Effective Genomic Selection in a Narrow-Genepool Crop with Low-Density Markers: Asian Rapeseed as an Example

Christian R. Werner, Kai P. Voss-Fels, Charlotte N. Miller, Wei Qian, Wei Hua, Chun-Yun Guan, Rod J. Snowdon, and Lunwen Qian*

ABSTRACT
Genomic selection (GS) has revolutionized breeding for quantitative traits in plants, offering potential to optimize resource allocation in breeding programs and increase genetic gain per unit of time. Modern high-density single nucleotide polymorphism (SNP) arrays comprising up to several hundred thousand markers provide a user-friendly technology to characterize the genetic constitution of whole populations and for implementing GS in breeding programs. However, GS does not build upon detailed genotype profiling facilitated by maximum marker density. With extensive genome-wide linkage disequilibrium (LD) being a common characteristic of breeding pools, fewer representative markers from available high-density genotyping platforms could be sufficient to capture the association between a genomic region and a phenotypic trait. To examine the effects of reduced marker density on genomic prediction accuracy, we collected data on three traits across 2 yr in a panel of 203 homozygous Chinese semiwinter rapeseed (Brassica napus L.) inbred lines, broadly encompassing allelic variability in the Asian B. napus genepool. We investigated two approaches to selecting subsets of markers: a trait-dependent strategy based on genome-wide association study (GWAS) significance thresholds and a trait-independent method to detect representative tag SNPs. Prediction accuracies were evaluated using cross-validation with ridge-regression best linear unbiased predictions (rrBLUP). With semiwinter rapeseed as a model species, we demonstrate that low-density marker sets comprising a few hundred to a few thousand markers enable high prediction accuracies in breeding populations with strong LD comparable to those achieved with high-density genotyping. Our results are valuable for facilitating routine application of cost-efficient GS in breeding programs.

Core Ideas
- Efficiency in genomic selection is not particularly based on detailed genotype profiling facilitated by maximum marker density.
- Extensive genome-wide linkage disequilibrium is a common characteristic of breeding pools in many crop species.
- Every quantitative trait locus across the genome can be captured by one or a few representative markers.
- Fewer representative markers selected in respect of linkage disequilibrium (LD) can capture the association between a genomic region and a phenotypic trait.
- Low-density marker sets enable genomic prediction accuracies in breeding populations with strong LD comparable to those achieved with high-density genotyping.

Genomic selection has fundamentally revolutionized breeding for quantitative traits in plants and animals. In field crops, the efficacy of GS has been demonstrated in several important species, including

C.R. Werner, K.P. Voss-Fels and R.J. Snowdon, Dep. of Plant Breeding, Justus Liebig Univ., 35392 Giessen, Germany; C.N. Miller, John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK; W. Qian, College of Agronomy and Biotechnology, Southwest Univ., Chongqing 400716, P.R. China; W. Hua, C.-Y. Guan, L. Qian, Collaborative Innovation Center of Grain and Oil Crops in South China, Hunan Agricultural Univ., Changsha 410128, P.R. China. Received 30 Sep. 2017. Accepted 21 Feb. 2018.
*Corresponding author (qianlunwen@163.com).

Abbreviations: BLUP, best linear unbiased prediction; GS, genomic selection; GSL, glucosinolate content in the seed; GWAS, genome-wide association study; H², broad-sense heritability; LD, linkage disequilibrium; MAS, marker-assisted selection; QTL, quantitative trait locus or loci; PHT, plant height; SNP, single-nucleotide polymorphism; TP, training population.
maize (Zea mays L.), wheat (Triticum aestivum L.), barley (Hordeum vulgare L.), rice (Oryza sativa L.) and rapeseed (e.g., Battenfield et al., 2016; Crossa et al., 2014; Gorjanc et al., 2016; Jan et al., 2016; Spindel et al., 2015; Sun et al., 2017; Thorwarth et al., 2017; Werner et al., 2017; Zhao et al., 2012, 2014), thus offering great potential to optimize breeding programs. Many traits of agronomic importance are controlled by multiple genes of small effect, and their phenotypic expression is strongly affected by nongenetic factors as well as genotype × environment interaction. The complex interdependency of these determinants often results in low to moderate trait heritability, impeding reliable estimation of genetic values based on phenotype only, especially during early breeding cycles.

Marker-assisted selection (MAS) strategies applying classical quantitative trait loci (QTL) mapping approaches exploit only a limited number of markers associated with major-effect loci and have proven inadequate in unraveling the genetic background of highly polygenic traits (Ingvarsson and Street, 2011). Consequently, only a limited proportion of the genetic variance can be explained by the identified markers (Lande and Thompson, 1990; Meuwissen et al., 2001) and QTL effects are often inflated.

