Genomic Prediction Using Prior Quantitative Trait Loci Information Reveals a Large Reservoir of Underutilised Blackleg Resistance in Diverse Canola (Brassica napus L.) Lines

Mulusew Fikere,* Denise M. Barbulescu, Michelle M. Malmberg, Fan Shi, Joshua C.O. Koh, Anthony T. Slater, Iona M. MacLeod, Phillip J. Bowman, Phillip A. Salisbury, German C. Spangenberg, Noel O.I. Cogan, and Hans D. Daetwyler

Abstract
Genomic prediction is becoming a popular plant breeding method to predict the genetic merit of lines. While some genomic prediction results have been reported in canola, none have been evaluated for blackleg disease. Here, we report genomic prediction for seedling emergence, survival rate, and internal infection, using 532 Spring and Winter canola lines. These lines were phenotyped in two replicated blackleg disease nurseries grown at Wickliffe and Green Lake, Victoria, Australia. A transcriptome genotyping-by-sequencing approach revealed 98,054 single nucleotide polymorphisms (SNPs) after quality control. We assessed various genomic prediction scenarios based on Genomic Best Linear Unbiased Prediction (GBLUP), BayesR and BayesRC, which can make use of prior quantitative trait loci information, via cross-validation. Clustering based on genomic relationships showed that Winter and Spring lines were genetically distinct, indicating limited gene flow between sets. Genetic correlations within traits between Spring and Winter lines ranged from 0.68 and 0.90 (mean = 0.76). Based on GBLUP in the whole population, moderate to high genomic prediction accuracies were achieved within environments (0.35–0.74) and were reduced across environments (0.28–0.58). Prediction accuracy within the Spring set ranged from 0.30–0.69, and from 0.19–0.71 within the Winter set. The BayesR model resulted in slightly lower accuracy to GBLUP. The proportion of genetic variance explained by known blackleg quantitative trait loci (QTL) regions was < 30%, indicating that there is a large reservoir of genetic variation in blackleg traits that remains to be discovered, but can be captured with genomic prediction. However, providing prior information of known QTL in the BayesRC method resulted in an increased prediction accuracy for the traits survival and internal infection.

Core Ideas
- Genomic prediction accuracy was moderate to high for blackleg resistance within and low across locations indicating substantial genotype-by-environment effects.
- Known blackleg resistance quantitative trait loci (QTL) regions explained at most 30% of the genetic variance for survival rate and internal infection.
- Incorporating known QTL regions in the genomic prediction method BayesRC increased prediction accuracy for the traits survival and internal infection.
- Strict prediction from Winter into Spring lines (and vice versa) had limited prediction accuracy, and combining both sets in one reference population was beneficial only if reference population size was low.
- Selection for improved blackleg resistance based on genomic breeding values is now feasible and will capture a greater proportion of the genetic variance than marker-assisted selection on known QTL alone.
**A LLOTETRAPLOID CANOLA** (Brassica napus L.) originated from a natural hybridization between cabbage (Brassica oleracea CC, 2n = 18) and turnip rape (Brassica rapa; AA, 2n = 20) between 400 and 500 yr (Go’mez-Campo and Prakash, 1999) and approximately 7500 yr ago (Chalhoub et al., 2014). It is the second largest global oil-seed crop in terms of area of production after soybean and the third largest crop in Australia after wheat and barley (GRDC, 2009). The demand of canola has grown rapidly for food, feed, biodiesel, and other industrial purposes. Phenotypic plant breeding schemes have played a vital role in improving agronomic traits (e.g., yield, pod shattering, lodging), quality traits (e.g., reduced glucosinolate and erucic acid content, improved oil content and shelf life) and developing blackleg disease resistant varieties.

Blackleg disease is caused by the fungal pathogen Leptosphaeria maculans and can result in up to 90% yield loss in unprotected crops (Sosnowski et al., 2004). L. maculans is a hemibiotrophic Dothideomycete fungus that also has a saprophytic lifestyle feeding on the canola stubble (Howlett, 2004). Sexual reproduction occurs in canola stubble with a large number of ascospores released from fruiting bodies in the following sowing season’s autumn rainfall. Ascospores are carried by wind and infect the new crop via colonization of the apoplast in the asexual growth stage. The fungus creates leaf lesions and eventual growth down the stem, destroying the cells and creating a blackened basal stem canker that may completely girdle the stem at the crown level. The infection and life cycle of the pathogen is described extensively by Van De Wouw et al. (2016). Reports show that in Australia, Europe, and Canada over $900 million dollars are lost annually due to blackleg (Fitt et al., 2006; Howlett, 2004). Thus, developing blackleg disease resistant varieties is of utmost importance for canola breeding programs. In Australia, screening for blackleg disease resistance started in the 1970s. Subsequently, breeding programs expanded the canola gene pool through multi-purpose hybridization and developed disease resistance varieties (Salisbury et al., 1995; Sprague et al., 2006; Cowling, 2007). Most resistant cultivars contain major genes conferring resistance to specific blackleg pathotypes, which is qualitative resistance. While very effective initially, the pathogen has proven adept at overcoming the different major gene resistances in areas where high disease pressure was created by the intense cultivation of varieties with the same major gene (e.g., Rlm1 in France, (Rouxel et al., 2003); and LepR3 in Australia, (Sprague et al., 2006). In contrast, quantitative resistance, which is due to many minor genes, offers incremental resistance improvements. Initial blackleg leaf lesions may appear similar to those on susceptible varieties, but later in the season no or less severe stem cankers are observed on varieties with improved quantitative resistance (Delourme et al., 2014). Phenotypic and pedigree selection can potentially exploit both types of resistance to breed varieties with durable field resistance to L. maculans (Brun et al., 2010; Delourme et al., 2014). Phenotyping for both qualitative and quantitative resistance can be done in field disease nurseries against a full fungal population (Kaur et al., 2009; Light et al., 2011) as well as in controlled environment conditions against specific single spore fungal isolates (Rouxel et al., 2003; Marcroft et al., 2012; Brun et al., 2010).

In the last 20 yr, genomic data has been incorporated into breeding and quantitative trait loci (QTL) have been mapped that confer mainly qualitative resistance using a variety of statistical designs, such as F2-crosses (Ansan-Melayah et al., 1998), doubled haploid populations (Delourme et al., 2004), backcrosses (Long et al., 2011) and Winter and Spring diversity panels (Uzunova et al., 1995; Kaur et al., 2009; Jestin et al., 2011; Raman et al., 2012a; Delourme et al., 2014; Van De Wouw et al., 2016). These efforts have identified a growing list of L. maculans resistance genes (e.g., LepR1, LepR2, LepR3, LepR4, BLMR2, Rlm1, Rlm2, Rlm3, Rlm4, Rlm5, Rlm6, Rlm7, Rlm8, Rlm9, Rlm10, Rlm11, Rlm12, and LmFr1) in Brassica species (Yu et al., 2005; Delourme et al., 2006; Yu et al., 2008; Marcroft et al., 2012; Raman et al., 2012b; Balesdent et al., 2013; Larkan et al., 2014; Fomeju et al., 2014; Raman et al., 2016). Consequently, irrespective of some limitations, marker assisted selection has been used for mapping and validating resistance QTL to L. maculans using progenies/populations derived from breeding programs. However, studies in canola and other species show that genetic variance explained by known QTL tend to be low (Yang et al., 2010; Huo. et al., 2016).

