quality tests in the laboratory) required for measuring the target trait. Cost for genotyping one candidate (C_G), expressed in units of PE, include costs for DNA extraction, the genomic assay, and all necessary bioinformatic analyzes of the genomic data. For given values of the budgets allocated to the training set (B_T) and the prediction set (B_P), the costs for producing a candidate (C_C) and C_G, we find the size of the training set to be

$$N_\mathrm{T} = \frac{B_\mathrm{T}}{C_\mathrm{C} + C_\mathrm{G} + E} \tag{8}$$

and for the size of prediction set, we find

$$N_\mathrm{P} = \frac{B_\mathrm{P}}{C_\mathrm{C} + C_\mathrm{G}}, \tag{9}$$

where E refers to the number of environments. Putting these results together, we find for the selected fraction in the training set $\alpha_\mathrm{T} = \theta \frac{N_\mathrm{s}}{N_\mathrm{T}}$ and for the prediction set $\alpha_\mathrm{P} = (1 - \theta) \frac{N_\mathrm{s}}{N_\mathrm{P}}$, where θ is the fraction of selected candidates originating from π_T.

Example trait

For a numerical example, we chose maize grain yield at the testcross level as a model trait for detailed investigation with values of the variance components taken from Lorenzana and Bernardo (2008): $\sigma^2_\mathrm{g} = 0.65$, $\sigma^2_\mathrm{ge} = 0.15$, and $\sigma^2_\varepsilon = 1.11$. The ratios of variance components were therefore $\kappa_1 = \frac{\sigma^2_\mathrm{ge}}{\sigma^2_\mathrm{g}} = 0.231$ and $\kappa_2 = \frac{\sigma^2_\varepsilon}{\sigma^2_\mathrm{g}} = 1.708$. C_C was set to 1 if not explicitly stated otherwise. C_G was set to a rather low value of 0.3 to account for the decline in genotyping costs, especially, when used in a streamlined commercial breeding program. A summary of all parameter assumptions is given in Table 1.

Optimization of the total selection gain under GS and PhS

Constraint optimization approach

Our goal was to find the optimum values $\mathbf{V}^* = (b^*, E^*, \theta^*)$ that maximize the total genetic gain ΔG_GS for given values of $\mathbf{U}_\mathrm{GS} = (\kappa_1, \kappa_2, M_\mathrm{e}, B, C_\mathrm{C}, C_\mathrm{G}, N_\mathrm{s})$:

$$\Delta G_\mathrm{GS}\left(\mathbf{V}^* | \mathbf{U}_\mathrm{GS}\right) = \max\{\Delta G_\mathrm{GS}(\mathbf{V}|\mathbf{U}_\mathrm{GS})\}.$$

In addition, we determined the optimum allocation of resources for maximizing the genetic gain under phenotypic selection by finding the optimum value E° that maximizes

$$\Delta G_\mathrm{PhS} = i(\alpha) \times h \text{ with } \alpha = \frac{N_\mathrm{s}}{N} \text{ and } N = \frac{B}{C_\mathrm{C} + E}$$

for given values of $\mathbf{U}_\mathrm{PhS} = (\kappa_1, \kappa_2, B, C_\mathrm{C}, N_\mathrm{s})$:

$$\Delta G_\mathrm{PhS}(E^\circ | \mathbf{U}_\mathrm{PhS}) = \max\{\Delta G_\mathrm{PhS}(E|\mathbf{U}_\mathrm{PhS})\}. \tag{10}$$

Thus, finding $\mathbf{V}^*$ and E° reduces to numerically solving nonlinear constraint optimization problems. Here, we used the interior point algorithm implemented in the Find Maximum function in Mathematica 9.0 (Wolfram Research Inc., 2013).

Relative efficiency of GS

For comparison of our GS scheme with pure phenotypic selection, we calculated the relative efficiency of GS for given values of $\mathbf{U}_\mathrm{GS}$ and $\mathbf{U}_\mathrm{PhS}$ by the ratio:

$$\mathrm{RE} = \frac{\Delta G_\mathrm{GS}\left(\mathbf{V}^* | \mathbf{U}_\mathrm{GS}\right)}{\Delta G_\mathrm{PhS}\left(E^\circ | \mathbf{U}_\mathrm{PhS}\right)}. \tag{11}$$

Results

Optimum resource allocation for GS under variable budget

Selection gain under GS (ΔG_GS) showed a substantial increase from 0.5 to 2.0 when increasing the budget B from 100 to 1,500 plot equivalents (PE), but further increase in B resulted in diminishing additional selection

