



When less can be better: How can we make genomic selection more cost-effective and accurate in barley?

Amina Abed^{1,2} · Paulino Pérez-Rodríguez³ · José Crossa⁴ · François Belzile^{1,2}

Received: 13 November 2017 / Accepted: 24 May 2018 / Published online: 1 June 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

Key message We were able to obtain good prediction accuracy in genomic selection with ~2000 GBS-derived SNPs. SNPs in genic regions did not improve prediction accuracy compared to SNPs in intergenic regions.

Abstract Since genotyping can represent an important cost in genomic selection, it is important to minimize it without compromising the accuracy of predictions. The objectives of the present study were to explore how a decrease in the unit cost of genotyping impacted: (1) the number of single nucleotide polymorphism (SNP) markers; (2) the accuracy of the resulting genotypic data; (3) the extent of coverage on both physical and genetic maps; and (4) the prediction accuracy (PA) for six important traits in barley. Variations on the genotyping by sequencing protocol were used to generate 16 SNP sets ranging from ~500 to ~35,000 SNPs. The accuracy of SNP genotypes fluctuated between 95 and 99%. Marker distribution on the physical map was highly skewed toward the terminal regions, whereas a fairly uniform coverage of the genetic map was achieved with all but the smallest set of SNPs. We estimated the PA using three statistical models capturing (or not) the epistatic effect; the one modeling both additivity and epistasis was selected as the best model. The PA obtained with the different SNP sets was measured and found to remain stable, except with the smallest set, where a significant decrease was observed. Finally, we examined if the localization of SNP loci (genic vs. intergenic) affected the PA. No gain in PA was observed using SNPs located in genic regions. In summary, we found that there is considerable scope for decreasing the cost of genotyping in barley (to capture ~2000 SNPs) without loss of PA.

Communicated by Marcos Malosetti.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00122-018-3120-8>) contains supplementary material, which is available to authorized users.

✉ François Belzile
francois.belzile@fsaa.ulaval.ca

¹ Département de Phytologie, Université Laval, Quebec City, QC, Canada

² Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC, Canada

³ Programa de Estadística y Cómputo, Colegio de Postgraduados, CP 56230 Montecillos, Edo. de México, Mexico

⁴ Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 Mexico City, Mexico

Introduction

Three selection strategies are currently practiced in the field of plant breeding: phenotypic selection, marker-assisted and genomic selection (Ortiz Ríos 2015). The impetus for developing new breeding procedures has always been the desire to make breeding more efficient, quicker and less costly. During centuries, plant breeding was mainly achieved by crossing parents with the desired traits to generate genetic variation through recombination, and then selecting the best segregating offspring based on phenotypic evaluation via extensive field and/or greenhouse trials throughout generations, across locations, and over time (Walsh 2001; Cattivelli et al. 2011; Ortiz Ríos 2015). Phenotypic selection tends to achieve relatively lower genetic gain for complex traits with low heritability compared with traits with high heritability (Bhat et al. 2016; Rajsic et al. 2016). For complex traits, this process can be extremely time-consuming (Sallam and Smith 2016) and the logistics of implementing can be resource-intensive endeavor and very costly (Spindel et al. 2015).

Since the end of the 1980s, advances in molecular genetic research opened new avenues for obtaining genotypic information. From that point on, phenotypic breeding methods have been used along with novel technologies and tools such as molecular markers (Ortiz Ríos 2015). Selection for some traits, previously performed via phenotypic selection, could now be performed via marker-assisted selection (MAS) (Graner et al. 2011; Bhat et al. 2016). Marker-assisted selection, where molecular markers are used to tag genes of interest, has been very useful for manipulating genes with large effects and known association with a marker, offering new opportunities for a more efficient and faster breeding process (Steffenson and Smith 2006; Lorenz et al. 2011; Spindel et al. 2015). However, the efficiency of MAS has nonetheless been limited, as most traits of interest to breeders are controlled by many genes with small effects and/or by a combination of major and minor genes with epistatic interactions (Jiang 2013; Zhang et al. 2013; Zhao et al. 2014).

Along with advances in high-throughput sequencing technologies, the development of statistical methods linking genome-wide genetic information with phenotype has given rise to a new approach for improving quantitative traits: genomic selection (GS). Since GS was first proposed (Meuwissen et al. 2001), several genomic prediction models have been developed and applied in plant breeding for different traits (e.g., Bernardo and Yu 2007; Zhong et al. 2009; de los Campos et al. 2009b, 2010; Crossa et al. 2010, 2011, 2016; Burgueño et al. 2012; Heslot et al. 2012; Pérez-Rodríguez et al. 2012, 2017; Massman et al. 2013; Sousa et al. 2017). In a GS scheme, genome-wide markers such as single nucleotide polymorphisms (SNPs) are used to predict the breeding values of both parents and segregating offspring for traits of interest. These predicted values are derived from a statistical model of the relationship between genotypes and phenotypes in a training population (TP), applied to genotypes of selection candidates (SCs) (de los Campos et al. 2009b; Crossa et al. 2010). In principle, including all markers in the model regardless of the size of the associated effect allows the selection for small or large effect genes/quantitative trait loci (QTLs) in complex quantitative traits (Spindel et al. 2015).

The efficiency of GS is directly related to the accuracy of predictions, which itself depends strongly on gene effects, trait complexity, size of the TP, structure of populations and relatedness, as well as marker density but less on the method adopted to compute marker contributions (de los Campos et al. 2009b; Crossa et al. 2010, 2011; González-Camacho et al. 2012; Hickey et al. 2012; Lorenz et al. 2012; Pérez-Rodríguez et al. 2012, 2017; Riedelsheimer et al. 2012). Lorenz et al. (2012) discussed the impact of TP size and marker density and highlighted a complex non-linear relationship, dependent on QTL number and trait heritability. Prediction accuracy seems to be inversely related to the

number of QTLs and positively correlated to the heritability of traits (Zhong et al. 2009; Ornella et al. 2012). Moreover, the relative efficiency of statistical models often depends on the genetic architecture of a trait (Howard et al. 2014). For example, nonparametric methods such as reproducing kernel Hilbert spaces regression (RKHS) would be better suited for non-additive genetic architecture (i.e., dominance and epistasis) as it accounts for additive and epistatic effect and predict genetic value (the total performance of an individual) (Gianola 2006; Gianola and van Kaam 2008; de los Campos et al. 2009a, 2010; Pérez-Rodríguez et al. 2012). However, as suggested by several studies, parametric methods based on an additive framework predict breeding value (the expected performance of an individual's progeny), can be better than nonparametric methods in the case of additive genetic architectures, so the use of nonparametric models may not give the expected accuracy (Desta and Ortiz Ríos 2014; Howard et al. 2014).

Breeding for resistance in barley (*Hordeum vulgare* L.) is a very complex task, and the identification of desired recombinants by classical selection has almost reached the limits of manageability (Friedt 2011). More recently, MAS has been successfully employed to control resistance to stem rust and spot blotch, through the introduction of major resistance genes (Steffenson and Smith 2006). However, this is not the case with Fusarium head blight (FHB), caused primarily by *Fusarium graminearum*, one of the most destructive fungal diseases of barley (Paulitz and Steffenson 2011; Lorenz et al. 2012; Mamo and Steffenson 2015). The genetic basis of FHB resistance has been studied intensively through QTL mapping, indicating a complex quantitative trait with low heritability in elite germplasm, highly influenced by the environment. So far, no major resistance gene against FHB has been reported in barley, but minor-effect QTLs that act additively to confer partial resistance to FHB have been reported (Horsley et al. 2006; Liu et al. 2006; Steffenson and Smith 2006; Mamo and Steffenson 2015). As is the case for FHB, grain yield is also controlled by many genes (Wang et al. 2014). QTLs associated with yield have been found on almost all barley chromosomes, but the number of QTLs, their additive effects, and their localization on chromosomes differ from one population to another (Mikołajczak et al. 2016). The potential of GS to improve quantitative traits in barley, in addition to these two major traits, has been assessed in several studies (Zhong et al. 2009; Iwata and Janink 2011; Lorenz et al. 2012; Zhang et al. 2013; Lorenz and Smith 2015; Sallam et al. 2015; Nielsen et al. 2016; Sallam and Smith 2016; Schmidt et al. 2016).

Genomic selection offers great potential for increasing rates of genetic progress in plant breeding (Crossa et al. 2011; Hickey et al. 2012). It can improve breeding efficiency and be cost-effective by: (i) increasing the accuracy of estimated breeding values as molecular markers allow tracing

Mendelian segregation; and (ii) reducing the breeding cycle time by limiting field evaluation (Daetwyler et al. 2013; Sallam and Smith 2016; Pérez-Rodríguez et al. 2017). The costs of GS are associated with genotyping and phenotyping the TP and genotyping the SCs (Sallam and Smith 2016). A wide variety of SNP genotyping systems have been recently developed and the number of SNP loci can exceed millions (Lorenz et al. 2011; Gorjanc et al. 2017). In barley, SNP genotyping arrays such as Barley Oligonucleotide Pool Assays (BOPA1 and BOPA2) (Close et al. 2009) and 9 K barley chip (Comadran et al. 2012) have been employed in multiple applications including GS. Despite their success, these SNP marker platforms are being replaced by methods that exploit next-generation sequencing technologies (NGS). Methods based on reducing the complexity of the genome, such as genotyping-by-sequencing (GBS) (Elshire et al. 2011), provide very high-density genotyping at an extremely low cost per data point (Davey et al. 2011; Waugh et al. 2014; Gorjanc et al. 2017). Such marker resources will be highly valuable for dissecting the genetic architecture of complex agronomic traits and facilitating GS (Lorenz et al. 2011).

Genomic selection is based on linkage disequilibrium (LD) between QTLs and specific alleles of SNPs; a stronger LD leads to higher accuracy of prediction (Meuwissen, et al. 2001). Additionally, Wientjes et al. (2013) stated that the reliability of genomic predictions can be strongly influenced by family relationships. A relatively small number of markers would be sufficient to cover the genome if LD was extensive enough (Wientjes et al. 2013). High, long-range, genome-wide LD exists in cultivated barley populations (Kraakman 2004; Caldwell 2005; Rostoks et al. 2006; Zhong et al. 2009; Hamblin et al. 2010; Iwata and Jannink 2011; Lamara et al. 2013; Ramsay et al. 2014). However, the extent of LD varies across chromosomes (Rostoks et al. 2006); within the same elite/cultivated gene pool, LD may extend from hundreds of kilobases in genomic region with high rate of recombination to hundreds of megabases in rarely recombining regions (e.g., centromere-proximal regions) (Waugh et al. 2014). Thus, other than empirically, it is difficult to determine the number of markers needed to achieve genome-wide coverage.

Using populations of spring barley, the objectives of the present study were to explore, first, how different variations in the genotyping-by-sequencing protocol used to generate the genotypic data (multiplexing and genotype filtering) affected: (1) the number of SNP markers available; (2) the accuracy of the resulting genotypic data; (3) the extent of coverage of both the physical and genetic maps; and (4) the accuracy of GS predictions; second, how modeling epistasis in GS impacted the accuracy of prediction for six important traits in barley and third how the localization of SNP loci in different functional regions of the barley genome affected the prediction accuracy of the six traits.

