




Article

# Evaluation of Genomic Prediction for PasmO Resistance in Flax

Liqliang He <sup>1,2</sup>, Jin Xiao <sup>2</sup>, Khalid Y. Rashid <sup>3</sup>, Gaofeng Jia <sup>4</sup>, Pingchuan Li <sup>3</sup>, Zhen Yao <sup>3</sup>,  
Xiue Wang <sup>2</sup>, Sylvie Cloutier <sup>1,\*</sup> and Frank M. You <sup>1,2,\*</sup> 

<sup>1</sup> Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, ON K1A 0C6, Canada; liqliang.he@canada.ca

<sup>2</sup> State Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Agriculture, Nanjing Agricultural University/JiangSu Collaborative Innovation Center for Modern Crop Production, Nanjing 210095, China; xiaojin@njau.edu.cn (J.X.); xiuew@njau.edu.cn (X.W.)

<sup>3</sup> Morden Research and Development Centre, Agriculture and Agri-Food Canada, Morden, MB R6M 1Y5, Canada; khalid.rashid@canada.ca (K.Y.R.); lipingchuan@gmail.com (P.L.); zhen.yao@canada.ca (Z.Y.)

<sup>4</sup> Crop Development Centre, University of Saskatchewan, Saskatoon, SK S7N 5A8, Canada; gaofeng.jia@usask.ca

\* Correspondence: sylvie.cloutier@canada.ca (S.C.); frank.you@canada.ca (F.M.Y.); Tel.: +1-613-759-1744 (S.C.); +1-613-759-1539 (F.M.Y.)

Received: 28 November 2018; Accepted: 11 January 2019; Published: 16 January 2019



**Abstract:** PasmO (*Septoria linicola*) is a fungal disease causing major losses in seed yield and quality and stem fibre quality in flax. PasmO resistance (PR) is quantitative and has low heritability. To improve PR breeding efficiency, the accuracy of genomic prediction (GP) was evaluated using a diverse worldwide core collection of 370 accessions. Four marker sets, including three defined by 500, 134 and 67 previously identified quantitative trait loci (QTL) and one of 52,347 PR-correlated genome-wide single nucleotide polymorphisms, were used to build ridge regression best linear unbiased prediction (RR-BLUP) models using pasmo severity (PS) data collected from field experiments performed during five consecutive years. With five-fold random cross-validation, GP accuracy as high as 0.92 was obtained from the models using the 500 QTL when the average PS was used as the training dataset. GP accuracy increased with training population size, reaching values >0.9 with training population size greater than 185. Linear regression of the observed PS with the number of positive-effect QTL in accessions provided an alternative GP approach with an accuracy of 0.86. The results demonstrate the GP models based on marker information from all identified QTL and the 5-year PS average is highly effective for PR prediction.

**Keywords:** genomic selection; genomic prediction; genotyping by sequencing; pasmo resistance; pasmo severity; quantitative trait loci; single nucleotide polymorphism; *Septoria linicola*; flax

## 1. Introduction

Flax (*Linum usitatissimum* L.) is an important food and fibre crop cultivated and grown in cooler regions of the world, such as Canada [1]. PasmO, elicited by the fungus *Septoria linicola*, is one of the most widespread diseases of flax, causing reductions in seed and oil yield, as well as fibre quality and durability [2]. Developing resistant cultivars is the most viable and effective option to control this disease that has become widespread in all flax production areas of North America and other parts of the world. Resistance to pasmo has a low heritability [3] and is quantitatively inherited [4]. Large variations in pasmo disease severity were observed in the flax core collection, which can be capitalized upon to develop resistant cultivars [3]. Phenotypic recurrent selection is typically used to develop cultivars with improved resistance and selection is usually carried out based on phenotypic

assessments of resistance in field conditions [5]. However, field assessment of pasmo severity (PS) in germplasm and breeding lines is costly and, is heavily influenced by the environments due to strong genotype  $\times$  environment (G  $\times$  E) interactions [3,4].

With the advancements in molecular marker development over the last decade, efforts to use marker-assisted breeding strategies have been pursued. One such strategy involves identifying quantitative trait loci (QTL) in biparental mapping populations and using markers to efficiently backcross QTL into elite breeding materials [6]. This so-called marker-assisted recurrent selection (MARS) or simply marker-assisted selection (MAS) characterizes many breeding programs that employ molecular markers to select non-phenotyped individuals for crossing and downstream selection of segregating populations [7]. This method is suitable for the selection of monogenic or oligo-genic architectures but has limited use for quantitative traits controlled by many genes of smaller effects [8]. Genomic selection (GS) or prediction (GP) is an alternative marker-assisted breeding strategy better suited to polygenic quantitative traits, especially those with low heritability, because it makes use of all marker effects across the entire genome to calculate genomic estimated breeding values (GEBVs) [9] for individual plant selection [9,10].

In GP, a training population (TP) is genotyped with genome-wide markers and phenotyped for the trait(s) under selection; statistical models that best predict the breeding values from the marker data are then applied to select non-phenotyped germplasm. GP has been used to select for disease resistance in several crops such as *Fusarium* head blight (FHB) in wheat, a typically quantitatively inherited trait with predominantly additive genetic variation, where GP had a significantly higher accuracy than pedigree-based information alone [11]. GP feasibility has also been studied for selection of wheat rust resistance and was found particularly effective when validation lines had at least one which is close to the reference lines [12]. The implementation of GP on northern leaf blight, a complex genetic architecture trait in maize, resulted in superior gains and reduced breeding cycle time to  $\leq 80\%$  of the phenotypic cycle [13]. Despite the many successful examples, the use of GP to improve disease resistance in crops has been challenging for two reasons: (i) selection for major resistance genes can be ephemeral due to changes in pathogen races; and (ii) breeding for minor resistance genes with small effects may face the remarkable complexities encountered in GP [14].

The fast-evolving genotyping platforms have been a game-changer in the implementation of GP, allowing the production of large numbers of genome-wide markers, whereas progresses in phenotyping were not associated with similar cost reduction or quantum leaps in throughput. Given the number of markers ( $p$ ) and sample size ( $n$ ) in a given population, there are many more  $p$  effects to be estimated than the  $n$ , leading to an infinite number of possible marker effect estimates [15], that is, the so-called “large  $p$ , small  $n$  problem” ( $p \gg n$ ) when applying markers to predict phenotypes [11]. Several GP statistical models have been proposed to address this issue [16]. For example, the ridge-regression best linear unbiased prediction (RR-BLUP) is a mixed linear model that considers markers as random effects. Covariance between markers is considered to be zero and the marker variance is assumed to be the total genetic variance divided by the number of markers. The variance is assumed to be equal for all markers, allowing many more marker effects to be estimated than there are phenotypic records [17]. Unlike RR-BLUP, the Bayesian LASSO (BL) assumes markers to have unequal variances and, performs continuous shrinkage and variable selection simultaneously, with small-effect markers shrinking more severely than larger-effect loci. In the  $p \gg n$  setting, LASSO will select at most  $n - 1$  variables and set the effects of the remaining predictors at zero [18]. Although the problem is solved statistically in these models, improving the accuracy and efficiency of GP by reducing the number of genome-wide markers would be advantageous because any increment in the TP size comes at a cost [19–22]. Genome-wide association study (GWAS) is an approach to identify genome-wide markers linked to QTL, resulting in a limited number of favourable genetic loci responsible for traits of interest [23]. For example, GP of crown rust resistance in *Lolium perenne* demonstrated GWAS’s ability to identify and rank markers, which enabled the identification of a small subset of single nucleotide polymorphisms (SNPs) that

could achieve predictive abilities close to that attained using the complete marker set [24]. Utilization of GWAS removes a large proportion of unrelated markers and in the construction of prediction models.

The only GP empirical study published to date in flax, which used bi-parental populations for yield, oil content and fatty acid composition traits, indicated that GP could increase genetic gain per unit time in linseed breeding. The GP results significantly exceeded those from direct phenotypic selection, especially for traits with low broad-sense heritability [25]. Resistance to flax pasmo is polygenic. Our previous study reported 500 non-redundant QTL for PR from 370 diverse flax accessions of a core collection based on five-year pasmo field assessments; of those, 134 QTL were statistically stable in all five years and 67 had relatively stable and large effects [4].

The objective of this study was to evaluate the potential of QTL markers in GP and compare the GP efficiency affected by different markers, including genome-wide SNPs and QTL markers, to provide a realistic and highly accurate model for germplasm evaluation and parent selection in pasmo resistance breeding.

## 2. Results

### 2.1. Evaluation of Pasma Resistance

PS ratings at green boll stage or maturity across five consecutive years were similar but on average PS ratings in 2014 and 2016 were higher than those in other years (Table 1). They had single peak distributions but skewed towards high PS ratings except for those in 2014 (Figure 1). Scatter plots of PS ratings between years indicated strong genotype  $\times$  year interaction even though statistically significant correlations of PS ratings between years were observed (Figure 1), as shown in the variance analysis results in the previous study [4]. However, the Pearson correlations of 5-year averages of PS ratings (PS-mean) with those in individual years ( $r = 0.72$ – $0.83$ ) were much higher than the Pearson correlations between individual years ( $r = 0.31$ – $0.62$ ) (Figure 1), implying that the mean PS ratings over multiple years or environments were a more suitable data set than individual year's data sets for model construction of genomic prediction.