To overcome these constraints, Meuwissen et al. (2001) proposed a method that allowed simultaneous use of extensive whole-genome marker data, drastically exceeding the number of phenotypic observations (a large number regression coefficients and a small number of observations or genotypes). This strategy, known as GS, removes the potential bias caused by setting an arbitrary significance threshold to separate a few markers with an allegedly major impact from the remainder, which are assigned zero effect. The general procedure of GS is relatively straightforward. A training population (TP) that is representative of the breeding population under investigation is used to generate a statistical model. The allelic effects of all marker loci are calculated by regressing the phenotypic observations on their respective genotypes. Subsequently, this model can be used to predict the phenotype of untested individuals, based only on their genotypic profile. By summing up the allelic effects, the genomic estimated breeding value of an individual is calculated. In breeding programs, this estimate can be used, for example, for preselection of promising genotypes, resulting in more efficient organization of field trials and an increased selection gain per unit of time (Heffner et al., 2009, 2010; Marulanda et al., 2016).

High-density genotyping arrays provide a user-friendly device for characterizing the genetic constitution of individuals and unraveling allelic variation in whole populations. Recent arrays can be loaded with more than a million polymorphic sites (Haraksingh et al., 2017), and SNP chips comprising a few thousand up to several hundred thousand SNPs have become a routinely used tool in plant, animal, and human genetics. Given their extremely simple application and high reproducibility, genotyping arrays provide a plug-and-play technology for the implementation of GS in plant breeding programs. However, in contrast to application areas like genome-wide association studies (GWAS) and genetic mapping, or structural analyses of genomic regions, which build particularly on detailed genotype profiling, GS does not necessarily require maximum marker density. Instead, every QTL across the genome should be captured by one or a few representative markers (Habier et al., 2009; Lorenz et al., 2011; Solberg et al., 2008), meaning that the LD structure in the chromosome regions containing those QTL is the critical feature that needs to be considered. Although genotyping costs have dropped massively within the last few years, screening whole breeding populations with thousands of selection candidates still constitutes a significant economical factor. Evaluation of extensive breeding germplasm requires deliberate allocation of resources, and the integration of genomic selection in a well-established breeding flow might be associated with financial risk. However, with high-density SNP arrays already being available for many crop species, service providers distributing genotyping technology can offer low-density chips, derived from the same set of markers, for a much lower price. Given a simple, straightforward strategy for targeted identification and assembly of pool-specific marker subsets, breeders could easily design their own customized SNP arrays as a cost-efficient alternative to current high-density panels.

Rapeseed is an allotetraploid crop species (AACC, 2n = 38) that originated from a very limited number of spontaneous inter-specific hybridization events between Brassica rapa L. (AA, 2n = 20) and Brassica oleracea L. (CC, 2n = 18), not more than a few thousand years ago (Chalhoub et al., 2014). Except for synthetic B. napus generated artificially from new crosses between the two diploid progenitor species, only cultivated forms are known and intense selection for key agronomic and quality traits has led to a severe genetic bottleneck in this species (Becker et al., 1995). Breeding for geographically adapted varieties has resulted in three distinct major gene pools, which differ primarily with regard to their vernalization requirements and flowering characteristics (Schiessl et al., 2014). The Asian semiwinter type gene pool constitutes an intermediate between European winter-type oilseed rape and spring-sown canola. With a production capacity of more than 14 million tons per year, semiwinter rapeseed is one of China’s most important oilseed crops. Extensive highly conserved chromosome regions have been detected throughout the genome in all major gene pools of B. napus (Qian et al., 2014; Schiessl et al., 2015; Voss-Fels and Snowdon, 2016), reflecting the very recent formation of this species coupled with strong adaptive and agronomic selection during its intense cultivation and breeding history. Although genetic diversity in the A-subgenome is somewhat higher than in the C-subgenome (Qian et al., 2014) and can be further improved with introgressions from related Brassica species, especially in the Asian gene pool (Zou et al., 2017), the overall degree of diversity in B. napus is
extremely low compared with the highly diverse primary
gene pools of other major global crops like rice (Caicedo
et al., 2007), maize (Navarro et al., 2017), barley (Russell
et al., 2016), or sorghum [Sorghum bicolor (L.) Moench]
(Mace et al., 2013; Morris et al., 2013).

Studies on the application of GS have been con-
ducted in winter-type oilseed rape (Würschum et al.,
2014; Werner et al., 2017), spring-sown canola (Jan et
al., 2016), and a biparental population based on a cross
between a European winter cultivar and a Chinese semi-
winter cultivar (Zou et al., 2016; Liu et al., 2017). How-
ever, the effect of marker density on prediction accuracy
has remained disregarded so far. Considering the lack
of comprehensive genetic variation, resulting in large
blocks of long-range LD, Asian semiwinter rapeseed is
particularly suitable for examining the effect of reduced
marker numbers in the species B. napus and other plant
breeding populations characterized by relatively narrow
variation. Starting with 24,336 high-quality SNPs from
the Illumina Brassica 60k genotyping array (Illumina,
San Diego, CA), we investigated the value of two marker
selection approaches in Asian rapeseed, comparing a
trait-dependent strategy based on GWAS significance
thresholds and a trait-independent method of detecting
representative tag SNPs (de Bakker et al., 2005). Marker
numbers were gradually reduced, and the respective
marker subsets were applied for genomic prediction of
three traits in a diversity set encompassing 203 Chinese
semiwinter rapeseed lines. Our results illustrate that high
prediction accuracies for polygenic traits can be achieved
even with low marker density, provided that the repre-
sentative markers are selected with regard to the genome-
wide LD structure in a population.

