Published March 20, 2017

GBS-Based Genomic Selection for Pea Grain Yield under Severe Terminal Drought

Paolo Annicchiarico,* Nelson Nazzicari, Luciano Pecetti, Massimo Romani, Barbara Ferrari, Yanling Wei, and E. Charles Brummer

Abstract

Terminal drought is the main stress that limits pea (*Pisum sativum* L.) grain yield in Mediterranean-climate regions. This study provides an unprecedented assessment of the predictive ability of genomic selection (GS) for grain yield under severe terminal drought using genotyping-by-sequencing (GBS) data. Additional aims were to assess the GS predictive ability for different GBS data quality filters and GS models, comparing intrapopulation with interpopulation GS predictive ability and to perform genome-wide association (GWAS) studies. The yield and onset of flowering of 315 lines from three recombinant inbred line (RIL) populations issued by connected crosses between three elite cultivars were assessed under a field rainout shelter. We defined an adjusted yield, which is associated with intrinsic drought tolerance, as the yield deviation from the value expected as a function of onset of flowering (which correlated negatively with grain yield). Total polymorphic markers ranged from approximately 100 (minimum of eight reads per locus, maximum 10% genotype missing data) to over 7500 markers (minimum of four reads, maximum 50% missing rate). Best predictions were provided by Bayesian Lasso (BL) or ridge regression best linear unbiased prediction (rrBLUP), rather than support vector regression (SVR) models, with at least 400–500 markers. Intrapopulation GS predictive ability exceeded 0.5 for yield and onset of flowering in all populations and approached 0.4 for the adjusted yield of a population with high trait variation. Genomic selection was preferable to phenotypic selection in terms of predicted yield gains. Interpopulation GS predictive ability varied largely depending on the pair of populations. GWAS revealed extensive colocalization of markers associated with high yield and early flowering and suggested that they are concentrated in a few genomic regions.

Published in Plant Genome
Volume 10. doi: 10.3835/plantgenome2016.07.0072

Core Ideas

- GBS-based genomic predictions of pea grain yield and phenology are accurate and cost-efficient.
- Genomic areas related to high yield and early flowering colocate under severe terminal drought.
- Cross-population genomic predictions have quite variable predictive ability.

LEGUME CROPS are expected to assume a pivotal role in future farming systems—to increase their sustainability in terms of soil fertility, energy efficiency, greenhouse gas emissions, and crop diversity, while satisfying the increasing demands for high-protein feed and nutritious food (Jensen and Hauggard-Nielsen, 2003; Schneider and Huyghe, 2015). Field pea (*Pisum sativum* L.) is the most-grown grain legume in Europe, where it displays higher yield potential than other cool-season grain legumes in western (Carrouée et al., 2003) and southern Europe (Annicchiarico, 2008). High protein and energy value for animal nutrition and remarkable flexibility of utilization (as grain, hay, or silage) are further assets of this crop (Carrouée et al., 2003).

P. Annicchiarico, N. Nazzicari, L. Pecetti, M. Romani and B. Ferrari, Council for Agricultural Research and Economics (CREA), Research Centre for Fodder Crops and Dairy Productions, 29 viale Piacenza, 26900 Lodi, Italy; Y. Wei and E.C. Brummer, Plant Breeding Center, Dep. of Plant Sciences, Univ. of California, Davis, CA 95616. Received 22 July 2016. Accepted 23 Jan. 2017. *Corresponding author (paolo.annicchiarico@crea.gov.it).

Abbreviations: BL, Bayesian Lasso; BLUP, best linear unbiased prediction; GBS, genotyping-by-sequencing; GS, genomic selection; GWAS, genome-wide association study; LD, linkage disequilibrium; LG, linkage group; QTL, quantitative trait locus; RIL, recombinant inbred line; rrBLUP, ridge regression BLUP; SNP, single-nucleotide polymorphism; SVR, support vector regression.

Drought is the main environmental factor limiting agricultural production worldwide. The Mediterranean climate is characterized by wet, mild winters and dry, hot summers. Drought that occurs during spring tends to coincide with the critical phases of grain setting and filling, thereby seriously affecting the crop yield. The ability of crops to yield satisfactorily in these conditions may be achieved through different mechanisms that provide either drought escape or drought tolerance (Fang and Xiong, 2015). Escape via an early phenology is very important in severely drought-prone Mediterranean-climate environments (Turner et al., 2001). These environments are expected to become common throughout southern Europe and northern Africa and to expand northward and eastward into central Europe as a consequence of climate change (Alessandri et al., 2014). Crop genetic improvement is a main means to adapt to climate change and mitigate its effects (Ceccarelli et al., 2010). Breeding for harsh Mediterranean environments implies selection under severe drought (Ceccarelli, 1989). Managed selection environments can be an important asset for coping with the year-to-year rainfall variation that is typical of rainfed Mediterranean environments, by reproducing faithfully the genotype adaptive responses that are observed under ordinary stress in the target region (Annicchiarico and Piano, 2005; Bänziger et al., 2006).

Genomic selection enables breeders to predict complex, polygenic traits by means of a statistical model constructed from genome-wide marker information (Meuwissen et al., 2001). Genomic selection has shown potential for increasing the accuracy of traditional marker-assisted selection in terms of gain per selection cycle and per unit cost (Heffner et al., 2010). Earlier applications of GS to pea on the basis of single-nucleotide polymorphism (SNP) array data displayed high accuracy for prediction of flowering time and two grain-yield component traits (Burstin et al., 2015; Tayeh et al., 2015b). Genotyping-by-sequencing (Elshire et al., 2011) is a recent method to derive genome-wide marker genotypes from sequence data at a lower cost than many SNP array platforms, albeit with large amounts of missing data. Various factors, such as the chosen genomic selection model (Lorenz et al., 2011) and the method of missing data imputation (Nazzicari et al., 2016), deserve specific investigation, because they may influence the GS predictive ability (e.g., Annicchiarico et al., 2015; Burstin et al., 2015).

An issue of great practical interest is the ability of GS models to predict traits in germplasm/reference populations that differ from the population in which the models were defined. Transferability of models would obviously decrease the cost of model development and impact the strategies of GS implementation in breeding programs. For crop yield, interpopulation predictions were reportedly poor in wheat (Charmet et al., 2014) and moderate in alfalfa (Annicchiarico et al., 2015). Moderately high interpopulation prediction ability was reported in pea for onset of flowering and seed weight (Tayeh et al., 2015b).

The main objective of our study was to assess the ability of GS based on GBS data to predict pea grain yield under severe terminal drought. Grain yield was assessed in one growing season under managed drought stress conditions. In addition to analyzing the actual yield results, we also analyzed data corrected to remove the positive effect of early flowering, to minimize the impact of a stress escape strategy. The newly obtained adjusted grain yield allowed for investigating yield responses associated with stress tolerance mechanisms. Phenotypic data were obtained from 315 inbred lines derived from three connected crosses between elite parent genotypes of European or Australian origin that displayed excellent adaptation to south European environments in prior, extensive variety testing. Additional aims of our study were (i) providing information on the extent of polymorphic SNP markers obtainable by GBS for pea inbred line populations; (ii) investigating the effect on GS predicting ability of different GBS data quality filters and GS models; (iii) assessing the decrease of predictive ability of GS models trained on lines from one cross when applied for predicting lines from another cross that share one common parent. Finally, we also used GBS-generated markers for GWAS studies for grain yield and phenology traits to get a better understanding of the genetic control underlying these traits.

MATERIALS AND METHODS

Plant Material

Our study included 105 genotypes for each of three connected RIL populations originated from paired crosses between 'Attika' (a European cultivar described as a spring-type), 'Isard' (a French winter-type cultivar), and 'Kaspa' (an Australian cultivar). These cultivars displayed fairly similar phenology and cycle duration along with high and stable grain yield and other positive agronomic characteristics across environments of northern and southern Italy (Annicchiarico, 2005; Annicchiarico and Iannucci, 2008). The RIL populations are coded hereafter as A×I, K×A, and K×I from the initials of their respective parents. Four F_6 plants per line were previously grown in an unheated glasshouse to collect DNA samples for line genotyping and to produce seed, which underwent one additional generation of multiplication before being used for phenotyping.