Currently, many plant genetic improvement schemes are transitioning from phenotype and pedigree based selection to genetic marker based genomic selection. Genomic selection uses a reference population that is genotyped and phenotyped to predict the genetic value for individuals that are only genotyped using genome-wide markers (Meuwissen et al., 2001). Several studies argue that genomic selection provides better accuracies than might be achieved on the basis of pedigree information alone (Crossa et al., 2010; Jia and Jannink, 2012; Perez-Enciso et al., 2015; Tsai et al., 2015; Vela-Avitu et al., 2015). This paradigm shift is significantly improving genetic gain by reducing the breeding cycle, increasing selection accuracy in un-phenotyped germplasm, and increasing selection intensity. Furthermore, studies in a wide variety of plant species confirm the realized and potential benefits of using genomic selection (Heffner et al., 2011; He et al., 2016; Bentley et al., 2014; Grenier et al., 2015). Various genomic prediction models have been proposed such as Genomic Best Linear Unbiased Prediction (GBLUP) (Nejati-Javaremi et al., 1997; Habier et al., 2007; VanRaden, 2008), Ridge Regression Best Linear Unbiased Prediction (RR-BLUP), Bayes A and B (Meuwissen et al., 2001), BayesR (Erbe et al., 2012). BayesR has been extended to incorporate prior information obtained from variant, gene, or regulatory regions (BayesRC) (MacLeod et al., 2016). Genomic selection in canola was first highlighted by Wurschum et al. (2014), who reported prediction accuracies for agronomic traits using RR-BLUP and BayesB in a lower diversity set of 391 doubled haploid...
lines genotyped with 251 SNPs. A more diverse set of 475 Spring canola lines was used by Jan et al. (2016) to examine the performance of testcross combinations using 24,403 SNPs markers based on the RR-BLUP model.

However, further questions remain with respect to the application of genomic selection in canola. Can genomic selection be effectively applied to blackleg resistance? Are predictions possible from Winter into Spring type varieties? Are genomic predictions accurate across environments? Can we increase the accuracy of genomic selection by incorporating prior information such as previously published QTL results in our models? How much of the genetic variance is explained by known blackleg QTL?

In this study, we intend to address these questions using three different genomic selection methods (GBLUP, BayesR, and BayesRC) in a diverse set of 532 Winter and Spring canola lines genotyped for 98,054 SNPs and phenotyped for adult plant survival and stem internal infection in replicated field disease nurseries at two locations.

**MATERIAL AND METHODS**

**Phenotyping**

Infield phenotyping was conducted at two sites, Wickliffe (WL; 37.665839°S, 142.754126°E) and Green Lake (GL; 36.768420°S, 142.264679°E), Victoria, Australia during the 2015 growing season. The weather data for these two regions is presented in Table S2 and Figure S1. A starting total of 600 canola lines (of which 326 Winter and 206 Spring lines remained after genotype quality control), including lines with identified major resistance genes (Rlm), were planted in canola stubble (cultivar ATR-Gem, Rlm1) from the previous year’s crop. The disease nursery was sown in single rows (5 m long by 0.75 m apart) at 150 seeds per row/genotype, using a randomized block design with a check variety grid (Trigold) and two replications per site. The trials were monitored regularly and managed to minimize plant death from factors other than blackleg infection. Three phenotypes were recorded: emergence count (number of plants emerged 6 wk after sowing), adult plant survival rate (survival count at maturity/emergence count × 100) and percentage internal infection of the stem at maturity. To determine the internal infection of the stem, a maximum of 20 randomly selected plants per row were cut with secateurs at the crown and the cross-section of the stem was examined. The area of the stem infected by L. maculans was recorded as a percentage of affected area of the total stem cross-section, as described in the Spring Blackleg Management Guide (GRDC, 2015). Averaged internal infection values for each line tested were used in the statistical analysis. Phenotypes were spatially adjusted with autocorrelation error (i.e., AR1 × AR1) models for field condition variability to generate Best Linear Unbiased Estimates (BLUEs) using ASReml (Gilmour et al., 2009), as described in Burgueno et al. (2000) and Burgueño et al. (2012). Applying the following model for a line i sown at row j and column k,

\[ y_{ijk} = \mu + g_i + r_j + c_k + e_{ijk} \]

where \( y_{ijk} \) is the phenotype, \( \mu \) is the overall mean, \( g_i \) is the fixed effect for variety \( i \), \( r_j \) is the random effect for row \( j \), \( c_k \) is the random effect for column \( k \) fit as spline, and \( e_{ijk} \) is the residual. Log Likelihoods and uniformity of residual variogram plots were used to judge the best model. BLUE summary statistics for the whole population, Spring, and Winter lines are presented in Table 1. Analyzed survival rate BLUE values were also converted to blackleg resistance rating classes as defined in the Australian Accreditation System for canola protocol (1999, revised 2014) (where \( R = >75\% \) survival rate, MR = 61–75\%, MS = 41–60\% S = 11–40\%, VS. = < 10\% survival rate).

**Genotyping**

A total of 578 Spring and Winter type canola lines were successfully genotyped using the protocol described in Malmberg et al. (2017). Briefly, mRNA was extracted from young leaf tissue (Dynabeads: Life Technologies) and used for library preparation for RNA sequencing (SureSelect: Agilent Technologies). Circa 3 million reads were generated per sample using either an Illumina HiSeq 3000 or a NextSeq 1500. The resulting sequence data was adaptor and quality trimmed before aligning to the CDS reference genome Darmor-bzh: (Chalhoub et al., 2014) using the BWA-MEM algorithm (Li, 2013) and SNP were identified with SAMtools mpileup (Li et al., 2009).

**Quality Control**

The original dataset consisted of 578 genotyped canola lines and 310,575 SNP markers. We then excluded SNP without a unique known chromosome position according to Darmor-bzh reference sequence (Chalhoub et al., 2014), reducing the set to 209,474 markers. The SNP genotypes with < 5 read depth were set to missing and SNP were removed if they had: minor allele frequency < 0.01, > 50% missing genotypes, and > 0.4 heterozygosity. Lines were removed if they had > 80% missing genotypes, dropping the set of samples and SNP to 532 and 149,676, respectively. We then used the ability to impute SNP genotypes as a proxy for SNP quality, where imputation is the inference of missing genotypes from a reference set of known SNP genotypes. To do so, for each SNP, the genotypes for 50 randomly chosen accessions were masked and subsequently imputed using FIImpute (Sargolzaei et al., 2011). Pearson’s correlation coefficients of observed values and imputed values of the 50 test samples were calculated to assess imputation accuracy. The random simulation was repeated 10 times and the mean imputation accuracy was calculated for every SNP. Subsequently, we removed SNP with a mean imputation accuracy < 0.5 and imputed the remaining real missing genotypes in the filtered data with FIImpute (Figure S2). It is of course possible that SNP with an imputation accuracy < 0.5 were still true SNP, but they may have a high missing rate or be of low minor allele frequency. Finally, we removed SNP with identical genotypes across all samples and filtered again for minor allele frequency.