Material and Methods

Plant Material and Phenotype Data

A diverse panel of 203 homozygous Chinese semiwinter
inbred lines, broadly encompassing allelic variability in
the Asian B. napus gene pool, was selected from a breed-
ing program of Southwest University in Chongqing,
China, and tested at two locations in Germany. Strong
(>95%) homozygosity resulting from maintenance via
self-pollination was confirmed by SNP genotyping (Qian
et al., 2014). Field trials were conducted by applying an
unrelicated completely randomized design in Gross
Gerau (49.941000°N, 8.501391°E) and Rauischholzhausen
(50.760932°N, 8.880663°E) in 2013 and 2014, respectively.
Phenotypic data were collected for plant height (PHT),
along with the essential seed quality characteristics oil
content and glucosinolate content (GSL) (Table 1). Both
seed quality parameters were measured by near-infrared
spectroscopy. Phenotypic distributions and correlations
among traits and environments were analyzed and illus-
trated with the R packages HMISC (Harrell and Dupont,
2016) and psych (Revelle, 2017). Variance components
were estimated via restricted maximum likelihood
implemented in SPSS Statistics version 22.0 (IBM Corp.,
2013), and broad sense heritability (H2) for all traits was
calculated via the following equation:

\[ H^2 = \frac{\sigma^2_t}{\sigma^2_t + \sigma^2_e / n} \]

where the genotypic and residual variance components
are represented by \( \sigma^2_t \) and \( \sigma^2_e \), respectively, and estimates of the residual variance were divided by the number or
environments \( n \).

Genotypic Data

High-density genotypic characterization of the genotype
panel was performed by the service provider TraitGenetics
(Gatersleben, Germany) with the Illumina Brassica
60k SNP genotyping array (Clarke et al., 2016; Mason
et al., 2017). We applied a filtering for single-copy BLAST
hits that were physically assignable to the B. napus “Dar-
mor-bzh” reference genome assembly version 4.1 (Chal-
houb et al., 2014). Only markers that exhibited at least
95% sequence identity and no gaps in their 50-bp probe
sequence were retained, resulting in 28,698 unique,
locus-specific SNPs. In an additional preprocessing step,
all markers with more than 10% missing values and a
minor allele frequency below 5% were excluded. The final
marker set used in our analyses comprised 24,338 high-
quality SNPs. A detailed description of the genetic com-
position and population structure of the diversity set was
provided by Qian et al. (2014).

Selection of Marker Subsets
for Genomic Prediction

Two methods were examined for the reduction of marker
numbers for genomic prediction. The first involved the
selection of “tag SNPs” (de Bakker et al., 2005); the second
was based on a GWAS-assisted marker selection approach.
Although the GWAS-based selection method was trait-
specific, tag SNP identification was independent of the
trait under investigation. A sample of randomly selected
markers, corresponding to the respective number of tag
SNPs or the number of SNPs detected via GWAS across
cross-validations, was used as a reference set to assess the
efficiency of the two reduction approaches. For both meth-
ods, several subsets with varying marker numbers were
generated (see Supplemental File S1, Supplemental File S2)
and assessed for their predictive ability compared with the
complete set of 24,338 SNPs. Furthermore, if highly signif-
ificant markers were detected in a population-wide GWAS,
these were included as fixed effects in the complete marker
set as well as in the different tag SNP subsets.

Tag SNP Selection

Subsets of markers were assembled by applying the tag SNP
selection algorithm of de Bakker et al. (2005) implemented
in Haploviz version 4.2 (Barrett et al., 2005). The pairwise
mode was executed to select a minimal set of 9793 mark-
ers for which all alleles were correlated at \( r^2 > 0.8 \). Further
reduced subsets were generated by setting a maximum
number of tag-SNPs (Supplementary Material S1).
Genome-Wide Association Study-Assisted Marker Selection

Genome-wide association studies were conducted separately for all three traits in both environments, resulting in six individual analyses. Examination of marker–trait associations was performed by applying the software GenABEL (GenABEL Project Developers, 2013) with the P + K mixed linear model as proposed by Zhao et al. (2007), closely following the Q + K method of Yu et al. (2006):

\[
y = \mu + So + Pr + Zu + e
\]  

[2]

