frontiers in Genetics

ORIGINAL RESEARCH
published: 23 February 2022
doi: 10.3389/fgene.2022.835781

Classification and Regression Models for Genomic Selection of Skewed Phenotypes: A Case for Disease Resistance in Winter Wheat (*Triticum aestivum* L.)

Lance F. Merrick[1], *Dennis N. Lozada*[2], *Xianming Chen*[3] *and Arron H. Carter*[1]*

[1]*Department of Crop and Soil Sciences, Washington State University, Pullman, WA, United States,* [2]*Department of Plant and Environmental Sciences, New Mexico State University, Las Cruces, NM, United States,* [3]*USDA-ARS Wheat Health, Genetics and Quality Research Unit and Department of Plant Pathology, Washington State University, Pullman, WA, United States*

OPEN ACCESS

Edited by:
Muhammad Sajjad, COMSATS University, Islamabad Campus, Pakistan

Reviewed by:
Pengtao Ma, Yantai University, China
Yue Hao, Arizona State University, United States

***Correspondence:**
Arron H. Carter
ahcarter@wsu.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *14 December 2021*
Accepted: *19 January 2022*
Published: *23 February 2022*

Citation:
Merrick LF, Lozada DN, Chen X and Carter AH (2022) Classification and Regression Models for Genomic Selection of Skewed Phenotypes: A Case for Disease Resistance in Winter Wheat (Triticum aestivum L.). Front. Genet. 13:835781. doi: 10.3389/fgene.2022.835781

Most genomic prediction models are linear regression models that assume continuous and normally distributed phenotypes, but responses to diseases such as stripe rust (caused by *Puccinia striiformis* f. sp. tritici) are commonly recorded in ordinal scales and percentages. Disease severity (SEV) and infection type (IT) data in germplasm screening nurseries generally do not follow these assumptions. On this regard, researchers may ignore the lack of normality, transform the phenotypes, use generalized linear models, or use supervised learning algorithms and classification models with no restriction on the distribution of response variables, which are less sensitive when modeling ordinal scores. The goal of this research was to compare classification and regression genomic selection models for skewed phenotypes using stripe rust SEV and IT in winter wheat. We extensively compared both regression and classification prediction models using two training populations composed of breeding lines phenotyped in 4 years (2016–2018 and 2020) and a diversity panel phenotyped in 4 years (2013–2016). The prediction models used 19,861 genotyping-by-sequencing single-nucleotide polymorphism markers. Overall, square root transformed phenotypes using ridge regression best linear unbiased prediction and support vector machine regression models displayed the highest combination of accuracy and relative efficiency across the regression and classification models. Furthermore, a classification system based on support vector machine and ordinal Bayesian models with a 2-Class scale for SEV reached the highest class accuracy of 0.99. This study showed that breeders can use linear and non-parametric regression models within their own breeding lines over combined years to accurately predict skewed phenotypes.

Keywords: generalized linear model, non-parametric, ordinal regression, rrBLUP, stripe rust, support vector machines, transformations

1 INTRODUCTION

Genomic selection (GS) is posed to increase genetic gain and reduce cycle time for complex agronomic traits that are difficult to phenotype and analyze (Meuwissen et al., 2001). With the advent of high-throughput genotyping, it is now feasible to develop and implement GS models for categorical/ordinal phenotypes that are common in most breeding programs and often difficult to analyze. The difficulty in phenotyping and analysis can be due to the traits' genetic complexity, environmental dependency to display variation, and the inability of statistical models to model phenotypes adequately. Most GS models are linear regression models that assume continuous and normally distributed phenotypes (Montesinos-López et al., 2015c).

When faced with data that do not follow the assumption of a linear model, researchers have several options. They may either ignore the lack of normality, transform the phenotypes, use generalized linear models (GLMs), or use machine learning (ML) algorithms and classification models. Machine learning models have no restriction on the distribution of response variables, which are less sensitive when modeling ordinal scores (Montesinos-López et al., 2015a; González-Camacho et al., 2018). Most GS models treat disease resistance as continuous values and utilize regression models and transformations for prediction whereas only a few studies have used classification methods (Ornella et al., 2012; Ornella et al., 2014; Rutkoski et al., 2014; Muleta et al., 2017).