"FIKERE ET AL.: GENOMIC SELECTION FOR BLACKLEG DISEASE IN CANOLA"
frequency > 0.01. This final dataset contained 532 canola lines and 98,054 SNP markers. The overall workflow is illustrated in Figure S3 and quality control parameters are presented in supplementary Table S3.

### Linkage Disequilibrium

Linkage disequilibrium for *B. napus* genomes (A and C) was estimated using PLINK (Purcell et al., 2007). The genotypes were coded 0, 1, or 2 to represent the number of alternative variant alleles. The squared correlation linkage disequilibrium coefficient ($R^2$) between the genotypes was estimated for each pair of 98,054 SNPs for Spring and Winter types separately. Genetic similarities between 532 canola lines were estimated using a genomic relationship matrix as described in Yang et al. (2010). The heatmap plot showing the genetic relatedness among lines was generated using the ggplot2 R package (Fig. 1).

### Heritability, Phenotypic and Genetic Correlations

Broad-sense and genomic heritabilities were estimated for the whole population and within seasonality via Restricted Maximum Likelihood (REML) in ASReml (Gilmour et al., 2009) by fitting BLUEs in the following mixed model:

$$ y = Xb + Zg + e $$

where $y$ a vector of BLUEs, $X$ and $Z$ are design matrices for fixed ($b$) and random effects ($g$), respectively, $b$ only contained the overall mean, $g$ is a vector of random genetic effects, $e$ are random residual effects $~ N(0, \mathbf{I}\sigma^2_e)$, where $\sigma^2_e$ is the residual variance and $\mathbf{I}$ is an identity matrix. When estimating broad-sense heritability ($H^2$),

$$ g \sim N(0, \mathbf{I}\sigma^2_g) $$

and when estimating genomic heritability ($h^2$) in the GBLUP model

$$ g \sim N(0, \mathbf{G}\sigma^2_g), $$

where $\sigma^2_g$ is the variance due to lines and $\sigma^2_g$ is the total additive genetic variance captured by the markers, and $\mathbf{G}$ is the genomic relationship matrix between individuals calculated as in Yang et al. (2010). We calculated genetic correlations between traits across sites within seasonality and in the whole population, as well as within site between Spring and Winter lines using bivariate GBLUP models, where $y$ was now a $1 \times 2n$ vector of BLUEs, $n$ was the number of lines in each trait, and the mixed model equations were expanded accordingly (e.g., Mrode, 2014). Phenotypic correlations were calculated as the Pearson correlation of BLUEs across sites and traits. Heritabilities were calculated using the following ratios:

$$ 	ext{Heritability} = \frac{V_g}{V_p}, $$

where $V_g$ is the phenotypic variance (calculated as $V_g + $ residual variance / $r$), where $r$ is the number of replicates per trial and $V_g$ is the variance due to lines in $H^2$ or the variance due to SNPs in $h^2$ (Holland et al., 2002).

### Genomic Prediction

A variety of genomic prediction models have been suggested for the estimation of genomic breeding values (de Los Campos et al., 2013). We chose GBLUP and BayesR (Erbe et al., 2012) because these models implement different assumptions, particularly in the distribution of QTL effects. For instance, GBLUP assumes marker effects are small, non-zero and normally distributed across loci.
having a common variance, whereas BayesR models allow for heterogeneity among marker variances. In addition, we have tested a genomic prediction model that incorporates prior QTL information in the analysis BayesRC. (MacLeod et al., 2016). Other Bayesian methods such as BayesB (Meeuwissen et al., 2001) would likely produce similar results to BayesR, as shown in Daetwyler et al. (2014). Detailed descriptions of models are presented below:

**Genomic Best Linear Unbiased Prediction (GBLUP)**

The GBLUP method was implemented using REML in ASReml (Gilmour et al., 2009) using the mixed model described in Eq. [1]. Additionally, the basic GBLUP model was expanded by including a genotype-by-environment (G×E) effect to estimate the proportion of the variance captured by the G×E interaction. Specifically, this expanded Eq. [1] to:

\[ y = Xb + Zg + Z_i g + e \]

where \( y \) is a vector of BLUEs, \( X \) and \( W \) are design matrices connecting phenotypes to the fixed effect (\( b \), including the overall mean) and marker effects (\( y \)), respectively. Marker effects (\( y \)) were distributed as a mixture of the four distributions as listed above and genotypes were centered and standardized, \( e \) was a vector of residual errors, distributed as \( N(0, 1\sigma_e^2) \): with \( \sigma_e^2 = \text{error variance} \). More detail is provided in Erbe et al. (2012) and in Kemper et al. (2015). For all BayesR models and traits we implemented four replicate chains of the Gibbs sampler, and each chain ran for 20,000 iterations with 10,000 iterations discarded as burn-in. Final parameter estimates were derived from the means across the four replicate MCMC chains.

**BayesRC Model**

The BayesRC model uses the same approach as BayesR, however this model incorporates independent biological prior information into the analysis (MacLeod et al., 2016). The prior biological information may be different variant annotations or classes or variants from functional assays. In this study we used QTL regions or genes as a prior even though QTL are not biological priors in the strictest sense as they are statistical associations. Several classes of variants can be defined and the mixture distribution of effects is estimated independently for each class. If for example, there is evidence from the data that one class is more enriched for large effects, then the mixture distribution in this class will differ to the other classes.

The same parameters were applied in BayesRC as in BayesR. We utilized the SNP in known blackleg QTL regions (Table S4) and categorized them (Table S4) as a first prior (class I), whereas all SNP in other regions were allocated to the second prior (class II). The prior probability for the mixture distribution was the same for each class as described in (MacLeod et al., 2016). The classes were given flat priors and the posterior mixture of SNPs effect distributions were driven by the data, especially if there is enrichment for real QTL in any given class. Therefore, if a class is found to be enriched for QTL, this increases the probability that a true QTL effect in this class will be included in the model. We compared the BayesR and BayesRC models for emergence count, survival rate and average internal infection across sites and we estimated the number of QTL effects per variance distributions. In addition, we compared marker effect estimates between BayesR and BayesRC models across the 19 B. napus chromosomes (Fig. 2). For Manhattan plots, marker effects were de-standardized to account for the fact that genotypes had been centered to a mean of zero and standardized to a variance of 1:

\[ (\alpha_j^d)^2 = \frac{(\alpha_j)^2}{2p_j(1-p_j)} \]

where \( \alpha_j^d \) is the de-standardized effect of SNP \( j \), \( \alpha_j \) is the mean effect of SNP \( j \), and \( p_j \) is the minor allele frequency of SNP \( j \) calculated from the genotypes of reference individuals.
This makes $\alpha_j$ independent of the allele frequency in the sample and is therefore comparable across analyses.