**Table 1.** Pasma severity of 370 flax accessions across five years in the field condition.

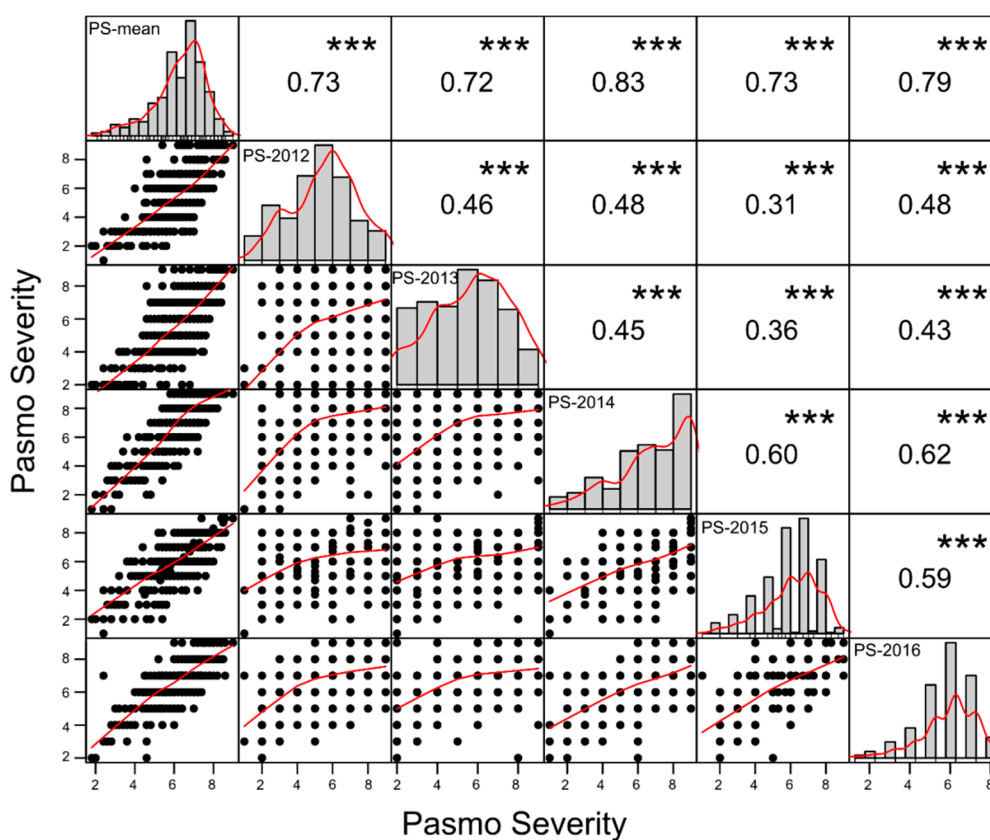
Data Set	$\bar{x} \pm s$	Range	CV (%)
PS-2012	5.57 $\pm$ 1.86	1.00–9.00	32.76
PS-2013	5.69 $\pm$ 1.91	2.00–9.00	33.20
PS-2014	6.86 $\pm$ 2.07	1.00–9.00	29.41
PS-2015	6.11 $\pm$ 1.55	1.00–9.00	25.44
PS-2016	6.72 $\pm$ 1.37	2.00–9.00	20.39
PS-mean	6.22 $\pm$ 1.32	1.80–9.00	21.27

$\bar{x}$ : average pasmo severity across five years;  $s$ : standard deviation; CV: coefficient of variation.

### 2.2. Evaluation of Marker Sets Used in Genomic Prediction

Four marker sets were used for GP of pasmo resistance. The first marker set contained 52,347 genome-wide SNPs (SNP-52347) that were correlated to the five-year average PS and the PS of the five individual years at a  $10^{-5}$  probability level [4]. The other three marker sets were the 500 unique QTL (SNP-500QTL), the 134 QTL statistically stable over five consecutive years (SNP-134QTL) and the 67 stable and relatively large-effect QTL (SNP-67QTL) sets previously identified [4]. The SNP-500QTL dataset comprises markers for all small- or large-effects, including QTL stable across environments and environment-specific QTL identified using three single-locus and seven multi-locus statistical models and all six phenotypic datasets (Figure 2). The SNP-134QTL dataset is a subset of the SNP-500QTL dataset whereas SNP-67QTL is a subset of the former; all SNP-500QTL markers were included in SNP-52347. These four marker sets explained 54%, 72%, 27% and 29% of the phenotypic variation of the five-year PS average (PS-mean), respectively; these values exceeded those of the individual year PS

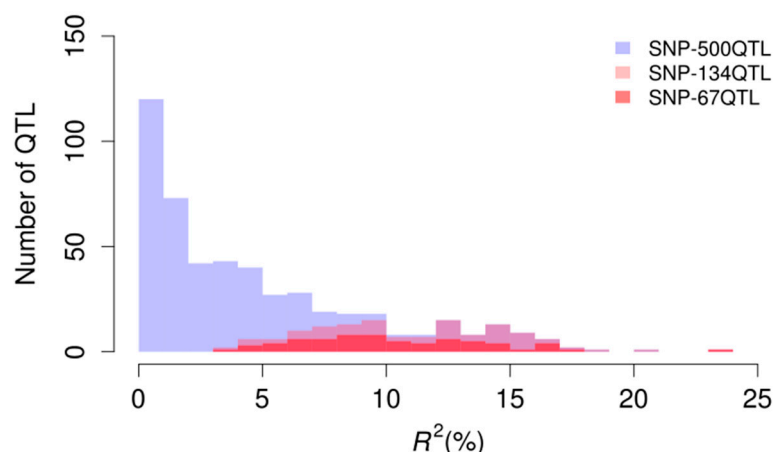
data (Table 2). Although SNP-500QTL was a subset of SNP-52347, this marker set explained a greater percentage of the phenotypic variation for PS than SNP-52347 for all datasets.



**Figure 1.** Dot plots (lower triangle), histograms (diagonal) and Pearson correlations (upper triangle) between six pasmo severity datasets. Best curves are fitted in dot plots and histograms. \*\*\* represents significance at the <0.001 probability level.

**Table 2.** Phenotypic variation of pasmo severity (PS) ( $h^2 \pm s$ ) explained by the four marker sets.

PS Dataset	Marker Set			
	SNP-500QTL	SNP-134QTL	SNP-67QTL	SNP-52347
PS-mean	0.72 ± 0.04	0.27 ± 0.05	0.29 ± 0.05	0.54 ± 0.07
PS-2012	0.64 ± 0.06	0.18 ± 0.05	0.16 ± 0.04	0.43 ± 0.08
PS-2013	0.63 ± 0.06	0.12 ± 0.04	0.12 ± 0.04	0.38 ± 0.08
PS-2014	0.65 ± 0.06	0.23 ± 0.05	0.20 ± 0.05	0.45 ± 0.08
PS-2015	0.56 ± 0.06	0.20 ± 0.05	0.17 ± 0.04	0.44 ± 0.09
PS-2016	0.53 ± 0.06	0.18 ± 0.05	0.18 ± 0.05	0.38 ± 0.07



**Figure 2.** Distribution of  $R^2$  (%) (phenotypic variation explained by individual QTL) in the three QTL marker sets.

### 2.3. Accuracy of Genomic Prediction in Relation to Marker Sets and Pasma Severity Datasets

Genomic prediction models were constructed using RR-BLUP with pairwise combinations of the four marker sets and the six PS datasets. Statistical models for the 24 combinations were generated and evaluated for their accuracy ( $r$ ) and relative efficiency ( $RE$ ) using a five-fold random cross-validation scheme (Table 3).  $RE$  represents the relative efficiency of GP over direct phenotypic selection which depends on the heritability of a selective trait. Direct phenotypic selection for a trait was considered to have a baseline efficiency of 1. Thus,  $RE$  values greater than 1 indicate GP models more efficient than direct phenotypic selection in one selection cycle [25–27]. Analysis of variance (ANOVA) (Table S1) indicated that  $r$  and  $RE$  both significantly differed among the four marker sets and the six PS datasets; there was also a significant interaction effect between marker sets and PS datasets (Table S1). Owing to the significant marker  $\times$  phenotype dataset interaction, multiple comparisons of the 24 combinations were performed. For all marker sets, the PS-mean models significantly outperformed those based on individual year datasets (Table 3). The SNP-500QTL marker set models generated significantly higher  $r$  and  $RE$  values than any other marker sets (Figure 3). Interestingly, the SNP-67QTL derived models produced slightly but significantly higher values of  $r$  and  $RE$  than SNP-134QTL models. The highest  $r$  and  $RE$  values were obtained for models combining the SNP-500QTL and PS-mean datasets (Table 3, Figure 3). Intriguingly, the SNP-52347 models yielded the lowest  $r$  and  $RE$  values despite including all QTL markers (Table 3, Figure 3); both BL and Bayesian ridge regression (BRR) corroborated this finding (Figure S1). No significant differences in  $r$  and  $RE$  values were observed among the three statistical models: RR-BLUP, BL and BRR (Figure S1).