When the number of categories is large and the data follow more of a normal distribution, the ordinality of data can be ignored (Montesinos-López et al., 2015b). However, ignoring the lack of normality and using linear regression models imposes various problems. Linear regression models are limited to modeling additive effects only, whereas machine learning models account for both non-additive and epistatic genetic effects (Riedelsheimer et al., 2012). Modeling only additive effects on quantitative resistance to stripe rust is not a major issue, nonetheless, due to previous studies showing mainly additive effects of high-temperature adult-plant (HTAP) resistance to stripe rust (Chen et al., 1995a; Chen et al., 1995b). Ultimately, linear regression models assume continuous and normally distributed phenotypes, whereas machine learning models are not restricted to a certain distribution of response variable and this causes an issue on the analysis of traits (González-Camacho et al., 2018).

Data transformation is another approach used to deal with skewed and ordinal trait information. Logarithmic or square root transformations are commonly implemented to transform data for small sample sizes (Montesinos-López et al., 2015c), where they are considered standard procedures to stabilize variance, but fail to normalize inflated count data (O'Hara and Kotze, 2010; Montesinos-López et al., 2015b). Moreover, transforming data results in a loss of accuracy and power in models, especially in a small sample size (Montesinos-López et al., 2015a). When transformations are used on count data with a high number of zeros causing overdispersion, transformations may not be able to create a normal distribution (Montesinos-López et al., 2016). Another issue with using transformations is the resulting negative predicted values which are not plausible for disease resistance scores.

Another approach is to use GLMs, which accommodate non-normal data with heterogenous variance and correlated observations (Montesinos-López et al., 2015a; Montesinos-López et al., 2015b). GLMs provide more sensible results and have greater power to identify model effects as statistically significant (Montesinos-López et al., 2015b). Poisson and negative binomial regression models are the most common GLMs used for count and ordinal data (Montesinos-López et al., 2015c). GLMs model a function of the response mean as a linear function of the coefficients rather than modeling y as a linear function. These models have advantages over linear models due to their ability to model a skewed non-negative discrete distribution towards lower numbers as seen in disease resistance phenotypes (Montesinos-López et al., 2016). Several studies have shown the feasibility of integrating GLM parametric approaches into GS models such as Bayesian logistic ordinal regression (BLOR), threshold genomic best linear unbiased predictor (TGBLUP), and Bayesian mixed-negative binomial (BMNB) genomic regression (Montesinos-López et al., 2015a; Montesinos-López et al., 2015b; Montesinos-López et al., 2015c, Montesinos-López et al., 2016) and observed that the ordinal models present a viable alternative for predicting ordinal traits.

The last approach is to use machine learning algorithms, and classification models with no restriction on the distribution of response variables are less sensitive when modeling ordinal scores while also accounting for epistatic effects (Ornella et al., 2014; González-Camacho et al., 2018). Support vector machines (SVMs) previously displayed higher performance for relative efficiency and Cohen's kappa coefficient than traditional regression models such as Bayesian LASSO, Ridge Regression, and Reproducing Hilbert spaces (Ornella et al., 2014; González-Camacho et al., 2018). For the classification models, Ornella et al. (2014) further showed the superiority of SVM as the best-performing model compared to random forest (RF). Additionally, classification models displayed an advantage in selecting the top performing lines.

Resistance to diseases, such as stripe rust (caused by *Puccinia striiformis* Westend. f. sp. *tritici* Erikss.) in wheat (*Triticum aestivum* L.) is commonly recorded in ordinal scales and percentages that do not follow the assumptions of linear regression models (Montesinos-López et al., 2015a; González-Camacho et al., 2018). The unbalanced, skewed distribution of resistant phenotypes is another issue for disease resistance traits in breeding programs. For example, in most wheat breeding programs, disease resistance is selected early (i.e., headrow selection before yield trials) in the breeding process. Consequently, this early selection and screening process skews the lines in disease nurseries and yield trials towards mostly resistant lines. Therefore, not only are disease-resistant traits commonly expressed in ordinal and categorical scales, but they can also be very skewed towards resistance and no longer follow a normal distribution.

TABLE 1 | Study populations for stripe rust infection type and disease severity for the diversity panel (DP) and breeding line (BL) training populations phenotyped from 2013 to 2016 and 2016–2020, respectively.