**Cross Validation and Genomic Prediction Scenarios**

A workflow is illustrated in Figure S3. Ten-fold cross validation was used for the whole population and for the winter set, whereas fivefold cross validation was applied for the spring set due to smaller sample size. In each cross validation, the reference set was randomly divided into $k$ ($k = 10$ or $5$) approximately equal-sized subsets. Each of the $k$ subsets was in turn chosen as the validation set by masking their phenotypes and was subsequently predicted using the other $k - 1$ subsets as the reference set. Each cross-validation was repeated 10 times and the average prediction accuracy across all subsets and cross-validations was calculated. We also investigated genomic prediction using only the spring lines as the reference population to predict the winter lines and vice versa. Finally, the accuracy for each genomic prediction scenario is presented as the Pearson's correlation coefficient between Genomic Estimated Breeding Values (GEBVs) and BLUEs ($r = \text{cor}[\text{GEBV}, \text{BLUE}]$). We have assessed the bias of our GEBV estimates, which is the regression coefficient of BLUEs on GEBV ($b = \text{reg}[\text{BLUE}, \text{GEBV}]$) using R (R Core Team, 2014). If bias is less than 1 it indicates overestimation of GEBVs (inflation), while a coefficient larger than 1 indicates underestimation of GEBVs (deflation).
RESULTS

SNP Distribution and Linkage Disequilibrium (LD)

We genotyped 578 lines using a transcriptome genotyping-by-sequencing assay (Malmberg et al., 2017). After quality control on the whole dataset, 532 samples and 98,054 SNP markers remained. The SNP distribution across sub genomes (A and C) was uniform forming a good basis for genomic selection to capture all genomic regions (Fig. 3). The LD varied across sub genomes and seasonality (winter and spring types) (Figure S4). We estimated the average $R^2$ for each chromosome in 10kbp windows, which revealed that for the spring types chromosome 8 (A08) and 11 (C01) had the highest mean LD followed by chromosome 2 (A02) and 13 (C13), while the least LD was observed for chromosome 10 (A10) and 1 (A01). Winter lines had the highest mean LD on chromosome 2 (A02) and 11 (C01), followed by chromosome 16 (C06) (Figure S4).

Trait Heritability, Correlations, and Population Relatedness

Phenotypic and genetic correlations were analyzed for three traits for the whole population along with broad-sense and genomic heritability at both sites. Summary statistics for BLUEs of the three traits along with their heritabilities within seasonal type are presented in Table 1. The overall in-field performance of spring and winter lines at WL was better than the GL site due to lower rainfall at GL (Figure S5). Broad-sense heritability was estimated as the ratio of total genetic variance to phenotypic variance and ranged from 0.37 at GL to 0.76 at WL sites. Similarly, genomic heritabilities (i.e., the proportion of the phenotypic variance explained by the markers) calculated using GBLUP ranged in the whole population from 0.24 for emergence count at GL to 0.70 and 0.68 for survival rate and average internal infection, respectively. The lower heritabilities observed at GL are in part due to environmental factors such as low rainfall before and during the growing season (Table 1, Table S2, Figure S1). Similar heritability trends were observed within spring and winter lines. We calculated phenotypic correlations between emergence count, survival rate and average internal infection across sites in the whole population (Fig. 4). Low correlation coefficients were found between emergence count and survival rate as well as internal infection at both locations. In contrast, highly negative correlations ($> -0.90$) were estimated between survival rate and average internal infection, as expected. Bivariate GBLUP models in a whole population were used to estimate genetic correlations across traits, which followed similar trends and supported a highly antagonistic relationship between survival and internal infection. We further assessed the genetic correlations in between Spring and Winter lines (Table 2). Across the three traits the mean genetic correlation was relatively high at 0.76 (range 0.68–0.90) and, as expected, the highly antagonistic relationship between survival and internal infection was confirmed within seasonality. Genetic correlations within seasonal type and trait across locations were generally also high at 0.83 (range 0.73–0.90, Table S1).

The genetic relatedness between 532 canola lines was investigated with a genomic relationship matrix calculated as in Yang et al. (2010). Two clearly distinguished large populations are apparent which follow the Winter and Spring seasonality pattern (Fig. 1). A third subgroup has higher relationship with Spring lines, but clustered separately. These lines had Spring seasonality but had been derived from Winter lines. We combined this set with Spring lines for all subsequent genomic prediction analysis.

Fig. 3. SNP distribution in the B. napus sub genomes. Regions in red are known to contain blackleg QTL (more detail provided in Table S4). Class I = known QTL regions, Class II = Non-QTL regions.
Table 1. Summary statistics based on spatially adjusted phenotypes (BLUES), broad sense and genomic heritability for emergence count, survival rate and average internal infection for whole population and within spring and winter sets at two sites.

<table>
<thead>
<tr>
<th>Sites</th>
<th>Traits</th>
<th>Min</th>
<th>Max</th>
<th>Mean ± SD</th>
<th>H²†</th>
<th>h²‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole population WL</td>
<td>Emergence count</td>
<td>1</td>
<td>79</td>
<td>31.17 ± 11.08</td>
<td>0.63</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>Survival rate</td>
<td>0</td>
<td>85</td>
<td>23.05 ± 17.46</td>
<td>0.76</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Av. Internal infection</td>
<td>22</td>
<td>100</td>
<td>83.4 ± 13.31</td>
<td>0.74</td>
<td>0.68</td>
</tr>
<tr>
<td>GL</td>
<td>Emergence count</td>
<td>7</td>
<td>29</td>
<td>14.37 ± 3.32</td>
<td>0.38</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>Survival rate</td>
<td>20</td>
<td>85</td>
<td>55.97 ± 15.09</td>
<td>0.54</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Av. Internal infection</td>
<td>18</td>
<td>89</td>
<td>56.52 ± 16.92</td>
<td>0.69</td>
<td>0.52</td>
</tr>
<tr>
<td>Spring WL</td>
<td>Emergence count</td>
<td>1</td>
<td>64</td>
<td>27.36 ± 11.45</td>
<td>0.48</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>Survival rate</td>
<td>4</td>
<td>85</td>
<td>36.44 ± 18.41</td>
<td>0.72</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Av. Internal infection</td>
<td>42</td>
<td>98</td>
<td>78.27 ± 12.87</td>
<td>0.54</td>
<td>0.37</td>
</tr>
<tr>
<td>Winter WL</td>
<td>Emergence count</td>
<td>6</td>
<td>79</td>
<td>32.62 ± 10.6</td>
<td>0.58</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>Survival rate</td>
<td>0</td>
<td>80</td>
<td>17.91 ± 14.08</td>
<td>0.75</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Av. Internal infection</td>
<td>22</td>
<td>100</td>
<td>85.34 ± 12.67</td>
<td>0.68</td>
<td>0.59</td>
</tr>
<tr>
<td>GL</td>
<td>Emergence count</td>
<td>7</td>
<td>29</td>
<td>14.89 ± 3.21</td>
<td>0.39</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>Survival rate</td>
<td>20</td>
<td>84</td>
<td>54.71 ± 14.74</td>
<td>0.59</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>Av. Internal infection</td>
<td>18</td>
<td>89</td>
<td>57.35 ± 16.97</td>
<td>0.66</td>
<td>0.58</td>
</tr>
</tbody>
</table>

†H² = genomic heritability.
‡H² = broad-sense heritability.
§WL = Wickliffe.
¶GL = Green Lake.