Phenotyping

The 315 lines were evaluated for onset of flowering and grain yield under severe terminal drought under a field rainout shelter equipped with a double-rail irrigation boom at the Research Centre for Fodder Crops and Dairy Productions, Lodi, Italy. Sowing took place in late winter (25 Feb. 2015) to avoid any confounding effect of genetic variation for tolerance to low winter temperatures. The RILs were sown in plots that were 0.160 m^2 (0.8 m × 0.2 m) and included two rows of eight plants each, with 0.1 m spacing between rows and between plants (resulting in an agronomic seed rate of 100 seeds m^{-2}). The experimental design was an α lattice with four replications: each replication included 21 incomplete blocks of 15 plots. By enabling better control of the experimental error, the lattice design provided better-quality phenotypic data for BLUP computation and subsequent analyses. The soil was sandy-loam (FAO classification; 53% sand, 35% clay, and 12% silt), with a field capacity of 17.2% at 15 cm and 13.9% at 30 cm depth. Plot harvest spanned from 26 May to 3 June 2015. A total of 120 mm irrigation was provided throughout the crop cycle at progressively lower amounts (60 mm in March, 35 mm in April, 25 mm in May). This irrigation scheme, which mimicked the Mediterranean-climate rainfall pattern observed in the driest areas of southern Italy (Del Monte et al., 1995), provided a severe terminal drought for the crop, when considering the late sowing and the fact that the potential evapotranspiration from March to May amounted to 239 mm. Field pea is a cool-season species whose optimal mean air temperatures for growth and production range between 12 and 18°C. In this study, daily mean temperatures averaged 14.4°C in April 2015 and 19.3°C in May 2015. A slight heat stress may have occurred during May, when daily maximum temperatures averaged 24.9°C. Temperatures recorded under the rainout shelter were approximately 1°C higher than in the surrounding fields. We recorded onset of flowering as the number of days after 1 April when 50% of plants in the plot had at least one fully open flower. At maturity, all the plants from each plot were harvested and hand-threshed, dry grain yield was recorded in g per plot after oven-drying the seeds at 90°C for 4 d and then converting the yield to tons per hectare.

Statistical Analysis of Phenotypic Data

After adjusting line means according to the α lattice design, we assessed each RIL population individually to determine whether genotypic grain yield was affected by onset of flowering; we used a regression analysis performed on mean yield values of each line. We verified the significance of linear and curvilinear responses and used analysis of covariance to assess the occurrence of variation among regression slopes of the three RIL populations. In the presence of a significant inverse linear response, for each population we estimated for each line an "adjusted" grain yield on a plot basis as the deviation of its actual yield from the yield value expected for the line as a function of its onset of flowering in the linear regression model for the population. Thus, adjusted grain yields (which had negative or positive values according to the deviation direction and averaged zero across plot values of all lines within each RIL population) enabled comparison of the RILs for grain yield after removing the mean effect of drought escape as determined by differences in phenology, thereby focusing on grain yield as affected essentially by drought tolerance mechanisms.

An analysis of variance (ANOVA) assessed the variation among genotypes within each RIL population for grain yield, onset of flowering, and adjusted grain yield. Components of variance relative to variation among lines (s^2_1) and experimental error (s^2_e) were estimated for each RIL population by a restricted maximum likelihood method and used to compute broad sense heritability (H^2) values on a line mean basis as $H^2 = s^2_1/(s^2_1 + s^2_e/r)$, where r (number of replicates) = 4. Another ANOVA including the fixed factor RIL population and the random factor line within population aimed to compare the populations for grain yield and onset of flowering using the average within-population variation as the error term. We used best linear unbiased prediction (BLUP) values computed according to DeLacy et al. (1996) for GS and GWAS analyses. All the statistical analyses were performed using Statistical Analysis System (SAS) or CropStat software programs.

DNA Isolation, GBS Library Construction, and Sequencing

Green tissue was collected from bulked stipules of four F_6 plants per line, flash frozen in liquid nitrogen, stored at −80°C, and then ground for genomic DNA isolation. DNA was extracted using 400 mg of tissue using a CTAB method described by Rogers and Bendich (1985) and then checked on 1% agarose gel to assess yield and quality.

The DNA was quantified with a Quant-iT PicoGreen dsDNA assay kit (Life Technologies). We used the protocol by Elshire et al. (2011) with modifications. Each DNA sample (100 ng) was digested with *Ape*KI (NEB) and then ligated to a unique barcoded adapter plus a common adapter. Equal volumes of the ligated product were pooled and cleaned up with QIAquick PCR Purification Kit (QIAGEN) for subsequent amplification. In the polymerase chain reaction, 50 ng template DNA was mixed with two primers and KAPA Library Amplification Readymix (KAPA Biosystems). Amplification was performed on a thermocycler for 10 cycles with 10 sec of denaturation at 98°C followed by 30 sec of annealing at 65°C and a 30-sec extension at 72°C. Each library was sequenced in two lanes on Illumina HiSeq 2000 at the Genomic Sequencing and Analysis Facility at the University of Texas at Austin.

Genotype SNP Calling and Data Filtering

We used the UNEAK pipeline (Lu et al., 2013) for SNP discovery and genotype calling. The raw reads (100 bp, single-end read) obtained from sequencing were first quality-filtered and de-multiplexed. All reads beginning with the expected barcodes and cut-site remnant were trimmed to 64 bp. Identical reads were grouped into one tag. Tags with 10 or more reads across all individuals were retained for pairwise alignment, which aimed to find tag pairs that differed by 1 bp. For each SNP marker, the read distribution of the paired tags in each individual was used for SNP genotype calling. A further quality filter, implemented through ad hoc Python scripts, removed markers with fewer than four, six, or eight aligned reads. The resulting three data sets were filtered for increasing levels of allowed missing values, excluding markers whose missing rate over genotypes was greater than a fixed threshold of 10, 20, 30, 40, and 50%. Markers that were monomorphic or had a minor allele frequency below 2.5% were removed. Following Nazzicari et al. (2016), we estimated missing data using the K-nearest neighbors imputation algorithm (K = 4) coupled with the simple matching coefficient distance function (Schwender, 2007), as implemented in the R package Scrime (Schwender and Fritsch, 2013).

Genomic Regression Models

Four regression models were evaluated for genomic predictions: rrBLUP, BL, and two SVR models.

Ridge regression BLUP assumes a linear mixed additive model in which each marker is assigned an effect as a solution of the equation

$$\mathbf{y} = \mu + \mathbf{G}\,\mathbf{u} + \varepsilon\ ,$$

where $\mathbf{y}$ is the vector of observed phenotypes, μ is the mean of $\mathbf{y}$, $\mathbf{G}$ is the genotype matrix (e.g., {0,1,2} for biallelic SNPs), $\mathbf{u} \sim N(0,\mathbf{I}\sigma_u^2)$ is the vector of marker effects, and $\varepsilon \sim N(0,\mathbf{I}\sigma_\varepsilon^2)$ is the vector of residuals. Solving with the standard ridge-regression method, the solution is

$$\hat{\mathbf{u}} = \mathbf{G}'(\mathbf{G}\,\mathbf{G}' + \lambda I)^{-1}(\mathbf{y} - \mu)\ ,$$

where $\lambda = \sigma_\varepsilon^2 / \sigma_u^2$ is the ridge parameter, representing the ratio between residual and markers variance (Searle et al., 2009). Given the vector of effects, it is then possible to predict phenotypes and estimate genetic breeding values. Ridge regression BLUP analysis was performed through the R software package rrBLUP (Endelman, 2011), estimating λ in a restricted maximum likelihood scheme implemented by a spectral decomposition algorithm (Kang et al., 2008) and solving the resulting linear model.

Bayesian-based models assign prior densities to markers effects, thereby inducing different types of shrinkage. The solution is obtained by sampling from the resulting posterior density through a Gibbs sampling approach (Geman and Geman, 1984; Casella and George, 1992). We selected the BL (Park and Casella, 2008) as implemented by the R software package BGLR (De los Campos and Perez Rodriguez, 2014).

Support vector regression models are based on the computation of a linear regression function in a high-dimensional feature space where input data are mapped via a kernel function (Schölkopf and Smola, 2002). We considered two major kernel functions: linear (SVR-lin) and radial base (SVR-rbf). We used the ε-insensitive regression present in the Weka framework (Hall et al., 2009).

Methodological Evaluation

Different minimum read thresholds per locus may be considered for self-pollinated crop material. While a four-read threshold might be suitable for perfectly inbred lines, higher thresholds could be considered for material with some expected level of heterozygosity (as the current one, in which genotyping concerned the pooled DNA from various F_6 plants). We considered all possible combinations of three minimum numbers of reads per locus (four, six, or eight), five genotype missing data thresholds (10, 20, 30, 40, and 50%), and four regression models (rrBLUP, BL, SVR-lin, SVR-rbf). For each of these 60 combinations, we evaluated the predictive ability as Pearson's correlation between true and predicted phenotypes in a 10-fold stratified cross-validation scheme (where training and validation hold 90 and 10% of data, respectively). Each cross-validation experiment was repeated 100 times, averaging the results to ensure numerical stability.

Unaccounted population structure may affect the predictive ability of GS models (Guo et al., 2014). For rrBLUP and BL models, population structure can be taken into account by adding a RIL population fixed factor, as a 3 × n matrix (n = number of samples) of zeros and ones (coupling each sample to one of the three populations). We compared cross-validation-based predictive abilities in the absence and the presence of imputed structure information.