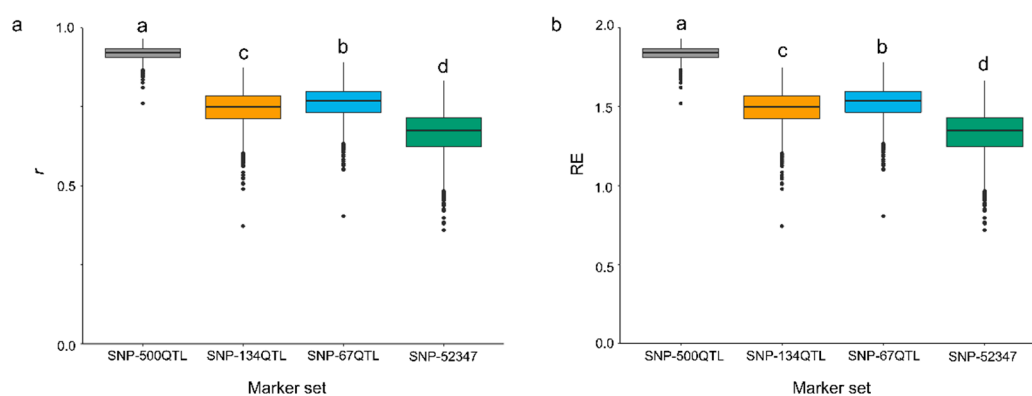
### 2.4. Sample Size of Training Populations versus Genomic Prediction Accuracy

To find an optimal size for the TP, the relationship between TP size and prediction accuracy was analysed. TPs of various sizes from 18 to 351, corresponding to 5% to 95% of the total 370 accessions, were used to build models with the SNP-500QTL marker set and the PS-mean phenotypic dataset. The prediction accuracy significantly increased for TP sizes up to 100, followed by smaller accuracy gains with every additional TP size increments (Figure 4). A GP accuracy  $>0.9$  was obtained once the TP size reached 185 (Figure 4).

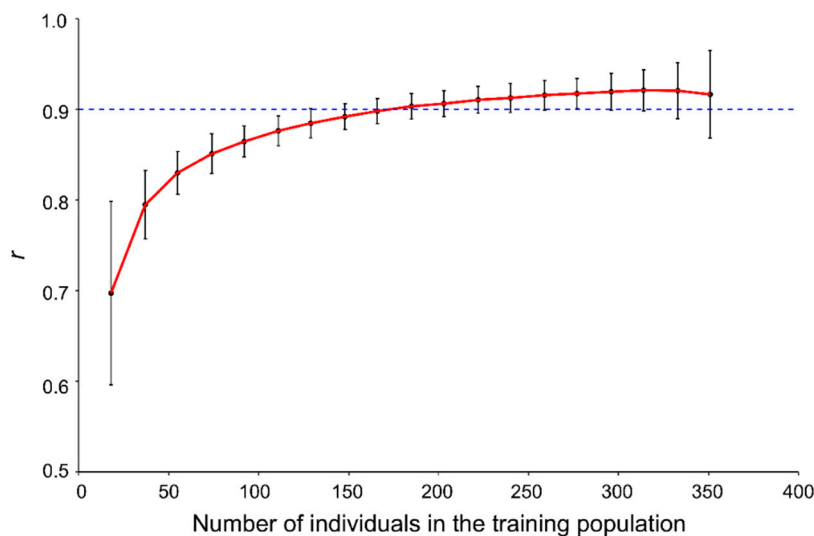
**Table 3.** Accuracy ( $r$ ) and relative efficiency ( $RE$ ) values of the 24 combinations representing the four marker sets and six pasmo severity (PS) datasets using RR-BLUP obtained using a random five-fold cross-validation.

Marker Set	PS Dataset	$r (\bar{x} \pm s)$ <sup>1</sup>	$RE (\bar{x} \pm s)$ <sup>1</sup>
SNP-500QTL	PS-mean	$0.92 \pm 0.02a$	$1.84 \pm 0.04a$
	PS-2012	$0.84 \pm 0.03b$	$1.68 \pm 0.06b$
	PS-2013	$0.81 \pm 0.04c$	$1.62 \pm 0.07c$
	PS-2014	$0.82 \pm 0.04c$	$1.63 \pm 0.07c$
	PS-2015	$0.76 \pm 0.05d$	$1.52 \pm 0.09d$
	PS-2016	$0.76 \pm 0.05d$	$1.52 \pm 0.11d$
SNP-134QTL	PS-mean	$0.75 \pm 0.06e$	$1.49 \pm 0.11e$
	PS-2012	$0.68 \pm 0.06f$	$1.36 \pm 0.11f$
	PS-2013	$0.60 \pm 0.07ij$	$1.19 \pm 0.14ij$
	PS-2014	$0.60 \pm 0.07i$	$1.21 \pm 0.14i$
	PS-2015	$0.47 \pm 0.09o$	$0.94 \pm 0.18o$
	PS-2016	$0.56 \pm 0.09l$	$1.12 \pm 0.17l$
SNP-67QTL	PS-mean	$0.76 \pm 0.05d$	$1.53 \pm 0.1d$
	PS-2012	$0.67 \pm 0.06g$	$1.35 \pm 0.11g$
	PS-2013	$0.60 \pm 0.07ij$	$1.20 \pm 0.14ij$
	PS-2014	$0.60 \pm 0.07ij$	$1.20 \pm 0.14ij$
	PS-2015	$0.50 \pm 0.09n$	$1.00 \pm 0.17n$
	PS-2016	$0.59 \pm 0.08k$	$1.17 \pm 0.17k$
SNP-52347	PS-mean	$0.67 \pm 0.07g$	$1.33 \pm 0.14g$
	PS-2012	$0.63 \pm 0.06h$	$1.27 \pm 0.12h$
	PS-2013	$0.59 \pm 0.07jk$	$1.19 \pm 0.14jk$
	PS-2014	$0.53 \pm 0.08m$	$1.06 \pm 0.17m$
	PS-2015	$0.38 \pm 0.09q$	$0.77 \pm 0.17q$
	PS-2016	$0.46 \pm 0.09p$	$0.93 \pm 0.18p$

<sup>1</sup> Different letters represent multiple test significance among the 24 combinations at the 0.05 probability level.



**Figure 3.** Accuracy ( $r$ ) (a) and relative efficiency ( $RE$ ) (b) of RR-BLUP prediction models built with combinations of four marker sets using the five-year average PS dataset (PS-mean) and random five-fold cross-validations. Letters above box plots indicated statistical significance ( $p < 0.05$ ) for  $r$  and  $RE$  among marker sets.



**Figure 4.** Relationship between the genomic prediction accuracy ( $r$ ) and the size of the training population based on the SNP-500QTL marker set, the PS-mean dataset and the RR-BLUP models. The dash line represents a prediction accuracy of 0.9.

### 2.5. Prediction Models of Pasmu Resistance

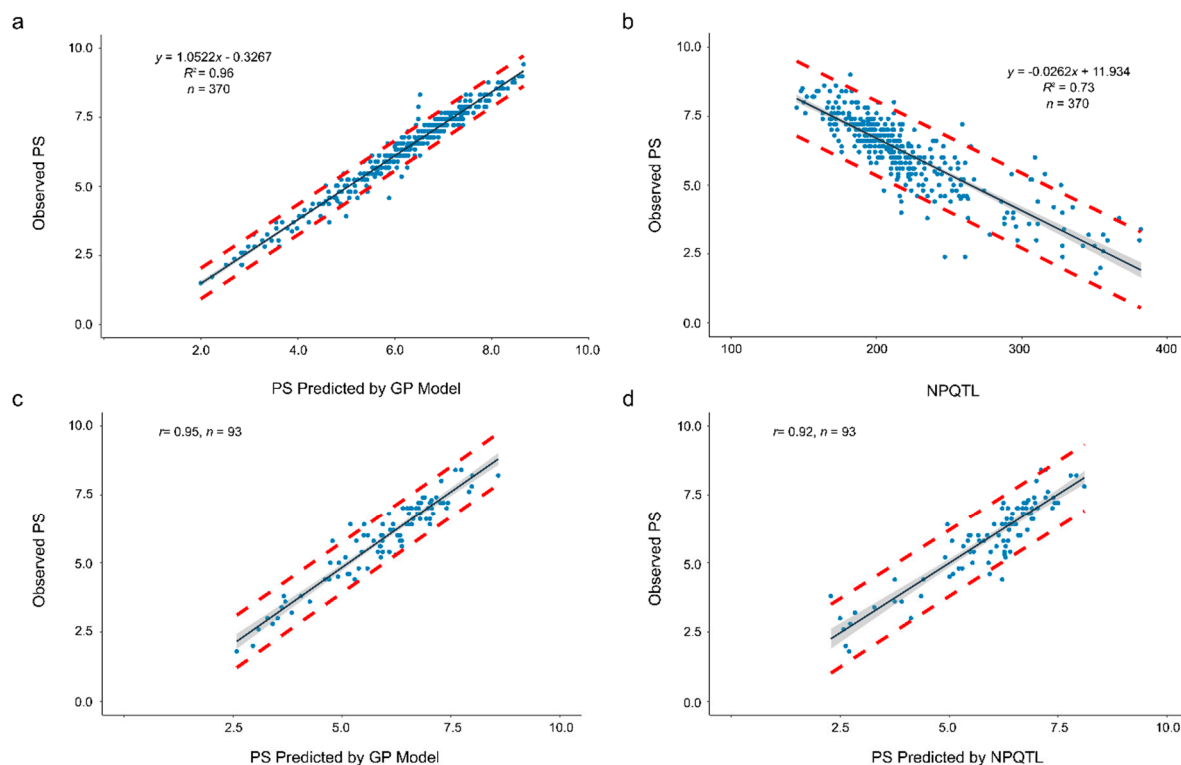
All 370 accessions were used as a training population to build a prediction model using the SNP-500QTL genotypic dataset and the PS-mean phenotypic dataset because this combination outperformed all other models. The model was then employed to predict PS in each year (Table 4). Prediction accuracies ( $r$ ) ranging from 0.71 to 0.81 and  $RE$  values of 1.42 to 1.62 were obtained when predicting PS for individual years (Table 4).