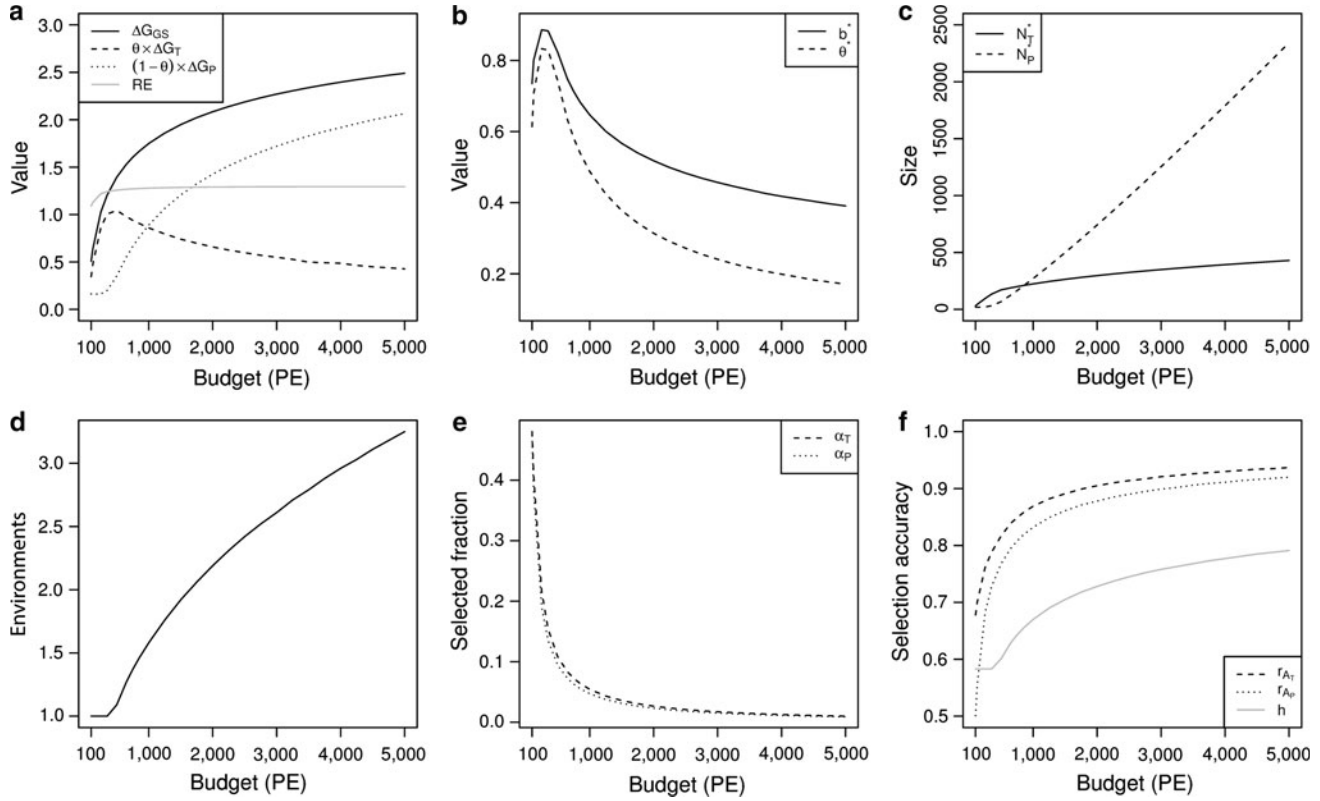


Fig. 1 Influence of the available budget in plot equivalents (PE) on (**a**) the relative efficiency (RE) of GS and the weighted selection gains in the training and prediction sets (which sum up to ΔG_{GS}), (**b**) optimum values θ^* and b^*, (**c**) the optimum size of training set (N_T^*) and the prediction set (N_P^*), (**d**) the optimum number of environments (E^*), (**e**) the selected fractions in the training set (α_T) and the prediction set (α_P), and (**f**) the selection accuracies in the training set (r_{A_T}) and prediction set (r_{A_P}). The square root of the heritability (h) in the training set is shown for comparison. Assumptions were $\kappa_1 = 0.231$, $\kappa_2 = 1.708$, $C_C = 1$, $C_G = 0.3$, $N_s = 25$, $M_e = 30$

gain (Fig. 1a). Relative efficiency (RE) increased to 1.25 at $B = 500$ PE and flattened off to a maximum of 1.29 at $B = 5{,}000$ PE. The relative contribution of the prediction set $((1-\theta) \times \Delta G_P)$ to the total selection gain increased convexly with increasing B. In contrast, the relative contribution of the training set $(\theta \times \Delta G_T)$ to the total selection gain increased until a maximum at $\sim$1.0 was reached at $B \sim 500$ PE and declined afterwards. At this point, the fraction of the budget allocated to the training set (parameter b) reached the maximum of 0.82 (Fig. 1b) and declined afterwards in a concave curve. The trajectory of parameter θ followed the one of b closely, but at a reduced level. Whereas the optimum size of the training set (N_T^*) increased convexly with a decreasing gradient as a function of B, the optimum size of the prediction set (N_P^*) increased in a concave curve until a budget of $B \sim 1{,}000$ PE and linearly afterwards (Fig. 1c). The optimum number of environments (E^*) was 1 for $B \leq 500$ PE and increased in a concave curve afterwards and exceeded three only for $B > 4{,}000$ PE (Fig. 1d). The fractions α_T and α_P selected from the training and prediction sets declined rapidly with increasing budget where α_T remained marginally higher than α_P (Fig. 1e). Selection accuracy in the training set (r_{A_T}) was higher than in the prediction set over the whole budget range (Fig. 1f). Both selection accuracies increased rapidly to >0.8 with a budget of $\sim$1,000 PE but thereafter, the gradient of the curves rapidly decreased.

The response curve of ΔG_{GS} as a function of b and E was commonly fairly flat in the vicinity of the optimum parameter combination $\mathbf{V}^*$ (Fig. 2). Irrespective of the budget, the values of b could deviate up to 0.2 from b^* without much reduction in ΔG_{GS}. The level, at which values for E could deviate from E^* without substantially compromising ΔG_{GS}, depended on the budget and was larger for large values of B.

Influence of costs, N_s, and M_e on resource allocation and relative efficiency of GS

Costs for producing a candidate (C_C) had a strong influence on ΔG_{GS} and RE of GS as well as the ratio of the optimum training set size (N_T^*) to the optimum prediction set size (N_P^*) (Table 2). With a budget of 500 PE (5,000 PE), reducing C_C from 1 to 0.01 increased ΔG_{GS} by 32 %

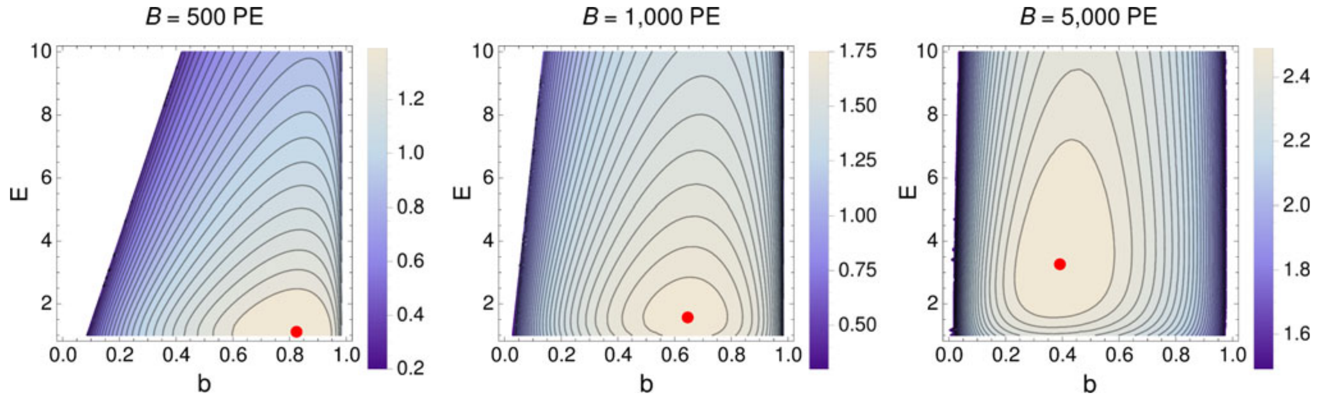


Fig. 2 Contour plots of ΔG_{GS} as a function of the number of test environments E and the proportion b of the budget allocated to the training set for different values of the available budget (B). The optimum parameter combination V^* is indicated as a *red dot*. Assumptions were $\kappa_1 = 0.231$, $\kappa_2 = 1.708$, $C_C = 1$, $C_G = 0.3$, $N_s = 25$, $M_e = 30$ (color figure online)