Materials and methods

Barley populations

We used two populations: a TP and a validation population (VP). The TP was composed of 258 advanced lines and varieties chosen to represent the genetic diversity of the six-row barley breeding program at Université Laval (Quebec, Canada; <https://www.ulaval.ca>). The VP comprised 30 advanced lines and varieties representing the germplasm of a private breeding program (Céréla Inc.; <http://cerela.ca/>) located in the same broad geographic area. These populations are representative of the genetic diversity present within breeding programs active in eastern Canada.

Experimental phenotypic data

For the TP, phenotypic data were recovered from registration trials carried out in 14 different locations in Quebec from 2004 to 2014, for a total of between 10 and 25 environments (location \times year) depending on the trait, as displayed in Table 1. For the VP, we obtained data from the same registration trials until 2014 and performed additional phenotyping in 2015 and 2016 for a total of between 4 and 15 environments.

Fusarium head blight tolerance was evaluated by quantification of deoxynivalenol (DON) accumulated in the kernels. In inoculated FHB nurseries, barley lines were grown as two-row plots 0.65–1 m in length, spaced 17 cm apart, at a planting density of 375 plants m^{-2} in a randomized complete block design with two replications. Artificial inoculation was performed with maize (*Zea mays* L.) kernels infected by *Fusarium graminearum* mycelium (as per Prom et al. 1996). The inoculum consisted of a pool of four aggressive and virulent *F. graminearum* strains belonging to the same chemotype (3Ac-DON) and representing the molecular diversity present within a large collection of *F. graminearum* isolates from five agricultural locations in Quebec. Three to four weeks before anthesis, the inoculum was spread on the ground between the rows of each plot, approximately 45 g per row. Automatic irrigation with sprinklers was performed during non-rainy days for 5 h per day (a rotation of 5 mn per row during 5 h) until the maturity stage (Zadoks growth scale 83–87). The quantification of DON (in ppm) was performed using a commercial ELISA test (Veratox, Neogen Corporation, Saint Hyacinthe, QC, Canada) on 10- to 20-g samples of harvested kernels (Tangni et al. 2011).

Field trials for agronomic traits were conducted in a randomized complete block design with two replications

Table 1 Number of years, locations and environments from which phenotypic data were obtained for six traits for both the training and validation populations

	Traits	Years		Number of locations	Number of environments
		Number	Range		
Training population	DON	9	2004–2013	5	18
	HTM	10	2004–2013	1	10
	MAT	9	2006–2013	9	28
	TKW	8	2006–2013	5	19
	GYP	8	2006–2014	7	18
	PHT	9	2006–2014	8	25
Validation population	DON	8	2004–2015	5	15
	HTM	2	2015–2016	3	4
	MAT	3	2014–2016	7	12
	TKW	4	2013–2016	7	13
	GYP	3	2014–2016	7	12
	PHT	3	2014–2016	6	11

The six traits were deoxynivalenol (DON) content in kernels, heading time (HTM), days to maturity (MAT), thousand-kernel weight (TKW), grain yield (GYD) and plant height (PHT)

of four-row plots (4–5 m in length, 17-cm row spacing and a planting density of 375 plants m⁻²). Five different agronomic traits were evaluated: heading time (HTM), days to maturity (MAT), thousand-kernel weight (TKW), grain yield (GYD) and plant height (PHT). Heading time was expressed in days from seeding; it was scored when 50–80% of ears in the plot had emerged from the sheath. Maturity (days from seeding date) was reached when 50–80% of ears had kernels (in the central part of the spike) at the soft dough to early ripening stages (Zadoks growth scale 86–90). Thousand-kernel weight was measured in grams on a sample of 1000 seeds obtained using a seed counter. For each plot, grain yield was measured and converted to kg/ha. Plant height (in cm) was measured on two randomly selected plants (from the middle of the plot) as the distance from the ground to the top of the ear (without awns).

Phenotypic data analysis

Using the META-R program v. 6.01 (<http://hdl.handle.net/11529/10201>) (Alvarado et al. 2015), we performed an analysis of variance and estimated the broad-sense heritability H_e^2 in each environment. Broad-sense heritability was estimated as $H_e^2 = \sigma_G^2 / (\sigma_G^2 + \sigma_e^2/r)$, where σ_G^2 is the genetic variance, σ_e^2 is the error (residuals) variance, and r is the number of replicates. As all lines were not evaluated in all environments, best linear unbiased predictions (BLUPs) were computed using META-R following the model $y_{ijk} = \mu + Env_i + Rep_j(Env_i) + Line_k + Env_i Line_k + e_{ijk}$, where y_{ijk} is the observed phenotype, μ is the overall mean, Env_i is the random effect of the i th environment (a location-year combination), $Rep_j(Env_i)$ is the random effect of j th block nested within the i th environment, $Line_k$ is the random

effect of the k th line, $Env_i Line_k$ is the random effect of the interaction between the i th environment and k th line and e_{ijk} is the random error term. The broad-sense heritability H^2 on an entry mean basis was computed for each trait. The BLUP values were considered as the observed performances of each line across all environments and used in GS analysis.

Genotypic data

Genomic DNA was extracted from 5 mg of dried young leaves using a CTAB-based protocol. The DNA concentration (ng/ μ L) in each sample was measured using a fluorometric quantification method (PicoGreen). A total of 200 ng per sample was used for the preparation of 96-plex *PstI/MspI* GBS libraries as per the methods of Mascher et al. (2013); the optimized protocol is detailed in Abed et al. (2017). After amplification and purification, each of the three GBS libraries was sequenced on two PI chips on an Ion Torrent Proton sequencer at the Plateforme d'analyses g enomiques (IBIS, Universit e Laval). As controls to assess the quality and reproducibility of SNP calls, we included DNA from cv. Morex, the cultivar used to build the barley reference genome, and three lines were analyzed in duplicate on different plates.

SNP calling and procedures for varying the number of SNPs obtained

To estimate the impact of various depths of sequencing of the GBS libraries, we extracted three subsets of reads (from the original FASTQ file obtained following sequencing) containing 1/2, 1/4 and 1/8 of the reads to simulate 192-plex, 384-plex and 768-plex libraries, respectively. Informative SNPs were identified and called for each set of reads using

the Fast-GBS pipeline (Torkamaneh et al. 2017) and IBSC reference genome (Ensembl Plant, Barley genome v. 35). SNPs were called using reads ≥ 50 nucleotides in length and if supported by ≥ 2 reads. Additionally, SNP loci having more than 10% heterozygous genotypes were excluded. To assess the impact of missing data (N), for all four degrees of multiplexing described above (96-plex, 192-plex, 384-plex, 768-plex), we separately applied four different thresholds for N: $\leq 80\%$ (N80), $\leq 50\%$ (N50), $\leq 20\%$ (N20) and $\leq 15\%$ (N15). Finally, missing data were imputed using Beagle v. 4.1 (<https://faculty.washington.edu/browning/beagle/beagle.html>) (Browning and Browning 2007) and only SNPs with a minor allele frequency (MAF) $\geq 5\%$ were used. In total, 16 different SNP-calling conditions were thus used (4 levels of sequencing depth and 4 N thresholds). For each of the resulting SNP sets, we estimated both SNP accuracy and reproducibility. Accuracy of SNP calls was measured as the degree of concordance between the GBS-derived genotype for cv. Morex and the genotype at the same physical position in the reference genome; based on only imputed genotypes, we additionally computed the imputation accuracy. Reproducibility was the degree of concordance between the GBS-derived genotypes for three lines analyzed in duplicate.

SNP distribution on the physical and genetic maps and their functional impact

To assign a position to each SNP on the genetic map, we extracted the reads underlying each SNP in all 16 SNP sets using an R in-house script (<https://www.r-project.org/>) (R Core Team 2016) and determined their position on the IBSC map (2012) using Barleymap (<http://foresta.eead.csic.es/barleymap/>) (Cantalapiedra et al. 2015). PhenoGram (<http://visualization.ritchielab.psu.edu/phenograms/plot>) (Wolfe et al. 2013) was used to represent the physical distribution of the SNPs on chromosomes. Using the SNPeff program v. 4.3 (<http://snpeff.sourceforge.net/>) (Cingolani et al. 2012) and based on gene annotation information (GFT format, Barley genome v. 35), we analyzed the distribution of SNP loci according to their location in different functional regions of the genome. We sub-divided the whole SNP set into four major categories: (1) an intergenic region which corresponded to SNPs located within 5 kb upstream and within 5 kb downstream of an open reading frame (ORF), in addition to SNPs located in intergenic regions outside the upstream and downstream regions; (2) a genic region corresponding to SNPs present in exons, introns or untranslated regions (UTRs); (3) a coding region where SNPs resided only in exonic regions; and (4) a non-coding region corresponding to SNP in introns, 5' UTRs and 3' UTRs. For this classification of SNPs, only the largest SNP set obtained under the most permissive conditions (96-plex, N80) was used.

Methods for estimating the accuracy of predicted phenotypes

The prediction accuracy was measured as Pearson's correlation between the predicted and the observed performance (BLUPs) across environments. In a first assessment of prediction accuracy, we performed 80:20 cross-validations by dividing the TP into two groups: 80% of lines used to train the model and 20% used to validate the model. Cross-validation was repeated 50 times (by randomly selecting lines assigned to each subset) and the same subsets were used to analyze different GS models. The mean and standard errors across the 50 iterations were computed for each trait. In the second assessment, we performed an inter-population validation in which the VP, an independent set of lines not present within the TP, was used. All lines present in the TP were used to fit the statistical models and to predict the performance of lines belonging to the VP. The prediction accuracy was measured by computing Pearson's correlation between predicted performances and phenotypes observed in the field for these lines. As a final comparison, the correlation was computed between the observed and predicted phenotype based on a line's rank rather than its phenotypic value.

Genomic-enabled prediction models

The statistical models used in this study were fitted using the Bayesian framework. Three models were chosen: (1) GBLUP, an RKHS model used as a basic model including only additive SNP effects; (2) GBLUPe, an RKHS model with a variance–covariance matrix that captures epistasis; and (3) RKHSg, an RKHS model with a variance–covariance matrix based on a Gaussian kernel. The two latter models aim to capture both additive and non-additive epistatic effects. In the RKHS approach, the regression function is provided by the Reproducing Kernel (RK), which is an $n \times n$ matrix whose entries are functions of marker profiles of pairs of lines. The RK must be semi-positive definite (Pérez-Rodríguez and de los Campos 2014; Jiang and Reif 2015). Based on the work of Pérez-Rodríguez and de los Campos (2014), we implemented the three models using single- and multi-kernel methods with the BGLR statistical package. For computations, we used 100,000 iterations of Gibbs sampling and a burn-in of 10,000, and a thin of 10.

Model with linear kernel (GBLUP)

In GBLUP, we employed a single kernel method with a linear reproducing kernel (RK). The RK is a function that maps from pairs of points in input space (e.g., pairs of individuals or pairs of vectors of marker genotypes) (Pérez-Rodríguez and de los Campos 2014). We used the following model:

$$y = 1\mu + u_1 + \varepsilon \quad (1)$$

where y is the vector of phenotypic records (response variable), μ is the general mean and considered a fixed parameter, u_1, ε are random parameters with $\varepsilon \sim MN(0, \sigma_\varepsilon^2 I)$ is the error term and $u_1 \sim MN(0, \sigma_{u_1}^2 K)$ is a vector of random additive effect. We chose $K = G$, where G is a genomic relationship matrix among all lines. The G matrix was computed following the method described by Cuevas et al. (2016) using the equation:

$$G = ZZ' / p \quad (2)$$

where Z is a matrix of SNPs coded for additive effects and p is the number of markers centered and standardized. In this setting, the diagonal values of G are around one, so that $\sigma_{u_1}^2$ is defined in the same scale as σ_ε^2 . By choosing $K = G$, this model is equivalent to genomic BLUP (GBLUP) (Jiang and Reif 2015).