A prediction accuracy as high as 0.98 and a  $RE$  value of 1.96 were obtained when the model was used to predict PS-means of the 370 accessions (Table 4). A linear relationship was observed between the observed ( $y$ ) and predicted PS ( $x$ ):  $y = 1.0522x - 0.3267$  ( $R^2 = 0.96$ ) (Figure 5a). Based on this equation, the average prediction interval between the two red dashed lines, representing the 95% confidence interval, was only less than 1 (an average of 0.97) on the PS ratings (Figure 5a).

NPQTL in the 370 accessions for the 500 QTL set was tallied. Significant linear correlation between PS-mean and NPQTL ( $r = 0.86$  or  $R^2 = 0.73$ ) was observed (Figure 5b). This correlation was less than but close to the accuracy of the GP model with SNP-500QTL and higher than the GP models using other marker sets (Table 3). However, the single linear regression equation ( $y = -0.0262x + 11.934$ ) of the observed PS ( $y$ ) to NPQTL ( $x$ ) had a large standard deviation for each prediction value, with an average prediction interval width of 2.70, nearly three times the average prediction interval width of the GP model; that is, the NPQTL model had a higher prediction error than the GP model.

**Table 4.** Accuracy ( $r$ ) and relative efficiency ( $RE$ ) of genomic prediction for pasmo severity in different years using the RR-BLUP model built with the SNP-500QTL marker set and the PS-mean phenotypic data using all 370 accessions as training data set.

PS Dataset for Prediction	$r$	$RE$
PS-mean	0.98	1.96
PS-2012	0.73	1.46
PS-2013	0.71	1.42
PS-2014	0.81	1.62
PS-2015	0.71	1.43
PS-2016	0.77	1.55



**Figure 5.** Relationship of observed pasmo severity (PS) with PS predicted by a GP model (a,c) or with PS predicted by the number of QTL with positive-effect alleles (NPQTL) (b,d). (a) Linear regression of observed PS ( $y$ ) to predicted PS ( $x$ ) using the genomic prediction model built with the PS-mean dataset and the SNP-500QTL marker set of all 370 accessions as training data set. (b) Linear regression of observed PS ( $y$ ) to NPQTL ( $x$ ) in the 370 flax accessions. (c) Relationship of observed PS of 93 randomly chosen accessions with the PS predicted by the genomic model constructed with the SNP-500QTL marker set and PS-mean dataset when a random subset of 277 accessions was used as training population. (d) Relationship of observed PS of 93 randomly chosen accessions with the PS predicted by NPQTL (Figure S2) The red dashed lines represent upper and lower boundaries of the 95% prediction intervals, that is, it is expected that the value of a sample lies within that prediction interval in 95% of the samples. The grey band represents the 95% confidence interval, that is, 95% of those intervals include the true value of the population mean.

## 2.6. A Case Study of Genomic Prediction

To assess GP prediction accuracy, a training-testing partition was generated with random assignment of breeding lines to either training or testing subsets. Considering the different improvement status of accessions in the population (cultivars, breeding lines, landraces or unknown types) and different levels of resistance, we randomly chose 20% of the 370 accessions in the population, that is, 93 accessions (52 cultivars, 21 breeding lines, 3 landraces and 17 unknown types) as validation dataset, that is, a five-fold random cross-validation set. To predict the PS of these 93 accessions, a RR-BLUP model using the SNP-500QTL set and the PS-mean of the remaining 277 accessions as TP set was built to predict PS. The predicted results are shown in Figure 5c and Table S2. The prediction accuracy was as high as 0.95 ( $r$  between observed and predicted PS). Similarly, a linear regression model of observed PS ( $y$ ) to NPQTL ( $x$ ) of the 277 accessions (the same TP as GP) produced  $y = -0.026x + 11.902$  (Figure S2), which was similar to the regression equation previously obtained with the complete accession set (Figure 5b). Using this prediction model, predicted PS and intervals were calculated (Figure 5d, Table S2). The prediction accuracy of 0.92 for NPQTL was slightly inferior to that of the GP model. The observed PS values all fell within prediction intervals (Table S2).



### 3. Discussion

Cross-validation remains the most popular method to evaluate GP accuracy [14,28]. Our RR-BLUP model prediction accuracy of 0.92 for PR is the highest of all published GP models for plant disease resistance traits [14]. This model is especially valuable because PR has low heritability and high inheritance complexity [3,4]. The QTL markers, multi-year phenotypic data and the genetic diversity and size of the population likely contributed positively to this high prediction accuracy [29].

#### 3.1. All Detected QTL Used as Markers in Genomic Prediction

Three sets of QTL markers (SNP-500QTL, SNP-134QTL and SNP-67QTL) and a genome-wide SNP marker set (SNP-52347) were evaluated here. GP models built using SNP-500QTL consistently outperformed models derived with any of the other three marker sets (Table 3, Figure S1), lending credence to the robustness and reliability of the QTL identified using multiple single-locus and multi-locus GWAS statistical methods [4]. Most GWAS aim to detect large-effect QTL, such as the SNP-67QTL set. While potentially useful in MAS, these tend to explain a reduced portion of the phenotypic variation compared to more comprehensive models (Table 2). Consequently, the GP models built with such marker sets have lower GP accuracies. Therefore, using all potential QTL associated with the selective trait to build GP models is advantageous because it greatly improves prediction accuracy. Prediction accuracies of models obtained with SNP-134QTL and SNP-67QTL data sets were comparable (Table 3, Figure S1) and they explained a similar proportion of the phenotypic variation for PS (Table 2), confirming the redundancy or overlap between the two datasets. Removal of redundant QTL from SNP-134QTL to produce SNP-67QTL produced slightly higher accuracy models (Figure 3). Simplifying GP models by removal of redundant and unrelated markers will ease the practical implementation of GP in breeding programs.

#### 3.2. Superior Performance of Genomic Prediction Combined with GWAS

Surprisingly, the GP models built using SNP-52347 generated a lower prediction accuracy than the models with SNP-500QTL (Table 3, Figure S1), regardless of the statistical methods (Figure S1). Similarly, SNP-52347 explained a lower percentage of the phenotypic variation for PS than SNP-500QTL (Table 2). Besides interaction between SNPs, introduction of noise from genome-wide markers [30], the low prediction accuracy may also be owing to some of the erroneously called SNPs and imputation of missing SNP data. SNP-500QTL includes all or nearly all QTL potentially associated with PS; additional markers, not only failed to increase but actually reduced the prediction accuracy, further emphasizing the effectiveness of the QTL identification methodology adopted in our previously published GWAS study [4]. Similar findings were found for FHB in wheat where deoxynivalenol (DON) concentration QTL-linked markers significantly improve prediction accuracy compared to random genome-wide markers [30]. Markers linked to QTL underlying important traits are deemed more useful for prediction strategies because genome-wide markers may introduce noise, thereby reducing accuracy [30]. Using QTL for GP models may be beneficial to balance genetic backgrounds along with maximum gain of breeding value [31]. Genome-wide prediction models based on ~5000 SNPs from de novo GWAS for tropical rice improvement were as effective for prediction as the full marker set of 108,005 SNPs, indicating that the relationship between marker number and prediction accuracy is neither strict nor linear [32]. To sum up, combined applications of the QTL discovered via GWAS and the accelerated breeding cycles through GP facilitate the full use of genome-wide markers in crop disease resistance breeding [10,33]. Removal of redundant markers has the potential to alleviate the effect of the “large  $p$ , small  $n$ ” issue.

#### 3.3. Accuracy of GP Modelling by Environment, Training Population and Statistical Methods

G  $\times$  E interactions, which affects the accuracy of trait assessment, are common for plant traits. A strong G  $\times$  E interaction was observed in flax PR [4]. As a consequence, different PS QTL were

identified for individual years and for the 5-year average [4]; similarly, GP efficiencies differed when individual yearly and average PS data sets were used as training sets (Table 3). The highest accuracies were obtained when the 5-year mean phenotypic data was used as training data (Table 4), suggesting that the average phenotypic data across multiple environments should be used for GP model construction. Because phenotypic values of genotypes in each year had one replication, the average phenotypic data across multiple years is actually equivalent to the best linear unbiased prediction values (BLUPs) or the best linear unbiased estimators (BLUEs). Therefore, the means across multiple environments estimate or reflect the true breeding values of a trait.