Table 1 Summary of assumptions on and treatments of parameters for GS of grain yield at the testcross level

Parameter	Assumption/treatment
π_T Training set	DH lines originating from the same biparental cross as π_P
π_P Prediction set	DH lines originating from the same biparental cross as π_T
κ_1 Variance ratio σ^2_{ge}/σ^2_g	0.231
κ_2 Variance ratio $\sigma^2_{\varepsilon}/\sigma^2_g$	1.708
M_e Effective number of loci	30
B Budget	Range of 100–5,000 PE
C_C Costs for producing a candidate	1 PE
C_G Costs for genotyping	0.3 PE
C_{Ph} Costs related to phenotyping	1 PE
θ Fraction of selected candidates originating from π_T	Optimized
b Fraction of B allocated to π_T	Optimized
E No. of environments in which $\pi_T$is phenotyped	Optimized
N_s No. of selected genotypes	25
N_T No. of genotypes in π_T	Inferred from C_C, C_G, b, B, and L
N_P No. of genotypes in π_P	Inferred from C_C, C_G, b, and B
α_T Selected fraction in π_T	$\theta \frac{N_s}{N_T}$
α_P Selected fraction in π_P	$(1-\theta)\frac{N_s}{N_P}$
r_{A_T} Selection accuracy in π_T	Correlation between optimum index combining phenotypic and genomic predicted genotypic values across E environments, and the true genotypic value
r_{A_P} Selection accuracy in π_P	Formula proposed by Daetwyler et al. (2008) extended to account for genotype × environment interactions
Number of selection cycles	One
Number of replications per environment in phenotyping	One

(16 %) and RE by 16 % (13 %) by expanding the prediction set size ~10 (4.5) fold and increasing the training set size only marginally. With $B = 500$ PE, N_P^* reached zero at $C_C = 2$. Further increase in C_C led to a further decrease of N_T^* while increasing E^* slightly. Increasing costs for genotyping (C_G) lead to a nearly linear decline of RE (Fig. 3). The gradient of the decline strongly depended on the available budget. The smaller the budget, the stronger the influence of C_G on RE was. Whereas under $B = 500$ PE, equal efficiency of GS and phenotypic selection

Table 2 Selection gain (ΔG_{GS}) and relative efficiency (RE) as well as optimum resource allocation of GS for yield at the testcross level for different costs of producing a candidate (C_C) under two budgets (B) in plot equivalents (PE)

C_C	$B = 500$ PE					B = 5,000 PE				
	ΔG_{GS}	RE	N_T^*	N_P^*	E^*	ΔG_{GS}	RE	N_T^*	N_P^*	E^*
0.01	1.83	1.46	175	701	1.3	2.90	1.46	456	10,533	3.5
0.05	1.80	1.44	175	610	1.3	2.87	1.44	454	9,324	3.5
0.1	1.76	1.42	174	521	1.3	2.83	1.43	452	8,148	3.5
0.5	1.55	1.32	171	199	1.2	2.63	1.35	439	3,979	3.3
1	1.39	1.26	172	68	1.1	2.49	1.29	430	2,343	3.3
1.5	1.29	1.22	174	5	1.0	2.39	1.26	424	1,606	3.2
2	1.20	1.20	147	0	1.1	2.31	1.23	419	1,186	3.1
2.5	1.13	1.18	126	0	1.2	2.25	1.22	416	914	3.1
3	1.11	1.17	110	0	1.3	1.20	1.20	413	724	3.0

Assumptions were $\kappa_1 = 0.231$, $\kappa_2 = 1.708$, $M_e = 30$, $C_G = 0.3$, $N_s = 25$

N_T^* Optimum number of genotypes in the training set

N_P^* Optimum number of genotypes in the prediction set

E^* Optimum number of environments for phenotyping

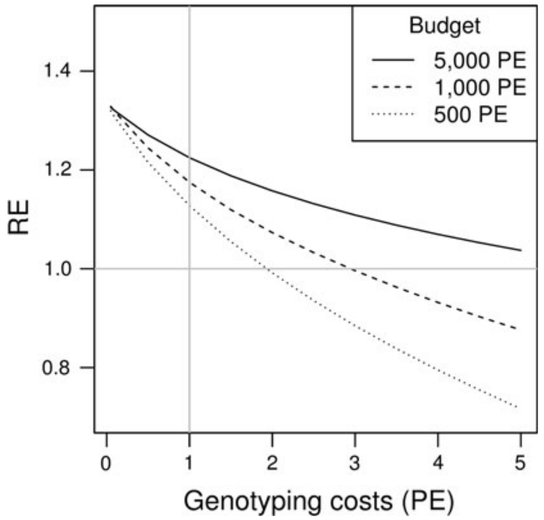


Fig. 3 Relative efficiency (RE) of GS compared to phenotypic selection (PhS) depending on genotyping costs and the available budget. *Vertical and horizontal gray lines* show the point at which genotyping costs equal phenotyping one plot and at which selection gain of GS equals that of PhS. Assumptions were $\kappa_1 = 0.231$, $\kappa_2 = 1.708$, $C_C = 1$, $N_s = 25$, $M_e = 30$

(RE = 1) was reached at $C_G \sim 2$ PE, this value was reached at $C_G \sim 3$ PE with $B = 1{,}000$ PE.

Relative efficiency declined in a concave curve with the number of independent loci (M_e, Fig. 4). The extent of this decline was largely influenced by the available budget. From $M_e = 3$ to 150, RE declined from 1.51 to 1.07 under a budget of 500 PE and from 1.38 to 1.18 under a budget of 5,000 PE.

Increasing the number of selected genotypes (N_s) led to an increasing RE by reducing the size of the training set and the number of environments while enlarging the prediction set (Table 3). Under a small budget, N_s had a stronger influence on the RE and the size of N_T^* and N_P^* than under a large budget. The optimum number of environments E^* was only marginally reduced by increasing values of N_s.