Intrapopulation versus Interpopulation Predictive Ability

The best-performing combination of minimum reads per locus, threshold for genotype missing data, and regression model in the methodological study was selected for comparing the predictive ability of GS models within versus across RIL populations. For within-population predictions, both training and test sets came from the same cross (A×I, K×A, or K×I). We adopted a 10-fold stratified cross-validation approach to avoid overfitting. Also, a second evaluation was performed in which the model was trained on the joint data set containing the three populations. Cross-validations were repeated 100 times. This procedure resulted in six assessments for each trait (one per cross, with training done on single or all crosses). For interpopulation predictions, all data from a single RIL population were used to train a model that was then tested on the two other populations. The predictions were not repeated, since all available data were used for training (hence, with no variability in data selection). This procedure resulted in six assessments for each trait (three models, each tested on two populations).

Prediction accuracy (i.e., the correlation between genome-based predicted values and true breeding values) is more meaningful than predictive ability (i.e., the correlation between genome-based predicted values and observed values) for assessing GS gains. However, the ordinary estimation of prediction accuracy as the ratio of predictive ability to the square root of the broad-sense heritability on a line mean basis (Heffner et al., 2011) may introduce a bias when cross-validations are applied to data of the same experiment (Lorenz et al., 2011), as in the current case. Therefore, we preferred to use predictive ability in place of prediction accuracy for comparing GS with phenotypic selection in terms of predicted yield gains. This choice implied a prudent assessment of the relative value of GS, since predictive ability has a lower value than prediction accuracy (the two terms being coincident only when heritability reaches unity).

Genome-Wide Association Studies

A GWAS study was performed for each trait (onset of flowering, grain yield, and adjusted grain yield) on pooled data of the three RIL populations. We preferred GWAS to QTL mapping analysis (which is also possible for biparental populations) because of the opportunity that it offered to use the entire set of lines in a single study, thereby maximizing the statistical power of the assessment. GWAS analyses were performed using the *egscore* function in the R package GenABEL (Aulchenko et al., 2007). The *egscore* function is based on a linear mixed model and handles population structure and relatedness in the data by adjusting for principal components of the genomic kinship matrix, as described by Price et al. (2006). After inspecting the eigenvalues of the kinship matrix (obtained from the GenABEL function *ibs* using weights based on allelic frequency), we opted to correct for the first two principal components. A further genomic control correction of the obtained P values was performed using the inflation factor (λ). In addition, we performed GWAS studies on each individual RIL population (for a total of three traits $\times$ three population studies) to verify the consistency of its results with those of the pooled analysis (albeit on the basis of reduced linkage detection power relative to the pooled analysis).

No reference genome is available yet for *Pisum sativum*. Its closest relative having a mature reference genome is *Medicago truncatula* L. We used the Bowtie 2 tool (Langmead and Salzberg, 2012) to query the consensus sequence of each tag pair containing a SNP against the *M. truncatula* reference genome version 4.1 using the *-very-sensitivelocal* preset, assessing the proportion of aligned markers. We set an association score threshold of 2 (i.e., $P < 0.01$ for significance of the association), sorted significant markers by their association value produced by GWAS (thus creating one ranking per trait), and investigated the markers associated with more than one trait.

Ferrari et al.'s (2016) pea trait-marker study included a consensus map derived from 206 SNP markers obtained by an Illumina array platform. Some 130 polymorphic loci distributed across the seven pea chromosomes covered approximately 1094 cM overall. These markers were observed on 270 lines (90 per RIL population), of which 180 (60 per population) were common to this study. We measured the linkage disequilibrium (LD) between our significantly associated markers and those used for the consensus map. We selected the squared correlation coefficient (r^2) as a measure of LD, classifying the results into highly linked ($r^2 > 0.8$), moderately linked ($0.4 < r^2 \leq 0.8$), and low/no linked markers ($r^2 \leq 0.4$). Concurrently, we assessed the LD between GBS-generated markers that showed no linkage with Illumina array-generated markers. Computation was performed with the R package genetics (Warnes et al., 2013).

Table 1. Mean and range values for grain yield, onset of flowering, and adjusted grain yield for three pea recombinant-inbred-line populations including 105 lines each.

Population†	Onset of flowering	Grain yield	Adjusted grain yield‡
	d from 1 April	t ha^{-1}	
Mean			
A × I	33.3 b§	0.339 a	—
K × A	35.0 a	0.321 a	—
K × I	35.9 a	0.293 a	—
Range			
A × I	26.6–39.2 **	0.097–0.658 **	−0.210–0.260 **
K × A	29.5–48.0 **	0.030–0.599 **	−0.261–0.210 *
K × I	27.4–50.0 **	0.028–0.708 **	−0.184–0.247 **

*Variation among lines significant at $P < 0.05$.

**Variation among lines significant at $P < 0.01$.

† Populations coded according their parent cultivars Attika (A), Isard (I), and Kaspa (K).

‡ As yield deviation from the yield value expected according to onset of flowering; see Fig. 1.

§ Means followed by different letters differ at $P < 0.05$ according to Newman and Keuls test.

RESULTS

Phenology and Yield under Drought Stress

The occurrence of severe terminal stress in the phenotyping experiment was confirmed by the mean grain yield, which averaged only 0.317 t ha^{-1}, the yield of the top-performing line being 0.708 t ha^{-1}, and the presence in each population of lines that approached yield failure (Table 1). The variation among lines for grain yield and onset of flowering was significant ($P < 0.01$) within each RIL population. On average, A×I exhibited an approximately 2-d earlier onset of flowering than the other populations, along with a narrower range of variation for this trait (~ 14 d) relative to K×A (~ 21 d) and K×I (~ 25 d) (Table 1). The population K×I displayed a wider range of line values also for grain yield (Table 1). Populations did not differ for mean grain yield (Table 1).

Regression analysis indicated a significant ($P < 0.001$) inverse linear response of grain yield as a function of onset of flowering, as well as no curvilinear response ($P > 0.05$), within every RIL population (Fig. 1). The estimated regression slopes of the three populations were fairly similar (Fig. 1) and differed just at $P < 0.10$ in the covariance analysis. However, r^2 values indicate that the dependency of grain yield from phenology was very high in the populations K×A and K×I ($r^2 > 0.70$) and moderately high in the population A×I ($r^2 = 0.41$), which is characterized by earlier and less variable phenology.

Adjusted grain yield values (computed as deviations from the yield value expected according to phenology) displayed significant variation at $P < 0.01$ for A×I and K×I and at $P < 0.05$ for K×A. P values and range values (Table 1) indicated, altogether, that A×I featured larger variation for this trait than the other populations. Broad-sense heritability averaged across the three populations was very high for grain yield ($H^2 = 0.90$) and onset of flowering ($H^2 = 0.92$) and moderately high for the adjusted grain yield ($H^2 = 0.57$).

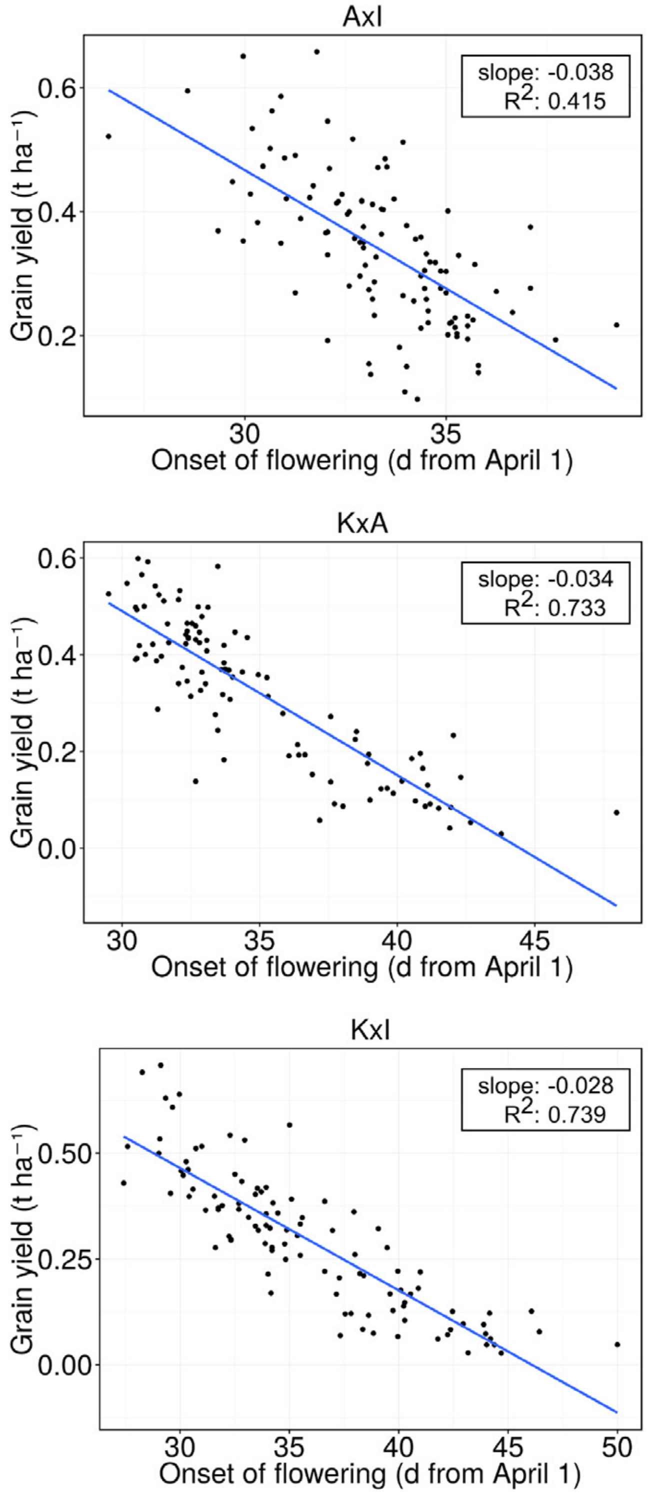

Figure 1. Linear regression of grain yield as a function of onset of flowering for three pea recombinant-inbred-line populations that include 105 lines each.