Location	Trial[a]	Year	Individuals	IT[b] 1	SEV[c] 1	IT 2	SEV 2	IT 3	SEV 3
Central Ferry	DP	2013	475	X	X	X	X	X	X
Pullman	DP	2013	475	X	X	X	X	-	-
Central Ferry	DP	2014	475	X	X	-	-	-	-
Pullman	DP	2014	475	X	X	X	X	X	X
Central Ferry	DP	2015	475	X	X	X	X	X	X
Pullman	DP	2015	475	X	X	X	X	X	X
Central Ferry	DP	2016	475	X	X	X	X	X	X
Pullman	DH	2016	136	X	X	X	-	-	-
Pullman	F5	2016	173	X	X	X	-	-	-
Lind	F5	2017	171	X	X	X	-	-	-
Lind	DH	2017	29	X	X	X	-	-	-
Pullman	DH	2017	34	X	X	X	X	X	X
Pullman	F5	2017	506	X	X	X	X	-	-
Lind	DH	2018	448	X	X	-	-	-	-
Pullman	DH	2018	732	X	X	X	X	X	X
Pullman	F5	2018	65	X	X	X	X	X	X
Lind	DH	2020	373	X	X	-	-	-	-

[a]*Trial: DP: Diversity panel; DH: Doubled-haploid.*
[b]*IT: Infection type.*
[c]*SEV: Disease severity.*
X: Indicates measurement recorded.
-: No measurement recorded.

Stripe rust is one of the most devastating diseases of wheat worldwide (Chen, 2020) and is especially destructive in the western United States (Chen et al., 1995b; Rutkoski et al., 2014; González-Camacho et al., 2018; Liu et al., 2019) causing more than 90% yield losses in fields planted with susceptible cultivars (Liu et al., 2020). The use of resistant varieties and fungicide applications are the primary methods to control stripe rust (Chen et al., 1995b; Liu et al., 2020). Quantitative stripe rust resistance, also known as adult-plant resistance (APR) or HTAP resistance, is usually a non-race specific resistance associated with durable resistance with some genes being effective for more than 60 years (Klarquist et al., 2016). APR is conferred by different numbers of loci with varying effects and often displays partial resistance, which makes it difficult to incorporate into new cultivars (Liu et al., 2019). Therefore, APR must be improved over multiple cycles of selection and can be approached similarly to other agronomic traits (Rutkoski et al., 2014; Poland and Rutkoski, 2016; González-Camacho et al., 2018). GS approaches would be able to capture the additive effects of APR and are therefore relevant for accumulating favorable alleles for rust resistance (Rutkoski et al., 2014; Michel et al., 2017).

However, most GS studies treat disease resistance as continuous values and utilize regression models and transformations for prediction whereas only a few studies have used classification methods (Ornella et al., 2012; Ornella et al., 2014; Rutkoski et al., 2014; Muleta et al., 2017). Therefore, this study presents empirical research to 1) evaluate GS methods using all transformations, GLMs, and non-parametric models for handling ordinal categorical phenotypes; and 2) implement these methods into selected and unselected training populations for predicting stripe rust resistance. This study identified the most accurate methods for dealing with complex phenotypes in the context of disease resistance in winter wheat.

2 MATERIALS AND METHODS

2.1 Phenotypic Data

The Washington State University (WSU) Winter Wheat Breeding Program takes stripe rust notes every year to select for stripe rust-resistant lines. Two training populations were used to compare the different methods. The first training population consists of $F_{3:5}$ breeding lines (BL) and doubled-haploid (DH) unreplicated trials in Pullman and Lind, WA planted in 2016–2018 and 2020 growing seasons evaluated for stripe rust responses (**Table 1**). Due to the unreplicated nature of the single plots, each trial in the BL consisted of unique lines, which resulted in a total of 2,634 lines (1,009 in Lind and 1,625 in Pullman) over all years and locations. The BL population was subjected to stripe rust resistance screening and culling in headrows the previous year in unreplicated trials and therefore represents our prior selected population. The second training population consisted of a diverse association mapping panel (DP) with 475 lines evaluated in unreplicated trials in Central Ferry and Pullman, WA from 2013 to 2016. The DP consisted of varieties from various

TABLE 2 | Regression and classification genomic selection models for stripe rust infection type (IT) and disease severity (SEV) in winter wheat.