In this expanded mixed model, \( y \) is the vector of phenotypic observations, \( \mu \) is the overall mean, \( a \) is a vector of the fixed allelic effects associated with the SNP under investigation, \( v \) is a vector of fixed population effects, \( u \) is a vector of random genetic background effects (relationships among individuals), and \( e \) is the vector of residuals. The matrix \( P \) contains information on population structure using the first five principle components as fixed factors to adjust the model for population stratification, and \( S \) and \( Z \) are incidence matrices relating \( y \) to \( \alpha \) and \( u \), respectively. The variances of the random effects \( u \) and \( e \) are assumed to be normally distributed with \( u \sim N(0, G \sigma_a^2) \) and \( e \sim N(0, \sigma_e^2) \), where \( G \) is a 203 × 203 genomic relationship matrix and \( R \) is a 203 × 203 matrix in which the off-diagonal elements are 0 and the diagonal elements are the reciprocal of the number of phenotypic observations per individual. The \( G \) matrix was computed according to the first method proposed by VanRaden (2008). Since every genotype was tested only once per environment and GWAS were performed for all environment–trait combinations separately, \( R \) corresponds to the identity matrix \( I \).

Subsets of markers for genomic prediction were selected based on their association with a trait (Supplementary File S2). Beginning with the complete set of 24,338 markers, \(-\log(p)\) values were set to 0.05, 0.1, and 0.2, and then gradually increased by 0.2 until no markers were left in more than 10% of the 100 cross-validation cycles per trait–environment setting. All SNPs below a particular threshold were excluded from prediction. This procedure was conducted in the TP only.

For determining potential fixed-effect markers with major impact on a trait, we applied a strategy to identify significantly associated markers closely related to the method described by Spindel et al. (2016). To correct for multiple testing, we calculated a Bonferroni adjusted \( p \)-value for \( p = 0.05 \) and used a relatively stringent significance threshold of \(-\log_{10}(0.05 \div 24,338) = 5.69\) as a reference for association. Groups of adjacent markers were binned into haplotypes, and only the most significant marker of a haplotype was considered. Haplotype blocks were defined with Haplovlew version 4.2 (Barrett et al., 2005), implementing a method proposed by Gabriel et al. (2002) that calculates the LD between all marker pairs and combines SNPs in a common block if the one-sided upper 95% confidence bound on \( D^* \) is >0.98 and the lower bound lies above 0.7. Significant marker–trait associations were detected only for GSL in both years (Supplementary Material S3). Subsequently, we matched our GWAS results with known QTL regions to assemble a set of fixed-effect markers. In contrast to the subset selection described above, fixed factors were identified using the whole diversity set. Therefore, it must be noted that this compilation was based on the assumption that the detected major-effect markers were representative for the whole breeding population, rather than solely for the training populations (TPs) or for only a subset of individuals. Our intention was not to validate the method itself but to investigate the relevance of including previously discovered, pool-specific SNPs with substantial effects on reduced-density arrays. Use of the whole set rather than the TP for identification and subsequent validation of fixed-effect candidates would obviously not be justifiable.

Prediction Models

Genomic predictions for the six trait–environment combinations were conducted via the following mixed linear model:

\[
y = X\beta + Zu + e,
\]  

[3]

in which \( y \) is a \( n \times 1 \) vector of phenotypic observations and \( n \) is the number of individuals. \( X \) is an incidence matrix relating fixed effects to individuals; \( \beta \) is the corresponding vector containing the respective effects. When no fixed marker effects have been included, \( X \) only assigns the overall phenotypic mean \( \mu \) to all lines (then \( X\beta \) can be replaced by \( I_n \mu \)). \( Z \) is the \( n \times m \) design matrix defining the allelic state of all \( m \) marker loci (AA, AB, or BB) as \(-1, 0, 1\). The vector \( u \) contains the average effects of an allele substitution of the marker alleles related to \( Z \), and \( e \) is the \( n \times 1 \) vector of residual effects. Best linear unbiased estimators of the fixed effects and BLUPs of random effects were calculated using a ridge regression BLUP implemented in the R package rrBLUP (Endelman, 2011), assuming that all marker and residual effects were normally distributed with \( u \sim N(0, \sigma_u^2) \) and \( e \sim N(0, \sigma_e^2) \), respectively. Prediction accuracies for every marker subset within a specific trait–environment setting were assessed by running 100 cross-validations. In each cycle, the diversity set was randomly subdivided into 80% TP and 20% validation population. The average prediction accuracies were determined as the mean Pearson correlation coefficient (\( r \)) between observed and predicted values.

Results

Phenotypic Variation and Pairwise Trait Correlations

Phenotypes for oil content, GSL, and PHT and their distribution within the genotype panel were assessed in each of the two environments separately (Table 1, Fig. 1). In both years, oil content and PHT were approximately normally distributed, whereas GSL exhibited a strong
negative skew caused by several genotypes with relatively high GSL values. Within-trait correlations ($r$) between the two environments ranged from 0.46 for oil content and 0.71 for PHT to 0.95 for GSL, and were positive and highly significant ($p < 0.001$) in all cases. Values for $H^2$ ranked in same order, with 0.42, 0.70, and 0.86 for oil content, PHT, and GSL, respectively, suggesting a substantial effect of nongenetic factors on oil content in this genotype panel across the two environments. Correlations between different traits were, in general, low and only significant for PHT and GSL, where weak and negative correlations from $-0.18$ to $-0.25$ were observed.