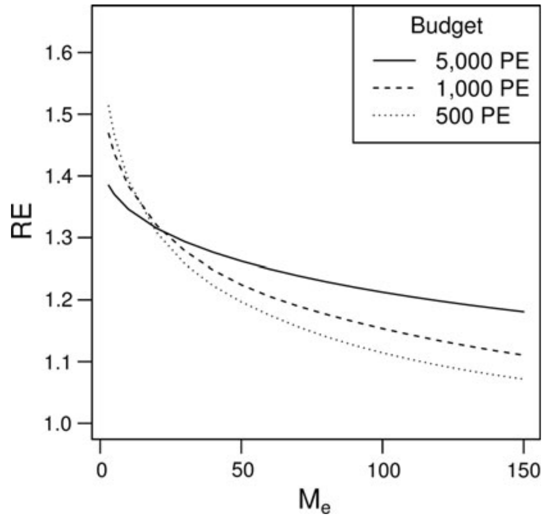


Fig. 4 Relative efficiency (RE) of GS compared to phenotypic selection (PhS) depending on the assumed number of independent loci (M_e) and budget. Assumptions were $\kappa_1 = 0.231$, $\kappa_2 = 1.708$, $C_C = 1$, $C_G = 0.3$, $N_s = 25$

Resource allocation for traits with different relative size of masking variances

Traits other than grain yield were modeled by changing one of two variance component ratios while keeping the other

Table 3 Selection gain (ΔG_{GS}) and relative efficiency (RE) as well as optimum resource allocation of GS for grain yield at the testcross level for different numbers of selected genotypes (N_s) under two budgets (B) in plot equivalents (PE)

N_s	$B = 500$ PE					$B = 5{,}000$ PE				
	ΔG_{GS}	RE	N_T^*	N_P^*	L^*	ΔG_{GS}	RE	N_T^*	N_P^*	L^*
3	2.12	1.13	184	0	1.4	3.09	1.20	474	2,067	3.6
5	1.96	1.16	193	0	1.3	2.96	1.22	465	2,128	3.5
10	1.73	1.19	189	23	1.2	2.76	1.25	451	2,216	3.4
15	1.58	1.22	182	42	1.2	2.65	1.27	442	2,270	3.3
20	1.48	1.24	176	56	1.1	2.56	1.28	435	2,310	3.3
25	1.39	1.25	172	68	1.1	2.49	1.29	430	2,343	3.3
30	1.32	1.27	168	78	1.1	2.43	1.30	425	2,370	3.2
35	1.26	1.29	164	87	1.1	2.39	1.31	421	2,393	3.2
40	1.21	1.30	161	95	1.0	2.34	1.32	418	2,413	3.2
45	1.16	1.31	158	102	1.0	2.30	1.33	415	2,432	3.1
50	1.11	1.33	156	109	1.0	2.27	1.34	412	2,449	3.1

Assumptions were $\kappa_1 = 0.231$, $\kappa_2 = 1.708$, $M_e = 30$, $C_C = 1$, $C_G = 0.3$

N_T^* Optimum number of genotypes in the training set

N_P^* Optimum number of genotypes in the prediction set

E^* Optimum number of environments

one at the value assumed for grain yield (Fig. 5). Increasing the ratio of the genotype × environment to the genotypic variance component (κ_1) led to a stronger increase in the number of environments E^* under GS than E° under PhS (Fig. 5a). Increasing κ_1 from 0.02 to 0.77 led to a linear decrease of the population size under PhS from 898 to 742 while the size of the training set under GS declined rapidly from 1,093 to 262 (Fig. 5b). With this change in resource allocation, RE stayed nearly constant at ~1.3 and the difference between selection accuracy in the training set (r_{A_T}) and in the prediction set (r_{A_P}) became increasingly larger with increasing κ_1 (Fig. 5c). For PhS, increasing the ratio of the error variance to the genotypic variance component (κ_2) led to a strong increase in the number of environments (Fig. 5d) and a rapid decline in the population size (Fig. 5e). In contrast, the number of environments increased only slightly under GS while the training set was slightly increased with increasing κ_2. RE increased strongly with increasing κ_2 and difference between r_{A_T} and r_{A_P} became smaller (Fig. 5f).

Discussion

In this study, we provide the theory for a constraint-based approach to analytically optimize the allocation of resources in genomic selection (GS) for one breeding cycle with DH lines of one population under consideration of genotype × environment interactions. Besides providing practitioners of GS with an easy-to-use software tool to make better use of their money, our objective was also to give answers to several important but still open questions related to resource allocation and selection strategy in GS.

Optimum resource allocation under variable budget

An important issue in the implementation of GS concentrates on how to optimally split the total budget between expenditure for the training set on the one hand and the prediction set on the other hand. The changes in the optimal resource allocation with increasing budget (Fig. 1) clearly show that once a certain amount of money is spent for producing, genotyping, and especially phenotyping a sufficiently large training set with high precision, further selection gain can only be achieved by increasing the selection intensity in the prediction set through massively expanding its size. With our model trait, this change occurred with a budget of $B \sim 500$ PE, for which the optimum fraction of B allocated to the training set (b^*) reached its maximum (Fig 1b). From this point on, the training set warrants a sufficiently high prediction accuracy so that proportionally more money can be spent on the prediction set. As the optimum fraction of the budget allocated to the training set (b^*) decreases with $B > 500$, so does θ^*. Thus, as the proportion of money invested in the prediction set further increases above this threshold, so does the fraction of the total number of candidates to be selected from the prediction set. Since with an increasing budget, the bulk of money is best invested by expanding the prediction set, cutting down on the costs for producing a candidate is central, once the training set warrants high enough selection accuracy in the prediction set (r_{A_P}).