Model with linear kernel capturing epistasis: GBLUPe

The GBLUPe is a linear mixed model that can be written as follows:

$$y = 1\mu + u_1 + u_2 + \varepsilon \quad (3)$$

where y, μ, u_1 and ε are as in Eq. (1), $u_2 \sim MN(0, \sigma_{u_2}^2 H)$,

where $H = G \# G$, and $\#$ stands for the Haddamart product or cell-by-cell product. With the addition of this random term, the model is able to capture epistatic effects (Henderson 1985).

Model with Gaussian kernel capturing epistasis: RKHSg

The RKHSg model is based on a multi-kernel method. Following Pérez-Rodríguez and de los Campos (2014), we used Gaussian RK evaluated in the (average) squared Euclidean distance between lines. The RKHSg model can be written as:

$$y = 1\mu + u_1 + u_2 + u_3 + \varepsilon \quad (4)$$

The notations are the same as in Eq. (1) and the assumptions are $u_1 \sim MN(0, \sigma_{u_1}^2 K_1)$, $u_2 \sim MN(0, \sigma_{u_2}^2 K_2)$, $u_3 \sim MN(0, \sigma_{u_3}^2 K_3)$, where K_1, K_2, K_3 are $n \times n$ semi-positive definite kernel matrix based on Euclidean distance between lines and different values of the bandwidth parameter h . It is computed as follows:

$$K(x_i, x_{i'}) = \exp \left\{ -h \times \frac{\sum_{k=1}^p (x_{ik} - x_{i'k})^2}{p} \right\} \quad (5)$$

where $x_i, x_{i'}$ are pairs of vectors of markers genotypes. The choice of h can be performed by applying a cross-validation or a Bayesian approach definite (Pérez-Rodríguez and de los Campos 2014; Jiang and Reif 2015). de los Campos et al. (2010) proposed using a multi-kernel approach or kernel averaging (KA) by defining a sequence of kernels based on a set of values of h and fitting a linear mixed model given by different values of the bandwidth parameter $h_1 = 1/M$ for K_1 , $h_2 = \frac{1}{5M}$ for K_2 and $h_3 = \frac{5}{M}$ for K_3 , where M is the median squared Euclidean distance. To tackle the problem of selection of h , different approaches were used. We performed kernel based models where we do not need to specify the h (as base for comparison, $K=G$ which is our first model). These approaches deal with the problem of bandwidth selection.

From each model, we reported the variance component: additive genotypic, non-additive genotypic and residual variance. To choose the best statistical model, we performed cross-validation as described above using the largest SNP set obtained with the most permissive conditions of call (Condition 1: 96-plex, N80).

Population structure and relatedness between lines

Using the SNP set obtained in Condition 1 (96-plex, N80), the genetic structures of the TP and the VP were evaluated using principal component analysis (PCA) with the software TASSEL v. 5.2.31 (<http://www.maizegenetics.net/tassel>) (Bradbury et al. 2007). A matrix P of uncorrelated variables called principal components (PCs), capturing most of the variation present in the original data, was generated. The structure of the population is subsequently defined by creating a scatter plot of the first two PC vectors. Additionally, the same SNP set was used to assess the relatedness between the lines by computing a G matrix as in Eq. (2). The first and the second eigen vectors of the eigen decomposition of the G matrix were plotted.

Impact of the number of SNPs on prediction accuracy

We selected six SNP sets that covered the complete range of number of SNPs and that were approximately evenly spaced within this range. For each SNP set, predicted phenotypes were calculated using the GBLUPe model, as it was found to be the most accurate overall for the six traits under study. Validation of predictions was performed using both a cross-validation and an inter-population approach, as described above.

Impact of the functional effect of SNPs on prediction accuracy

Based on gene annotation information, we targeted functional regions to sample four SNP sets corresponding to: (1) an intergenic region; (2) a genic region; (3) a coding region; and (4) a non-coding region. We previously normalized the size of each SNP set to an equivalent number of SNPs in order to avoid any additional impact due to the varying number of SNPs. In each region, ten randomizations were performed to choose two sets of 500 and 2000 SNPs. In this part, we used a statistical model based directly on the SNP effect: the Bayesian ridge regression (BRR) model (Pérez-Rodríguez and de los Campos 2014). The BRR model can be written as follows:

$$y = 1\mu + Z\beta + \varepsilon \quad (6)$$

where y , μ , ε are the same as in Eq. (1), Z is the matrix of SNP markers as in Eq. (2) and $\beta|\sigma_\beta^2 \sim MN(0, I\sigma_\beta^2)$ is a vector of the random effect of the markers. Validation of predictions was performed using an inter-population approach (described above) for the six traits (DON, HTM, MAT, TKW, GYD, PHT).

Results

Impact of sequencing depth and SNP filtering on the number of SNPs and the accuracy of genotype calls

We first wanted to explore how the degree of multiplexing used in GBS library preparation and sequencing as well as the tolerance towards missing data would impact both the number of polymorphic SNP loci identified and the accuracy of the genotype calls. In a first phase focused on the impact of sequencing depth on the number of SNPs called in a training population composed of 258 barley lines, different subsets of raw reads were extracted to simulate four multiplexing levels: 96-plex (all reads), 192-plex (1/2 of all reads), 384-plex (1/4 of all reads) and 768-plex (1/8 of all reads). This procedure yielded a mean number of 1860, 931, 465 and 233 K reads per line, respectively. As shown in Table 2, using a uniform tolerance for missing data ($\leq 80\%$, “N80”), the number of informative SNP loci called using a subset of all reads was reduced by 28% (192-plex), 44% (384-plex) and 56% (768-plex) relative to the initial set of 96-plex (100%). In addition to reducing the number of SNP loci, increased multiplexing also reduced the mean depth of coverage per SNP and per line for each level of multiplexing; it decreased from 16 reads/SNP at 96-plex, to 11 reads at 192-plex, 7 reads at 384-plex and 4 reads at 768-plex. In

Table 2 Number, accuracy and reproducibility of SNP calls obtained at four depths of sequencing per line (multiplexing) and four levels of tolerance for missing data

Condition	SNP-calling conditions		Number of SNPs	Relative to Condition1 (%)	Mean missing data (%)	SNP quality	
	Multiplexing	Missing data (%)				Accuracy (%)	Reproducibility ^a (%)
1	96	≤ 80	35,121	100	39	95	98
2	192	≤ 80	25,354	72	36	95	99
3	384	≤ 80	19,552	56	35	96	99
4	768	≤ 80	15,348	44	40	95	99
5	96	≤ 50	21,543	61	19	97	99
6	192	≤ 50	17,180	49	18	98	99
7	384	≤ 50	14,362	41	21	98	99
8	768	≤ 50	10,994	31	29	97	99
9	96	≤ 20	12,582	36	6	99	100
10	192	≤ 20	10,398	30	8	99	100
11	384	≤ 20	7619	22	11	98	100
12	768	≤ 20	2106	6	15	98	100
13	96	≤ 15	11,002	31	5	99	100
14	192	≤ 15	8892	25	6	99	100
15	384	≤ 15	5591	16	9	99	100
16	768	≤ 15	490	1	11	98	100

^aMean reproducibility as assessed using three lines analyzed in duplicate

a second phase, we explored how a decreasing tolerance regarding the allowable proportion of missing data at an SNP locus would impact the number of informative SNP loci. At all four levels of sequencing depth described above, we used four different missing data thresholds for retaining SNP loci: $\leq 80\%$ (N80), $\leq 50\%$ (N50), $\leq 20\%$ (N20) and $\leq 15\%$ (N15). As expected, considering each threshold, the mean missing data were always higher for the highly multiplexed subset. Compared to the data obtained under the most permissive conditions (96-plex, N80), at N50, the number of retained SNPs was reduced by between 39 and 69%; at N20, it was reduced by 64–94%; and at N15, we obtained an SNP subset reduction of between 69 and 99%. The 16 SNP sets used in this study are displayed in Online Resource 1.

To evaluate the accuracy of SNP calls made under the various conditions described above, we compared the GBS-derived SNP data for cv. Morex with the barley reference genome produced by sequencing this same accession. For all multiplexing levels, the accuracy of SNP calls was greatest (98–99%) when tolerating fewer missing data (≤ 20 and $\leq 15\%$), although it remained high (95–96%) even for SNP datasets with higher missing data thresholds ($\leq 50\%$ and $\leq 80\%$). The accuracies of the SNP sets obtained using the same threshold for missing data were highly similar across the different multiplexing levels, differing by no more than 1%. As three lines were analyzed in duplicate (in different GBS libraries), we were also able to assess the technical reproducibility of SNP genotyping. Under all conditions studied, reproducibility proved very high (98–100%). Thus, using the two filtering criteria together (sequencing depth and missing data), we obtained SNP catalogs differing greatly in the number of markers (490–35,121) while maintaining a high level of accuracy and reproducibility. The accuracy of imputation was about 88% on average.

SNP distribution and functional impact

Next, we wanted to explore the extent of genome coverage and marker density on both the physical and genetic maps of barley. On the physical map, SNPs showed an uneven but consistent distribution for the 16 SNP conditions (Online Resource 2). The gene-rich distal portions of the chromosomes showed the highest marker density, whereas the gene-poor pericentromeric regions were more sparsely populated. Additionally, as illustrated in Table 3, SNP density on each chromosome was proportional to its length, except for 4H and 6H in Condition 16, where an important reduction in the number of SNPs was detected. To assess the SNP distribution on the genetic map, which is more relevant for genomic selection, we assigned each SNP a position on the

Table 3 Distribution of SNPs across the barley chromosomes in three highly contrasting sets of SNP markers (conditions)

Chromosomes	Length in Mb	Condition 1 SNP ^a	Condition 15 SNP	Condition 16 SNP
1H	464	3407	595	64
2H	628	5222	969	90
3H	564	6782	966	78
4H	544	2871	613	59
5H	561	5676	912	79
6H	539	5177	714	40
7H	602	5986	822	80
Total ^b		35,121	5591	490