GBS Data, SNP Data Filtering, and Model Selection

Sequencing produced an average of 551,210 reads per sample. The UNEAK pipeline produced a gross total of 95,740 SNP markers. Increasingly relaxed requirements on the minimum number of reads per locus and allowed missing rate resulted in progressively more polymorphic markers available for regression models (Fig. 2), with total marker value varying from approximately 100 (minimum eight reads per locus, maximum 10% missing rate) to 7,521 markers (minimum four reads per locus, maximum 50% missing rate).

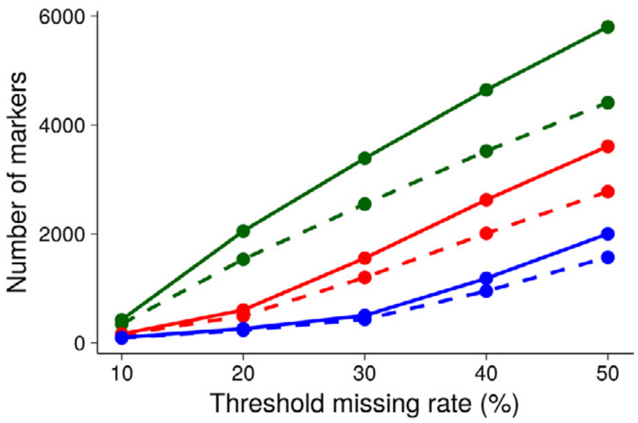

Figure 2. Number of polymorphic markers available in three pea recombinant-inbred-line populations for combinations of three minimum numbers of reads per locus (four, green lines; six, red lines; eight, blue lines) by five genotype missing data thresholds (10, 20, 30, 40, and 50%). Solid lines, total number of markers; dashed lines, average number of markers in each population.

Averaged across all crosses, high predictive ability was generally obtained when imposing a threshold of four reads per locus (with limited variation across missing data thresholds), six reads per locus with missing data thresholds of 20% or more, and eight reads with missing data thresholds of 40 or 50% (Fig. 3). The SVR-lin regression model displayed lower predictive ability than the other GS models, whereas BL and rrBLUP were top-performing in nearly all cases (with a slight, negligible advantage of the former). Results for the single RIL populations agreed well with results averaged across populations (data not reported). Following this study, we selected six reads per locus, the 20% genotype missing data threshold, and the BL model as the GS configuration for successive analyses. These parameters provided 617 polymorphic SNP markers overall, with values for individual RIL populations that ranged from 479 (A×I) to 514 (K×I) (Table 2).

Predictive Ability of Genomic Regression

Training the GS model on data of all of the RIL populations (in the absence of structure information), compared with training on individual populations, produced a modest gain in predictive ability for onset of flowering and grain yield, and a distinct predictive gain (19%) for the adjusted grain yield, which was characterized by lower predictive ability (Table 3). Including population structure information in the model trained on all populations provided inconsistent results, namely, a slight advantage for phenology and grain yield but a sizable decrease of predictive ability for the adjusted grain yield (–5%) (Table 3).

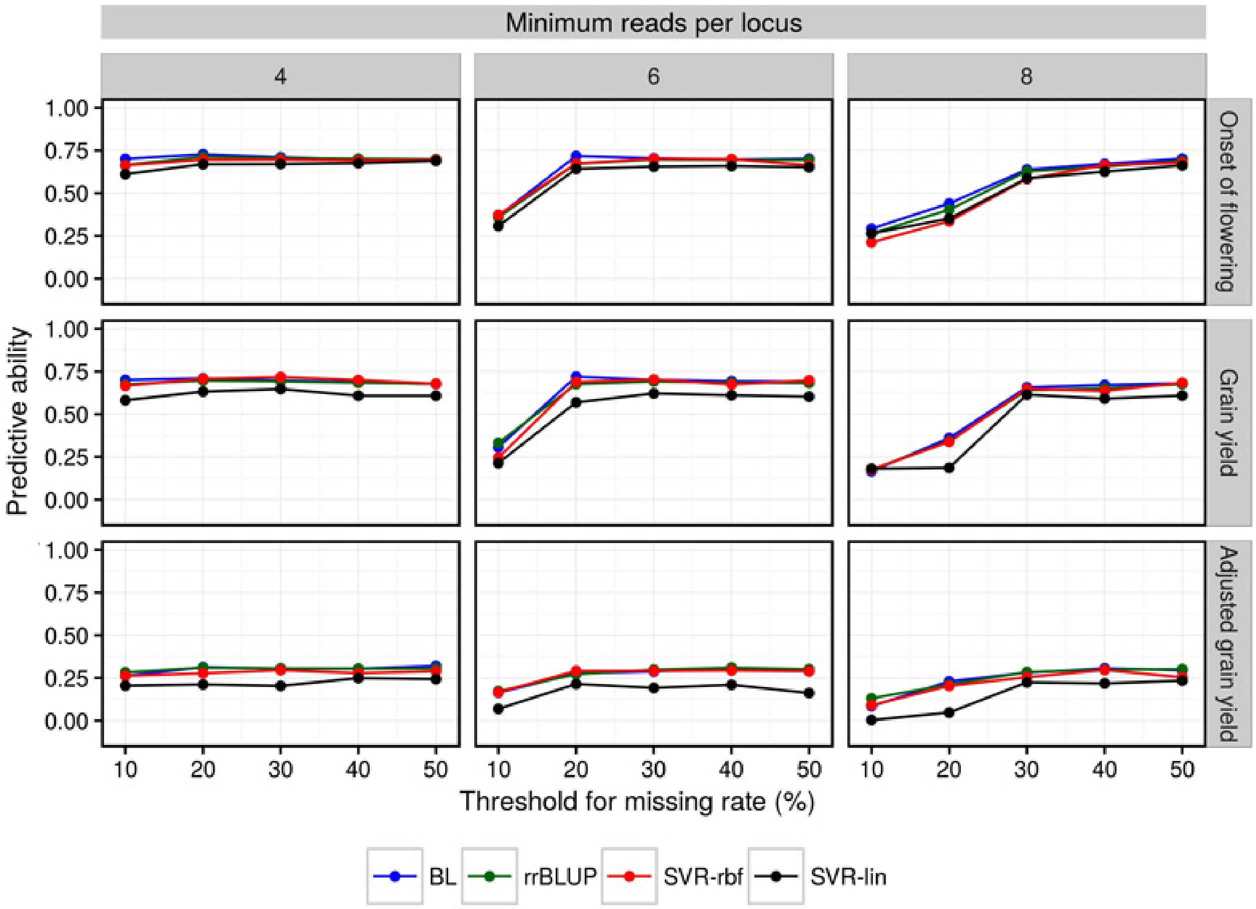

Figure 3. Predictive ability for all combinations of four genomic selection models (line type) by three minimum numbers of reads per locus (four, six, and eight) by five genotype missing-data thresholds (10, 20, 30, 40, and 50%) for onset of flowering, grain yield, and adjusted grain yield (as yield deviation from the yield value expected according to onset of flowering), averaged across three pea recombinant-inbred-line populations. Results for each combination and population are averages of 100 ten-fold stratified cross-validation repetitions. BL, Bayesian Lasso; rrBLUP, ridge regression best linear unbiased prediction; SVR-rbf, support vector regression, radial base function; SVR-lin, support vector regression, linear.

Table 2. Number of polymorphic single-nucleotide-polymorphism markers available in three pea recombinant-inbred-line populations for the selected genomic selection configuration of minimum six reads per locus and maximum 20% genotype missing data per marker.