Model	Type	Description	References
rrBLUP	Regression	Linear ridge regression model using untransformed phenotypes	Endelman (2011)
SQRT rrBLUP	Regression	Linear ridge regression model using square-root (SQRT) transformation	Endelman (2011)
LOG rrBLUP	Regression	Linear ridge regression model using logarithmic (LOG) transformation	Endelman (2011)
BC rrBLUP	Regression	Linear ridge regression model using Box-Cox (BC) transformation	Endelman (2011)
GLM	Regression	Generalized linear model (GLM) with a Poisson distribution	Hastie et al. (2016)
SVMR	Regression	Non-parametric regression support vector machine (SVMR) using a radial kernel	Karatzoglou et al. (2019)
BOR	Classification	Bayesian ordinal regression (BOR) model using the full-scale IT (0–9) and SEV (0–100%)	Pérez and de los Campos, (2014)
BOR 3-Class	Classification	Bayesian ordinal regression (BOR) model using the reduced three class scale IT (0–2) and SEV (0–2)	Pérez and de los Campos, (2014)
BOR 2-Class	Classification	Bayesian ordinal regression (BOR) model using the reduced two class scale IT (0–1) and SEV (0–1)	Pérez and de los Campos, (2014)
SVM	Classification	Non-parametric classification support vector machine (SVM) using a radial kernel using the full- scale IT (0–9) and SEV (0–100%)	Karatzoglou et al. (2019)
SVM 3-Class	Classification	Non-parametric classification support vector machine (SVM) using a radial kernel using the reduced three class scale IT (0–2) and SEV (0–2)	Karatzoglou et al. (2019)
SVM 2-Class	Classification	Non-parametric classification support vector machine (SVM) using a radial kernel using the reduced two class scale IT (0–1) and SEV (0–1)	Karatzoglou et al. (2019)

breeding programs in the Pacific Northwest region of the US and represented our unselected population.

The disease traits measured were stripe rust infection type (IT) and stripe rust disease severity (SEV). The IT was based on a 0–9 scale (resistant: 0–3; intermediate: 4–6; susceptible: 7–9) (Line and Qayoum, 1992), whereas SEV was measured as the percentage of the total area of the leaf infected using a modified Cobb Scale (Peterson et al., 1948). Stripe rust data were dependent on natural infection and incidence at the time of observation. Some trials had three observations and were identified with sequential numbers. The trials with only one observation were recorded right after anthesis to measure stripe rust responses at the adult-plant stage. The reason there was only one observation was that stripe rust was not present in the field at earlier growth stages. If there were three observations, stripe rust was present in the field at earlier growth stages where the first, second, and third scores were taken soon after flag leaf emergence, after anthesis, and at early milk stage, respectively. Entries with a high infection type in the first observation, but a low infection type in the following observations may indicate that they have a HTAP resistance (Chen, 2013). However, due to the nature of APR being effective in the adult stage and that not all trials had multiple recordings, only the last observation for each trial was used to measure the stripe rust response.

2.2 Phenotypic Adjustments

In order to compare the regression and classification strategies, we used multiple methods of phenotypic adjustments. For the regression models, standard adjusted means were calculated considering the field design used. The ability of ridge regression best linear unbiased prediction (rrBLUP), GLM, and SVM regression (SVMR) to predict the standard and transformed [square root (SQRT), LOG, and boxcox (BC) transformed] adjusted means was then compared (**Table 2**). For the classification models, Bayesian and SVM classification (SVM) models were used to predict the full-scale categories for IT and SEV with the standard adjustments for field design as our control values (**Table 2**). We then reduced both traits using multiple number of classes to determine the scenario resulting in the highest accuracy for breeding program implementation.

For the field design adjustment for controls for both the regression and classification phenotypic adjustments, a two-step adjusted means method was used, in which a linear model was implemented to adjust both IT and SEV means within and across environments. Then, a GS model was used to calculate genomic estimated breeding values (GEBVs; Ward et al., 2019). Adjusted means from the stripe rust data collected in the unreplicated trials were adjusted using residuals calculated for the unreplicated genotypes in individual environments and across environments using the modified augmented complete block design model (ACBD; Federer 1956; Goldman 2019). The adjustments were made following the method implemented in Merrick and Carter (2021), as follows:

$$\boldsymbol{Y_{ij}} = \mu + \boldsymbol{Block_i} + \boldsymbol{Check_j} + \boldsymbol{\varepsilon_{ij}} \quad (1)$$