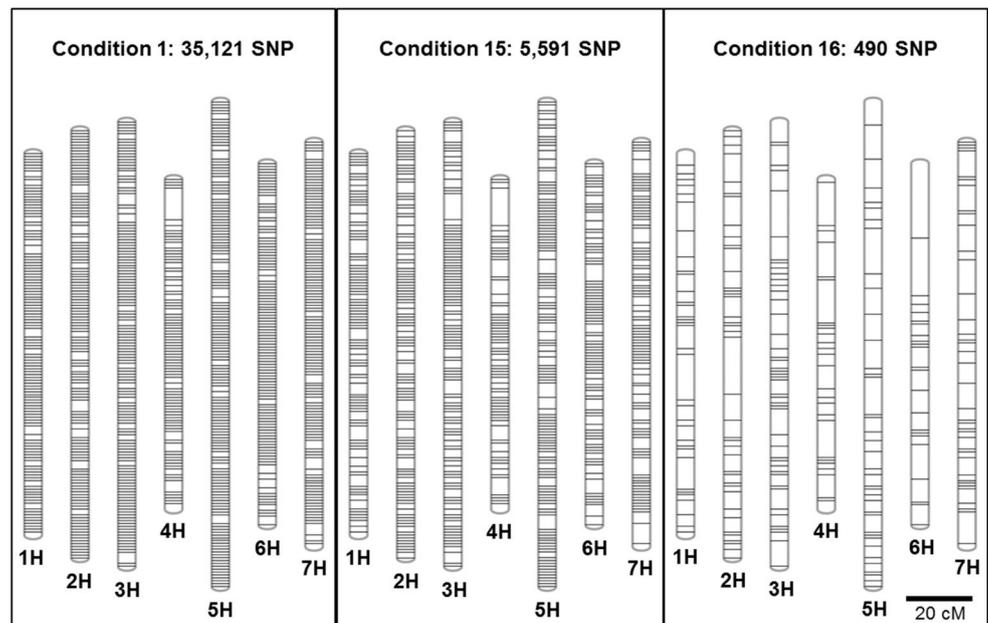
^aThe number of SNPs mapped on each chromosome based on the IBSC physical map (IBSC 2012)

^bThe total number of markers called in each condition

barley consensus map; the distribution of markers on the genetic map is shown in Fig. 1. The resulting genetic maps displayed a much more uniform SNP distribution along the chromosomes with few gaps. On average, the distance between neighboring markers was 0.04, 0.20 and 2.30 cM, respectively, for the SNPs obtained under conditions 1, 15 and 16 (35, 5.5 and 0.5 K SNPs, respectively). With more than 5000 SNP markers on a map (conditions 1 and 15), a single gap exceeding 10 cM was seen and, in both cases, it resided in the same region of chromosome 4H. As there was no lack of mapped reads in this segment of the genome (data not shown), it likely reflects the simple absence of polymorphism within this portion of chromosome 4H. As expected, the much smaller set of markers obtained under condition 16 (490 SNPs) resulted in a total of ~18 gaps exceeding 10 cM and 3 gaps larger than 20 cM. Again, the previously mentioned segment on 4H presented a gap. As we can see, the distribution of SNPs along each chromosome remained uniform but became clearly unbalanced (large gaps) and potentially problematic from ~500 SNP.

In addition to the distribution of SNPs on a genomic scale, we assessed their distribution in major functional regions according to their location based on gene annotation. Using the largest SNP set (35,121 SNP), 12.9% were located within the upstream region, 10.4% within the downstream region and 60.8% of these SNPs resided in intergenic regions outside the upstream and downstream regions; thus, a total of 84.1% of SNPs could be categorized as lying in the intergenic space. The remaining SNPs (15.9%) were distributed in exons (9.2%), introns (2.27%), 3' UTRs (2.16%), and 5' UTRs (2.30%). Of the SNPs located in the coding region (exons), these were sub-divided into synonymous (57.4%) and non-synonymous (42.6%).

Fig. 1 Distribution of SNP loci on the genetic map in three highly contrasted conditions. SNP loci were mapped on the IBSC consensus genetic map (IBSC 2012)



Phenotypic evaluation

Genomic prediction models were constructed based on extensive phenotypic characterization of a training population (TP) exploiting a wealth of historical data (2004–2014, a total of 14 different locations). Models were built for six traits: deoxynivalenol content in kernels (DON), heading time (HTM), days to maturity (MAT), thousand-kernel weight (TKW), grain yield (GYD) and plant height (PHT). For each environment, descriptive statistics are summarized in the table (Online Resource 3) for the six traits. For all traits and environments, we obtained moderate to high broad-sense heritability H_e^2 and differences between lines were significant (p value < 0.05 ; Online Resource 4). For each trait, BLUP values exhibited a normal distribution in the population (Online Resource 5). As expected, the extent of variability between BLUP values depended on the trait. Indeed, for DON and GYD the ratio of the standard deviation to the mean was, respectively, 0.23 and 0.13 compared with the other traits (HTM, MAT, TKW, PHT) for which this ratio ranged between 0.02 and 0.07. As was the case for the TP, the VP was phenotyped extensively for the same six traits (2004–2016; 11 locations) and descriptive statistics are summarized in Online Resource 6. The highest variability among lines was obtained for DON, with a ratio of the standard deviation to the mean of 0.29, while it ranged between 0.01 and 0.07 for the remaining traits. The phenotypic data used in this study are given in Online Resource 7.

Accuracy of different genomic selection models

To predict phenotypes for the six traits under study, we explored the accuracy of three models: (1) linear RKHS without epistasis (GBLUP); (2) linear RKHS with epistasis (GBLUPe); and (3) Gaussian RKHS (RKHSg). Using the largest SNP set ($> 35,000$ SNPs) to perform cross-validation (80:20), the prediction accuracies ranged, on average, from 0.44 for DON to 0.67 for TKW, and they were broadly correlated with the heritability of each trait (Fig. 2). In upper triangles, we observed slight but significant ($p < 0.05$) differences in the observed mean prediction accuracies for five of the six traits, with only PHT not showing differences in accuracy between the three models tested. Moreover, as displayed in the lower triangles, the RKHSg and GBLUPe models were strongly correlated and, in general, these models proved slightly superior in terms of accuracy.

To further refine the comparison between the models used, we analyzed the variance component of each statistical model. The models used make different assumptions and model different parts of the genetic variance (additivity, epistasis). As shown in the table (Online Resource 8), the estimated residual variance (σ_e^2) decreased from the GBLUP to RKHSg model for all six traits, whereas the magnitude of the estimated genetic variance (σ_G^2) changed according to the trait. Based on the overall performance of the three statistical models, GBLUPe was selected for subsequent analyses for its highest accuracy of prediction and shortest computing time.

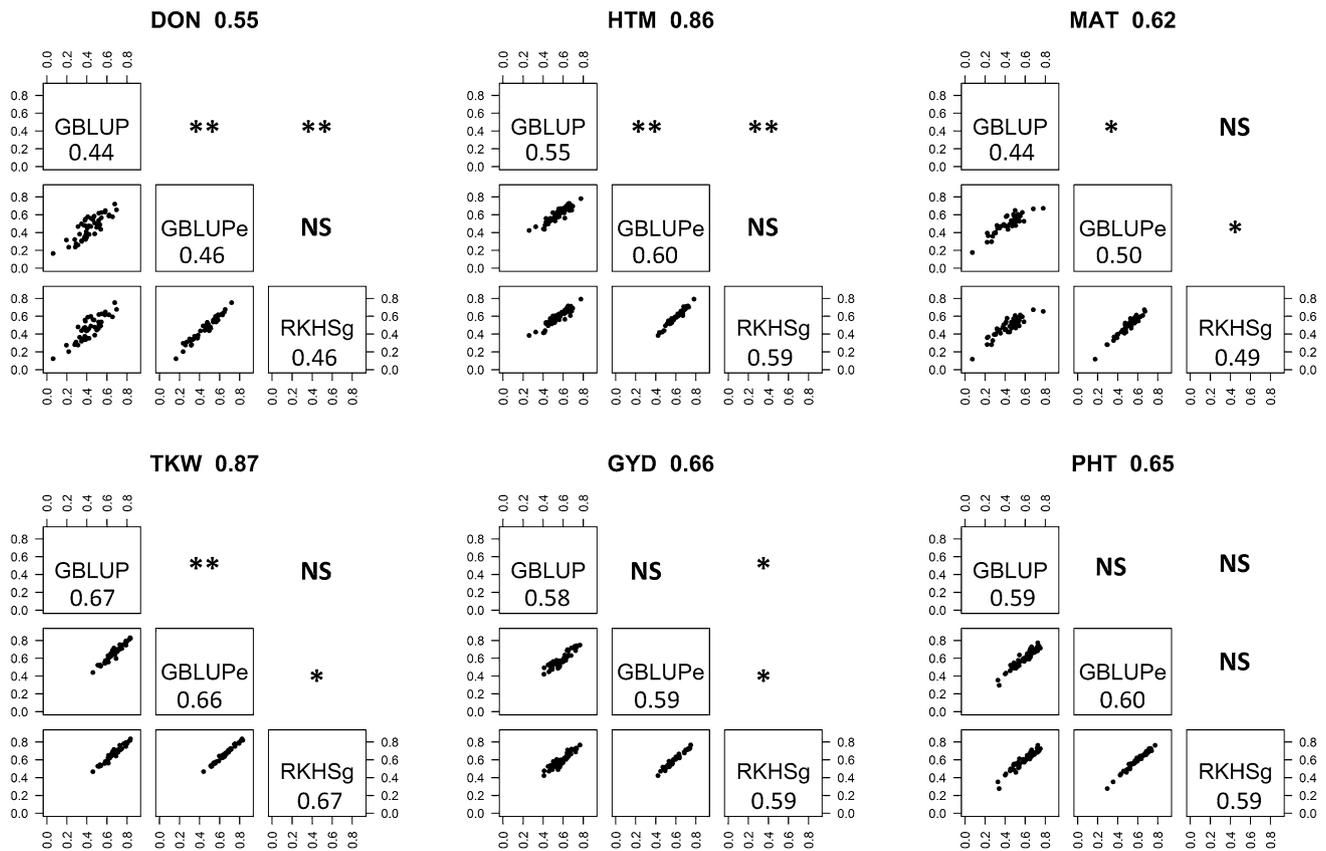


Fig. 2 Pairwise scatterplot matrix of the accuracy of three prediction models assessed through cross-validation (80:20). Three models (GBLUP, GBLUPe and RKHSg) were used to predict the phenotype of barley lines for six traits (DON, HTM, MAT, TKW, GYD, and PHT). The broad-sense heritability H^2 is displayed next to each trait. Accuracy was measured as Pearson correlations between predicted and observed performance in 50 validation subsets randomly chosen from the training population. The mean accuracy for each model

and trait is presented in the boxes on the diagonal with the name of the model that generated these accuracies. In the boxes below the diagonal, we show the pairwise comparison of accuracies obtained using two different models for each of the 50 subsets of the cross-validation. The results of Tukey's multiple comparison test between the various mean accuracies obtained for a trait are shown above the diagonal (** $P \leq 0.01$, * $P \leq 0.05$, NS not significant)

Impact of the number of SNPs on the accuracy of predictions

Having chosen a statistical model that provided the most accurate predictions using the largest set of SNPs, we wanted to investigate how the number of SNPs (and the resulting genome coverage) would affect prediction accuracy. For this purpose, we selected a sample of six SNP sets, covering the complete range from 490 to 35,121 markers obtained through the 16 SNP-calling conditions characterized above. We first employed a cross-validation approach to test the accuracies obtained with six SNP sets using the GBLUPe model. As illustrated in Fig. 3, the impact of the number of SNPs differed depending on the trait. The number of SNPs had no effect on the prediction accuracy for DON, MAT and GYD. For the remaining traits (HTM, TKW and PHT), when using the second smallest subset (6% of all SNPs, 2106 of

35,121 SNPs), the prediction accuracy decreased by 5–10% depending on the trait, whereas prediction accuracies were significantly affected when as many as 99% of the original SNP were removed from the prediction model. More detailed results by trait are presented in the table (Online Resource 9).