Subset	Number
Markers in at least one population	617
Markers in A × I	479
Markers in K × A	497
Markers in K × I	514
Markers in A × I and K × A	362
Markers in A × I and K × I	378
Markers in K × A and K × I	412
Markers in A × I, K × A and K × I	279

On average, the intrapopulation predictive ability of the best-performing GS models was very high (around 0.7) for onset of flowering and grain yield and low for the adjusted yield (Table 3). However, the predictive ability varied between RIL populations in a fashion related to their within-population variation for the target trait. The population K×I, characterized by wider variation

Table 3. Predictive ability of intrapopulation genomic selection with different Bayesian Lasso model training and account of population structure for onset of flowering, grain yield, and adjusted grain yield of three pea recombinant-inbred-line populations.†

Trait	Training‡	Structure§	A × I	K × A	K × I	Average predictive ability
Onset of flowering	Single	No	0.537	0.755	0.802	0.698
	All	No	0.492	0.790	0.823	0.702
	All	Yes	0.489	0.796	0.830	0.705
Grain yield	Single	No	0.538	0.694	0.834	0.689
	All	No	0.540	0.744	0.836	0.707
	All	Yes	0.563	0.766	0.843	0.724
Adjusted grain yield¶	Single	No	0.391	0.079	0.218	0.229
	All	No	0.398	0.179	0.242	0.273
	All	Yes	0.396	0.158	0.221	0.259

† Data averaged across 100 repetitions of ten-fold stratified cross-validation.

‡ Single, model trained on the specific population; all, model trained on all populations joined in a single data set.

§ No, no structure information; yes, structure information as population fixed factor.

¶ As yield deviation from the yield value expected according to onset of flowering.

Table 4. Predictive ability of interpopulation genomic selection for onset of flowering, grain yield, and adjusted grain yield of three pea recombinant-inbred-line populations.

Trait	Training†	A×I	K×A	K×I	Average predictive ability
Onset of flowering	A × I	—	−0.194	0.396	0.101
	K × A	−0.182	—	0.694	0.256
	K × I	0.316	0.761	—	0.538
Grain yield	A × I	—	−0.065	0.410	0.173
	K × A	−0.059	—	0.705	0.323
	K × I	0.279	0.667	—	0.473
Adjusted grain yield‡	A × I	—	0.186	0.211	0.198
	K × A	0.278	—	0.044	0.161
	K × I	0.250	0.033	—	0.141

† Training column reports the population used for Bayesian Lasso model training.

‡ As yield deviation from the yield value expected according to onset of flowering.

Table 5. Number of significant markers (association score ≥ 2) in genome-wide association studies for three traits in pooled data of three pea recombinant-inbred-line populations and in each individual population. Common markers between pooled and individual populations are reported in parentheses.

Material	Onset of flowering	Grain yield	Adjusted grain yield
All populations	26	26	21
A × I	1 (0)	4 (0)	18 (8)
K × A	19 (17)	15 (15)	14 (5)
K × I	22 (21)	15 (15)	8 (0)

for onset of flowering and grain yield (Table 1) (as well as by a somewhat higher number of polymorphic markers: Table 2), exhibited top values of predictive ability for these variables, whereas the population A×I displayed the opposite pattern (Table 3). However A×I, which displayed the widest genetic variation for adjusted grain yield (Table 1), was the only population that achieved moderate predictive ability for this trait (around 0.4: Table 3).

Interpopulation predictive abilities are reported in Table 4. Imputing structure information was irrelevant here, since training was always performed using data from a single population. K×A and K×I, which were the population pair with highest number of common polymorphic markers (Table 2) and featured large variation for onset of flowering and grain yield, provided good models for each other for these traits. Their average interpopulation predictive ability for each other was 0.728 for onset of flowering and 0.686 for grain yield (Table 4). Compared with an average intrapopulation predictive ability of 0.779 for onset of flowering and 0.764 for grain yield for the same populations (Table 3), this result implied an approximately 7% prediction loss for onset of flowering and a 10% loss for grain yield when substituting interpopulation prediction for intrapopulation prediction. The populations A×I and K×I displayed moderate values of interpopulation predictive ability for each other, as they displayed approximately 50% loss for onset of flowering and grain yield and 24% loss for the adjusted grain yield when substituting interpopulation prediction for intrapopulation prediction (based on data in Tables 3 and 4). A×I and K×A displayed mostly no predicting ability for each other (Table 4).

Genome-Wide Association Studies

The GWAS performed on pooled data of the three RIL populations identified 26 GBS-generated markers associated with onset of flowering and grain yield and 21 markers associated with the adjusted grain yield, as summarized in Table 5. Association scores are graphically summarized in Figure 4, whereas detailed information on marker ranking and association score, sequence, and useful polymorphisms for each significant marker-trait association is reported in Supplemental Table S1. Twenty-five markers were associated with both onset of flowering and grain yield, confirming the strong genetic association between these traits under severe terminal drought. Several of these markers exhibited moderately high association score values, whereas marker-trait associations for the adjusted grain yield tended to be weaker.

The GWAS performed on the individual RIL populations confirmed to a large extent the findings obtained for the pooled populations (Table 5). Inconsistencies were mainly due to markers with a low but significant association score in the pooled GWAS, which failed to achieve significance in GWAS for individual populations (Supplemental Table S1), as expected from smaller detection power in the latter analyses. For example, of the 26 markers significantly associated with grain yield in the pooled GWAS, the 16 that featured higher association scores were also significant in K×I and/or K×A, whereas no remaining marker achieved significance in the A×I population (where, however, a few other markers emerged as significant, albeit with a low association score) (Supplemental Table S1). Likewise, association scores were lower in individual populations than in the pooled set of RIL material (Supplemental Table S1).

Only approximately 15% of the GBS-generated markers aligned to the *M. truncatula* genome, reinforcing the scope for exploring the linkage of these markers with Illumina array markers in Ferrari et al.'s (2016) consensus map to gain some clue about the possible position on pea linkage groups of these markers. Linkage analysis results are reported in Supplemental Table S1 for each marker. Inspection of highly ($r^2 > 0.8$) and moderately linked markers ($0.4 < r^2 \leq 0.8$) in the pooled set of RIL populations suggests that (i) 23 GBS markers that associated with both onset of flowering and grain yield, and 1 marker associated with the former trait (but close to significant association with grain yield), are in linkage with Illumina array markers of a genomic region in linkage group (LG) II, where an important QTL colocated for onset of flowering and grain yield in Ferrari et al. (2016);

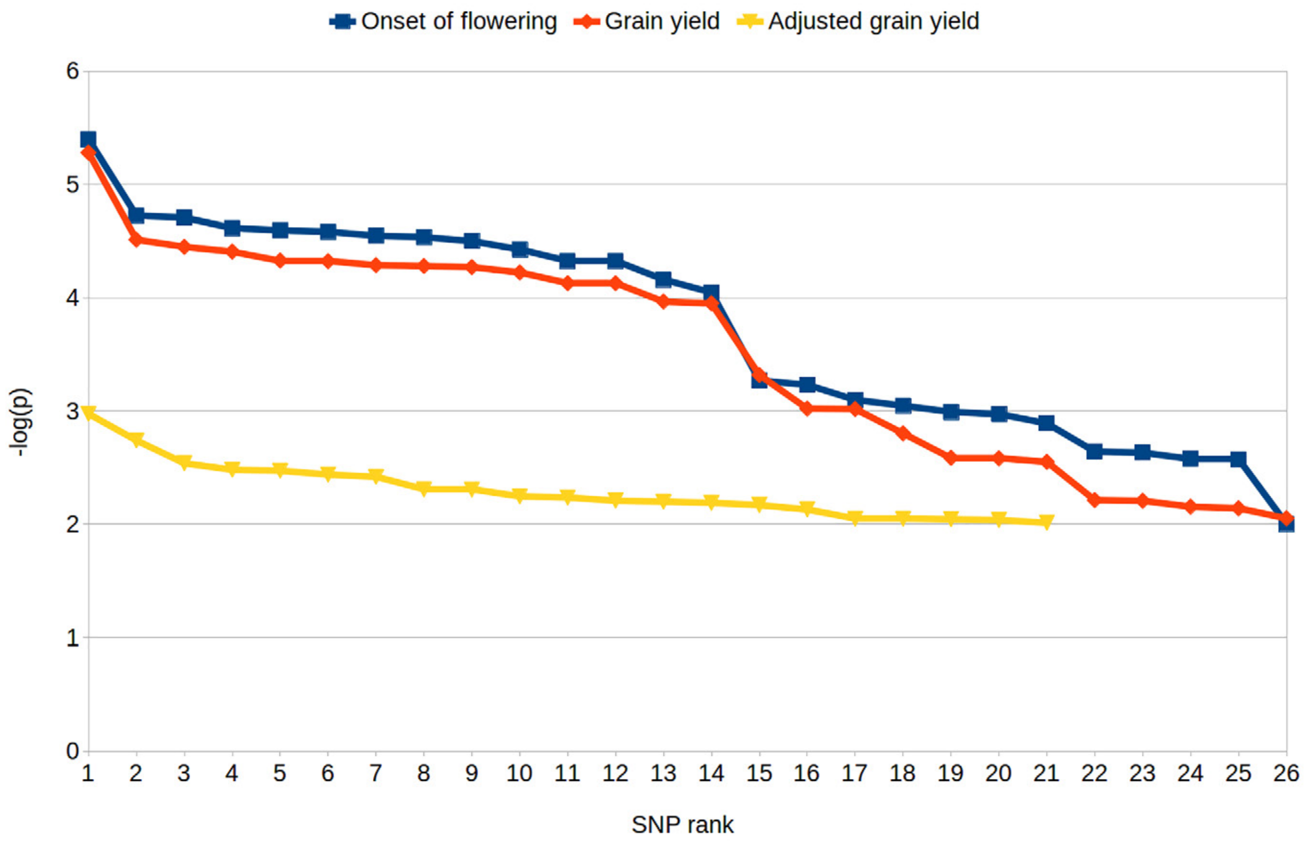

Figure 4. Genome-wide association results for markers associated with onset of flowering, grain yield and adjusted grain yield (as yield deviation from the yield value expected according to onset of flowering) in pooled data of three pea recombinant-inbred-line populations. SNP, single-nucleotide polymorphism.