Some studies report that prediction accuracy of GP is highly affected by the size of the TP. In general, the prediction accuracy increases with TP size [21,28,29,34–36]. In the GP of seed weight in soybean, for example, prediction accuracy was sensitive to changes in TP size, which may have led to changes of relatedness between training and validation sets [21]. Lorenzana and Bernardo observed that, in an Arabidopsis family, prediction accuracy improved by 0.10 when TP size increased from 48 to 96, by an additional 0.07 when TP size was increased to 192 and by a further 0.05 with a TP size of 332 [37]. Here GP accuracy >0.9 was observed when the TP size reached 185 which slightly increased to 0.921 with a TP size of 314 (Figure 4). Large TPs provide the statistical power needed to improve prediction accuracy [38], especially for traits with low heritability [34,39]. When TP size is sufficiently large, even low heritability traits can be accurately predicted [28,40], including the low heritability PS studied therein. Diversity of the population also affect prediction accuracy [21,29,34,41–43]. A diverse TP may contain more QTL associated with selective traits and increase the correlation of the TP with validation populations (VPs) or test/prediction populations (PPs), resulting in a subsequent increase in prediction accuracy. Although some breeding lines [11,30,44] and bi-parental derived lines [25,41,45,46] are used for TPs, many studies have opted for a more diverse TP germplasm [29,41–43]. Our core collection TP preserves the variation present in the world collection of 3378 accessions maintained by Plant Gene Resources of Canada (PGRC) and represents a broad range of geographical origins, different improvement statuses (landraces, historical and modern cultivars, breeding lines) and two morphotypes (linseed and fibre types) [1,3]. This collection also contains most parents of modern Canadian flax cultivars [25]. Therefore, diverse phenotypic and genetic variabilities within the flax core collection render it useful as a resource for breeding and as a TP for GP model construction.

A variety of statistical methods have been proposed to estimate marker effects for GP. In general, GP methods are based on additive genetic models and their accuracies may vary depending on genetic architecture of target traits. According to the assumptions for statistical distributions of the marker effects, two groups of GP models have been proposed. The first group of models, such as RR-BLUP, genomic BLUP (GBLUP) and BRR, assume that all markers have some effects on the target trait and the same variance, that is, all markers contribute to the variation of the trait. The second group of models, including BayesA, BayesB, BayesC and BL, assume a specific variance for each marker. Some of these models such as BayesB, BayesC and BL, also allow variable (marker) selection when some of markers have very small or no effects. Based on these assumptions, the first group of models are expected to be useful for complex quantitative traits that have a polygenic architecture, while the second group of models are suitable for traits that controlled by a small number of genes or QTL with large effects. Several studies have shown better performance of BayesB for traits controlled by a few of genes with large effect [47–50]. Some simulation studies have also shown that BayesB outperformed GBLUP that is equivalent to RR-BLUP, when the number of QTL underlying a trait are small [47,51]. However, BayesB, RR-BLUP and other models had a similar prediction accuracy under the infinitesimal model [51] or for some complex traits [19,49]. In this study, no difference among RR-BLUP, BRR and BL was observed (Figure S1), primarily because flax pasmo resistance is a complex and polygenic trait and most of QTL associated with it had similar and small effects (Figure 2). RR-BLUP is most commonly used because of some superior features [11,14,42,52–54]. For example, RR-BLUP successfully recognized complex patterns with additive effects and delivered good GP in wheat disease resistance [55]. RR-BLUP also has a clear-cut computational efficiency compared with any other statistical models [11,54,56,57]. Here

the RR-BLUP model with the 500 QTL markers and the 5-year mean PS produced high prediction accuracy and is therefore recommended for the prediction of PR in flax.

#### 3.4. Pasmus Severity Prediction Using Number of Positive-Effect QTL

A highly significant correlation ( $r = 0.86$  or  $R^2 = 0.73$ ) between NPQTL and PS (Figure 5b) provides an alternative approach to directly predict PS phenotypes. The prediction accuracy using the linear regression equation of PS to NPQTL was inferior to the GP model (Figure 5) because the QTL effects were variable (Figure 2), whereas the linear regression equation considered only the number of QTL but not their individual effects. However, NPQTL is advantageous because it can be readily calculated based on the genotyping by sequencing (GBS) or other genotyping data for the QTL markers [14] and the prediction accuracy based on the NPQTL is comparable to most GP models. Thus, the NPQTL-based prediction equation provides a simple alternative model for PS prediction.

#### 3.5. Breeding Application of Genomic Prediction

Plant breeding is to pyramid favourite alleles from distinct parents using different approaches such as conventional crossing, mutation or transgenic methods to develop new varieties. However, most traits of agronomic importance are genetically controlled by polygenes and have a low heritability such as seed yield and horizontal resistance to diseases. Conventional phenotype selection for these traits is usually inefficient because assessment for them must be performed in multiple environments to obtain breeding values of individuals and thus it is very costly, time consuming and inaccurate; and also because of difficulty of evaluation in fields, greenhouses or laboratories. GS or GP provides an efficient approach to increase selection efficiency by not only increasing selection accuracy but also shortening breeding cycles [58]. In this study, we demonstrate a good example of GP for flax pasmo resistance that is environment-sensitive, costly and difficult for field evaluation. As high as 0.92 of prediction accuracy was obtained for PR, corresponding to 1.84 of relative efficiency over the direct phenotypic selection (Table 3), demonstrating efficiency of GP for low heritability traits. Because the training population underlying the GP models is a diverse germplasm collection that contains more than 90 breeding lines and 245 varieties from different breeding programs [3], the GP models developed in this study are expected to be used for germplasm evaluation, parent selection and individual selection of segregation populations for PR.

## 4. Materials and Methods

### 4.1. Population

A total of 370 diverse flax accessions from the core collection [1] were used to evaluate different GP models. This subset of the core collection collected from 38 countries in 12 geographic regions has been used to identify the QTL associated with PS used in our PS models [4].

### 4.2. Pasmus Resistance Data

All flax accessions were assessed for PS in the same pasmo nursery from 2012 to 2016 at the Morden Research and Development Centre, Agriculture and Agri-Food Canada (AAFC), Morden, Manitoba, Canada [4]. A type-2 modified augmented design (MAD2) [59,60] was used for the field trials [3]. Accessions were seeded during the second or third week of May every year. Approximately 200 g of pasmo-infested chopped straw from the previous growing season was spread between rows as inoculum when plants were approximately 30-cm tall. A misting system was operated for 5 min every half hour for 4 weeks, except on rainy days, to ensure conidia dispersal and disease infection and development. Field assessments were conducted at the early (P1) and late flowering stages (P2, 7–10 days after P1), the green boll stage (P3, 7–10 days after P2) and the early brown boll stage (P4, 7–10 days after P3). In 2014 and 2015, only the first three field assessments were conducted because early maturity of the plants did not allow for a fourth rating. The PS observed at green boll stage or

maturity was used for GP as previously described [4]. PS was assessed on leaves and stems of all plants in a single row plot using a 0–9 scale (0 = no sign of infection and 9 = > 90% leaf and stem area infected) [4]. Six sets of PS, including five individual year datasets and the 5-year average, were used for GP modelling. The function “chart.Correlation” of the R package PerformanceAnalytics (v1.5.2, <https://cran.r-project.org/web/packages/PerformanceAnalytics/index.html>) was used to analyse correlations between different PS datasets and draw histograms and scatter plots.

#### 4.3. Genomic Data

A total of 258,873 SNPs were obtained from the 370 accessions after pruning by removing redundant SNPs [4]. The missing data of SNPs (on average 14.13% of a missing data rate) were imputed using Beagle v.4.2 with default parameters [61]. Our previous GWAS analyses of PS in flax were conducted separately for combinations of the five individual year and the 5-year average datasets with ten statistical methods [4]. The statistical methods for GWAS included three single locus models (GLM [62], MLM [63] and GEMMA [64]) and seven multi-locus models (FarmCPU [65], mrMLM [66], FASTmrEMMA [67], ISIS EM-BLASSO [68], pLARmEB [69], pKWmEB [70], FASTmrMLM [71]). For GLM, MLM and FarmCPU, the first six principal components (PCs), accounting for 33.04% of the total variation, were chosen as covariates to measure population structure, while Frappe (<http://med.stanford.edu/tanglab/software/frappe.html>) was used to estimate the population structure of the 370 accessions for other six multi-locus models. GEMMA does not require a Q matrix. The threshold of significant associations for all three single-locus methods (GLM, MLM and GEMMA) and the multi-locus method FarmCPU was determined by a critical  $p$  value ( $\alpha = 0.05$ ) subjected to Bonferroni correction, that is, the corrected  $p$  value =  $1.93 \times 10^{-7}$  ( $0.05/258,873$  SNPs), while a log of odds (LOD) score of three was used to detect robust association signals for the remaining six multi-locus models. The R package MVP (<https://github.com/XiaoleiLiuBio/MVP>) was used for GWAS analyses for the GLM, MLM and FarmCPU, the GEMMA software (<https://github.com/genetics-statistics/GEMMA>) for GEMMA and the R package mrMLM (<https://cran.r-project.org/web/packages/mrMLM/index.html>) for the additional six multi-locus models. The details of GWAS analyses were described in Reference [4]. A total of 500 non-redundant QTL for PS were identified from 370 diverse flax accessions, including 134 QTL that statistically stable in all five years and 67 QTL with relatively stable and large effects [4]. These three QTL datasets (500 unique QTL, 134 statistically stable QTL and 67 stable and large-effect QTL) were used for GP model construction. In addition, we performed Pearson’s  $\chi^2$  test with Yate’s continuity correction to detect all SNPs significantly associated with PS using a  $10^{-5}$  probability level. The three QTL sets and the genome-wide SNP set were used to construct the GP models. Thus, GP models with the 24 combinations of the four marker sets and the six phenotypic datasets were built and compared.