where $\boldsymbol{Y_{ij}}$ is the phenotypic value for the trait of interest of the *i*th block and *j*th replicated check cultivar (i = 1, . . . ,I,j = 1, . . . ,J); μ is the mean effect; $\boldsymbol{Block_i}$ is the fixed effect of the *i*th block; $\boldsymbol{Check_j}$ is the fixed effect of the *j*th replicated check cultivar; and $\boldsymbol{\varepsilon_{ij}}$ are the residual errors with a random normal distribution of $\boldsymbol{\varepsilon} \sim N(0, \sigma^2_{\varepsilon})$. For adjusted means across environments, the model is as follows:

$$\boldsymbol{Y_{ijk}} = \mu + \boldsymbol{Block_i} + \boldsymbol{Check_j} + \boldsymbol{Env_k} + \boldsymbol{Block_i}\text{:}\ \boldsymbol{Env_k} + \boldsymbol{Check_j}\text{:}\ \boldsymbol{Env_k} + \boldsymbol{\varepsilon_{ik}} \quad (2)$$

where $\boldsymbol{Y_{ij}}$ is the phenotypic value for the trait of interest of the *i*th block and *j*th replicated check cultivar in the *k*th environment (i = 1, . . . ,I, j = 1, . . . ,J, k = 1, . . . , K); μ is the mean effect; $\boldsymbol{Block_i}$ is the fixed effect of the *i*th block; $\boldsymbol{Check_j}$ is the fixed effect of the *j*th replicated check cultivar; $\boldsymbol{Env_k}$ is the fixed effect of the *k*th

environment; and $\boldsymbol{\varepsilon_{ijk}}$ are the residual errors with a random normal distribution of $\boldsymbol{\varepsilon} \sim N(0, \sigma_{\varepsilon}^2)$.

The BLUPs for heritability were calculated for each trial and across trials using a mixed linear model for the full augmented randomized complete block design in a single environment and is as follows:

$$\boldsymbol{Y_{ijk}} = \mu + \boldsymbol{Block_i} + \boldsymbol{Check_j} + \boldsymbol{Gen_{l(j)}} + \boldsymbol{\varepsilon_{ijk}}, \quad (3)$$

where $\boldsymbol{Y_{ijk}}$ is the phenotypic value for the trait of interest of the *l*th unreplicated genotype nested in the *j*th replicated check cultivar of the *i*th block (i = 1, . . . ,I, j = 1, . . . ,J, k = 1, . . . ,K); μ is the mean effect; $\boldsymbol{Block_i}$ is the random effect of the *i*th block with the distribution $\boldsymbol{Block} \sim N(0, \sigma_{Block}^2)$; $\boldsymbol{Check_j}$ is the fixed effect of the *j*th replicated check cultivar; $\boldsymbol{Gen_{l(j)}}$ is the unreplicated genotype *l* in the *j*th check with the distribution $\boldsymbol{Gen} \sim N(0, \sigma_{Gen}^2)$; and $\boldsymbol{\varepsilon_{ijk}}$ are the residual errors with a random normal distribution of $\boldsymbol{\varepsilon} \sim N(0, \sigma_{\varepsilon}^2)$. The full model across environments is as follows:

$$\boldsymbol{Y_{ijkl}} = \boldsymbol{\mu} + \boldsymbol{Block_i} + \boldsymbol{Check_j} + \boldsymbol{Gen_{l(j)}} + \boldsymbol{Env_k} + \boldsymbol{Block_i}\text{:}\ \boldsymbol{Env_k} + \boldsymbol{Check_j}\text{:}\ \boldsymbol{Env_k} + \boldsymbol{Gen_{l(j)}}\text{:}\ \boldsymbol{Env_k} + \boldsymbol{\varepsilon_{ijkl}} \quad (4)$$

where $\boldsymbol{Y_{ijkl}}$ is the phenotypic value for the trait of interest of the *l*th unreplicated genotype nested in the *j*th replicated check cultivar of the *i*th block and in the *k*th environment (i = 1, . . . ,I, j = 1, . . . ,J,k = 1, . . . ,K, l = 1, . . . ,L); μ is the mean effect; $\boldsymbol{Block_i}$ is the random effect of the *i*th block with the distribution $\boldsymbol{Block} \sim N(0, \sigma_{Block}^2)$; $Check_j$ is the fixed effect of the *j*th replicated check cultivar; $\boldsymbol{Gen_{l(j)}}$ is the random effect of the genotype *l* in the *j*th replicated check cultivar with the distribution $\boldsymbol{Gen} \sim N(0, \sigma_{Gen}^2)$; $\boldsymbol{Env_k}$ is the random effect of the *k*th environment with the distribution $\boldsymbol{Env} \sim N(0, \sigma_{Env}^2)$; and $\boldsymbol{\varepsilon_{ijkl}}$ are the residual errors with a random normal distribution of $\boldsymbol{\varepsilon} \sim N(0, \sigma_{\varepsilon}^2)$. After adjustments were made, values outside of the 0–9 (IT) and 0–100 (SEV) scales were rounded back to 0–9 and 0–100, respectively, to avoid negative values for log transformations or Poisson distributions and to have the standard adjusted means for all comparisons.