(ii) the remaining 2 GBS markers associated with onset of flowering, and two other markers associated with grain yield, were not linked to any Illumina array marker but displayed high or moderate linkage with each other, suggesting colocalization with at least another QTL for onset of flowering and grain yield in an unknown genomic area; (iii) 1 GBS marker associated with grain yield stood on its own (no association with Illumina array or other relevant GBS markers); (iv) 15 GBS markers that associated with the adjusted grain yield displayed linkages with Illumina array markers, which, depending on the pattern and extent of linkage values, suggested the presence of at least one QTL on LG III, two QTLs on LG V (in genomic regions distant ∼ 50 cM), and two QTLs on LG VII (in genomic regions distant ∼ 35–40 cM); (v) 6 GBS markers that associated with the adjusted grain yield pointed to three QTLs in separate, unknown genomic regions (given the extent and pattern of linkage between each other, and the lack of linkages with Illumina array markers).

Kaspa was the main donor of alleles associated with late flowering and lower grain yield. Out of the 26 markers significant for onset of flowering in the joint-populations study, Kaspa was the parent donor for all 26, and Isard for 5. All parent cultivars provided some useful SNPs associated with higher adjusted grain yield (Supplemental Table S1).

DISCUSSION

We achieved high GS predictive ability ($r > 0.5$) for grain yield in the three RIL populations. Contributing reasons for this result include (i) accurate yield phenotyping and experimental design (as indicated by the high broad-sense heritability obtained for this trait relative to various field studies, e.g., Singh and Singh, 2006; Annicchiarico and Iannucci, 2008; and Georgieva et al., 2016); (ii) the high correlation of grain yield with onset of flowering in two populations, since the relatively simple genetic control of the latter trait could simplify genomic predictions for the former trait; (iii) the modest genetic variation and limited number of relevant QTLs, and the slow LD decay, that are expected for a biparental population relative to a germplasm collection. The selection of elite parent lines, in which many positive alleles may already be fixed, probably contributed to further decrease the diversity for useful alleles in the RIL populations. Our focus on RIL populations issued by elite parent lines was motivated by the crucial importance of this material in breeding of self-pollinated crops such as pea. However, Tayeh et al. (2015b) reported high GS prediction accuracy for two grain yield components—individual seed weight and number of seeds per plant—in a genetically broad germplasm collection of pea that included recent and old cultivars, landraces, and wild genotypes. They used a high number of Infinium Array markers but showed that relatively high prediction accuracy could be maintained when using a subset of 369 well-distributed markers. In an earlier study performed on that germplasm set using 331 markers, the best-performing GS models exhibited high prediction accuracy for individual seed weight and moderate accuracy for number of seeds per plant (Burstin et al., 2015). Overall, these studies and the current one indicate that GS prediction of pea grain yield can be a feasible objective.

Our study confirmed the importance of early flowering as a key stress escape mechanism for adaptation to environments with severe terminal drought. Its impact on line variation for grain yield was overwhelming for two populations that featured a later mean value and greater variation in the phenology trait: K×I and K×A. In this situation, the genotype adaptive responses for grain yield were ecologically and genetically simple, as confirmed by the extensive colocalization of SNP markers for grain yield and onset of flowering in the GWAS study. Indirect selection for grain yield based on early onset of flowering could provide an alternative option to GS-based selection for yield in this case. However, GS offers the advantages of somewhat greater speed and lower cost relative to the phenotyping assessment, as well as the potential for selecting also for other traits (including intrinsic drought tolerance, notwithstanding its modest impact on genotype grain yield in these two populations).

For the population A×I, which was somewhat earlier and less variable for onset of flowering than the other populations, our study also detected sizable variation for the adjusted yield, that is, yield unrelated to stress escape mechanisms. The possibility to select genomically for this trait is important, owing to the greater practical interest of intrinsic drought tolerance relative to drought-stress escape and the difficulty to select phenotypically for this trait. In many Mediterranean areas (e.g., southern Italy), the exploitation of early flowering is limited by greater susceptibility of early material to frost events, whose effect on de-hardened or insufficiently hardened plants may cause high plant mortality (Annicchiarico and Iannucci, 2007). Our current adoption of late-winter sowing purposely aimed to avoid the confounding effects of drought and cold stress and to concentrate only on the former. Incidentally, late-winter sowing exacerbated the effect of drought stress relative to autumn sowing, as shown by lower crop grain yields relative to those observed in autumn-sown crops under comparable spring rainfall amounts (Annicchiarico and Iannucci, 2008). The current GS model for predicting grain yield under severe terminal drought, and a GS model for predicting grain yield in autumn-sown, cold-prone environments that we are developing in northern Italy, could be applied jointly for selection by assigning relatively greater weight to predictions of the model whose target stress has relatively greater expected severity and frequency in the target region.

Bayesian Lasso and rrBLUP displayed higher predictive ability, whereas SVR-lin tended toward lower predictive ability, among the four tested GS models. Earlier studies on different species suggest that results of model comparisons vary depending on the data set. For example, the BL model outperformed rrBLUP in simulation studies by Ogutu et al. (2012), but it was outperformed by a model analogous to rrBLUP (GBLUP) in two studies on pea (Burstin et al., 2015; Tayeh et al., 2015b). SVR models, especially SVR-rbf, tended to outperform BL in Long et al.'s (2011) study and confirmed this result in an alfalfa study (Annicchiarico et al., 2015), whereas it tended to the opposite pattern here. However, we found modest differences in accuracy between better-predicting models (BL, rrBLUP, and SVR-rbf), which is in agreement with most previous studies.

Requiring a higher number of reads per locus implies a statistically lower number of heterozygotes erroneously called as homozygotes. Our results indicated, however, a lower impact on GS model predictive ability of this kind of error relative to the low number of polymorphic markers that are available when requiring high read depths. The differences in predictive ability between different thresholds for the number of reads per locus emerged only when applying relatively stringent thresholds for genotype missing data, which resulted in few markers for the higher minimum number of reads (particularly for the eight-read threshold). In general, approximately 400–500 polymorphic markers proved sufficient for achieving good GS predictions, probably because of the slow LD decay and the relatively narrow genetic variation that characterize biparental populations. These characteristics would also facilitate the correct estimation of genotype missing data based on polymorphism from nearby markers, thereby accounting for the high predictive ability displayed by GS models with as many as 50% missing data. Nevertheless, we preferred the six-read threshold to the four-read one for final GS analyses because of its greater expected reliability for genotyped material that is still characterized by some degree of heterozygosity.

The value of including population structure information in GS models may vary depending on the specific combination of populations and the target trait (Zhong et al., 2009; Janss et al., 2012; Guo et al., 2014). In this study, including structure information was not useful, possibly because of the narrow genetic base of our material. However, Burstin et al. (2015) found no increase in GS predictive ability when including structure information in the wider diversity panel represented by a pea germplasm collection.

A simple comparison of phenotypic selection with GS in terms of yield gain per cycle (considering same selection intensity) can be performed by comparing the square root of the estimated broad-sense heritability (which is proportional to phenotypic selection gain) to the estimated GS prediction accuracy as indicated by predictive ability values (which is proportional to GS gain) (Heffner et al., 2010). For grain yield in the current phenotyping platform, H^2 is 0.90, thus allowing a comparison of $H =$ 0.95 with the prediction accuracy of 0.72 (for best GS configuration, results averaged across populations: Table 3).

An advantage for GS would arise in terms of yield gains per unit time when envisaging two selection cycles per year for GS and one for phenotypic selection, which would double the value of the GS-based prediction accuracy. A slight GS advantage would appear also for the adjusted grain yield of the population A×I that displayed sizable variation for this trait, by comparing $H = 0.75$ with the prediction accuracy of 0.40 and then doubling the latter figure to account for double number of selection cycles per unit time. Such a comparison holds true even when considering that our estimates of GS prediction accuracy might be overestimated by the use of cross-validations on data of the same environment rather than different environments, because the broad-sense heritability estimate suffers from the same limitation, that is, its probable overestimation due to lack of account for genotype × environment interaction in its formula (whose extent would be estimated from experiments repeated in different environments). A comparison of phenotypic selection with GS for grain yield that takes account of their different selection costs could be performed according to the closest scenario among those reported in Rajsic et al.'s (2016) Table 2. For the current heritability ($H^2 = 0.90$), a ratio of phenotypic to GS estimated cost per genotype around 2 (€80 versus €40), and an effective number of chromosome segments below 25 (as suggested by GWAS results), GS would be more convenient economically than phenotypic selection. The same result would apply to the adjusted grain yield of the population A×I (considering $H^2 = 0.57$ and the same relative costs and effective number of chromosome segments). In addition, Rajsic et al.'s (2016) study for these traits under the current scenarios indicates that GS training sets per population as small as 96 RILs would be economically preferable for both traits. Indeed, our increase in the training genotype set obtained by combining the lines of the three populations resulted in modest accuracy gains.