#### 4.4. Genomic Prediction Models

Three statistical methods RR-BLUP [9,17,20], Bayesian LASSO (BL) [20,25,33] and Bayesian ridge regression (BRR) [25,72] were used to build GP models for PS. These predictive models estimate marker effects by modelling markers as random effects. No fixed effects were fitted in the models. The statistical models and their computation procedures are described in detail elsewhere [40,73]. The R package rrBLUP [56] was used to fit the RR-BLUP model and the R package BLR [74] was used to fit the BL and BRR models. The parameters used to fit BL and BRR were determined based on suggestions of de los Campos et al. [74]. Broad-sense heritability (0.25) of PS estimated in the population [3] was used. When preparing QTL marker data for model construction, the positive-effect allele of the tag SNP of a QTL was coded ‘1’ and the alternative allele ‘-1’. Similarly for the SNP marker set, the reference allele of an SNP was coded ‘1’ and the alternative allele ‘-1’. Missing data were coded ‘0’. The EM algorithm implemented in the R package rrBLUP [56] was used to impute the missing marker data because missing marker data were not allowed in the model construction.

#### 4.5. Evaluation of Prediction Models

Two validation methods were used to evaluate prediction models generated from combinations of statistical models, marker sets and PS datasets. The first method was a five-fold random cross-validation. The 370 flax accessions were randomly partitioned into five subsets. For a given partition, each subset was in turn used as validation or test data and the remaining four subsets made the training dataset. This partitioning was repeated 500 times. In this manner, a total of 2500 training data sets were created to build GP models and estimate marker effects. These were used to predict the breeding values of the individuals in the corresponding 2500 test/validation datasets. The accuracy of the genomic predictions ( $r$ ) was defined by the Pearson's simple correlation coefficient between the genetic values predicted by GP and the observed phenotypic values. The relative efficiency of genomic prediction over phenotypic selection ( $RE$ ) was estimated using  $|r|/H^2$  [26,27], where  $H^2$  refers to the broad-sense heritability of PS, estimated to be 0.25 [3].  $RE$  was used as a criterion to compare the response to one cycle of genome-wide selection versus one cycle of phenotypic selection. Means of  $r$  and  $RE$  of the 500 samplings for each marker set, GP model and PS dataset were used to describe the prediction accuracy of GP and the efficiency of one GP cycle relative to one phenotypic selection cycle, respectively. To compare different marker and PS datasets, a joint analysis of variance with Tukey multiple pairwise-comparisons was performed to test the statistical significance of differences in  $r$  and  $RE$  using R. As a case study, we randomly selected 20% of all 370 accessions as validation dataset and used the remaining 277 accessions as training dataset to build a GP model for genomic prediction of unknown germplasm.

The second cross-validation approach involved comparisons across different PS datasets, that is, each of the six complete PS phenotypic datasets were used as training datasets to build GP models that were applied to itself and to the other five phenotypic datasets. The same set of markers for all 370 accessions was used for training and validation. This method tests the relevance of models built based on single year phenotypic data to predict phenotypes measured in different years.

#### 4.6. Phenotypic Variation Explained by Markers

The phenotypic variation explained by all markers in various marker sets, denoted  $h_{SNP}^2$ , was estimated for all PS datasets based on the mixed linear model [75] implemented in the GCTA software [76]. The detailed calculation is described in Reference [77].

### 5. Conclusions

Using a diverse worldwide flax core collection of 370 accessions as a training and test population with 500 QTL identified by GWAS, the 5-year average PS data and the RR-BLUP statistical model, we developed a highly effective GP model with a prediction accuracy as high as 0.92 for pasmo, a low heritability and high inheritance complexity trait. This is the highest reported accuracy value of all GP models for plant disease resistance traits and comparable with previously published results. As an alternative, we developed a linear regression prediction model based on NPQTL that also produced a high prediction accuracy of 0.86. The GP model and the NPQTL-based regression equation were validated and deemed to be applicable to the evaluation of flax germplasm including parent selection for PR. The use of all potential QTL associated with a target trait would be beneficial because the exclusion of a large proportion of unrelated markers would facilitate the construction of highly accurate GP models.

**Supplementary Materials:** Supplementary Materials can be found at <http://www.mdpi.com/1422-0067/20/2/359/s1>.

**Author Contributions:** Conceptualization, F.M.Y. and S.C.; Methodology, F.M.Y.; Software, F.M.Y.; Formal Analysis, F.M.Y., G.J., P.L. and L.H.; Resources, K.Y.R. (field phenotypic data), S.C. and F.M.Y.; Data Curation, F.M.Y., K.Y.R. and Z.Y.; Writing-Original Draft Preparation, F.M.Y., L.H. and J.X.; Writing-Review & Editing, F.M.Y., S.C. and X.W.; Visualization, Z.Y.; Supervision, F.M.Y., S.C. and X.W.; Funding Acquisition, S.C., K.R. and F.M.Y.

**Funding:** This work was part of the Total Utilization Flax GENomics (TUFGEN) project funded by Genome Canada and other stakeholders, the A-base project (J-001004) funded by Agriculture and Agri-Food Canada and

the flax cluster project funded by the Western Grains Research Foundation (WGRF) and the Canada-China science and technology and innovation action plan (2017ZJGH0106002).

**Acknowledgments:** We thank the China Scholarship Council for their financial support of L.H. for his research at Agriculture and Agri-Food Canada (AAFC).

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analysis or interpretation of the data; in the writing of the manuscript; and in the decision to publish the results.

## Abbreviations

ANOVA	Analysis of variance
BL	Bayesian LASSO
BRR	Bayesian ridge regression
DON	Deoxynivalenol
FHB	<i>Fusarium</i> head blight
G × E	Genotype by environment interaction
GBS	Genotyping by sequencing
GEV	Genomic estimated breeding value
GP	Genomic prediction
GS	Genomic selection
GWAS	Genome-wide association study
MARS	Marker-assisted recurrent selection
MAS	Marker-assisted selection
NPQTL	Number of QTL with positive-effect alleles
PGRC	Plant Gene Resources of Canada
PP	Test/prediction population
PR	Pasmo resistance
PS	Pasmo severity
QTL	Quantitative trait locus/loci
RE	Relative efficiency
RR-BLUP	Ridge regression best linear unbiased prediction
SNPs	Single nucleotide polymorphisms
TP	Training population
VP	Validation population

## References

1. Diederichsen, A.; Kusters, P.M.; Kessler, D.; Baines, Z.; Gugel, R.K. Assembling a core collection from the flax world collection maintained by Plant Gene Resources of Canada. *Genet. Resour. Crop Evol.* **2012**, *60*, 1479–1485. [\[CrossRef\]](#)
2. Vera, C.L.; Irvine, R.B.; Duguid, S.D.; Rashid, K.Y.; Clarke, F.R.; Slaski, J.J. Pasmo disease and lodging in flax as affected by pyraclostrobin fungicide, N fertility and year. *Can. J. Plant Sci.* **2014**, *94*, 119–126. [\[CrossRef\]](#)
3. You, F.M.; Jia, G.; Xiao, J.; Duguid, S.D.; Rashid, K.Y.; Booker, H.M.; Cloutier, S. Genetic variability of 27 traits in a core collection of flax (*Linum usitatissimum* L.). *Front. Plant Sci.* **2017**, *8*, 1636. [\[CrossRef\]](#)
4. He, L.; Xiao, J.; Rashid, K.Y.; Yao, Z.; Li, P.; Jia, G.; Wang, X.; Cloutier, S.; You, F.M. Genome-wide association studies for pasmo resistance in flax (*Linum usitatissimum* L.). *Front. Plant Sci.* **2019**, *9*, 1982. [\[CrossRef\]](#)
5. Diederichsen, A.; Rozhmina, T.A.; Kudrjavceva, L.P. Variation patterns within 153 flax (*Linum usitatissimum* L.) genebank accessions based on evaluation for resistance to *fusarium* wilt, anthracnose and pasmo. *Plant Genet. Resour.* **2008**, *6*, 22–32. [\[CrossRef\]](#)
6. Collard, B.C.Y.; Mackill, D.J. Marker-assisted selection: An approach for precision plant breeding in the twenty-first century. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2008**, *363*, 557–572. [\[CrossRef\]](#)
7. Heslot, N.; Jannink, J.L.; Sorrells, M.E. Perspectives for genomic selection applications and research in plants. *Crop Sci.* **2015**, *55*, 1–12. [\[CrossRef\]](#)
8. Xu, Y.; Crouch, J.H. Marker-assisted selection in plant breeding: From publications to practice. *Crop Sci.* **2008**, *48*, 391–407. [\[CrossRef\]](#)