In a second experiment, we used an external validation population (VP), composed of a different set of barley lines from the same geographic area, to evaluate the quality of the predictions obtained using the six most contrasted SNP sets. First, we assessed the degree of genetic similarity between the TP and this VP. As displayed in Fig. 4, no evident separation between the two populations was detected based on PCA, the first principal component (PC1) explained ~13% of the total variation and the PC2 ~8%. This genetic proximity was also evident in the realized genomic relationship among TP and VP lines and its eigen decomposition (Online

Fig. 3 Prediction accuracies for different sizes of SNP sets assessed through cross-validation (80:20). The six SNP-calling conditions correspond to 35,121, 17,180, 10,994, 5591, 2106 and 490 SNP. The six SNP sets were used to predict the phenotypes of barley lines for six traits (DON, HTM, MAT, TKW, GYD, and PHT). Accuracy was measured as the mean of Pearson correlations between predicted and observed performance in 50 validation subsets randomly chosen from the training population

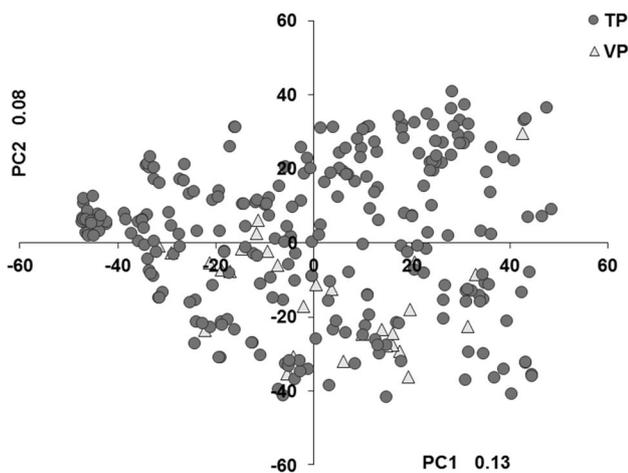
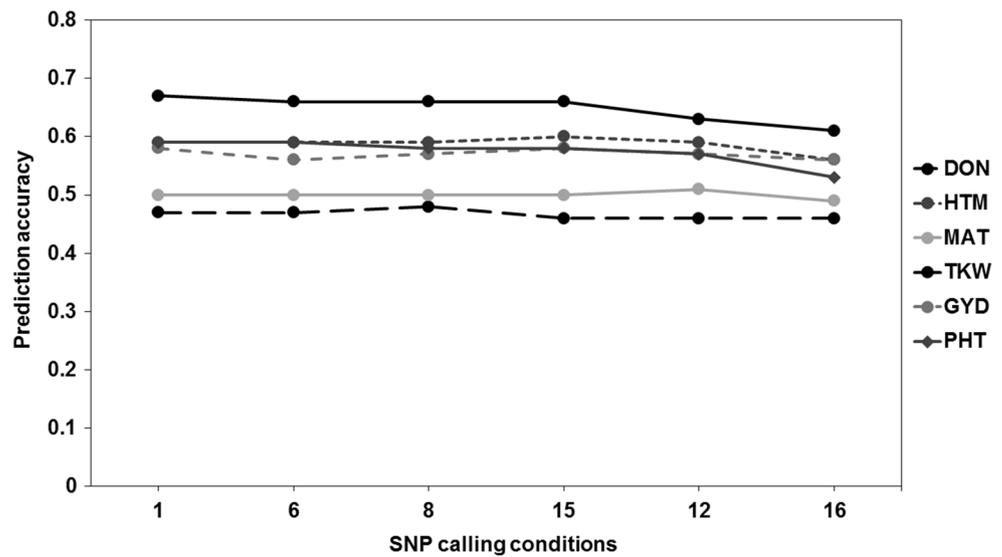


Fig. 4 Principal component analysis (PCA) of the training population (TP) lines (circles) and validation population (VP) lines (triangles)

of SNPs was not as strong as for MAT and PHT. Globally, using the first five conditions (35,121–2106 SNPs) the impact on the accuracy of prediction was not very important compared to 490 SNPs, at which point there seems to be a general decline in prediction accuracy. When comparing the two validation procedures, as displayed in the table (Online Resource 11), the pattern of accuracy was similar between the two procedures and the result found in cross-validation was confirmed by the second validation method; the latter being closer to a realistic genomic selection procedure. To provide a complementary view of the impact of the number of SNPs on prediction accuracy, we analyzed the ranks of lines instead of their performances. As shown in the figure (Online Resource 12), the results were in agreement with the results previously obtained using the phenotypic values.

Impact of the localization of SNPs on prediction accuracy

Resource 10). This result was not unexpected as several parental lines are in common between the two breeding programs and it suggested that using lines from the external program (VP) can provide a good basis for GS analysis without any reduction in prediction accuracy.

The observed prediction accuracies ranged from high to moderate, as shown in Fig. 5. TKW was the most accurately predicted trait, with an average accuracy of 0.73. HTM, MAT, GYD and PHT had intermediate accuracies of 0.43, 0.60, 0.53 and 0.47, respectively. With an average accuracy of 0.40, DON was the least accurately predicted trait. The impact of the number of SNPs depended on the trait, as had been seen earlier. Indeed, for DON, HTM, TKW and GYD, the impact of a reduction in the number

In the previous section, we randomly sampled SNPs to select different SNP subsets. Considering that an SNPs located in a specific region will potentially have a functional effect related on it; in this section, we wanted to test if SNPs located in genic regions could offer superior accuracy to an equivalent number of markers (500 or 2000) located in intergenic regions or in both. For sets of 2000 SNPs (Fig. 6), with the exception of DON, the accuracy of predictions produced using genic SNPs was actually lower or no different than that obtained using intergenic or a mixture of both types of markers for the studied traits. To further refine the analysis, SNPs located in genic regions were further categorized as being located in coding or non-coding portions of the gene. Prediction accuracies based on SNPs in coding regions were significantly higher in three cases (DON, MAT and TKW),

Fig. 5 Prediction accuracies with different sizes of SNP sets for six traits assessed through inter-population validation. The six SNP-calling conditions correspond to 35,121, 17,180, 10,994, 5591, 2106 and 490 SNP. The six SNP sets were used to predict the phenotype of barley lines for six traits (DON, HTM, MAT, TKW, GYD, and PHT). Accuracy was measured as the Pearson correlations between predicted and observed performance in the validation population

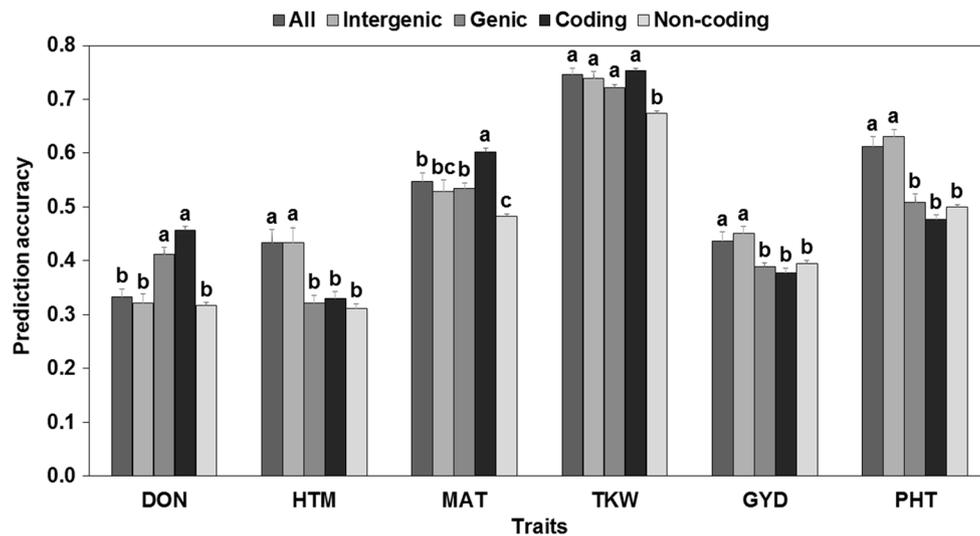
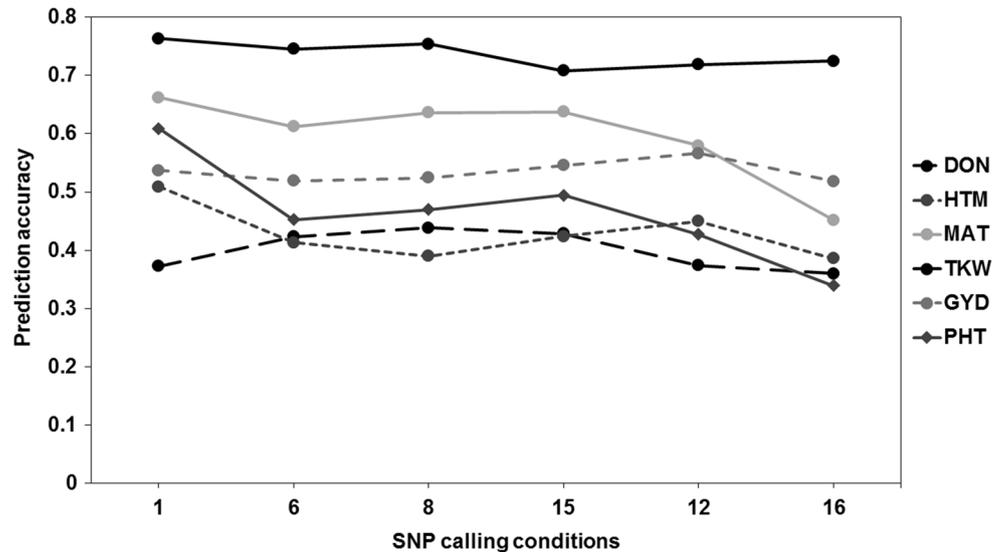


Fig. 6 Prediction accuracies with the different categories of SNP sets for six traits assessed through inter-population validation. The five SNP sets were normalized at 2000 SNPs, and the SNP sets were used to predict the phenotype of barley lines for six traits (DON, HTM, MAT, TKW, GYD, and PHT). Accuracy was measured as the mean

of Pearson correlations between predicted and observed performance in 10 sets of SNPs randomly sampled from the SNP set corresponding to each category. Means with the same letter are not significantly different according to Tukey's multiple comparison test at $\alpha=0.05$. Error bars represent the standard error of the mean

but in four of six traits the accuracies were not different or actually lower than those achieved with a non-selected set of markers (All) or only intergenic markers. The accuracy of predictions obtained with sets of 500 SNPs (Online Resource 13) showed a very similar profile. We were able to adequately predict the performance of lines even when the SNP set was exclusively located in the intergenic space, where the vast majority of GBS-derived SNPs are found in barley.

Discussion

Can we reduce the number of SNPs and genotyping costs?

Two criteria, the depth of sequencing and the tolerance towards missing data, were used jointly to alter the number of SNPs derived from GBS analysis of the TP. We first investigated the impact of four multiplexing levels and four filtering conditions on the number, quality and

genetic distribution of SNP loci. We obtained 16 SNP sets differing greatly in the number of markers from ~35 K to ~500 SNPs while maintaining a high level of SNP quality (accuracy and reproducibility). On average, higher levels of simulated multiplexing reduced the data per sample and increased the proportion of missing data, as reported in previous studies (Poland and Rife 2012; Huang et al. 2014). At all simulated levels of multiplexing, the quality of calls after imputation remained high (95–99%). The lowest quality of SNPs was obtained in two cases: (1) when tolerating up to 80% of missing data; and (2) with the lowest depth of coverage (768-plex). This finding has been reported by several authors, signaling that the quality of SNPs is influenced by the sequencing depth determined by the multiplexing level (Andolfatto et al. 2011; Huang et al. 2014). Clearly, we have to make a trade-off between the degree of multiplexing or depth of coverage, as it is important to identify stable and representative SNPs (He et al. 2014), the marker density, SNP quality and the cost of genotyping, and because it is possible to reduce the cost per sample by multiplexing many samples (e.g., 96, 384 or 768) (Huang et al. 2014; Gorjanc et al. 2015).