Genomic Selection in Whole Population

A total of 532 canola lines and 98,054 SNP markers were used for genomic selection predicting emergence count, survival rate and average internal infection at two sites (Wickliffe and Green Lake, Victoria, Fig. 5 and Table 3). Based on GBLUP, genomic prediction accuracy (Pearson’s correlation of BLUES and GEBVs) varied from 0.19 to 0.74. In the whole population, the highest accuracy was recorded for survival rate (r = 0.74) followed by average internal infection (r = 0.70) at WL and GL sites, while the lowest accuracy was observed for emergence count at GL. The mean BayesR accuracy was lower than that of GBLUP, except for emergence count at GL in the winter set, where it significantly outperformed GBLUP. The trends in genomic selection accuracy closely followed the trends in genomic heritability, confirming that it is a major factor affecting prediction accuracy. Similarly, we calculated the bias as the slope of the regression of corrected phenotypes on GEBVs, which has an expectation of 1. In the whole population, bias did not deviate > 0.33 from 1, indicating that GEBVs were neither over- or underestimated. One observed exception was BayesR for emergence count in the winter lines where the bias was 0.28 (Table S6).

We assessed the prediction accuracy across our two locations by predicting one location’s BLUES with a reference population consisting of BLUES from the second location. The results show that prediction accuracy was reduced to 0.28, 0.43, and 0.56 for emergence, survival, and internal infection at the GL location, respectively (Table 3). This 10 to 20% drop is expected and it indicates that additional multi-environment trials are required to investigate the full impact of G×E interaction on genomic prediction of genomic blackleg resistance. It is important to account for G×E interaction in genomic selection (Sprague et al., 2006; Cowling, 2007). We assessed the proportion of the phenotypic variance explained by G×E interactions in combined location data. The variance attributable to the G×E effect was substantial, but it was always less than 30% of the genetic variance (Table S5). While the number of environments we had available was only two allowing for only limited inference on G×E, this result confirms that G×E was likely one of the reasons for dropping prediction accuracy across locations (Table 2).

We also investigated the relationship between blackleg disease resistance classes (e.g., resistant R, moderately resistant MR) based on in-field phenotyping for survival rate at WL and predicted GEBVs to visualize the effectiveness of genomically predicting blackleg resistance (Fig. 6). We found that varieties classified as R based on in-field performance consistently had high survival GEBVs, which indicates that early selection based solely on genomic predictions would be effective.

Genomic Selection Within and Across Seasonality

Prediction accuracy was investigated with GBLUP, BayesR, and BayesRC models within and across seasonalities (i.e., 326 winter and 206 spring lines, Table 3). Within seasonality, in the traits survival and internal infection, the prediction accuracy was higher in winter lines across prediction models, in part because their reference population size was larger. However, using BayesR the accuracy in winter types was lower for emergence than in GBLUP and BayesRC especially at the GL site. For survival and internal infection BayesR resulted in improved prediction accuracy in spring but lower accuracy in winter lines. Furthermore, we have tested additional scenario by combining all lines in one reference population to predict spring and winter sets using GBLUP and found that combining lines benefitted only the Spring lines for survival and internal infection with no change in accuracy observed for Winter lines across all traits. The inclusion of winter lines in the reference population reduced prediction accuracy in spring lines for emergence count (Table 3).

We tested whether a reference population of only spring lines could be used to predict winter lines (whose phenotypes were masked) and vice versa. Using GBLUP, prediction accuracy across seasonal sets was much lower than within set (Table 3). The accuracy of predicting winter lines with spring lines was slightly higher than using a winter reference to predict spring. The genomic relationship heatmap (Fig. 1) shows a clear differentiation between seasonal groups supporting the conclusion that the two subsets are genetically different and only limited mixing of germplasm has occurred. The bias in
the GEBVs was also higher in these analyses (i.e., greater deviation from 1), because phenotypes of the seasonal set were not included leading to less optimal predictions.

**Genetic Variance Explained by Known Blackleg Resistance QTL Versus Rest of Genome**

Several genomic regions that harbor resistance to blackleg infection were identified in the literature (Table S4). We defined a list of QTLs with relatively large intervals on either side to ensure the potential causative mutations would be captured (Table S4). This resulted in regions encompassing all of chromosome- chr7 (A07) and chr10 (A10) and smaller regions on chromosome chr1 (A01), chr2 (A02), and chr6 (A06) (Fig. 3). A total of 18,886

<table>
<thead>
<tr>
<th>Traits</th>
<th>EmeWL winter</th>
<th>EmeGL winter</th>
<th>SurvWL winter</th>
<th>SurvGL winter</th>
<th>AvintinfWL winter</th>
<th>AvintinfGL winter</th>
</tr>
</thead>
<tbody>
<tr>
<td>EmeWL spring</td>
<td>0.68</td>
<td>0.85</td>
<td>0.09</td>
<td>0.09</td>
<td>-0.06</td>
<td>-0.23</td>
</tr>
<tr>
<td>EmeGL spring</td>
<td>0.74</td>
<td>0.90</td>
<td>0.23</td>
<td>0.13</td>
<td>-0.13</td>
<td>-0.28</td>
</tr>
<tr>
<td>SurvWL winter</td>
<td>0.05</td>
<td>0.12</td>
<td>0.74</td>
<td>0.73</td>
<td>-0.98</td>
<td>-0.82</td>
</tr>
<tr>
<td>SurvGL winter</td>
<td>0.09</td>
<td>0.22</td>
<td>0.87</td>
<td>0.75</td>
<td>-0.80</td>
<td>-0.77</td>
</tr>
<tr>
<td>AvintinfWL winter</td>
<td>-0.05</td>
<td>-0.07</td>
<td>-0.75</td>
<td>-0.80</td>
<td>0.78</td>
<td>0.88</td>
</tr>
<tr>
<td>AvintinfGL winter</td>
<td>-0.08</td>
<td>-0.06</td>
<td>-0.60</td>
<td>-0.73</td>
<td>0.88</td>
<td>0.73</td>
</tr>
</tbody>
</table>

†Eme = emergence count.
‡WL = Wickliffe.
§GL = Green Lake.
¶Surv = Survival rate.
#Avintinf = Average internal infection.