9. Meuwissen, T.H.; Hayes, B.J.; Goddard, M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **2001**, *157*, 1819–1829.
10. Lipka, A.E.; Kandianis, C.B.; Hudson, M.E.; Yu, J.M.; Drnevich, J.; Bradbury, P.J.; Gore, M.A. From association to prediction: Statistical methods for the dissection and selection of complex traits in plants. *Curr. Opin. Plant Biol.* **2015**, *24*, 110–118. [[CrossRef](#)]
11. Arruda, M.P.; Brown, P.J.; Lipka, A.E.; Krill, A.M.; Thurber, C.; Kolb, F.L. Genomic selection for predicting *Fusarium* head blight resistance in a wheat breeding program. *Plant Genome* **2015**, *8*. [[CrossRef](#)]
12. Daetwyler, H.D.; Bansal, U.K.; Bariana, H.S.; Hayden, M.J.; Hayes, B.J. Genomic prediction for rust resistance in diverse wheat landraces. *Theor. Appl. Genet.* **2014**, *127*, 1795–1803. [[CrossRef](#)]
13. Technow, F.; Burger, A.; Melchinger, A.E. Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. *G3 Genes Genomes Genet.* **2013**, *3*, 197–203. [[CrossRef](#)]
14. Poland, J.; Rutkoski, J. Advances and challenges in genomic selection for disease resistance. *Annu. Rev. Phytopathol.* **2016**, *54*, 79–98. [[CrossRef](#)]
15. Gianola, D. Priors in whole-genome regression: The Bayesian alphabet returns. *Genetics* **2013**, *194*, 573–596. [[CrossRef](#)]
16. Desta, Z.A.; Ortiz, R. Genomic selection: Genome-wide prediction in plant improvement. *Trends Plant Sci.* **2014**, *19*, 592–601. [[CrossRef](#)]
17. Whittaker, J.C.; Thompson, R.; Denham, M.C. Marker-assisted selection using ridge regression. *Genet. Res.* **2000**, *75*, 249–252. [[CrossRef](#)]
18. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [[CrossRef](#)]
19. Jiang, Y.; Zhao, Y.; Rodemann, B.; Plieske, J.; Kollers, S.; Korzun, V.; Ebmeyer, E.; Argillier, O.; Hinze, M.; Ling, J.; et al. Potential and limits to unravel the genetic architecture and predict the variation of *Fusarium* head blight resistance in European winter wheat (*Triticum aestivum* L.). *Heredity* **2015**, *114*, 318–326. [[CrossRef](#)]
20. Spindel, J.; Begum, H.; Akdemir, D.; Virk, P.; Collard, B.; Redona, E.; Atlin, G.; Jannink, J.L.; McCouch, S.R. Genomic selection and association mapping in rice (*Oryza sativa*): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genet.* **2015**, *11*, e1004982.
21. Zhang, J.; Song, Q.; Cregan, P.B.; Jiang, G.L. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). *Theor. Appl. Genet.* **2016**, *129*, 117–130. [[CrossRef](#)]
22. Li, Y.; Ruperao, P.; Batley, J.; Edwards, D.; Khan, T.; Colmer, T.D.; Pang, J.; Siddique, K.H.M.; Sutton, T. Investigating drought tolerance in chickpea using genome-wide association mapping and genomic selection based on whole-genome resequencing data. *Front. Plant Sci.* **2018**, *9*, 190. [[CrossRef](#)]
23. Yu, J.; Buckler, E.S. Genetic association mapping and genome organization of maize. *Curr. Opin. Biotechnol.* **2006**, *17*, 155–160. [[CrossRef](#)]
24. Arojju, S.K.; Conaghan, P.; Barth, S.; Millbourne, D.; Casler, M.D.; Hodkinson, T.R.; Michel, T.; Byrne, S.L. Genomic prediction of crown rust resistance in *Lolium perenne*. *BMC Genet.* **2018**, *19*, 35. [[CrossRef](#)]
25. You, F.M.; Booker, H.M.; Duguid, S.D.; Jia, G.; Cloutier, S. Accuracy of genomic selection in biparental populations of flax (*Linum usitatissimum* L.). *Crop J.* **2016**, *4*, 290–303. [[CrossRef](#)]
26. Dekkers, J.C. Prediction of response to marker-assisted and genomic selection using selection index theory. *J. Anim. Breed. Genet.* **2007**, *124*, 331–341. [[CrossRef](#)]
27. Ziyomo, C.; Bernardo, R. Drought tolerance in maize: Indirect selection through secondary traits versus genomewide selection. *Crop Sci.* **2013**, *53*, 1269–1275. [[CrossRef](#)]
28. Crossa, J.; Perez-Rodriguez, P.; Cuevas, J.; Montesinos-Lopez, O.; Jarquin, D.; de Los Campos, G.; Burgueno, J.; Gonzalez-Camacho, J.M.; Perez-Elizalde, S.; Beyene, Y.; et al. Genomic selection in plant breeding: Methods, models, and perspectives. *Trends. Plant Sci.* **2017**, *22*, 961–975. [[CrossRef](#)]
29. Gowda, M.; Das, B.; Makumbi, D.; Babu, R.; Semagn, K.; Mahuku, G.; Olsen, M.S.; Bright, J.M.; Beyene, Y.; Prasanna, B.M. Genome-wide association and genomic prediction of resistance to maize lethal necrosis disease in tropical maize germplasm. *Theor. Appl. Genet.* **2015**, *128*, 1957–1968. [[CrossRef](#)]
30. Rutkoski, J.; Benson, J.; Jia, Y.; Brown-Guedira, G.; Jannink, J.-L.; Sorrells, M. Evaluation of genomic prediction methods for *Fusarium* head blight resistance in wheat. *Plant Genome* **2012**, *5*, 51–61. [[CrossRef](#)]

31. Deshmukh, R.; Sonah, H.; Patil, G.; Chen, W.; Prince, S.; Mutava, R.; Vuong, T.; Valliyodan, B.; Nguyen, H.T. Integrating omic approaches for abiotic stress tolerance in soybean. *Front. Plant Sci.* **2014**, *5*, 244. [[CrossRef](#)]
32. Spindel, J.E.; Begum, H.; Akdemir, D.; Collard, B.; Redona, E.; Jannink, J.L.; McCouch, S. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* **2016**, *116*, 395–408. [[CrossRef](#)]
33. Kayondo, S.I.; Pino del Carpio, D.; Lozano, R.; Ozimati, A.; Wolfe, M.; Baguma, Y.; Gracen, V.; Offei, S.; Ferguson, M.; Kawuki, R.; et al. Genome-wide association mapping and genomic prediction for CBSD resistance in *Manihot esculenta*. *Sci. Rep.* **2018**, *8*, 1549. [[CrossRef](#)]
34. Wang, X.; Xu, Y.; Hu, Z.L.; Xu, C.W. Genomic selection methods for crop improvement: Current status and prospects. *Crop J.* **2018**, *6*, 330–340. [[CrossRef](#)]
35. Jarquin, D.; Kocak, K.; Posadas, L.; Hyma, K.; Jedlicka, J.; Graef, G.; Lorenz, A. Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genom.* **2014**, *15*, 740. [[CrossRef](#)]
36. Asoro, F.G.; Newell, M.A.; Beavis, W.D.; Scott, M.P.; Jannink, J.-L. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. *Plant Genome* **2011**, *4*, 132–144. [[CrossRef](#)]
37. Lorenzana, R.E.; Bernardo, R. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor. Appl. Genet.* **2009**, *120*, 151–161. [[CrossRef](#)]
38. Goddard, M. Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica* **2009**, *136*, 245–257. [[CrossRef](#)]
39. Nielsen, H.M.; Sonesson, A.K.; Yazdi, H.; Meuwissen, T.H.E. Comparison of accuracy of genome-wide and BLUP breeding value estimates in sib based aquaculture breeding schemes. *Aquaculture* **2009**, *289*, 259–264. [[CrossRef](#)]
40. Lorenz, A.J.; Chao, S.; Asoro, F.G.; Heffner, E.L.; Hayashi, T.; Iwata, H.; Smith, K.P.; Sorrells, M.E.; Jannink, J.L. Genomic selection in plant breeding. *Adv. Agron.* **2011**, *110*, 77–123.
41. Cuevas, J.; Crossa, J.; Montesinos-Lopez, O.A.; Burgueno, J.; Perez-Rodriguez, P.; de los Campos, G. Bayesian genomic prediction with genotype x environment interaction kernel models. *G3 Genes Genomes Genet.* **2017**, *7*, 41–53.
42. Dong, H.; Wang, R.; Yuan, Y.; Anderson, J.; Pumphrey, M.; Zhang, Z.; Chen, J. Evaluation of the potential for genomic selection to improve spring wheat resistance to Fusarium head blight in the Pacific Northwest. *Front. Plant Sci.* **2018**, *9*, 911. [[CrossRef](#)]
43. Isidro, J.; Jannink, J.L.; Akdemir, D.; Poland, J.; Heslot, N.; Sorrells, M.E. Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* **2015**, *128*, 145–158. [[CrossRef](#)] [[PubMed](#)]
44. Rutkoski, J.E.; Poland, J.A.; Singh, R.P.; Huerta-Espino, J.; Bhavani, S.; Barbier, H.; Rouse, M.N.; Jannink, J.-L.; Sorrells, M.E. Genomic selection for quantitative adult plant stem rust resistance in wheat. *Plant Genome* **2014**, *7*. [[CrossRef](#)]
45. McElroy, M.S.; Navarro, A.J.R.; Mustiga, G.; Stack, C.; Gezan, S.; Pena, G.; Sarabia, W.; Saquicela, D.; Sotomayor, I.; Douglas, G.M.; et al. Prediction of cacao (*Theobroma cacao*) resistance to *Moniliophthora* spp. diseases via genome-wide association analysis and genomic selection. *Front. Plant Sci.* **2018**, *9*, 343. [[CrossRef](#)]
46. Enciso-Rodriguez, F.; Douches, D.; Lopez-Cruz, M.; Coombs, J.; de Los Campos, G. Genomic selection for late blight and common scab resistance in tetraploid potato (*Solanum tuberosum*). *G3 Genes Genomes Genet.* **2018**, *8*, 2471–2481. [[CrossRef](#)]
47. Daetwyler, H.D.; Pong-Wong, R.; Villanueva, B.; Woolliams, J.A. The impact of genetic architecture on genome-wide evaluation methods. *Genetics* **2010**, *185*, 1021–1031. [[CrossRef](#)]
48. Jannink, J.L.; Lorenz, A.J.; Iwata, H. Genomic selection in plant breeding: From theory to practice. *Brief Funct. Genom.* **2010**, *9*, 166–177. [[CrossRef](#)]
49. Thavamanikumar, S.; Dolferus, R.; Thumma, B.R. Comparison of genomic selection models to predict flowering time and spike grain number in two hexaploid wheat doubled haploid populations. *G3 Genes Genomes Genet.* **2015**, *5*, 1991–1998. [[CrossRef](#)]
50. VanRaden, P.M.; Van Tassell, C.P.; Wiggans, G.R.; Sonstegard, T.S.; Schnabel, R.D.; Taylor, J.F.; Schenkel, F.S. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* **2009**, *92*, 16–24. [[CrossRef](#)]