### Marker Selection for Subset Compilation

Two selection approaches were examined for their usefulness in reducing marker numbers, including identification of tag SNPs and a method based on GWAS results. The initial set of 9793 tag SNPs was further reduced in a stepwise manner to test the threshold for prediction accuracy (Supplementary Material S1). Genome-wide association studies for marker selection were conducted prior to each prediction cycle only in the respective TP. Based on gradually increasing $-\log_{10}(p)$ thresholds, subsets of markers were selected to comprise only those SNPs that passed a defined significance value. This approach, resulting in an exponential decrease in marker number, was repeated in a similar manner for all traits (Fig. 2). The average threshold-specific number of SNPs per subset across cross-validation cycles is listed in Supplementary File S2.

### Identification of Fixed-Effect Markers

Markers exceeding the significance threshold of $-\log_{10}(0.05/24338) = 5.69$, which was used as an indicator for strong marker-trait associations, were detected only for GSL (Supplementary Material S3). Only the most significant SNP within a haploblock of adjacent markers entered into our model. We found markers on A09 in both years and on C07 and C09 in 2014. The involvement of these regions in GSL regulation had already been demonstrated in previous experiments (Harper et al., 2012; Li et al., 2014; Lu et al., 2014). Considering the major contribution of the two segregating QTL for GSL detected only in 2014, we assembled a pool-specific and environment-independent fixed-effect marker set for application in both years, comprising the three most significant SNPs from the aforementioned chromosome regions.

### Genomic Predictions

Both marker selection approaches were assessed for their applicability to reduce the number of SNPs for genomic prediction. All marker subsets were tested separately to assess the usefulness of the two methods with decreasing marker numbers.

For all three traits across both years, predictions based on the different tag SNP subsets remained relatively stable down to as few as 600 to 800 markers and delivered average prediction accuracies fluctuating only slightly from those obtained with the whole marker set (Fig. 3a–f). Similarly, the use of randomly selected markers led to comparable accuracies, although these generally remained below those of the tag SNP subsets and started to decrease earlier. For GSL, in both years, the inclusion of fixed-effect markers induced a substantial and consistent increase in prediction accuracy in the whole marker set and also in all tag SNP subsets (Fig. 3e–f).

As expected, predictions using the marker subsets generated on the basis of GWAS significance thresholds exhibited much greater trait dependency (Fig. 4). For oil content and PHT, prediction accuracy decreased with increasing $-\log_{10}(p)$ values (Fig. 4a–d). However, although PHT in 2013 and oil content in 2014 exhibited a direct and continuous reduction in accuracy, there was an initial improvement in PHT in 2014. Except for this trait–environment combination, GWAS-based marker subsets were predominantly outperformed by the randomly selected marker subsets in the other three scenarios; none of the subsets exceeded the prediction accuracies realized with the unfiltered marker set. For GSL, prediction accuracies exhibited a wide-ranging increase with significance thresholds and were constantly superior to the whole marker set (Fig. 4e–f). For both years, the highest accuracies were obtained with the most significant 20 to 40 markers, and were comparable only within this range to those achieved when including the fixed-effect markers in the whole set. Similar to GSL, oil content and PHT accuracies based on random markers

---

**Table 1. Statistical parameters describing the distribution of the three traits oil content in the seed (OIL), plant height (PHT), and glucosinolate content (GSL) in the panel of 203 Chinese semiwinter rapeseed lines.**

<table>
<thead>
<tr>
<th>Trait</th>
<th>Year</th>
<th>Genotypes tested</th>
<th>Min.</th>
<th>Max.</th>
<th>Mean</th>
<th>SD</th>
<th>$H^2$†</th>
</tr>
</thead>
<tbody>
<tr>
<td>OIL (%)</td>
<td>2013</td>
<td>168</td>
<td>32.00</td>
<td>45.95</td>
<td>38.72</td>
<td>2.68</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>174</td>
<td>33.50</td>
<td>44.36</td>
<td>37.78</td>
<td>1.98</td>
<td></td>
</tr>
<tr>
<td>PHT (cm)</td>
<td>2013</td>
<td>190</td>
<td>74.00</td>
<td>161.00</td>
<td>119.88</td>
<td>15.38</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>177</td>
<td>77.00</td>
<td>163.00</td>
<td>126.50</td>
<td>17.15</td>
<td></td>
</tr>
<tr>
<td>GSL (µmol g⁻¹)</td>
<td>2013</td>
<td>183</td>
<td>22.27</td>
<td>148.65</td>
<td>46.53</td>
<td>28.35</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>178</td>
<td>20.66</td>
<td>141.81</td>
<td>44.53</td>
<td>25.26</td>
<td></td>
</tr>
</tbody>
</table>

† $H^2$, broad-sense heritability.
Fig. 1. Phenotypic distributions and correlations among the diversity set of 203 Asian semiwinter rapeseed lines for oil content (%), plant height (cm), and glucosinolate content [μmol g⁻¹] in both experimental years. *, p ≤ 0.05; **, p ≤ 0.01; ***, p ≤ 0.001.