---

**Table 2. Genetic correlations estimated with bivariate GBLUP between spring and winter lines.**

<table>
<thead>
<tr>
<th>Traits</th>
<th><strong>Eme</strong>WL†</th>
<th><strong>Eme</strong>GL‡</th>
<th><strong>Surv</strong>WL‡</th>
<th><strong>Surv</strong>GL§</th>
<th><strong>Avintinf</strong>WL‡</th>
<th><strong>Avintinf</strong>GL§</th>
</tr>
</thead>
<tbody>
<tr>
<td>EmeWL winter</td>
<td>0.68</td>
<td>0.85</td>
<td>0.09</td>
<td>0.09</td>
<td>-0.06</td>
<td>-0.23</td>
</tr>
<tr>
<td>EmeGL winter</td>
<td>0.74</td>
<td>0.90</td>
<td>0.23</td>
<td>0.13</td>
<td>-0.13</td>
<td>-0.28</td>
</tr>
<tr>
<td>SurvWL winter</td>
<td>0.05</td>
<td>0.12</td>
<td>0.74</td>
<td>0.73</td>
<td>-0.98</td>
<td>-0.82</td>
</tr>
<tr>
<td>SurvGL winter</td>
<td>0.09</td>
<td>0.22</td>
<td>0.87</td>
<td>0.75</td>
<td>-0.80</td>
<td>-0.77</td>
</tr>
<tr>
<td>AvintinfWL winter</td>
<td>-0.05</td>
<td>-0.07</td>
<td>-0.75</td>
<td>-0.80</td>
<td>0.78</td>
<td>0.88</td>
</tr>
<tr>
<td>AvintinfGL winter</td>
<td>-0.08</td>
<td>-0.06</td>
<td>-0.60</td>
<td>-0.73</td>
<td>0.88</td>
<td>0.73</td>
</tr>
</tbody>
</table>

†Eme = emergence count.
‡WL = Wickliffe.
§GL = Green Lake.

---

**Table 3. Prediction accuracy of different cross-validation scenarios using genomic best linear unbiased prediction (GBLUP), BayesR and BayesRC.**

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Emergence count</th>
<th>Survival rate</th>
<th>Av. int. infection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WL†</td>
<td>GL‡</td>
<td>WL</td>
</tr>
<tr>
<td>Whole population, GBLUP</td>
<td>0.43</td>
<td>0.35</td>
<td>0.74</td>
</tr>
<tr>
<td>Whole population, BayesR</td>
<td>0.36</td>
<td>0.33</td>
<td>0.70</td>
</tr>
<tr>
<td>Whole population, BayesRC</td>
<td>0.34</td>
<td>0.34</td>
<td>0.75</td>
</tr>
<tr>
<td>Within Winter, GBLUP</td>
<td>0.40</td>
<td>0.19</td>
<td>0.67</td>
</tr>
<tr>
<td>Within Winter, BayesR</td>
<td>0.38</td>
<td>0.38</td>
<td>0.63</td>
</tr>
<tr>
<td>Within Winter, BayesRC</td>
<td>0.42</td>
<td>0.41</td>
<td>0.69</td>
</tr>
<tr>
<td>Within Spring, GBLUP</td>
<td>0.45</td>
<td>0.44</td>
<td>0.52</td>
</tr>
<tr>
<td>Within Spring, BayesR</td>
<td>0.41</td>
<td>0.35</td>
<td>0.66</td>
</tr>
<tr>
<td>Within Spring, BayesRC</td>
<td>0.39</td>
<td>0.30</td>
<td>0.59</td>
</tr>
<tr>
<td>All to predict Winter, GBLUP</td>
<td>0.40</td>
<td>0.19</td>
<td>0.65</td>
</tr>
<tr>
<td>All to predict Winter, BayesR</td>
<td>0.36</td>
<td>0.29</td>
<td>0.76</td>
</tr>
<tr>
<td>All to predict Winter, BayesRC</td>
<td>0.10</td>
<td>0.13</td>
<td>0.26</td>
</tr>
<tr>
<td>Winter to predict Spring, GBLUP</td>
<td>0.10</td>
<td>0.09</td>
<td>0.10</td>
</tr>
<tr>
<td>Winter to predict Spring, BayesR</td>
<td>0.09</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>Winter to predict Spring, BayesRC</td>
<td>0.35</td>
<td>0.54</td>
<td>0.58</td>
</tr>
<tr>
<td>WL to predict GL site, GBLUP</td>
<td>0.28</td>
<td>0.43</td>
<td>0.56</td>
</tr>
</tbody>
</table>

†WL = Wickliffe site.
‡GL = Green Lake site.

††Whole population; Prediction models are in subscript.
SNPs were located in known QTL regions. We calculated two separate genomic relationship matrices, one with SNPs in QTL regions (Class I) and another (Class II) with all remaining SNP \( (N = 79,168) \) (Table S4). Genetic variances were then estimated by simultaneously fitting the two genomic relationship matrices in a GBLUP model. The proportion of the genetic variance in survival and internal infection explained by the QTL regions was much smaller than the remaining regions ranging from 0.02 to 0.33 (Table S7). Thus, unknown or undiscovered regions explain the majority of the genetic variance for survival and internal infection in our study. Blackleg QTL regions explained no genetic variance in emergence count, as expected (Fig. 7 and Table S7).

**Genomic Selection Using Known QTL Regions as Priors**

We investigated whether classifying SNPs in known blackleg QTL regions as a separate class in BayesRC would improve prediction accuracy and SNP effect estimates in the whole population as well as Spring and Winter sets. In the traits survival and internal infection, BayesRC achieved a higher accuracy than either GBLUP or BayesR for Winter and Spring types. This trend was especially pronounced for Spring lines with up to a 0.16 (Spring, survival GL) advantage in accuracy observed (Fig. 5). The accuracy for emergence was not lifted by defining blackleg resistance QTL as a separate class in BayesRC, as expected, in some cases the accuracy for emergence was actually reduced by BayesRC (Table 3, Table S5a).

The average number of SNP estimated per non-zero distribution in the whole population (variances 0.0001σ², 0.001σ², and 0.01σ²) of BayesR and BayesRC models are presented in Table 4a,b and Table S8a,b. The utilization of SNPs in QTL regions as a separate class in BayesRC resulted in a larger number of SNPs in the large and small proportion of genetic variance explained distributions, but it generally reduced the number of SNPs in the medium distribution when compared to BayesR. The increase in SNPs with effects seems to have lifted genomic prediction accuracies from using BayesRC. Interestingly, while a substantial number of SNPs with effect seem to fall into QTL regions, they collectively do not seem to explain a large proportion of the genetic variance (Fig. 7). The general trend toward higher SNP effects and its potential to improve QTL mapping is also evident in Manhattan plots of scaled SNP effects across traits and sites (Fig. 2 and Figure S5).

---

**Fig. 5.** Comparison of genomic prediction accuracy between GBLUP, BayesR and BayesRC models based within spring and winter sets for survival and internal infection at Green Lake and Wickliffe locations.
DISCUSSION

We have investigated genomic prediction in a large diverse set of spring and winter type canola accessions. The prediction accuracy for survival rate and internal infection was moderate to high overall, especially when using the whole population as the reference population. Analyses within seasonality showed higher accuracy in winter than spring accessions in keeping with their respective reference population sizes. Strict prediction across seasonal types exhibited limited accuracy. However, spring line prediction accuracy was increased by including winter lines in the reference population. Known blackleg QTL were shown to explain less than a third of the genetic variance of the trait. Incorporating information about these QTL regions in genomic prediction has improved prediction accuracy using the BayesRC approach in our data set.