Our comparison of interpopulation with intrapopulation prediction accuracy is of special interest for pea breeding programs in which selection is performed on lines derived from many different crosses (where each cross provides relatively few tested lines) rather than many lines from a few crosses. In this case, the prediction power of GS models devised for partly different material is bound to decrease because of partly different useful alleles, epistatic effects, etc. Our results suggest that the decrease in predictive ability when using a GS model trained in one RIL population for another population (with one common parent line) may vary sharply depending on the specific set of populations, probably because of the relatively few useful alleles that featured the three biparental populations and the possibly marked effect of chance on pairs of parents showing polymorphism for these alleles. RIL populations with quite good predictive ability for each other may occur, as in the case of those sharing Kaspa as a parent, but some kind of preliminary assessment of different GS models based on phenotyping data of the target material is necessary to identify these cases. The only moderate predictive ability observed between most pairs of RIL populations sharing a common parent suggests the improbability of achieving good interpopulation predictions between populations having no common parent.

The GWAS study on pooled data of the three RIL populations confirmed at the genetic level the close relationship between onset of flowering and grain yield under severe terminal drought that emerged phenotypically. Also, it suggested that the many significant linked markers observed for each trait actually related to a small number of genomic regions, of which one on LG II already emerged as a key QTL colocating for flowering time and grain yield in the study by Ferrari et al. (2016). Interestingly, such colocalization emerged in that study even if its grain yield phenotyping was performed under quite different conditions—specifically, spring-sowing without rainout shelter, which resulted in much lower terminal drought (as indicated by a mean yield level of 2.56 t ha^{-1}). GWAS results for individual populations agreed largely with those of the pooled populations and confirmed Kaspa as the main donor of alleles linked to lateness and, therefore, lower grain yield. There are many pea flowering genes, which relate either to photoperiodic response or to temperature pattern (Weller and Ortega, 2015). The important QTL on LG II is probably related to temperature pattern, since the marked flowering delay conferred by photoperiodic genes contrasts with the modest flowering delay exhibited by the donor parent for later phenology (Kaspa) compared with the other parent lines (Annicchiarico, 2005). The current availability of many GBS-generated markers linked to this key QTL offers greater opportunity of exploitation relative to the few Illumina Array-linked markers detected by Ferrari et al. (2016). The results from GWAS could contribute to define marker-assisted selection procedures for grain yield. However, GS is expected to be more efficient than marker-assisted selection for polygenic traits, owing to its ability to take also account of minor genetic effects (Bernardo and Yu, 2007). Another advantage of GS is its simple and convenient weighing of the relative importance of major genetic effects within its model.

The lower association scores detected for adjusted grain yield relative to grain yield and onset of flowering were consistent with the markedly lower importance of yield responses associated with intrinsic drought tolerance relative to those associated with stress escape. Variation for traits conferring intrinsic drought tolerance is probably modest in elite modern pea cultivars (which are usually selected under favorable conditions) compared with landrace germplasm. This hypothesis is confirmed by the fact that several landraces displayed both higher grain yield and later phenology than modern varieties in a recent evaluation of a global pea collection in a drought-stress environment of Italy (P. Annicchiarico, M. Romani, G. Cabassi, and B. Ferrari, unpublished data). Even in the presence of modest genetic variation, GWAS results suggested a relatively complex genetic control for yield responses accounted for by the adjusted yield (which spanned several genomic regions), which is in agreement with its several possible contributing traits. Mechanisms of intrinsic drought tolerance may involve dehydration avoidance or dehydration tolerance, the former implying maintenance of water uptake or reduction of water loss, and the latter an osmotic adjustment (Sánchez et al., 1998; Turner et al., 2001). The distance of at least 35 cM on the same linkage group between genomic regions including different putative QTLs for this trait that was suggested by linkage analyses is consistent with LD decay for biparental populations (e.g., Tommasini et al., 2007). The completion of the ongoing pea-genome sequencing effort (Tayeh et al., 2015a) will offer new opportunities for locating putative QTLs on the genome and the possible discovery of genes underlying most important QTLs.

In conclusion, our study supports the potential value of GBS markers for developing accurate and low-cost GS models in pea. Some of our results can help optimize some steps of a GS analysis pipeline for this species, particularly for the ordinary context of selection performed on advanced lines issued by biparental crosses. Future work of ours will aim to assess the value of best-performing GS models in terms of actual pea yield gain in environments characterized by severe terminal drought.

Supplemental Information Available

Supplemental Table S1. Sequence, association scores, source of polymorphism of markers associated with higher values for onset of flowering, grain yield, and adjusted grain yield and linkage disequilibrium of these markers with SNP array markers in the consensus map by Ferrari et al. (2016).

Acknowledgments

We are grateful to S. Proietti, A. Passerini, and P. Gaudenzi for technical assistance. The research was carried out by CREA within the FP7-ArimNet project Resilient Water- and Energy-Efficient Forage and Feed Crops for Mediterranean Agricultural Systems (REFORMA) funded by the Italian Ministry of Agricultural and Forestry Policies, with the contribution of former staff from the Samuel Roberts Noble Foundation.

References

Alessandri, A., M. De Felice, N. Zeng, A. Mariotti, Y. Pan, A. Cherchi, J.-Y. Lee, B. Wang, K.-J. Ha, P. Ruti, and V. Artale. 2014. Robust assessment of the expansion and retreat of Mediterranean climate in the 21st century. Sci. Rep. 4:7211. doi:10.1038/srep07211

Annicchiarico, P. 2005. Scelta varietale per pisello e favino rispetto all'ambiente e all'utilizzo. Informatore Agrario 61(49):47–52.

Annicchiarico, P. 2008. Adaptation of cool-season grain legume species across climatically contrasting environments of southern Europe. Agron. J. 100:1647–1654. doi:10.2134/agronj2008.0085

Annicchiarico, P., and A. Iannucci. 2007. Winter survival of pea, faba bean and white lupin cultivars across contrasting Italian locations and sowing times, and implications for selection. J. Agric. Sci. 145:611–622. doi:10.1017/S0021859607007289

Annicchiarico, P., and A. Iannucci. 2008. Adaptation strategy, germplasm type and adaptive traits for field pea improvement in Italy based on variety responses across climatically contrasting environments. Field Crops Res. 108:133–142. doi:10.1016/j.fcr.2008.04.004

Annicchiarico, P., and E. Piano. 2005. Use of artificial environments to reproduce and exploit genotype × location interaction for lucerne in northern Italy. Theor. Appl. Genet. 110:219–227. doi:10.1007/s00122-004-1811-9

Annicchiarico, P., N. Nazzicari, X. Li, Y. Wei, L. Pecetti, and E.C. Brummer. 2015. Accuracy of genomic selection for alfalfa biomass yield in different reference populations. BMC Genomics 16:1020. doi:10.1186/s12864-015-2212-y

Aulchenko, Y.S., S. Ripke, A. Isaacs, and C.M. Van Duijn. 2007. GenABEL: An R library for genome-wide association analysis. Bioinformatics 23:1294–1296. doi:10.1093/bioinformatics/btm108

Bänziger, M., P.S. Setimela, D. Hodson, and B. Vivek. 2006. Breeding for improved abiotic stress tolerance in maize adapted to southern Africa. Agric. Water Manage. 80:212–224. doi:10.1016/j.agwat.2005.07.014

Bernardo, R., and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47:1082–1090. doi:10.2135/cropsci2006.11.0690

Burstin, J., P. Salloignon, M. Chabert-Martinello, J.-B. Magnin-Robert, M. Siol, F. Jacquin, A. Chauveau, C. Pont, G. Aubert, C. Delaitre, C. Truntzer, and G. Duc. 2015. Genetic diversity and trait genomic prediction in a pea diversity panel. BMC Genomics 16:105. doi:10.1186/s12864-015-1266-1

Carrouée, B., K. Crépon, and C. Peyronnet. 2003. Les protéagineux: Intérêt dans les systèmes de production fourragers français et européens. Fourrages (Versailles) 174:163–182.