51. Clark, S.A.; Hickey, J.M.; van der Werf, J.H. Different models of genetic variation and their effect on genomic evaluation. *Genet. Sel. Evol.* **2011**, *43*, 18. [[CrossRef](#)]
52. Rutkoski, J.; Singh, R.P.; Huerta-Espino, J.; Bhavani, S.; Poland, J.; Jannink, J.L.; Sorrells, M.E. Genetic gain from phenotypic and genomic selection for quantitative resistance to stem rust of wheat. *Plant Genome* **2015**, *8*. [[CrossRef](#)]
53. Gonzalez-Camacho, J.M.; Ornella, L.; Perez-Rodriguez, P.; Gianola, D.; Dreisigacker, S.; Crossa, J. Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. *Plant Genome* **2018**, *11*. [[CrossRef](#)]
54. Liabeuf, D.; Sim, S.C.; Francis, D.M. Comparison of marker-based genomic estimated breeding values and phenotypic evaluation for selection of bacterial spot resistance in tomato. *Phytopathology* **2018**, *108*, 392–401. [[CrossRef](#)]
55. Ornella, L.; Singh, S.; Perez, P.; Burgueño, J.; Singh, R.; Tapia, E.; Bhavani, S.; Dreisigacker, S.; Braun, H.-J.; Mathews, K.; et al. Genomic prediction of genetic values for resistance to wheat rusts. *Plant Genome* **2012**, *5*. [[CrossRef](#)]
56. Endelman, J.B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* **2011**, *4*, 250–255. [[CrossRef](#)]
57. Piepho, H.P. Ridge regression and extensions for genomewide selection in maize. *Crop Sci.* **2009**, *49*, 1165–1176. [[CrossRef](#)]
58. Bassi, F.M.; Bentley, A.R.; Charmet, G.; Ortiz, R.; Crossa, J. Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). *Plant Sci.* **2016**, *242*, 23–36. [[CrossRef](#)]
59. Lin, C.S.; Poushinsky, G. A modified augmented design (type 2) for rectangular plots. *Can. J. Plant Sci.* **1985**, *65*, 743–749. [[CrossRef](#)]
60. You, F.M.; Duguid, S.D.; Thambugala, D.; Cloutier, S. Statistical analysis and field evaluation of the type 2 modified augmented design (MAD) in phenotyping of flax (*Linum usitatissimum*) germplasm in multiple environments. *Aust. J. Crop Sci.* **2013**, *7*, 1789–1800.
61. Browning, S.R.; Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **2007**, *81*, 1084–1097. [[CrossRef](#)] [[PubMed](#)]
62. Price, A.L.; Patterson, N.J.; Plenge, R.M.; Weinblatt, M.E.; Shadick, N.A.; Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **2006**, *38*, 904–909. [[CrossRef](#)] [[PubMed](#)]
63. Yu, J.; Pressoir, G.; Briggs, W.H.; Vroh Bi, I.; Yamasaki, M.; Doebley, J.F.; McMullen, M.D.; Gaut, B.S.; Nielsen, D.M.; Holland, J.B.; et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **2006**, *38*, 203–208. [[CrossRef](#)]
64. Zhou, X.; Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **2012**, *44*, 821–824. [[CrossRef](#)] [[PubMed](#)]
65. Liu, X.; Huang, M.; Fan, B.; Buckler, E.S.; Zhang, Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* **2016**, *12*, e1005767. [[CrossRef](#)] [[PubMed](#)]
66. Wang, S.B.; Feng, J.Y.; Ren, W.L.; Huang, B.; Zhou, L.; Wen, Y.J.; Zhang, J.; Dunwell, J.M.; Xu, S.; Zhang, Y.M. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* **2016**, *6*, 19444. [[CrossRef](#)] [[PubMed](#)]
67. Wen, Y.J.; Zhang, H.; Ni, Y.L.; Huang, B.; Zhang, J.; Feng, J.Y.; Wang, S.B.; Dunwell, J.M.; Zhang, Y.M.; Wu, R. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* **2017**, *19*, 700–712. [[CrossRef](#)] [[PubMed](#)]
68. Tamba, C.L.; Ni, Y.L.; Zhang, Y.M. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* **2017**, *13*, e1005357. [[CrossRef](#)] [[PubMed](#)]
69. Zhang, J.; Feng, J.Y.; Ni, Y.L.; Wen, Y.J.; Niu, Y.; Tamba, C.L.; Yue, C.; Song, Q.; Zhang, Y.M. pLARmEB: Integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity* **2017**, *118*, 517–524. [[CrossRef](#)]
70. Ren, W.L.; Wen, Y.J.; Dunwell, J.M.; Zhang, Y.M. pKWmEB: Integration of Kruskal-Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* **2017**, *120*, 208–218. [[CrossRef](#)]

71. mrMLM. Available online: <https://cran.r-project.org/web/packages/mrMLM/index.html> (accessed on 25 August 2018).
72. De los Campos, G.; Naya, H.; Gianola, D.; Crossa, J.; Legarra, A.; Manfredi, E.; Weigel, K.; Cotes, J.M. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* **2009**, *182*, 375–385. [[CrossRef](#)] [[PubMed](#)]
73. De los Campos, G.; Hickey, J.M.; Pong-Wong, R.; Daetwyler, H.D.; Calus, M.P. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **2013**, *193*, 327–345. [[CrossRef](#)] [[PubMed](#)]
74. de Los Campos, G.; Perez, P.; Vazquez, A.I.; Crossa, J. Genome-enabled prediction using the BLR (Bayesian Linear Regression) R-package. *Methods Mol. Biol.* **2013**, *1019*, 299–320. [[PubMed](#)]
75. Yang, J.; Benyamin, B.; McEvoy, B.P.; Gordon, S.; Henders, A.K.; Nyholt, D.R.; Madden, P.A.; Heath, A.C.; Martin, N.G.; Montgomery, G.W.; et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **2010**, *42*, 565–569. [[CrossRef](#)] [[PubMed](#)]
76. Yang, J.; Lee, S.H.; Goddard, M.E.; Visscher, P.M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **2011**, *88*, 76–82. [[CrossRef](#)]
77. You, F.M.; Xiao, J.; Li, P.; Yao, Z.; Jia, G.; He, L.; Kumar, S.; Soto-Cerda, B.; Duguid, S.D.; Booker, H.M.; et al. Genome-wide association study and selection signatures detect genomic regions associated with seed yield and oil quality in flax. *Int. J. Mol. Sci.* **2018**, *19*, 2303. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).