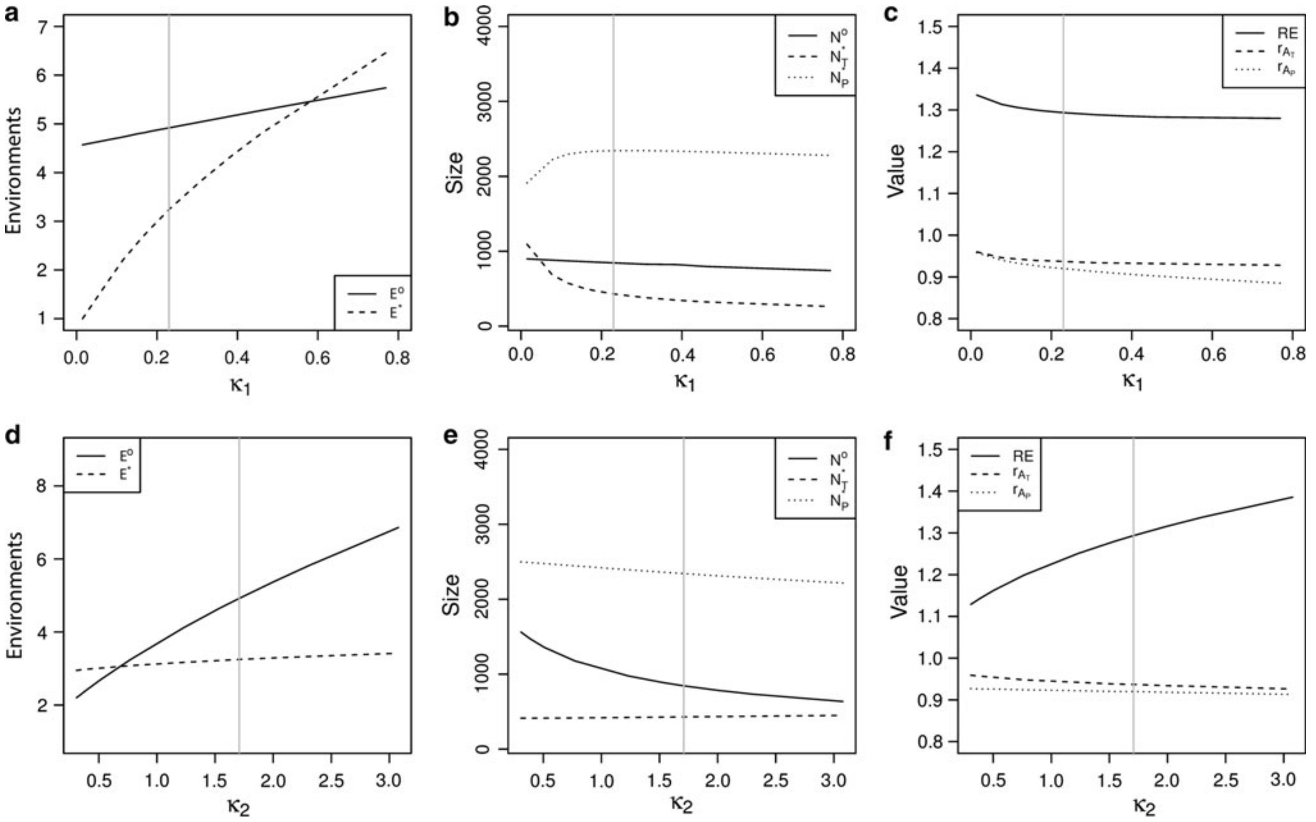


Fig. 5 Influence of the variance ratios $\kappa_1 = \frac{\sigma^2_{ge}}{\sigma^2_g}$ and $\kappa_2 = \frac{\sigma^2_e}{\sigma^2_g}$ on **a**, **d** the optimum number of environments (E°,E^*) under phenotypic selection (PhS) and genomic selection (GS); **b**, **e** the optimum size of the population under PhS (N°), the optimum size of the training set (N^*_T), and the optimum size of the prediction set (N^*_P), and **c**, **f** the relative efficiency (RE) of GS and the selection accuracy in the training set (r_{A_T}) and in the prediction set (r_{A_P}). The influence of κ_1 (κ_2) was analyzed while keeping the corresponding other variance ratio κ_2 (κ_1) constant at $\kappa_2 = 1.708$ and $\kappa_1 = 0.231$, respectively. Variance ratios for grain yield at the testcross level are shown as *grey vertical lines*. Assumptions were $B = 5{,}000$ PE, $C_G = 0.3$, $M_e = 30$, $N_s = 25$

Table 2 shows that moderate gains in *RE* can be achieved by reducing costs for producing a candidate (C_C). Although difficult to achieve in the short-term, this could be targeted in maize by developing inducer lines with a higher in vivo haploid induction rate (Prigge et al. 2012). Moreover, more efficient high-throughput screening techniques for identifying haploid kernels (Melchinger et al. 2013) as well as increasing the rate of fertile diploid plants in the D_0 generation can help to reduce the value for C_C.

However, ΔG_{GS} as a function of B follows a curve with rapidly diminishing marginal returns in selection gain (Fig. 1a). At a budget of $B = 3{,}000$ PE, ΔG_{GS} already reached more than 90 % of ΔG_{GS} achieved with the largest budget of $B = 5{,}000$ PE. Thus, it is important not to invest too much money in one population but to balance the number of populations with the budget spent for each population.

From our theoretical results in Eq. (3) and (6) it follows that r_{A_T} is always at least as high asr_{A_T}, independent of the genotyping and phenotyping costs. Hence, combining phenotypic and GS predicted values should always pay off when compared to pure GS. As illustrated by Fig. 1f, the difference between both selection accuracies is larger under a small budget and approaches almost zero under a budget $B = 5{,}000$ PE. This is in contrast to Lorenz (2013), who concluded from his simulation results that in the training set, selection accuracies were lower for the index than for GS predicted values alone. One difference could be that Lorenz (2013) considers only replications in a single environment and ignored genotype × environment interactions. Moreover, Endelman et al. (2013) proved under certain assumptions that the optimal index weight for the phenotypic score is zero, if the GS model is of the best linear predictor class. In this case, selection accuracy in the training set is the same as in the prediction set leading to slightly lower total selection gains ΔG_{GS}. It remains to be shown, however, if this holds true in general and for other frequently used GS models, such as BayesB or Bayesian LASSO. We therefore decided to adhere to the selection index framework. In the case of low budget and high costs for producing a candidate, the optimum size of the prediction set can reach zero, while still obtaining $\mathrm{RE} > 0$ owing to the use of the index in the training set (Table 2). Since using an index in the genotyped and phenotyped training set comes at no extra costs, our results suggest that phenotypic data of the training set should always be incorporated into an index. Breeding programs with limited budgets will profit most from this approach. In addition to the use of an index, the selection accuracy in the training set might also be higher than in the prediction set due to the generation of spurious LD between unlinked markers generated by the sampling effect when drawing genotypes from populations of finite size (Lorenz 2013).