Considering the 16 SNP sets obtained in the different SNP-calling conditions, we investigated the distribution of SNP loci on both the physical and genetic maps of barley. On the physical map, SNPs showed an uneven but consistent distribution along the seven chromosomes. The resulting genetic maps displayed a much more uniform SNP distribution with few gaps. The much smaller set of markers (~500 SNP) resulted in more gaps exceeding 10 cM. These gaps are regions with low read coverage or that have an important proportion of missing data; SNPs in these regions were rapidly eliminated as SNP-calling conditions became more stringent or relied on fewer reads. As for the large gaps found under all conditions (e.g., chromosome 4H), similar results were reported by Muñoz-Amatriaín et al. (2011) in a consensus linkage map.

Trade-off between marker density, genotyping cost and prediction accuracy

Using an empirical approach, we analyzed the impact of a reduction in the number of SNPs on the accuracy of GS through two validation procedures (cross-validation and inter-population validation). The prediction accuracies ranged from high to moderate depending on the trait. We obtained comparable and stable predictions with all SNP sets comprising at least 2000–5000 SNPs. For most traits, the smallest subset of SNPs (~500) resulted in a significant decrease in prediction accuracy. This was not surprising, as genome coverage was the least extensive with this set, resulting in more numerous and larger gaps on the genetic map. In other words, 2000 GBS-derived SNPs seemed sufficient to

achieve similar GS prediction accuracies as much larger sets of markers. In barley, Lorenz et al. (2012) sub-sampled 1023 array-derived SNPs (BOPA1 and BOPA2) into subsets of 384 and 768 SNPs following three different marker-selection strategies. They found that reducing the marker set down to 384 SNPs had little effect on the prediction accuracy for FHB tolerance and DON accumulation. Although this might appear to be in contradiction with our own findings, it must be noted that the range of values examined (384–768) is much narrower than the one studied here and might not have been sufficient to detect sizeable differences in prediction accuracy. In wheat, Arruda et al. (2015) randomly sampled marker subsets of 500, 1500, 3000, or 4500 SNPs from an original set of 5054 GBS-derived SNPs. In that case, accuracies increased with the number of SNPs and reached a maximum only when using all markers. The latter result can be considered highly comparable to our findings if one takes into account that the wheat genome is three times larger than the barley genome. Adjusting for genome size, it is as if they had used 167, 500, 1000 and 1500 SNPs in barley. These studies in barley and wheat have come to a similar conclusion, i.e., that a reduction in SNP number had no dramatic impact on GS accuracy as long as it is maintained above a critical level. Additionally, in a recent study on rapeseed, Werner et al. (2018) demonstrated that low-density marker sets of a few hundred to a few thousand markers enable high prediction accuracies in breeding populations with strong LD, similar conclusions to the ones we reached in this work. Admittedly, however, the scope of inference of the current study is possibly limited due to the relatively small size of the training population (258 lines). We cannot exclude the possibility that larger training sets could require a larger number of markers to ensure high prediction accuracies.

Even if the global impact of the number of SNPs on prediction accuracy was similar for the six traits, the pattern of the impact of each SNP set was trait specific. For DON, HTM and GYD, we obtained a slight reduction in prediction accuracies with the number of SNPs, less evident than for MAT and PHT, whereas for TKW, the prediction accuracies were stable regardless of the size of the SNP set. Several studies using simulation or a theoretical basis investigated the impact of the number of SNPs in relation with the trait and found that adequate marker density depends on trait architecture (QTL number) and heritability (Lorenz et al. 2011). Concentration of DON is an indirect trait to evaluate the FHB resistance, a complex quantitative trait with low heritability in elite germplasm and highly influenced by the environment. No major resistance gene against FHB has been reported but minor-effect QTL that act additively to confer resistance (Horsley et al. 2006; Liu et al. 2006; Steffenson and Smith 2006; Paulitz and Steffenson 2011; Mamo and Steffenson 2015). Heading time is complex and usually assumed to involve numerous genetic factors that

interact with environmental conditions; it is determined by the interaction of three genetic factors, vernalization response, photoperiodic response and earliness in the narrow sense (Cattivelli et al. 2011; Griffiths 2003; Campoli et al. 2012; Nishida et al. 2013). Genetic studies highlighted an epistatic interaction between genes from those genetic factors in barley. (Casao et al. 2011). Numerous studies dissected the genetic of earliness per se in spring barley where vernalization is not required for flowering. Several major genes called *Ea* or *Eam* (early maturity) were identified in barley (von Bothmer and Komatsuda 2011; Campoli et al. 2012; Komatsuda 2014; Pankin et al. 2014). Grain yield is controlled by many QTLs widely distributed on the barley chromosomes. Additionally, the QTL effects were affected by the QTL \times environment interaction (Mikołajczak et al. 2016; Wang et al. 2016). Grain weight is under strong genetic control but considerably affected by the environment (Zanke et al. 2015). Thousand kernel weight is one of the major yield components as it affects directly the final yield (Pasam et al. 2012; Wang et al. 2016). Plant height is influenced by many qualitative genes and QTL. In barley, plant height is controlled by more than 30 dwarfing, semidwarfing, and other plant height genes (Wang et al. 2014; Ren et al. 2016). As investigated in these studies, DON, HTM, GYD and TKW seem to be under the control of various QTLs; thus, even if we reduced the number of SNPs, the prediction accuracy did not decrease dramatically because we preserved a sufficient number of SNPs in LD with QTLs, allowing the GS models to accurately predict the trait. In contrast, MAT and PHT are determined by few QTLs; hence, when the number of SNPs is reduced, prediction can potentially be greatly reduced if we happen to lose coverage in the vicinity of some of these QTLs. Even if the concept of GS is based on LD between QTLs and markers, the accuracy of GS can be strongly influenced by family structure (Habier et al. 2013; Wientjes et al. 2013). Indeed, Wientjes et al. (2013) demonstrated that the level of relationship between selection candidates and TP individuals can have a higher effect on the accuracy of GS than LD. Habier et al. (2013) suggested that modeling polygenic effects via pedigree relationships matrix jointly with SNP effects can prevent the decline of GS prediction. Based on PCA and eigen decomposition of the realized genomic relationship (*G* matrix) among the lines comprising the TP and VP, no evident structure between the two populations was detected in our study, suggesting that the decline of accuracy is mainly caused by a lack of SNPs in LD with causal QTLs.

It is important to note that the prediction accuracies reported here are derived either from internal cross-validation or using an external VP that shares much in common in terms of genetic similarity with the TP. Thus, the correlation values obtained in this context are necessarily higher than would be seen on true selection candidates, i.e., the progeny

of a biparental cross, as has been widely reported in the past (Sallam et al. 2015). Nonetheless, the trend seen regarding the impact of the number of SNPs on the accuracy of predictions performed on such true selection candidates would be expected to be highly comparable to what we report in this study.

The incorporation of economic aspects into the evaluation of selection strategies is essential for a profitable and efficient GS. Strictly on a theoretical basis, it is difficult to give a benchmark for the suitable number of markers needed for a genome-wide representation. Using our empirical approach, we found that prediction accuracies in GS remained high across a broad range of sizes of the SNP catalogue used. A decrease in map coverage and prediction accuracies was only observed with the fewest SNPs (<2000). Thus, the cost of genotyping in GS can be significantly reduced (by increasing the level of multiplexing used) without suffering a decrease in the accuracy of the genomic predictions. Costs for GS are associated with genotyping and phenotyping. In our experience, genotyping one line with GBS (96-plex) can cost 3.75 times less than the use of the 9 K array and 4.25 times less than 50 k array. As pedigree information is not available in this study, it was not possible to compare SNP-based models with the pedigree-based ones. Indeed, we must not lose sight of the fact that the use of pedigrees remains an effective and economically interesting alternative (Crossa et al. 2010) if precise and accurate pedigrees information is available.

Is there an advantage to capturing epistasis in GS models?

Comparing the overall performance of the three statistical models, we found that GBLUPe exhibited a slightly better performance for DON, HTM, MAT and GYD, but performed equally well as a model capturing only additive variance for TKW and PHT. In some previous studies, a superiority of epistatic models has been reported. For example, Jiang and Reif (2015) reported that taking into account epistasis improved the prediction accuracy for grain yield in selfing species such as wheat. In others, however, modeling epistasis did not improve accuracy significantly. For example, Sallam et al. (2015) found that for four traits in barley (DON concentration, FHB resistance, yield, and plant height), GS models capturing epistasis performed similarly as models capturing only additive variance. Therefore, our results are largely consistent with previous work in that a significant, yet small, increase in accuracy was achieved only for some of the studied traits when using models that capture epistasis. As some epistatic effects will be lost due to recombination, using methods that capture and model epistatic interaction in regions of low recombination (as proposed by Akdemir and Jannink 2015) can lead to a gain in prediction accuracy, especially for complex traits and in species like

barley known to present a low recombination rate in general (Ramsay et al. 2014).

Do SNPs in genic regions lead to more accurate predictions?

It is often hypothesized, either explicitly or implicitly, that SNP markers located within genes are superior to SNPs located outside of the genic space. Here, we tested if using subsets comprising the same number of markers but differing in their location (e.g., genic vs intergenic) could impact the accuracy of predicted phenotypes. We could find no evidence for such an advantage. On the contrary, intergenic SNPs resulted in equally or more accurate predictions in all but one case (DON content). Several studies of GS in barley (e.g., Lorenz et al. 2012; Lorenz and Smith 2015; Sallam et al. 2015; Sallam and Smith 2016) have used array-derived SNPs (BOPA1 and BOPA2), mostly located in genic regions, to predict performances for some important and complex traits like FHB tolerance, DON accumulation, yield and height. Those studies exhibited, on average, comparable accuracies of prediction to the results we obtained when using the SNP set located in genic regions. It can thus be argued that genotyping via a GBS approach, where the distribution of SNPs more closely reflects the distribution of overall nucleotide diversity (Torkamaneh et al. 2017), results in an SNP catalogue that is better suited to GS. It may be that SNPs located in the intergenic space are slightly better at capturing the underlying haplotype diversity compared to SNPs located in the genic space that is likely subject to a greater selection pressure. Variation within the intergenic sequences can produce phenotypic variation between individuals (Barrett et al. 2012). As the intergenic space is home to important regulatory sequences, such as promoters and enhancers, it is conceivable that SNPs in these regions better reflect the alleles that are present.

Author contribution statement AA and FB conceived and designed the study. JC and PP supervised the statistical analysis and reviewed the manuscript. AA performed field and lab experiments, performed the bioinformatics and statistical analyses, and interpreted the results. AA and FB wrote the manuscript.