Heritability and Environmental Conditions

The broad-sense heritability for survival and internal infection was higher than it was for seedling emergence, suggesting that the latter had a larger environmental influence. Overall our estimates agreed with literature estimates (Hatzig et al., 2015; Larkan et al., 2016). Seedling emergence can be affected by age of seed, among many other factors. We attempted to control emergence by using seed that was less than 5 yr old, and, if seed was older than 5 yr, germination tests of accessions were conducted to ensure there was sufficient germination (> 70% germination) in field trials. Nevertheless, it was still possible that some seed age-based emergence variability might be affecting our estimates for emergence.

Heritabilities for the GL site were lower than for the WL trial, again suggesting a greater influence of environment. Indeed, Victoria experienced very dry growing conditions in 2015 and the GL site was especially dry during the growing season. The harsher GL environment would have contributed toward the increased variability at that site, especially for emergence inoculum load. Such heritability differences across regions have been documented in the literature (Hatzig et al., 2015; Bouchet et al., 2016; Larkan et al., 2016).

Similarly, we estimated genomic heritability to evaluate how much of the phenotypic variance can be explained by the genetic markers. We observed a higher genomic heritability for average internal infection and survival rate than for emergence count in the whole population and within seasonality (Table 1). Narrow sense heritabilities were generally close to broad-sense estimates, which indicates that our marker density was high enough to capture most genetic variation and that all three traits have a substantial additive genetic component. Nevertheless, a larger gap was observed for the GL location, this could indicate a greater proportion of interaction effects possibly due to unfavorable weather.

Genomic prediction

We have evaluated the following genomic prediction scenarios using GBLUP, BayesR, BayesRC and cross-validation: combined whole population and solely within seasonal types (spring and winter lines), whereas the GBLUP method were used for the remaining scenarios (as shown in Table 3). As expected, prediction accuracy followed the general trend of heritabilities, with higher
The accuracy observed at WL and for survival and internal infection (range 0.48–0.74 in whole population, Table 3, Fig. 5) in GBLUP and BayesRC methods. The reduced accuracy for emergence when compared to other traits was also observed by Jan et al. (2016). Using both seasonal types resulted in better genomic prediction for the spring types for blackleg traits. No advantage from adding spring types to the reference population was observed for winter accessions, this suggests that the advantage of adding less related lines diminishes as sample size within seasonality increases in this data set. Counterintuitively, the prediction of emergence for spring accessions was negatively impacted by adding winter germplasm to the reference. Further study is needed to understand whether this could be due to differences in genetic trait architecture for emergence between the two populations. Genetic correlations within traits between spring and winter pools were generally high (>0.68, Table 2), nevertheless, their deviation from one indicates that different genetic factors may be present in the pools. Across prediction models, when predicting purely within seasonal types, the spring set achieved lower accuracy than the winter population, which we expect to be primarily due to its lower samples size (206 versus 326 accessions). Furthermore, prediction accuracy with exclusive reference populations from the other seasonality was severely reduced. This is supported by the substantial genetic correlations observed between the seasonality pools. While strict across-pool prediction should be avoided in breeding programs because the populations are not sufficiently related (Fig. 1), supplementing additional reference populations with accessions from outside their seasonality may be of benefit to prediction accuracy. The effect of relatedness observed here, has been previously observed in plant and animal populations, and has highlighted the need to genomically assess populations prior to wide-spread use of genomic selection (Daetwyler et al., 2008; Hayes et al., 2009; de Roos et al., 2009; Saatchi et al., 2011; Lin et al., 2014). Our results have confirmed the well-known fact that spring and winter germplasm has been maintained primarily as quite separate populations. Establishing reference populations in both sets would allow for the mining of alleles from one set to introgress into the other, thereby accessing a larger pool of genetic diversity.

Most scenarios resulted in relatively small prediction bias (slope of the regression of phenotypes on GEBV), which is the deviation from its expectation of 1, indicating that our GEBVs were not substantially over or underestimated. When predicting winter lines with a spring reference population and vice versa, GEBVs were in some cases more biased. This is also expected as allele frequencies differ between the populations and, thus, the variance explained will also deviate. The importance of unbiased predictions is greatest when combining GEBVs with estimates from separate analyses (e.g., pedigree or simple phenotypes), as it will affect the ranking of individuals quite substantially. Few crop studies have reported genomic prediction bias, but it should be considered when evaluating genomic selection approaches (Garrick et al., 2009; Patry and Ducrocq, 2011).

We compared prediction accuracies obtained across three methods: GBLUP, BayesR, and BayesRC. For survival and internal infection, GBLUP and BayesR achieved similar prediction accuracy, but for emergence GBLUP performed better in the whole-population scenario. It is not uncommon that genomic selection methods rank differently across traits, but when traits are quite quantitative the differences in accuracy across models tend to be rather small. Wurschum et al. (2014) observed higher accuracy for oil content, protein and glucosinolate content using RR-BLUP (a method very similar to GBLUP) than BayesB. Overall, the accuracies observed were similar to that of other canola genomic selection studies, even though our dataset was likely quite a bit more genetically diverse (Wurschum et al., 2014; Zou et al., 2016). While our study is the first on

**Fig. 7.** Proportion of genetic variance explained by known QTL and non-QTL Regions in *B. napus.*
Table 4a. Average number of QTL estimated per non-zero distribution across replicate MCMC chains of BayesR models.

<table>
<thead>
<tr>
<th>Sites</th>
<th>Traits</th>
<th>Number of QTL per distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(N(0, 0.001 \sigma^2_s))</td>
</tr>
<tr>
<td>Wickliffe</td>
<td>Emergence count</td>
<td>3540</td>
</tr>
<tr>
<td></td>
<td>Survival rate</td>
<td>2487</td>
</tr>
<tr>
<td></td>
<td>Av. Internal infection</td>
<td>2099</td>
</tr>
<tr>
<td>Green Lake</td>
<td>Emergence count</td>
<td>2401</td>
</tr>
<tr>
<td></td>
<td>Survival rate</td>
<td>3298</td>
</tr>
<tr>
<td></td>
<td>Av. Internal infection</td>
<td>2881</td>
</tr>
</tbody>
</table>

Table 4b. Average number of QTL estimated per class of non-zero distribution across replicate MCMC chains of BayesRC models.

<table>
<thead>
<tr>
<th>Sites</th>
<th>Traits</th>
<th>Number of QTL Effects per Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(N(0, 0.001 \sigma^2_s))</td>
</tr>
<tr>
<td>Wickliffe</td>
<td>Emergence count</td>
<td>3228 (769)†</td>
</tr>
<tr>
<td></td>
<td>Survival rate</td>
<td>2097 (1042)</td>
</tr>
<tr>
<td></td>
<td>Av. Internal infection</td>
<td>2323 (975)</td>
</tr>
<tr>
<td>Green Lake</td>
<td>Emergence count</td>
<td>2455 (1776)</td>
</tr>
<tr>
<td></td>
<td>Survival rate</td>
<td>2171 (2268)</td>
</tr>
<tr>
<td></td>
<td>Av. Internal infection</td>
<td>1819 (2334)</td>
</tr>
</tbody>
</table>

† Italics (*) are SNPs in Class I which were defined as all SNP in previously known QTL regions.