Flexibility of resource allocation

In the discussion of his results, Lorenz (2013) emphasized the great amount of flexibility in terms of the training population size and the number of replications for achieving a high selection gain by GS. Our contour plots of ΔG_{GS} (Fig. 2) confirmed this conclusion by illustrating that the response surface of ΔG_{GS} as a function of b and E is rather flat in the vicinity of the optimum parameter combination $\mathbf{V}^*$. However, our results also show that the degree of flexibility strongly depends on the total budget B. The smaller B, the more restricted is the choice of b and E that yields values of ΔG_{GS} close to the maximum. Hence, the optimum allocation of resources is most critical for smaller budgets B.

Costs and parameters influencing the superiority of GS over PhS

Whether at all and under which circumstances a new approach should be implemented by breeders, depends largely on its efficiency in terms of the selection gain per unit time and money in comparison with the established breeding methods. Despite the current euphoria on GS, we have to emphasize that this does also apply to GS, because under certain situations ΔG_{GS} can be inferior to the selection gain ΔG_{PhS} under phenotypic selection. In the case of GS, this is illustrated in Fig. 3 which shows that the budget and the costs for genotyping predominately determine whether or not GS is superior to phenotypic selection. The declining costs for genotyping platforms like genotyping-by-sequencing (Elshire et al. 2011; Poland and Rife 2012) suggest that marker costs will not be a major issue in the future, at least not for major staple crops such as maize. As an alternative, genotype imputation techniques might be used if an appropriated reference set is available. In addition, biparental populations of DH lines with large parental linkage blocks in the selection candidates require only a limited number of SNPs to reach maximum selection accuracy in GS. Thus, cheap genotyping chips with a limited number of SNPs will be sufficient for this task. Nevertheless, especially under a small budget, detailed calculations are recommended to ensure that the implementation of GS yields additional selection gain on top of phenotypic selection.

Another crucial assumption is that the proposed formula of Daetwyler et al. (2008) holds true, upon which our modification for coping with genotype × environment interaction in plant breeding relies. Recently, Riedelsheimer et al. (2013) showed that if we move away from a single biparental population to a network of interconnected biparental populations, the design of the training set is important for the selection accuracy obtained for progeny of individual crosses. Further, the assumption on M_{e} is crucial, especially under a small budget (Fig. 4). Thus, methods need to be developed for comparing the theoretically proposed formulas for M_{e} with the actual number of independent loci. Provided that large amount of empirical data is available, it might be better to replace the formula of Daetwyler et al. (2008) with a function that estimates the relationship between $r(g, \hat{g}_{\mathrm{E}})$ and N_{T}, h_{E}, and M_{e} by fitting real data of the breeding material of interest. In this case, our theoretical framework would still be fully valid and applicable.

GS for traits other than grain yield

Like for every indirect selection method, our results also show that GS becomes increasingly superior when compared to PhS for traits which are difficult to phenotype because of a high $\frac{\sigma_{\varepsilon}^2}{\sigma_{\mathrm{g}}^2}$ ratio (Fig. 5f). Figure 5a also shows that with the level of genotype × environment interactions commonly encountered for most traits, the optimum number of environments under GS are below the number which one would optimally choose under PhS. This illustrates the paradigm shift when moving from phenotypic selection to GS (Lorenz 2013): While evaluating, the breeding value of each candidate is the basis for selection in the first approach, estimating and summing effects of alleles of all QTL captured by genome-wide markers are the foundation of the second approach. Theoretical and empirical studies in the field of QTL mapping already showed that given a minimum number of environments, the power to detect QTL is maximized by increasing population size at the expense of replications (Knapp and Bridges 1990; Schön et al. 2004). However, to assess the effect of alleles across multiple environments, it is necessary to evaluate the training set in multiple environments too. With a small to medium budget (Fig. 2) and moderate to low masking variances as reflected by κ_1 and κ_2 (Fig. 5), the optimum number of environments E^* can fall to two or even to one. Obviously, this entails a high risk of failure due to unpredictable environmental conditions. In this case, one could (i) exploit the amount of the flexibility in the choice of E and b (see Fig. 2) using a strategy which yields nearly equal selection gain, or (ii) work with an augmented phenotypic randomization layout spread out over several environments, where each genotype is tested in only two environments to enable testing a large number of genotypes but also capturing the allelic effects in a representative set of environments. Moreover, incorporating the phenotypic data of the training set together with the GS predicted values into an index becomes increasingly beneficial with increasing genotype × environment interactions (Fig. 5c).

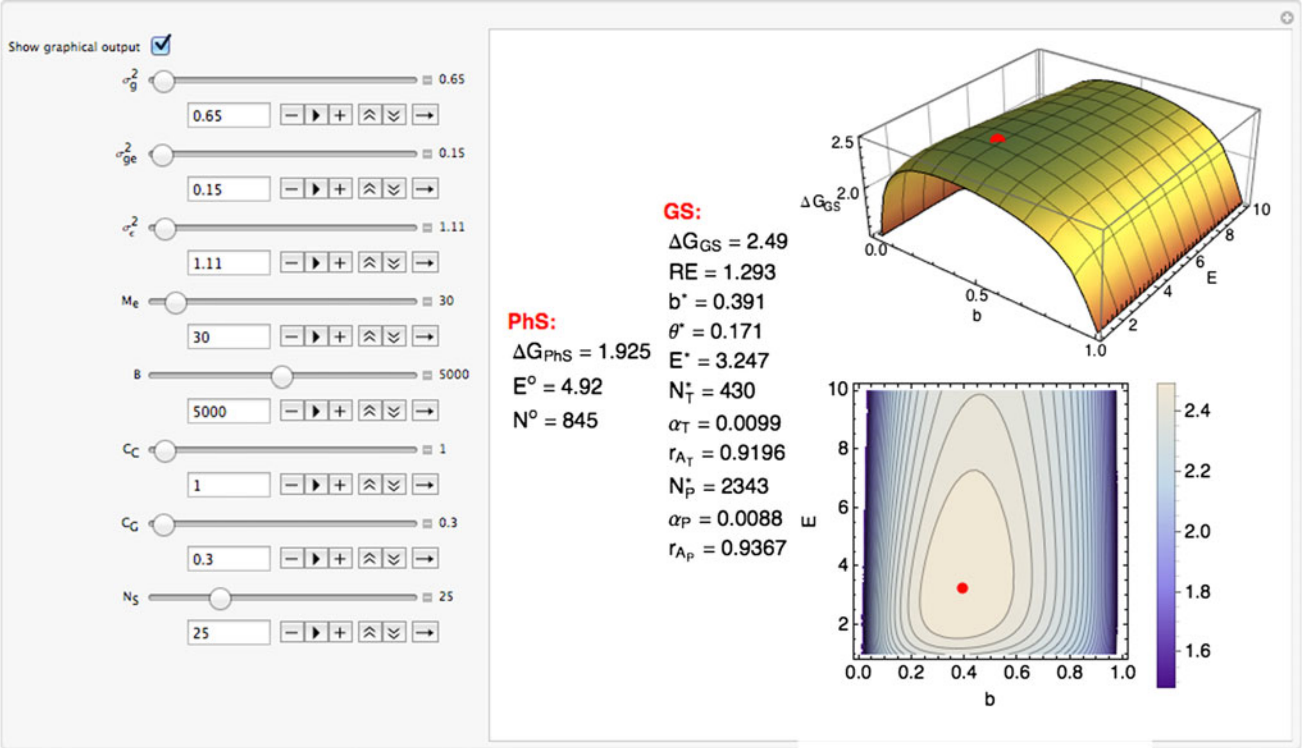
Extensions over multiple traits and multiple selection cycles