Fig. 2. Average number of single nucleotide polymorphism (SNPs) found in the training populations of Asian semiwinter rapeseed to pass the genome-wide association study significance thresholds, which are plotted along the abscissa. Single nucleotide polymorphism number was calculated as the mean value across the training populations sampled in 100 cross-validation cycles. The decrease in marker number is illustrated up to a significance threshold of –log₁₀(p) = 3.0, after which no more markers were significantly associated with oil content (OIL) or plant height (PHT) in more than 10% of the cross-validation cycles.
remained relatively stable until a $-\log_{10}(p)$ threshold of 1.2–1.4, but substantially declined afterward.

Even though the different scaling of the abscissa complicates direct comparison of the tag SNP and GWAS-based approaches, the curves representing the randomly assembled marker sets demonstrate that the former method enables higher and more stable prediction accuracies than the latter for oil content and PHT. In this context, the range between a $-\log_{10}(p)$ of 1.2 and 1.4 represents a key feature with marker numbers dropping below 1000 in all four scenarios. On the other hand, marker reduction using GWAS thresholds outperformed LD-based subsets of tag SNPs for GSL.

Discussion

In this study, we demonstrated that low-density subsets of markers comprising several hundred to a few thousand SNPs, representing a selection from a commercially
available high-density genotyping platform, can be sufficient for successful application of genomic prediction in polygenic traits in an Asian semiwinter rapeseed diversity panel. Promising results were obtained via identification of tag SNPs based on genome-wide LD patterns, using the Illumina Brassica 60k genotyping array (Illumina). Although this approach is not recommended as a stand-alone strategy for traits that are primarily controlled by a small number of major QTL, including previously detected, pool-characteristic SNPs exhibiting a profound marker–trait association as fixed effects not only considerably increased prediction accuracy but also performed better than exclusive application of the most significant markers.

The idea of GS to exploit large numbers of markers simultaneously across the whole genome emerged from the problem that the expression of polygenic traits, being controlled by an infinitesimally high number of
small-effect QTL, cannot be explained using classic MAS (Meuwissen et al., 2001). Nowadays, high-density SNP arrays carrying between a few thousand to several hundred thousand polymorphic sites have become routine tools for carrying out GS in plants. However, with extensive genome-wide LD being a common characteristic of plant breeding gene pools, we hypothesized that the assembly of marker panels containing only a small number of SNPs, carefully selected from already available high-density genotyping arrays, could be a cost-efficient approach to implementing and promoting GS in plant breeding programs.

To examine the effects of reduced marker density on genomic prediction accuracy, we collected data on oil content and PHT, both of which are polygenic traits of high agronomic importance. To further enable conclusions to be made regarding the efficacy of GS in traits known to be primarily controlled by few major QTL, we also included GSL in our predictions. For all three traits, high levels of phenotypic variation were observed across the 203 inbred lines in the genotype panel. Estimates of $H^2$ revealed that phenotypic variation in all traits was largely a result of genotypic variation, which is a fundamental prerequisite for the successful application of selection. As was expected for a trait mainly controlled by few major QTL (Harper et al., 2012; Li et al., 2014; Lu et al., 2014; Qian et al., 2014), the highest $H^2$ value was obtained for GSL (0.86). In contrast, the considerably lower heritability (0.42) observed for oil content indicated a substantial impact of nongenetic factors on the expression of this highly polygenic trait. Correlations among the traits were weak and mostly nonsignificant, suggesting different underlying genetic control mechanisms. Thus the three traits considered in this study were highly suitable for obtaining insights into the usefulness of reduced marker densities for genomic prediction of traits with different genetic architecture in modern breeding pools.