Broad-sense heritability on a genotype-difference basis was calculated using the variance components from the models (3) and (4) implemented by Merrick and Carter (2021) and using BLUP for both individual environments and across environments (Cullis et al., 2006):

$$H_{Cullis}^2 = 1 - \frac{\bar{v}_{\Delta..}^{BLUP}}{2\sigma_g^2} \quad (5)$$

where σ_g^2 and $\bar{V}^{\ \mathrm{BLUP}}$ are the genotype variance and mean variance of a difference between two BLUPs for the genotypic effect BLUPs, respectively (Schmidt et al., 2019). Trial evaluations were compared using general summary statistics, coefficient of variations (CV), skewness, kurtosis, and the non-parametric Kruskal–Wallis test using the R package "ggpubr" (R Core Team, 2018; Kassambara and Kassambara, 2020).

2.3 Genotypic Data

Wheat lines were genotyped using genotyping-by-sequencing (GBS; Elshire et al., 2011) through the North Carolina State University (NCSU) Genomics Sciences Laboratory in Raleigh, North Carolina (https://research.ncsu.edu/gsl/) using a two-enzyme (*Pst*I/*Msp*I) digestion protocol (Poland and Rife, 2012). Genomic DNA was isolated from individual seedlings at the one- to three-leaf stage using Qiagen BioSprint 96 Plant kits and the Qiagen BioSprint 96 workstation (Qiagen, MD, United States). Genotyping by sequencing was conducted using Illumina HiSeq® 2,500 and NovaSeq 6,000. Sequences were aligned to the Chinese Spring International Wheat Genome Sequencing Consortium (IWGSC) RefSeq v1.0 (Appels et al., 2018) using the Burrows-Wheeler Aligner (BWA) 0.7.17 (Li and Durbin, 2009). GBS-derived single-nucleotide polymorphism (SNP) markers were called using TASSEL-GBS v2 SNP calling pipeline in TASSEL v5.2.35 (Bradbury et al., 2007; Glaubitz et al., 2014). Markers with >20% missing data, minor allele frequency (MAF) <5%, and those that were monomorphic were removed. Imputation of missing genotypes was conducted using Beagle 5.0 (Browning et al., 2018) and markers with <5% MAF were further excluded. The remaining markers were binned together based on a linkage disequilibrium threshold value of 0.80 (Ward et al., 2019). The reduced genotype matrix was computed using JMP genomics version 9 (SAS Institute, Inc, 2011). Principal components analysis (PCA) using the SNP data was performed using "prcomp" and a biplot with *k*-mean clusters was created using the "autoplot" packages in R. Cluster number for *k*-means were calculated according to the elbow method using a scree plot with the optimal number of clusters identified when the total intra-cluster variation was minimized.

2.4 Regression Models

2.4.1 Transformations

Transformations using SQRT, LOG, and BC approaches were compared to determine the optimal method for phenotypic adjustment for skewed phenotypes (**Table 2**). The BC transformations were conducted using the "forecast" package (Hyndman and Khandakar, 2008) that identifies optimal lambda values using the "BoxCox.lambda" function in R.

2.4.2 rrBLUP Model

rrBLUP was used as the standard GS model for comparing the predictive ability of the adjusted means and transformed data. The rrBLUP was selected due to its high predictive performance for stripe rust resistance (**Table 2**; Rutkoski et al., 2014; Arruda et al., 2016; Poland and Rutkoski 2016; Muleta et al., 2017; Merrick et al., 2021). The model follows the basic mixed linear model that treats the effects of markers as random effects as described by Endelman (2011):

$$\boldsymbol{y_i} = \boldsymbol{WGu} + \boldsymbol{\varepsilon_i} \quad (6)$$

where $\boldsymbol{u} \sim N(0, \boldsymbol{I}\sigma_u^2)$ is a vector of marker effects; $\boldsymbol{y_i}$ is a vector of phenotypes; G is the genotype matrix; and W is the design matrix for **y**. The marker effects are then calculated using $\hat{\boldsymbol{u}} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{I})^{-1}\boldsymbol{Z}'\boldsymbol{y}$ with the ridge parameter of $\lambda = \sigma_{\varepsilon}^2/\sigma_u^2$, which is the ratio of the residual and marker variances.