Acknowledgements This study was funded jointly by the Natural Sciences and Engineering Research Council of Canada and by Céréla Inc. through a Collaborative Research and Development grant (RDCPJ 470998-14). The funders had no role in study design, data collection and analysis or preparation of the manuscript. We are grateful to Martin Lacroix, Suzanne Marchand and Eric Fournier (Université Laval), Annie Archambault and Samuel Ostiguy (Céréla) as well as the Biometrics and Statistics Unit (CIMMYT) for their assistance with various aspects of this work.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

- Abed A, Légaré G, Pomerleau S, St-Cyr J, Boyle B, Belzile F (2017) Genotyping-by-sequencing on the Ion Torrent platform in barley. In: Harwood W (ed) Barley: methods in molecular biology. Humana Press, New York
- Akdemir D, Jannink J-L (2015) Locally epistatic genomic relationship matrices for genomic association and prediction. *Genetics* 199(3):857–871. <https://doi.org/10.1534/genetics.114.173658>
- Alvarado G, López M, Vargas M, Pacheco A, Rodríguez F, Burgueño J, Crossa J (2015) META-R (Multi Environment Trial Analysis with R for Windows.) International Maize and Wheat Improvement Center. <http://hdl.handle.net/11529/10201>
- Andolfatto P, Davison D, Erezylmaz D, Hu TT, Mast J, Sunayama-Morita T, Stern DL (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res* 21(4):610–617. <https://doi.org/10.1101/gr.115402.110>
- Arruda MP, Brown PJ, Lipka AE, Krill AM, Thurber C, Kolb FL (2015) Genomic selection for predicting head blight resistance in a wheat breeding program. *Plant Genome*. <https://doi.org/10.3835/plantgenome2015.01.0003>
- Barrett LW, Fletcher S, Wilton SD (2012) Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell Mol Life Sci* 69(21):3613–3634. <https://doi.org/10.1007/s00018-012-0990-9>
- Bernardo R, Yu J (2007) Prospects for genome-wide selection for quantitative traits in Maize. *Crop Sci* 47(3):1082. <https://doi.org/10.2135/cropsci2006.11.0690>
- Bhat JA, Ali S, Salgotra RK, Mir ZA, Dutta S, Jadon V, Tyagi A et al (2016) Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Front Genet*. <https://doi.org/10.3389/fgene.2016.00221>
- Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* 23(19):2633–2635. <https://doi.org/10.1093/bioinformatics/btm308>
- Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81(5):1084–1097. <https://doi.org/10.1086/521987>
- Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Sci* 52(2):707. <https://doi.org/10.2135/cropsci2011.06.0299>
- Caldwell KS (2005) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. *Genetics* 172(1):557–567. <https://doi.org/10.1534/genetics.104.038489>
- Campoli C, Drosse B, Searle I, Coupland G, von Korff M (2012) Functional characterisation of HvCO1, the barley (*Hordeum vulgare*) flowering time ortholog of CONSTANS: functional characterisation of HvCO1 in Barley. *Plant J* 69(5):868–880. <https://doi.org/10.1111/j.1365-3113X.2011.04839.x>
- Cantalapiedra CP, Boudiar R, Casas AM, Igartua E, Contreras-Moreira B (2015) BARLEYMAP: physical and genetic mapping of nucleotide sequences and annotation of surrounding loci in barley. *Mol Breed*. <https://doi.org/10.1007/s11032-015-0253-1>

- Casao MC, Karsai I, Igartua E, Gracia MP, Veisz O, Casas AM (2011) Adaptation of Barley to mild winters a role for PPDH2. *BMC Plant Biol* 11:164. <https://doi.org/10.1186/1471-2229-11-164>
- Cattivelli L, Ceccarelli S, Romagosa I, Stanca M (2011) Abiotic stresses in barley: problems and solutions. In: Ullrich SE (ed) *Barley, production, improvement, and uses*. Wiley-Blackwell, Chichester, pp 282–306
- Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain W1118; Iso-2; Iso-3. *Fly* 6(2):80–92. <https://doi.org/10.4161/fly.19695>
- Close TJ, Bhat PR, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A et al (2009) Development and implementation of high-throughput SNP genotyping in Barley. *BMC Genom* 10(1):582. <https://doi.org/10.1186/1471-2164-10-582>
- Comadran J, Kilian B, Russell J, Ramsay L, Stein N, Ganai M, Shaw P et al (2012) Natural variation in a homolog of Antirrhinum CEN-TRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated Barley. *Nat Genet* 44(12):1388–1392. <https://doi.org/10.1038/ng.2447>
- Crossa J, de los Campos G, Pérez-Rodríguez P, Gianola D, Burgueno J, Araus JL, Makumbi D et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186(2):713–724. <https://doi.org/10.1534/genetics.110.118521>
- Crossa J, Pérez-Rodríguez P, de los Campos G, Mahuku G, Dreisigacker S, Magorokosho C (2011) Genomic selection and prediction in plant breeding. *J Crop Improv* 25(3):239–261. <https://doi.org/10.1080/15427528.2011.558767>
- Crossa J, de los Campos G, Maccaferri M, Tuberosa R, Burgueño J, Pérez-Rodríguez P (2016) Extending the marker × environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum Wheat. *Crop Sci* 56(5):2193. <https://doi.org/10.2135/cropsci2015.04.0260>
- Cuevas J, Crossa J, Soberanis V, Pérez-Elizalde S, Pérez-Rodríguez P, de los Campos G, Montesinos-López OA, Burgueño J (2016) Genomic prediction of genotype × environment interaction kernel regression models. *Plant Genome*. <https://doi.org/10.3835/plantgenome2016.03.0024>
- Daetwyler HD, Calus MPL, Pong-Wong R, Campos G, Hickey JM (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193(2):347–365. <https://doi.org/10.1534/genetics.112.147983>
- Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* 12(7):499–510. <https://doi.org/10.1038/nrg3012>
- de los Campos G, Gianola D, Rosa GJM (2009a) Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *J Anim Sci* 87(6):1883–1887. <https://doi.org/10.2527/jas.2008-1259>
- de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009b) Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182(1):375–385. <https://doi.org/10.1534/genetics.109.101501>
- de los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet Res* 92(4):295–308. <https://doi.org/10.1017/s0016672310000285>
- Desta ZA, Ortiz Ríos R (2014) Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci* 19(9):592–601
- 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(5):e19379. <https://doi.org/10.1371/journal.pone.0019379>
- Friedt W (2011) Barley breeding history, progress, objectives, and technology. In: Ullrich SE (ed) *Barley, production, improvement, and uses*. Wiley-Blackwell, Chichester, pp 160–220
- Gianola D (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173(3):1761–1776. <https://doi.org/10.1534/genetics.105.049510>
- Gianola D, van Kaam JBCHM (2008) Reproducing Kernel Hilbert Spaces Regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178(4):2289–2303. <https://doi.org/10.1534/genetics.107.084285>
- González-Camacho JM, de los Campos G, Pérez-Rodríguez P, Gianola D, Cairns JE, Mahuku G, Babu R, Crossa J (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor Appl Genet* 125(4):759–771. <https://doi.org/10.1007/s00122-012-1868-9>
- Gorjanc G, Cleveland MA, Houston RD, Hickey JM (2015) Potential of genotyping-by-sequencing for genomic selection in livestock populations. *Genet Sel Evol* 47(1):12. <https://doi.org/10.1186/s12711-015-0102-z>
- Gorjanc G, Dumasy J-F, Gonen S, Gaynor RC, Antolin R, Hickey JM (2017) Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in biparental segregating populations. *Crop Sci* 57(3):1404. <https://doi.org/10.2135/cropsci2016.08.0675>
- Graner A, Kilian A, Kleinhofs A (2011) Barley genome organization, mapping, and synteny. In: Ullrich SE (ed) *Barley, production, improvement, and uses*. Wiley-Blackwell, Chichester, pp 63–84
- Griffiths S (2003) The evolution of CONSTANS-like gene families in Barley, Rice, and Arabidopsis. *Plant Physiol* 131(4):1855–1867. <https://doi.org/10.1104/pp.102.016188>
- Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics* 194(3):597–607. <https://doi.org/10.1534/genetics.113.152207>
- Hamblin MT, Close TJ, Bhat PR, Chao S, Kling JG, Abraham KJ, Blake T et al (2010) Population structure and linkage disequilibrium in U.S. Barley germplasm: implications for association mapping. *Crop Sci* 50(2):556. <https://doi.org/10.2135/cropsci2009.04.0198>
- He J, Xiaoqing Z, Laroche A, Lu Z-X, Liu HK, Li Z (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front Plant Sci*. <https://doi.org/10.3389/fpls.2014.00484>
- Henderson CR (1985) Best Linear Unbiased Prediction of nonadditive genetic merits in noninbred populations. *J Anim Sci* 60(1):111–117
- Heslot N, Yang H-P, Sorrells ME, Jannink J-L (2012) Genomic selection in plant breeding: a comparison of models. *Crop Sci* 52(1):146. <https://doi.org/10.2135/cropsci2011.06.0297>
- Hickey JM, Crossa J, Babu R, de los Campos G (2012) Factors affecting the accuracy of genotype imputation in populations from several Maize breeding programs. *Crop Sci* 52(2):654. <https://doi.org/10.2135/cropsci2011.07.0358>
- Horsley RD, Schmierer D, Maier C, Kudrna D, Urrea CA, Steffenson BJ, Schwarz PB et al (2006) Identification of QTLs associated with fusarium head blight resistance in Barley accession CIho 4196. *Crop Sci* 46(1):145. <https://doi.org/10.2135/cropsci2005.0247>
- Howard R, Carriquiry AL, Beavis WD (2014) Parametric and non-parametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3-Genes Genom Genet* 4(6):1027–1046. <https://doi.org/10.1534/g3.114.010298>
- Huang Y-F, Poland JA, Wight CP, Jackson EW, Tinker NA (2014) Using genotyping-by-sequencing (GBS) for genomic