Starting with a set of 24,338 high-quality unique, polymorphic SNPs on the Brassica 60k genotyping array (Illumina), we compared two methods for reducing marker number: a strategy making use of GWAS significance thresholds and a tag SNP selection method. Although the GWAS-based approach builds on the extent of marker–trait associations and therefore involved analysis of each trait–environment setting separately, the identification of tag SNPs is solely based on LD structure and requires population-specific applications only. In all six individual association studies, the reduction of marker numbers using gradually increasing significance thresholds followed an exponential decay curve (Fig. 2). Consequently, for the vast majority of the 24,338 SNPs, virtually no or only very weak marker–trait associations were detected. Even though the three traits differ fundamentally with respect to the number of segregating QTL involved in their phenotypic expression, all six curves exhibited similar progression up to a threshold of $-\log_{10}(p) = 2.8 \to 3.0$, after which no more markers were significantly associated with OIL or PHT in more than 10% of the cross-validation cycles. These results were contrary to our expectations. When considering the differences in trait architecture, the intuitive assumption is that a lot more markers could be associated with oil content and PHT at low to moderate thresholds than with GSL, representing the highly polygenic control of oil content and PHT. On the other hand, markers significantly associated with GSL were found on chromosomes A09, C07, and C09, corresponding to known major-effect QTL regulating this trait in *B. napus* (Harper et al., 2012; Li et al., 2014; Lu et al., 2014). These results indicate that even though GWAS in general is a powerful method of tracking down major-effect QTL, it seems unsuitable for at least rudimentarily detangling different degrees of marker–trait associations at low significance thresholds. This might not be the case in large diversity sets that enable increased statistical power. However, it is probably a common feature of association studies in breeding populations, which often do not encompass more than a few hundred genotyped and deep-phenotyped individuals.

In contrast to the GWAS-based approach, the identification of tag SNPs is trait-independent, thereby requiring only one marker selection run per population. Basically, this haplotype-based tagging method samples a subset of variants in a way that ensures that every LD block is represented by the most informative marker within it, defined by either a minimum LD between adjacent SNPs or a maximum number (de Bakker et al., 2005). With 9793 markers being sufficient to cover the whole genome in this population at a minimum LD of 0.8 between marker alleles, the low overall genetic diversity in Chinese semiwinter rapeseed is clearly demonstrated (Qian et al., 2014). However, we subsequently further reduced the subset of markers by setting maximum SNP quantities rather than LD thresholds, a more intuitive approach that could be useful with regard to the design of pool-specific genotyping arrays, for example. A selection method implemented by setting a maximum number of markers is also more user-friendly than deciding on a reasonable significance threshold for every trait individually. Nonetheless, we expect that the trait-dependent method will be likely to enable superior prediction accuracies and thus justify the extra effort related to the GWAS-based method. This, however, did not hold true, as discussed in detail below.

Average prediction accuracies were obtained for each marker subset with a 100-fold cross-validation approach. Predictions based on randomly sampled markers, corresponding to the respective number of tag SNPs or markers detected via GWAS, served as a performance benchmark.

For the two highly polygenic traits oil content and PHT, tag SNP subsets generally performed best in the course of marker reduction, whereas prediction accuracies remained relatively stable in comparison to the unfiltered marker set, even with less than 1000 SNPs. These results clearly demonstrate the suitability of
this haplotype-based marker reduction approach. By adjusting marker spacing with regard to LD structure in a population, differences in the extent of conserved genome regions are taken into account by varying SNP densities. This strategy is especially useful in species like rapeseed with extensive, highly conserved haploblocks that allow for a relatively low number of polymorphisms to represent these regions in a GS model. The difference in prediction accuracy observed between 2013 and 2014 indicates the influence of specific genetic factors which may interact pleiotropically with oil content in specific environments. In the case of seed oil content in B. napus, which normally shows relatively high heritability, a negative correlation with other seed components (e.g., protein, seed coat fiber) can potentially lead to large indirect effects on oil from major loci influencing these associated factors under specific environmental conditions (e.g., high temperature). This might result in the increased complexity of genetic trait control and thereby impair prediction accuracy. Surprisingly, average prediction accuracies also stayed comparatively high and stable with decreasing marker density when SNPs were selected just by chance. The fact that random selection of even just a few thousand SNPs failed to cause severe losses in prediction accuracy further illustrates the great extent of conserved long-range LD and the strong association between adjacent markers within haploblocks in the test population. This situation may be typical for young breeding populations like that of rapeseed, which have undergone extreme genetic bottlenecks caused by adaptive and breeding selection. However, because of the generally inferior performance and higher SDs in prediction accuracies also stayed comparatively high and stable with decreasing marker density when SNPs were selected just by chance. The fact that random selection of even just a few thousand SNPs failed to cause severe losses in prediction accuracy further illustrates the great extent of conserved long-range LD and the strong association between adjacent markers within haploblocks in the test population. This situation may be typical for young breeding populations like that of rapeseed, which have undergone extreme genetic bottlenecks caused by adaptive and breeding selection. However, because of the generally inferior performance and higher SDs in prediction accuracy resulting from sampling variation, random marker selection cannot be recommended as a strategy to reduce marker density in a breeding population.