2.4.3 Generalized Linear Model

The GLM was implemented using "Glmnet" with a Poisson distribution (**Table 2**; Hastie et al., 2016). Glmnet fits a GLM *via* penalized maximum likelihood with the elastic net penalty computed at grid values on the log scale for the regularization parameter lambda. Glmnet solves the equation:

$$\min_{\beta_0,\beta} \frac{1}{N}\sum_{i=1}^{N} w_i l\left(\boldsymbol{y_i}, \boldsymbol{\beta}_0 + \boldsymbol{\beta}^T \boldsymbol{x_i}\right) + \boldsymbol{\lambda}\left[(1-\alpha)\boldsymbol{\beta}_2^2/2 + \boldsymbol{\alpha}\boldsymbol{\beta}_1\right], \quad (7)$$

over a grid of values of λ; $l(y_i - \eta_i)^2$ is the negative log-likelihood of i. The elastic net penalty is controlled by α and bridges the game between lasso regression ($\alpha = 1$) and ridge regression ($\alpha = 0$), with λ controlling the penalty. $\boldsymbol{y_i}$ is a vector of phenotypes; $\boldsymbol{\beta}$ is the genotype matrix; $\boldsymbol{x}$ is the design matrix for **y**. Poisson regression is used to model count data under the assumption of Poisson error, or otherwise non-negative data where the mean and variance are proportional. Like the Gaussian and binomial models, the Poisson distribution is a member of the exponential family of distributions. We model its positive mean on the log scale: $\log \mu(x) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}'\boldsymbol{x}$.

2.5 Classification Models

2.5.1 Factor Adjustments

We used a Bayesian ordinal model and an SVM to compare factor adjustments (**Table 2**). The adjusted means were used as control for categorical factors but rounded to discrete values, so they follow the initial ordinal scales for both IT and SEV. These scales are 0–9 for IT and 0–100 for SEV. The original 0–9 IT scale and 0–100 SEV scale were reduced to a three-class 0–2 scale (resistant/intermediate/susceptible), and a binary keep/discard scale of 0–1 in order to be more applicable to breeding programs and reduce the effect of unbalanced classes.

2.5.2 Bayesian Ordinal Regression Model

The Bayesian Ordinal Regression (BOR) model implemented in the BGLR package according to Pérez and de los Campos (2014) follows:

$$y_i = \sum_{k=1}^{p} \boldsymbol{x_{ik}}\boldsymbol{\beta_k} + \boldsymbol{\varepsilon_i} \quad (8)$$

where $\boldsymbol{y_i}$ is a vector of phenotypes; $\boldsymbol{x_{ik}}$ is the genotype of the kth marker and ith individual, p is the total number of markers, $\boldsymbol{\beta_k}$ is the estimated random marker effect of the kth marker; and $\boldsymbol{\varepsilon_i}$ is a vector of residuals with a random normal distribution of $\boldsymbol{\varepsilon} \sim N(0, \sigma_\varepsilon^2)$. Each version of the BOR model has its own conditional prior distribution and a scaled-inverse chi-squared density described in Pérez and de los Campos (2014) whose hyper-parameters are set internally by the software. The BOR model uses the probit link function in which the probability of each of the categories is linked to the linear predictor according to the link function outlined in Pérez and de los Campos (2014):

$$\boldsymbol{P}(\boldsymbol{y_i} = \boldsymbol{k}) = \Phi(\boldsymbol{\eta_i} = \boldsymbol{\gamma_k}) - \Phi(\boldsymbol{\eta_i} = \boldsymbol{\gamma_{k-1}}) \quad (9)$$

where $\Phi(.)$ is the standard normal cumulative distribution function, η_i is the linear predictor, and γ_k are threshold parameters, with $\gamma_0 = -\infty$, $\gamma_k \geq \gamma_{k-1}$, $\gamma_K = \infty$. The BOR model was implemented in the "BGLR" package in R with a burn-in rate of 10,000 and 80,000 iterations based on convergence of the models using trace plots (Pérez and de los Campos (2014); Merrick and Carter, 2021).