- discovery in cultivated Oat. *PLoS ONE* 9(7):e102448. <https://doi.org/10.1371/journal.pone.0102448>
- Iwata H, Jannink J-L (2011) Accuracy of genomic selection prediction in barley breeding programs: a simulation study based on the real single nucleotide polymorphism data of Barley breeding lines. *Crop Sci* 51(5):1915. <https://doi.org/10.2135/cropsci2010.12.0732>
- Jiang G-L (2013) Molecular markers and marker-assisted breeding in plants. In: Andersen SB (ed) *Plant breeding from laboratories to fields*. InTech, Rijeka, pp 45–83
- Jiang Y, Reif JC (2015) Modeling epistasis in genomic selection. *Genetics* 201(2):759–768. <https://doi.org/10.1534/genetics.115.177907>
- Komatsuda T (2014) Domestication. In: Kümlehn J, Stein N (eds) *Biotechnological approaches to Barley improvement*. Vol. 69. *Biotechnology in Agriculture and Forestry*. Springer, Berlin, pp 37–54
- Kraakman ATW (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring Barley cultivars. *Genetics* 168(1):435–446. <https://doi.org/10.1534/genetics.104.026831>
- Lamara M, Zhang LY, Marchand S, Tinker NA, Belzile F, Golding B (2013) Comparative analysis of genetic diversity in Canadian Barley assessed by SSR, DarT, and Pedigree data. *Genome* 56(6):351–358. <https://doi.org/10.1139/gen-2013-0048>
- Liu S, Zhang X, Pumphrey MO, Stack RW, Gill BS, Anderson JA (2006) Complex microcolinearity among Wheat, Rice, and Barley revealed by fine mapping of the genomic region harboring a major QTL for resistance to fusarium head blight in Wheat. *Funct Integr Genomics* 6(2):83–89. <https://doi.org/10.1007/s10142-005-0007-y>
- Lorenz AJ, Smith KP (2015) Adding genetically distant individuals to training populations reduces genomic prediction accuracy in Barley. *Crop Sci* 55(6):2657. <https://doi.org/10.2135/cropsci2014.12.0827>
- Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink J-L (2011) Genomic selection in plant breeding. In: Elsevier Inc. (ed) *Advances in Agronomy*, vol 110, pp 77–123
- Lorenz AJ, Smith KP, Jannink J-L (2012) Potential and optimization of genomic selection for fusarium head blight resistance in six-row Barley. *Crop Sci* 52(4):1609. <https://doi.org/10.2135/cropsci2011.09.0503>
- Mamo BE, Steffenson BJ (2015) Genome-wide association mapping of fusarium head blight resistance and agromorphological traits in Barley landraces from Ethiopia and Eritrea. *Crop Sci* 55(4):1494. <https://doi.org/10.2135/cropsci2014.06.0428>
- Mascher M, Wu S, St. Amand P, Stein N, Poland J (2013) Application of genotyping-by-sequencing on semiconductor sequencing platforms: a comparison of genetic and reference-based marker ordering in Barley. *PLoS ONE* 8(10):e76925. <https://doi.org/10.1371/journal.pone.0076925>
- Massman JM, Gordillo A, Lorenzana RE, Bernardo R (2013) Genome-wide predictions from Maize Single-cross data. *Theor Appl Genet* 126(1):13–22. <https://doi.org/10.1007/s00122-012-1955-y>
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4):1819–1829
- Mikołajczak K, Ogrodowicz P, Gudyś K, Krystkowiak K, Sawikowska A, Frohberg W, Górny A et al (2016) Quantitative trait loci for yield and yield-related traits in spring Barley populations derived from crosses between European and Syrian cultivars. *PLoS ONE* 11(5):e0155938. <https://doi.org/10.1371/journal.pone.0155938>
- Muñoz-Amatriaín M, Moscou MJ, Bhat PR, Svensson JT, Bartoš J, Suchánková P, Šimková H et al (2011) An Improved consensus linkage map of Barley based on flow-sorted chromosomes and single nucleotide polymorphism markers. *Plant Genome* 4(3):238. <https://doi.org/10.3835/plantgenome2011.08.0023>
- Nielsen NH, Jahoor A, Jensen JD, Orabi J, Cericola F, Edriss V, Jensen J (2016) Genomic prediction of seed quality traits using advanced Barley breeding lines. *PLoS ONE* 11(10):e0164494. <https://doi.org/10.1371/journal.pone.0164494>
- Nishida H, Ishihara D, Ishii M, Kaneko T, Kawahigashi H, Akashi Y, Saisho D et al (2013) Phytochrome C is a key factor controlling long-day flowering in Barley. *Plant Physiol* 163(2):804–814. <https://doi.org/10.1104/pp.113.222570>
- Ornella L, Singh S, Pérez-Rodríguez P, Burgueño J, Singh R, Tapia E, Bhavani S et al (2012) Genomic prediction of genetic values for resistance to wheat rusts. *Plant Genome* 5(3):136. <https://doi.org/10.3835/plantgenome2012.07.0017>
- Ortiz Ríos R (2015) *Plant breeding in the omics era*. Springer, Cham. <https://doi.org/10.1007/978-3-319-20532-8>
- Pankin A, Campoli C, Dong X, Kilian B, Sharma R, Himmelbach A, Saini R et al (2014) Mapping-by-sequencing identifies HvPHYTOCHROME C as a candidate gene for the Early Maturity 5 locus modulating the circadian clock and photoperiodic flowering in Barley. *Genetics* 198(1):383–396. <https://doi.org/10.1534/genetics.114.165613>
- Pasam RK, Sharma R, Malosetti M, van Eeuwijk FA, Haseneyer G, Kilian B, Graner A (2012) Genome-wide association studies for agronomical traits in a worldwide spring Barley collection. *BMC Plant Biol* 12(1):16
- Paulitz TC, Steffenson BJ (2011) Biotic stress in Barley: disease problems and solutions. In: Ullrich SE (ed) *Barley, production, improvement, and uses*. Wiley-Blackwell, Chichester, pp 307–354
- Pérez-Rodríguez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198(2):483–495. <https://doi.org/10.1534/genetics.114.164442>
- Pérez-Rodríguez P, Gianola D, González-Camacho JM, Crossa J, Manès Y, Dreisigacker S (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in Wheat. *G3-Genes Genom Genet* 2(12):1595–1605. <https://doi.org/10.1534/g3.112.003665>
- Pérez-Rodríguez P, Crossa J, Rutkoski J, Poland JA, Singh R, Legarra A, Autrique E, de los Campos G, Burgueño J, Dreisigacker S (2017) Single-step genomic and pedigree genotype × environment interaction models for predicting Wheat lines in international environments. *Plant Genome*. <https://doi.org/10.3835/plantgenome2016.09.0089>
- Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics *Plant. Genome* 5(3):92. <https://doi.org/10.3835/plantgenome2012.05.0005>
- Prom LK, Steffenson BJ, Salas B, Fetch TG Jr, Casper HH (1996) Evaluation of selected barley accessions for resistance to Fusarium head blight and deoxynivalenol concentration. In: Slinkard A, Scoles G, Rossnagel B (eds) *Proceeding of the 5th international Oat conference and the 7th International Barley Genetics Symposium*. University Extension Press, University of Saskatchewan, Saskatoon, Canada, pp 764–766
- R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>
- Rajsic P, Weersink A, Navabi A, Pauls KP (2016) Economics of genomic selection: the role of prediction accuracy and relative genotyping costs. *Euphytica* 210(2):259–276. <https://doi.org/10.1007/s10681-016-1716-0>
- Ramsay L, Colas I, Waugh R (2014) Modulation of meiotic recombination. In: Kümlehn J, Stein N (eds) *Biotechnological approaches to Barley improvement*. Vol. 69. *Biotechnology in Agriculture and Forestry*. Springer, Berlin, pp 311–329

- Ren X, Wang J, Liu L, Sun G, Li C, Luo H, Sun D (2016) SNP-based high density genetic map and mapping of *btwd1* dwarfing gene in Barley. *Sci Rep*. <https://doi.org/10.1038/srep31741>
- Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisek J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid Maize. *Nat Genet* 44(2):217–220. <https://doi.org/10.1038/ng.1033>
- Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT et al (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. *Proc Natl Acad Sci* 103(49):18656–18661
- Sallam AH, Smith KP (2016) Genomic selection performs similarly to phenotypic selection in Barley. *Crop Sci* 56(6):2871. <https://doi.org/10.2135/cropsci2015.09.0557>
- Sallam AH, Endelman JB, Jannink J-L, Smith KP (2015) Assessing genomic selection prediction accuracy in a dynamic Barley breeding population. *Plant Genome*. <https://doi.org/10.3835/plantgenome2014.05.0020>
- Schmidt M, Kollers S, Maasberg-Prelle A, Grober J, Schinkel B, Tomearius A, Graner A, Korzun V (2016) Prediction of malting quality traits in Barley based on genome-wide marker data to assess the potential of genomic selection. *Theor Appl Genet* 129(2):203–213. <https://doi.org/10.1007/s00122-015-2639-1>
- Sousa MB, Cuevas J, de Oliveira Couto EJ, Pérez-Rodríguez P, Jarquín D, Fritsche-Neto R, Burgueño J, Crossa J (2017) Genomic-enabled prediction in maize using kernel models with genotype × environment interaction. *G3-Genes Genom Genet*. <https://doi.org/10.1534/g3.117.042341>
- Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E, Atlin G, Jannink J-L, McCouch SR (2015) Genomic selection and association mapping in Rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical Rice breeding lines. *PLoS Genet* 11(2):e1004982. <https://doi.org/10.1371/journal.pgen.1004982>
- Steffenson BJ, Smith KP (2006) Breeding Barley for multiple disease resistance in the upper midwest region of the USA. *Czech J Genet Plant Breed* 42(3):79
- Tangni EK, Motte J-C, Callebaut A, Chandelier A, De Schrijver M, Pussemier L (2011) Deoxynivalenol loads in matched pair wheat samples in Belgium: comparison of ELISA VERATOX Kit against Liquid Chromatography. *Mycotoxin Res* 27(2):105–113. <https://doi.org/10.1007/s12550-010-0081-4>
- Torkamaneh D, Laroche J, Bastien M, Abed A, Belzile F (2017) FastGBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. *BMC Bioinform*. <https://doi.org/10.1186/s12859-016-1431-9>
- von Bothmer R, Komatsuda T (2011) Barley origin and related species. In: Ullrich SE (ed) *Barley, production, improvement, and uses*. Wiley-Blackwell, Chichester, pp 14–62
- Walsh B (2001) Quantitative genetics in the age of genomics. *Theor Popul Biol* 59:175–184
- Wang J, Yang J, Jia Q, Zhu J, Shang Y, Hua W, Zhou M (2014) A new QTL for plant height in Barley (*Hordeum vulgare* L.) showing no negative effects on grain yield. *PLoS ONE* 9(2):e90144. <https://doi.org/10.1371/journal.pone.0090144>
- Wang J, Sun G, Ren X, Li C, Liu L, Wang Q, Du B, Sun D (2016) QTL underlying some agronomic traits in Barley detected by SNP markers. *BMC Genet*. <https://doi.org/10.1186/s12863-016-0409-y>
- Waugh R, Thomas B, Flavell A, Ramsay L, Comadran J, Russell J (2014) Genome-wide association scans (GWAS). In: Kumlehn J, Stein N (eds) *Biotechnological approaches to Barley improvement*. Vol. 69. *Biotechnology in Agriculture and Forestry*. Springer, Berlin, pp 345–365
- Werner CR, Voss-Fels KP, Miller CN, Qian W, Hua W, Guan C-Y, Snowdon RJ, Qian L (2018) Effective genomic selection in a narrow-genepool crop with low-density markers: Asian rapeseed as an example. *Plant Genome*. <https://doi.org/10.3835/plantgenome2017.09.0084>
- Wientjes YCJ, Veerkamp RF, Calus MPL (2013) The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics* 193(2):621–631. <https://doi.org/10.1534/genetics.112.146290>
- Wolfe D, Dudek S, Ritchie MD, Pendergrass SA (2013) Visualizing genomic information across chromosomes with PhenoGram. *BioData Min* 6(1):18
- Zanke CD, Ling J, Plieske J, Kollers S, Ebmeyer E, Korzun V, Argillier O et al (2015) Analysis of main effect qtl for thousand grain weight in european winter Wheat (*Triticum Aestivum* L.) by genome-wide association mapping. *Front Plant Sci*. <https://doi.org/10.3389/fpls.2015.00644>
- Zhang G, Li C, Liu X (2013) *Advance in Barley sciences: Proceedings of 11th International Barley Genetics Symposium*. Springer Dordrecht Heidelberg New York London and Zhejiang University Press. <https://doi.org/10.1007/978-94-007-4682-4>
- Zhao Y, Mette MF, Gowda M, Longin CFH, Reif JC (2014) Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid Wheat. *Heredity* 112(6):638
- Zhong S, Dekkers JCM, Fernando RL, Jannink J-L (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a Barley case study. *Genetics* 182(1):355–364. <https://doi.org/10.1534/genetics.108.098277>