In both highly polygenic traits, marker reduction with GWAS results exhibited the weakest performance of all three methods. We initially assumed that for polygenic traits, low significance thresholds might be useful for removing markers that show virtually no trait associations. As a consequence, genetic variance could be allocated only to a reduced set of whole-genome markers truly associated with trait-related regions but also considering small-effect QTL, since the applied significance thresholds were chosen to be low. The observed results, however, could not reinforce our hypothesis. Genomewide association studies did not prove suitable for thoroughly dissecting genetic trait architecture or reliably separating the markers associated with small-effect QTL from markers in unassociated regions. It is likely that because of a lack of statistical power in the relatively small and highly conserved rapeseed population, sets of redundant markers within the same high-effect haplобlocks passed the selection thresholds and resulted in an over-representation of these regions, whereas a high number of small-effect QTL across the whole genome were not captured, so that these regions remained under-represented. Thus this approach does not facilitate extensive genome coverage and therefore is in contradiction to the principles underlying GS. However, it is possible that within a different context, the GWAS-based approach to marker reduction could be of higher value. This may, for example, be the case in studies comprising significantly more genotypes with increased levels of diversity, resulting in higher resolution and increased statistical power. However, validation of this hypothesis would require considerable investment to establish such a resource for plant breeding, which, in most cases, may make such an approach unrealistic.

In contrast to PHT and oil content, the GWAS-based marker reduction method was superior to the other two approaches in predicting GSL. Even though tag SNPs and randomly selected markers enabled GSL prediction accuracies comparable to those of the unfiltered set with low marker numbers, reducing the markers based on GWAS resulted in a considerable increase in prediction accuracy. This is not surprising for a trait that is mainly controlled by a small number of segregating QTL and basically leads back to classic MAS, which is now commonly used in breeding to transfer Mendelian traits or major QTL. In both years, however, the highest prediction accuracies were obtained with about 20 to 40 SNPs clustering around a few QTL. This illustrates the usefulness of haplобlocks encompassing several strongly associated SNPs as multiallelic markers in MAS. Comparably high prediction accuracies for GSL were also obtained by setting a few preselected, highly-significant markers as fixed factors in a GS model, still treating the remaining markers as random effects. These results also demonstrate that inclusion of marker information in addition to a few known major-effect SNPs can be advantageous even for traits whose inheritance basically proceeds in a Mendelian manner. The main advantage of combining classic MAS with GS is that defining the SNP markers that are strongly associated with the trait under investigation as fixed factors overcomes the restriction of the original ridge regression BLUP that even major-effect QTL are strongly reduced toward zero (Bernardo, 2014; Spindel et al., 2016). This is because the genetic grouping of individuals according to their fixed-effect markers is based on the assumption that genotypes exhibiting different marker profiles derive from distinct populations with different means. Several studies have already described the usefulness of this approach and various strategies of fixed-effect marker selection have been applied (Bentley et al., 2014; Bernardo, 2014; Spindel et al., 2016; van den Berg et al., 2016; Zhang et al., 2014).

In comparison to relating the differences in trait expression only to haplобlock structure around selected major-effect SNPs, a method capturing genetic dissimilarities in addition to strongly associated major-effect SNPs across the whole genome seems to be more direct and universally applicable. Given its simple implementation, this approach is also highly suitable to be incorporated in a GS framework with low marker densities. In this case, however, it has to be taken into account that...
trait-specific, fixed-effect SNPs common to a particular gene pool must be identified before and considered during the process of designing a new reduced-density marker array. In contrast to GS, a marker set encompassing a few hundred to a few thousand tag SNPs will not be suitable for a detailed dissection of trait architecture and therefore cannot guarantee that the most significant markers will be represented by the selected polymorphic sites. With numerous markers being known and routinely used for MAS in several traits in breeding programs, however, this should not be a limitation for selection of traits with comparable genetic architecture.

Conclusions

Using semiwinter rapeseed as a model, we demonstrate that low-density marker sets comprising only a few hundred to a few thousand markers enable high genomic prediction accuracies in breeding populations with strong LD, making them a cost-effective genotyping tool easily producible from already available high-density arrays. Although as few as 1000 tag SNPs enabled prediction accuracies comparable to 24,335 markers, it should also be noted that a breeder might still gain a strong advantage from selection by applying a chip with reduced marker density even when prediction accuracies are slightly worse than a high-density chip. The reason is because reduced marker densities allow a substantial reduction in costs, enabling field phenotyping resources to be focused on preselected materials that, even with somewhat lower selection accuracy, are still likely to span lines that significantly accelerate breeding progress. Our results are therefore of high importance to facilitating the routine application of GS in breeding programs and supplying breeders with a new source of information to make promising selection decisions.

Supplemental Information

Supplemental File S1. Prediction accuracies and descriptive statistics of tag SNP subsets.

Supplemental File S2. Prediction accuracies and descriptive statistics of SNP subsets selected on the basis of GWAS significance thresholds.

Supplemental File S3. Manhattan plots of GWAS for oil content, plant height and glucosinolate content.

Conflict of Interest Disclosure

The authors declare that there is no conflict of interest.

References


Bernardo, R. 2014. Genomewide selection when major genes are known. Crop Sci. 54:68–75. doi:10.2135/cropsci2013.05.0315


IBM Corp. 2013. IBM SPSS Statistics for Windows, version 22.0: IBM Corp., Armonk, NY