Genomic prediction of blackleg resistance in canola, biotic disease resistance have been successfully predicted in other crops (Daetwyler et al., 2014; Rutkoski et al., 2015; de Azevedo Peixoto et al., 2017).

Biological prior information could be based on functional annotations, such as certain classes of variants (e.g., missense mutations), amino acid substitutions affecting protein function, and expressed regions from RNA sequencing, or they could be just regions previously discovered to harbor important QTL, which technically are not true biological priors because they are based on statistical association (Perez-Enciso et al., 2015; MacLeod et al., 2016; VanRadan et al., 2017). We mined the literature for known blackleg resistance QTL and defined SNP within those regions as a separate set (Table S4). Fitting two genomic relationship matrices (Class I, SNPs within QTL regions; Class II, SNPs outside these regions, Fig. 3) in the same GBLUP model revealed that known QTL regions explained very little of the genetic variance across all traits for the WL trial, but these regions explained up to ~30% of the genetic variance at GL. This suggests that the GL site potentially had different L. maculans races than those present at WL. Both sites had the same canola variety grown the previous year as the stubble source for inoculum, in an attempt to standardize virulence across sites. This may have been only partly successful, which is supported by the genetic correlations across locations deviating from 1. Furthermore, the stubble source cultivar contained the Rlm1 major resistance gene. Consequently, the fungal population that developed on this stubble would have had a high proportion of isolates that would attack and erode the Rlm1 resistance in any cultivar the following year. Therefore, the Rlm1 containing lines in our core sample that would have been resistant on any other stubble type, could have displayed varied levels of susceptibility in our study. The impact of cultivar stubble type on blackleg genomic prediction accuracy will need further investigation.

The cultivar Darmor-bzh used to generate the B. napus reference assembly contains only the major resistance gene Rlm9 (Delourme et al., 2004). Consequently, we would not have mapped reads against the other 10 B. napus major resistance genes in our genotyping-by-sequencing assay and this could have affected our results. However, there is significant short-range LD within the data set and one would expect that SNP close to the potentially missing resistance genes would be in sufficient LD with the mutation to account for it. Our findings show that known blackleg QTL only explain a minimal or moderate proportion of the genetic variance in diverse populations, but that genomic selection can track additional variation and use it to accurately predict genomic breeding values for blackleg traits. This points to an important role that genomic selection will play in canola germplasm improvement as it makes use of many “undiscovered” mutations of smaller effect conferring blackleg resistance. These undiscovered mutations could also lead to more durable blackleg resistance, as the pathogen is unlikely to break quantitative resistance as it has with qualitative major resistance genes.

BayesRC is a BayesR extension that incorporates biological prior information such as QTL information into genomic prediction. BayesRC has been shown to increase the accuracy of genomic prediction compared to BayesR when there is adequate prior information (MacLeod et al., 2016). Defining SNP in known QTL regions as Class I and all other SNP as Class II in BayesRC increased the accuracy for survival and internal infection respective to GBLUP, especially in Spring lines. This highlights several learnings. First, when a proportion of the genetic variance can be captured by prior information (in our case QTL regions), BayesRC can utilize it and potentially increase prediction accuracy. Second, exploring whether features explain genetic variance with simple mixed models is a useful first step to demonstrate their utility (Sørensen et al., 2017). Third, our results confirm in canola that features may act in a location specific manner, especially for disease resistance with its close interactions between pathogen, host, and environment.

The accuracy of genomic prediction achieved in this diverse germplasm indicates its high utility for canola breeding programs. Resistant accessions consistently had the highest GEBVs and susceptible varieties consistently had the lowest GEBVs (Fig. 6). This allows for targeted use of GEBVs for selection early in the breeding cycle through elimination of susceptible lines. In addition to crossing best with best, moderately susceptible and resistant lines that potentially contain quantitative resistance could be retained and crossed to enrich for durable resistance over time. Furthermore, this quantitative resistance could be explicitly
targeted by calculating GEBVs that exclude known QTL effect regions, which has been termed residual resistance by Eckardt, (2002). However, this assumes that all resistance in those regions is due to major resistance QTL which is unlikely. Regardless of the strategy chosen, genomic selection early in the breeding cycle would speed up the germlasm resistance enrichment process substantially.

Supplemental Materials

Supplemental Figure S1. Weather data at Wickliffe and Green Lake area, 2015.

Supplemental Figure S2. Pearson’s correlation coefficient between observed values and imputed genotypes across 133,783 SNP markers when using FImpute. SNP with imputation accuracy <0.5 were removed.

Supplemental Figure S3. Workflow from genotype and phenotype quality control to genomic prediction.

Supplemental Figure S4. Pair-wise linkage disequilibrium (mean $R^2$ per 10kb window) for winter and spring types within the 19 Brassica napus chromosomes.

Supplemental Figure S5. Manhattan plot of marker effects estimated for emergence count using BayesR (BR) and BayesRC (BRC). The mean absolute value of marker effect calculated across replicates. WL = Wickliffe site and GL = Green Lake site.

Supplemental Table S1. Genomic correlation coefficient within spring (below diagonal) and winter canola lines (above diagonal) across sites.

Supplemental Table S2. Weather data at Wickliffe and Green Lake area, 2015.

Supplemental Table S3. Summary of quality control thresholds applied on canola genotype dataset.

Supplemental Table S4. List of resistant genes for Leptosphaeria maculans and genomic regions included in BayesRC Class I category.

Supplemental Table S5. Variance components for combined sites (Wickliffe and Green Lake sites).

Supplemental Table S6. A summary of prediction bias (slope) between different cross-validation scenarios using genomic best linear unbiased prediction (GBLUP).

Supplemental Table S7. Phenotypic and additive genetic variance and narrow-sense genomic heritability for emergence count, survival rate and average internal infection across sites by fitting GRMQTLS (18,886 SNPS) and GRMelse (79,168 SNPS) as a random effect.

Supplemental Table S8a. Proportion of non-zero marker effects estimated per distribution in BayesR.

Supplemental Table S8b. Proportion of non-zero marker effects estimated per distribution, within each class of the BayesRC.

Conflict of Interest Disclosure

The authors declare that there is no conflict of interest.

Acknowledgments

The authors thank Agriculture Victoria Services and Agriculture Victoria for funding, MF gratefully acknowledges PhD scholarship funding from La Trobe University and MM acknowledges her Australian Government Research Training Program Scholarship. The technical support from Sandra Mayber, Bruce Holding, and Glen Munro at Agriculture Victoria Horsham is gratefully acknowledged. We thank the Australian Grains GeneBank and, in particular, Sally Nordon for provision of canola germplasm.

References


R Core Team. 2014, Vienna, Austria, https://www.r-project.org/.

Brassica napus