Casella, G., and E.I. George. 1992. Explaining the Gibbs Sampler. Am. Stat. 46:167–174. doi:10.2307/2685208

Ceccarelli, S. 1989. Wide adaptation. How wide? Euphytica 40:197–205. doi:10.1007/BF00024512

Ceccarelli, S., S. Grando, M. Maatougui, M. Michael, M. Slash, R. Haghparast, M. Rahmanian, A. Taheri, A. Al-Yassin, A. Benbelkacem, M. Labdi, H. Mimoun, and M. Nachit. 2010. Plant breeding and climate changes. J. Agric. Sci. 148:627–637. doi:10.1017/S0021859610000651

Charmet, G., E. Storlie, F.X. Oury, V. Laurent, D. Beghin, L. Chevarin, A. Lapierre, M.R. Perretant, B. Rolland, E. Heumez, L. Duchalais, E. Goudemand, J. Bordes, and O. Robert. 2014. Genome-wide prediction of three important traits in bread wheat. Mol. Breed. 34:1843–1852. doi:10.1007/s11032-014-0143-y

DeLacy, I.H., K.E. Basford, M. Cooper, I.K. Bull, and C.G. McLaren. 1996. Analysis of multi-environment trials– an historical perspective. In: M. Cooper and G.L. Hammer, editors, Plant adaptation and crop improvement. CABI, Wallingford, UK. p. 39–124.

Del Monte, G., L. Perini, and A. Brunetti. 1995. Indici agroclimatici. Quantità attese di precipitazione ed evaporazione potenziale. UCEA, Rome.

De los Campos, G., and P. Perez Rodriguez. 2014. BGLR: Bayesian generalized linear regression (version 1.0.3). http://cran.r-project.org/web/packages/BGLR/index.html (accessed 28 June 2016).

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379. doi:10.1371/journal.pone.0019379

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250–255. doi:10.3835/plantgenome2011.08.0024

Fang, Y., and L. Xiong. 2015. General mechanisms of drought response and their application in drought resistance improvement in plants. Cell. Mol. Life Sci. 72:673–689. doi:10.1007/s00018-014-1767-0

Ferrari, B., M. Romani, G. Aubert, K. Boucherot, J. Burstin, L. Pecetti, M. Huart-Naudet, A. Klein, and P. Annicchiarico. 2016. Association of SNP markers with agronomic and quality traits of field pea in Italy. Czech J. Genet. Plant Breed. 52:83–93. doi:10.17221/22/2016-CJGPB

Geman, S., and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 6:721–741. doi:10.1109/TPAMI.1984.4767596

Georgieva, N., I. Nilkolova, and V. Kosev. 2016. Evaluation of genetic divergence and heritability in pea (*Pisum sativum* L.). J. Biosci. Biotechnol. 5:61–67.

Guo, Z., D.M. Tucker, C.J. Basten, H. Gandhi, E. Ersoz, B. Guo, Z. Xu, D. Wang, and G. Gay. 2014. The impact of population structure on genomic prediction in stratified populations. Theor. Appl. Genet. 127:749–762. doi:10.1007/s00122-013-2255-x

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. 2009. The WEKA data mining software: An update. ACM SIGKDD Explor. 11:10–18. doi:10.1145/1656274.1656278

Heffner, E.L., J.L. Jannink, H. Iwata, E. Souza, and M.E. Sorrells. 2011. Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci. 51:2597–2606. doi:10.2135/cropsci2011.05.0253

Heffner, E.L., A.J. Lorenz, J.L. Jannink, and M.E. Sorrells. 2010. Plant breeding with genomic selection: Gain per unit time and cost. Crop Sci. 50:1681–1690. doi:10.2135/cropsci2009.11.0662

Janss, L., G. de Los Campos, N. Sheehan, and D. Sorensen. 2012. Inferences from genomic models in stratified populations. Genetics 192:693–704. doi:10.1534/genetics.112.141143

Jensen, E.S., and H. Hauggard-Nielsen. 2003. How can increased use of biological N_2 fixation in agriculture benefit the environment? Plant Soil 252:177–186. doi:10.1023/A:1024189029226

Kang, H.M., N.A. Zaitlen, C.M. Wade, A. Kirby, D. Heckerman, M.J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723. doi:10.1534/genetics.107.080101

Langmead, B., and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359. doi:10.1038/nmeth.1923

Long, N., D. Gianola, G.J. Rosa, and K.A. Weigel. 2011. Application of support vector regression to genome-assisted prediction of quantitative traits. Theor. Appl. Genet. 123:1065–1074. doi:10.1007/s00122-011-1648-y

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, K.P. Smith, M.E. Sorrells, and J.-L. Jannink. 2011. Genomic selection in plant breeding: Knowledge and prospects. Adv. Agron. 110:77–123. doi:10.1016/B978-0-12-385531-2.00002-5

Lu, F., A.E. Lipka, J. Glaubitz, R. Elshire, J.H. Cherney, M.D. Casler, E.S. Buckler, and D.E. Costich. 2013. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. PLoS Genet. 9:E1003215. doi:10.1371/journal.pgen.1003215

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Nazzicari, N., F. Biscarini, P. Cozzi, E.C. Brummer, P. Annicchiarico. 2016. Marker imputation efficiency for genotyping-by-sequencing data in rice (*Oryza sativa*) and alfalfa (*Medicago sativa*). Mol. Breed. 36:69. doi:10.1007/s11032-016-0490-y

Ogutu, J.O., T. Schulz-Streeck, and H.-P. Piepho. 2012. Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions. BMC Proc. 6 (Suppl 2):S10. doi:10.1186/1753-6561-6-S2-S10

Park, T., and G. Casella. 2008. The Bayesian lasso. J. Am. Stat. Assoc. 103:681–686. doi:10.1198/016214508000000337

Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909. doi:10.1038/ng1847

Rajsic, P., A. Weersink, A. Navabi, and K.P. Pauls. 2016. Economics of genomic selection: The role of prediction accuracy and relative genotyping costs. Euphytica 210:259–276. doi:10.1007/s10681-016-1716-0

Rogers, S.O., and A.J. Bendich. 1985. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. Plant Mol. Biol. 5:69–76. doi:10.1007/BF00020088

Sánchez, F.J., M. Manzanares, E.F. de Andres, J.L. Tenorio, and L. Ayerbe. 1998. Turgor maintenance, osmotic adjustment and soluble sugar and proline accumulation in 49 pea cultivars in response to water stress. Field Crops Res. 59:225–235. doi:10.1016/S0378-4290(98)00125-7

Schneider, A., and C. Huyghe. 2015. Les légumineuses pour des systèmes agricoles et alimentaires durables. Editions Quae, Versailles, France.

Schölkopf, B., and A.J. Smola. 2002. Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press, Boston.

Schwender, H. 2007. Statistical analysis of genotype and gene expression data. https://eldorado.tu-dortmund.de/handle/2003/23306 (accessed 28 June 2016).

Schwender, H., and A. Fritsch. 2013. Scrime: Analysis of high-dimensional categorical data such as SNP data (version 1.3.3). http://cran.r-project.org/web/packages/scrime/index.html (accessed 28 June 2016).

Searle, S.R., G. Casella, and C.E. McCulloch. 2009. Variance components. John Wiley & Sons, New York.

Singh, J.D., and I.P. Singh. 2006. Genetic variability, heritability, expected genetic advance and character association in field pea (*Pisum sativum* L.). Legume Res. 29:65–67. doi:10.5958/j.2229-4473.26.2.072

Tayeh, N., G. Aubert, M.L. Pilet-Nayel, I. Lejeune-Hénaut, T.D. Warkentin, and J. Burstin. 2015a. Genomic tools in pea breeding programs: Status and perspectives. Front. Plant Sci. 6:1037. doi:10.3389/fpls.2015.01037

Tayeh, N., A. Klein, M.-C. Le Paslier, F. Jacquin, H. Houtin, C. Rond, M. Chabert-Martinello, J.-B. Magnin-Robert, P. Marget, G. Aubert, and J. Burstin. 2015b. Genomic prediction in pea: Effect of marker density and training population size and composition on prediction accuracy. Front. Plant Sci. 6:941. doi:10.3389/fpls.2015.00941

Tommasini, L., T. Schnurbusch, D. Fossati, F. Mascher, and B. Keller. 2007. Association mapping of *Stagonospora nodorum* blotch resistance in modern European winter wheat varieties. Theor. Appl. Genet. 115:697–708. doi:10.1007/s00122-007-0601-6

Turner, N.C., G.C. Wright, and K.H.M. Siddique. 2001. Adaptation of grain legumes (pulses) to water-limited environments. Adv. Agron. 71:194–233. doi:10.1016/S0065-2113(01)71015-2

Warnes, G., G. Gorjanc, F. Leisch, and M. Man. 2013. Genetics: Population genetics. https://CRAN.R-project.org/package=genetics (accessed 27 June 2016).

Weller, J.L., and R. Ortega. 2015. Genetic control of flowering time in legumes. Front. Plant Sci. 6:207. doi:10.3389/fpls.2015.00207

Zhong, S., J.C.M. Dekkers, R.L. Fernando, and J.L. Jannink. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. Genetics 182:355–364. doi:10.1534/genetics.108.098277