Until now, we considered selection for a single trait. In practice, selection is mostly exercised on an index of several traits. In this case, the costs for producing a DH line or genotyping in units of PE fall greatly. Recently, it was shown that selection accuracy of a lowly heritable trait can be significantly improved if a correlated, highly heritable trait is incorporated in a multiple-trait GS model (Jia and Jannink 2012; Calus and Veerkamp 2011). However, it is unclear whether a GS model should be built for each trait separately and the predicted values subsequently combined into an index, or if a GS model should be built directly for the index of all relevant traits. In addition, optimization strategies also need to consider the declining costs for genotyping relative to the costs needed to phenotype several instead of only one trait.

While breeders develop their material mainly for identifying superior parents to be used for creating commercial hybrids, they use their material also for recycling the best lines in future selection cycles or recurrent selection programs. In the latter case, it is unclear how selection gain per cycle is maintained under GS, i.e., after how many cycles the GS model has to be re-trained with fresh phenotypic data. For allogamous crops, a recent simulation study found a breeding scheme to be superior if re-trained after only several cycles, provided that the population size is sufficiently large (Yabe et al. 2013). It seems likely that a higher budget allocated to the training set will improve the stability of the selection accuracy over several cycles. Results from Massman et al. (2012) suggest diminishing or erratic gains from later cycles of selection for grain yield and stover-quality traits in maize. It is also important to note that phenotyping the training set is usually performed only in a single year so that environments refers to different locations. Under this scenario, the genotype × environment interaction variance disregards the genotype × year and genotype × year × location variances. This leads to inflated estimates of h^2 and consequently of the predicted selection accuracy. However, the expected upward bias in the expected selection gain should be similar for GS and PhS. Jannink (2010) further suggests putting an additional weight on low frequency favorable alleles to warrant higher long-term selection gain. The prediction model used to calculate genomic estimated breeding values seems to play an important role in this context. Methods such as BayesB, which try to precisely capture individual QTL, improved selection accuracies over generations in a study simulating the situation in animal breeding (Habier et al. 2007). Since economic optimizations rely on solid information about the expected predication accuracy, analytical functions describing the decline of the selection accuracy with the number of selection cycles need to be developed. Our approach to optimize the allocation of resources within one cycle could provide the basis for further extensions to cope with multiple selection cycles.

Constraint optimization of resource allocation as a planning tool

As an alternative to the model calculation approach followed in this study, one could have conducted simulations based on stochastically generated genotypic and phenotypic values. This approach requires a number of assumptions on the genome level including the number of chromosomes, loci and the distribution of genetic effects. Simulations might be best suited for investigating the long-term consequences of different selection schemes because the effects of selection on the gene pool composition can be followed in detail over time (Tomerius 2001; Jannink 2010). For one breeding cycle, the model calculation approach employed in this study has the advantage that it can be used as a planning tool early in the design of breeding programs. Based on quantitative genetic assumptions including Fisher's (1918) infinitesimal genetic model, the constraint optimization approach delivers directly the theoretical optimum allocation of resources given a priori known constrains (budget, costs for genotyping and phenotyping) and the ratios of the masking variances to the genotypic variance for the target trait. These ratios can initially be estimated from previous experiments or taken from the literature but deviations are possible for each cross, especially if the parents are related. However, given a reasonable high budget, the optimum allocation is quite robust with respect to these assumptions (Fig. 5). After phenotyping of the training population has been performed, calculation of the optimal weights in the index can be updated with the actual variance components measured for the target trait in the given population. In case that further practical limitations such as limitations in the number of available environments or the population size N emerge, they can be easily incorporated into the set of constraints of the model calculations.



Show graphical output ☑

σ_g^2 0.65

0.65

σ_{ge}^2 0.15

0.15

σ_ε^2 1.11

1.11

M_e 30

30

B 5000

5000

C_C 1

1

C_G 0.3

0.3

N_S 25

25

PhS:

$\Delta G_{PhS} = 1.925$

$E^{\circ} = 4.92$

$N^{\circ} = 845$

GS:

$\Delta G_{GS} = 2.49$

$RE = 1.293$

$b^* = 0.391$

$\theta^* = 0.171$

$E^* = 3.247$

$N_T^* = 430$

$\alpha_T = 0.0099$

$r_{A_T} = 0.9196$

$N_P^* = 2343$

$\alpha_P = 0.0088$

$r_{A_P} = 0.9367$

Fig. 6 Interactive graphical user interface of the optimization software. The software (Supplementary material 1) can be used with the freely available computable document format (CDF) player (Wolfram research Inc., www.wolfram.com/cdf-player). Parameters σ_g^2, σ_{ge}^2, σ_ε^2, M_e, B, C_C, C_G, and N_s can be adjusted using sliders at the *left panel*. The graphical output can be deactivated to speed up computation

We provide the optimization program in an interactive computable document format (CDF) usable with the freely available CDF player (www.wolfram.com/cdf-player) (Fig. 6; Online Resource 1). Users are able to adjust values of all parameters (σ_g^2, σ_{ge}^2, σ_ε^2, M_e, B, C_C, C_G, N_s,) on the left panel of the interface. The software not only immediately calculates the optimum allocation for PhS and GS, but also directly shows the amount of flexibility in resource allocation the breeder has at hand for achieving a selection gain close to the maximum. The value of ΔG_{GS} at the contour lines is directly given at the mouse pointer when moving over the contour plot.

Conclusions

Using grain yield at the testcross level as a model trait, we found the optimum resource allocation under GS and its flexibility to be highly dependent on the available budget. Under a small budget, the superiority of GS over PhS strongly depends on the genotyping costs and the flexibility in terms of budget splitting and the optimum number of environments is highly restricted. Using an index combining phenotypic values and GS predicted values in the training set especially pays off under a restricted budget and substantial contributions of genotype × environment interactions. With increasing budget, the relative efficiency of GS is increasingly more robust towards genotyping costs or the assumed number of effective loci. However, the focus can only start to switch from the training set towards the prediction set, once the training population warrants sufficiently high selection accuracy in the prediction set. Once selection accuracy r_{A_P} reaches $\sim$0.8, further selection gain can only be achieved by increasing the selection intensity in the prediction set through massively expanding its size. Reducing the costs for producing DH lines is crucial to accomplish this step. With increasing budget allocated to the prediction set, also a larger fraction of the total selected candidates needs to be drawn from the prediction set. However, since selection gain under GS quickly levels off with increasing budget, it is important to balance the number of populations with the budget spent for each population.

Our analysis of the efficiency of GS gives a promising perspective for increasing selection gain in line and hybrid development, especially in the light of decreasing costs for genotyping and further progress in the DH technology. Our analysis demonstrates that optimizing the allocation of resources must consider many interdependencies between quantitative genetic parameters of the trait and population, as well as economic parameters. Our software implementation can serve as a tool for assisting practitioners to analyze and optimize their implementation of GS.

Acknowledgments This publication is dedicated to Professor Chris-Carolin Schön, TU München, to acknowledge her outstanding role in initiating and coordinating the "Synbreed" project and promoting genomic selection in plant breeding. Funding for this research came from the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr "Synbreed—Synergistic Plant and Animal Breeding" (grant 0315528D) as well as from DuPont Pioneer under a Ph.D. fellowship for C.R.

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

Ethical standards The authors declare that all experiments comply with the current laws in Germany.

References

Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci 48:1649–1664

Bernardo R, Yu J (2007) Prospects for genome-wide selection for quantitative traits in maize. Crop Sci 47:1082–1090

Calus MPL, Veerkamp RF (2011) Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol 43:26

Combs E, Bernardo R (2013) Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. Plant Genome 6:1–7

Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3:e3395

Daetwyler HD, Calus MPL, Pong-Wong R, de los Campos G, Hickey JM (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347–365

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345

Dekkers JCM, Hospital F (2002) The use of molecular genetics in the improvement of agricultural populations. Nat Rev Genet 3:22–32

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379

Endelman JB, Atlin GN, Beyene Y, Semagn K, Zhang X, Sorrells ME, Jannink JL (2013) Optimal design of preliminary yield trials with genome-wide markers. Crop Sci. doi:10.2135/cropsci2013.03.0154

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. Pearson, Essex

Fisher RA (1918) The correlation between relatives on the supposition of mendelian inheritance. Phil Trans R Soc Edinb 52:399–433

Goddard M (2009) Genomic selection: prediction of accuracy and maximization of long term response. Genetica 136:245–257

Goddard ME, Hayes BJ, Meuwissen THE (2011) Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet 128:409–421

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397

Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1–12

Heslot N, Yang H-P, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146–160

Jannink J-L (2010) Dynamics of long-term genomic selection. Genet Sel Evol 42:35

Jia Y, Jannink J-L (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192:1513–1522

Knapp SJ, Bridges WC (1990) Using molecular markers to estimate quantitative trait locus parameters: power and genetic variances for unreplicated and replicated progeny. Genetics 126:769–777

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756

Lorenz AJ (2013) Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. G3: Genes, Genomes, Genetics 3:481–491

Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink JL (2011) Genomic selection in plant breeding: knowledge and prospects. Adv Agron 110:77–123

Lorenzana RE, Bernardo R (2008) Genetic correlation between corn performance in organic and conventional production systems. Crop Sci 48:903–910

Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet 120:151–161

Massman JM, Jung HJG, Bernardo R (2012) Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. Crop Sci 53:58–66

Melchinger AE, Utz HF, Schön CC (1998) Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. Genetics 149:383–403

Melchinger AE, Schipprack W, Würschum T, Chen S, Technow F (2013) Rapid and accurate identification of in vivo induced haploid seeds based on oil content provides a new tool for maize genetics and breeding. Sci Rep 3:2129

Meuwissen THE, Goddard M (2010) Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623–631

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

Moreau L, Charcosset A, Hospital F, Galais A (1998) Marker-assisted selection efficiency in populations of finite size. Genetics 148:1353–1365

Moreau L, Lemarié S, Charcosset A, Gallais A (2000) Economic efficiency of one cycle of marker-assisted selection. Crop Sci 40:329–337

Moreau L, Charcosset A, Gallais A (2004) Experimental evaluation of several cycles of marker-assisted selection in maize. Euphytica 137:111–118

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5:92–102

Prigge V, Schipprack W, Mahuku G, Atlin GN, Melchinger AE (2012) Development of in vivo haploid inducers for tropical maize breeding programs. Euphytica 185:482–490

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE

(2012a) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44:217–220

Riedelsheimer C, Technow F, Melchinger AE (2012b) Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. BMC Genomics 13:452

Riedelsheimer C, Endelmann JB, Stange M, Sorrells ME, Jannink JL, Melchinger AE (2013) Genomic predictability of interconnected biparental maize populations. Genetics 194:493–503

Schön CC, Utz HF, Groh S, Truberg B, Openshaw S, Melchinger AE (2004) Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics 167:485–498

Technow F, Riedelsheimer C, Schrag TA, Melchinger AE (2012) Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet 125:1181–1194

Tomerius AM (2001) Optimizing the development of seed-parent lines in hybrid rye breeding, Dissertation, University of Hohenheim

Utz HF (1969) Mehrstufenselektion in der Pflanzenzüchtung. Dissertation, University of Hohenheim

Villanueva B, Dekkers JCM, Woolliams JA, Settar P (2004) Maximizing genetic gain over multiple generations with quantitative trait locus selection and control of inbreeding. J Anim Sci 82:1305–1314

Xu YB, Crouch JH (2008) Marker-assisted selection in plant breeding: from publications to practice. Crop Sci 48:391–407

Xu Y, Lu Y, Xie C, Gao S, Wan J, Prasanna BM (2012) Whole-genome strategies for marker-assisted plant breeding. Mol Breeding 29:833–854

Yabe S, Ohsawa R, Iwata H (2013) Potential of genomic selection for mass selection breeding in annual allogamous crops. Crop Sci 53:95–105