2.5.3 Support Vector Machine

The SVM is a non-parametric model that can be used for both classification and regression (SVMR) with no specific phenotypic distribution requirement. The SVM performs well in a variety of settings due its use of a maximal margin classifier. The maximal margin classifier uses a hyperplane to classify and separate observations by computing the maximum distance of an observation to the hyperplane and then determining the class of the observation based on which side of the hyperplane it falls on (Gareth et al., 2013). Additionally, SVMs can enlarge the feature space of the data using kernels to accommodate non-linear boundaries between classes and simplify the inner product, which overcomes the dimensionality of the data. For classification, the radial basis function (RBF) was used due to its wide adaption and ability to be applied to any distribution of observations (Wang et al., 2018). Both SVM and SVMR were implemented using the "caret" package in R, with the RBF model using the "kernlab" function in R (Kuhn, 2008; Karatzoglou et al., 2019; Meyer et al., 2019). Furthermore, model tuning was completed using five replications of tenfold CV with resampling within the training set of the training fold of the cross-validation or validation sets. Additionally, for classification, the SVM model was tuned using up-sampling, which randomly samples the minority class to be the same size of the majority class in order to deal with class imbalances that can have significant negative impact on model fitting (Kuhn, 2008).

2.6 Prediction Accuracy and Scheme

Prediction accuracy for the regression models was reported using Pearson correlation coefficients (r) and prediction bias was reported using root mean square error (RMSE) between GEBVs and their respective adjusted means using the function "cor" in R. However, due to the unbalanced class type, the classification models were evaluated using overall class accuracy (R^2) using the "confusionMatrix" function in the "caret" package and reported as R^2 (Kuhn, 2008). Cohen's kappa coefficient (kappa) was used to evaluate classification model bias because it takes into account unbalanced classes (Ornella et al., 2014; González-Camacho et al., 2018).

In order to compare regression and classification models, relative efficiency (RE) was used. RE is based on expected genetic gain when individuals are selected by GS compared to the individuals selected by phenotypic selection. The model for RE according to Ornella et al. (2014) is:

$$\boldsymbol{RE} = \frac{\frac{\left(\sum_{\alpha'} y_i\right)}{N_{\alpha'}} - \frac{\left(\sum_{Test} y_i\right)}{N_{Test}}}{\frac{\left(\sum_{\alpha} y_i\right)}{N_{\alpha}} - \frac{\left(\sum_{Test} y_i\right)}{N_{Test}}}, \quad (10)$$

where α and α' are the 15% of individuals selected by the ranking of observed or predicted values, respectively. $N_\alpha = N_{\alpha'}$ is the

number of individuals selected; y_i is the observed phenotypic value of the ith individual; and $\frac{(\sum_{Test} y_i)}{N_{Test}}$ is the mean of the test population. The denominator is the selection differential of the individuals selected by phenotypic selection and the numerator is the selection differential of the individuals selected by GS. The 15% selection intensity was chosen due to its performance of RE when replacing phenotypic selection with GS (Ornella et al., 2014; González-Camacho et al., 2018).

The prediction accuracy was assessed using a fivefold cross-validation scheme and independent validation sets for IT and SEV in the DP and BL training populations (Merrick et al., 2021). The two populations were used to compare the effects of a selected and unselected population with varying degrees of resistance. Models for GS were conducted with fivefold cross-validation by including 80% of the samples in the training population and predicting the GEBVs of the remaining 20% (Merrick and Carter, 2021; Merrick et al., 2021). One replicate consisted of five model iterations, where the population was split into five different groups.

Independent validation sets were then performed according to Merrick and Carter (2021) on a yearly basis by combining the two locations for each training population and predicting the following year, which results in three continuous training scenarios for each population. For example, the combination of Pullman and Central Ferry trials for the DP in 2013 was used as a training population to predict the combination of Pullman and Central Ferry trials in the DP in 2014. Final validation set was completed by combining all years and locations within a training population and then predicting the combination of years and locations for the other training population. All trials in the BL in both Pullman and Lind combined across 2016 to 2020 were used to predict all trials in the DP in both Central Ferry and Pullman across 2013 to 2016. This allows the evaluation of models in a realistic breeding situation in which we combine all available data to build a training population. All cross-validations and independent validations were replicated 10 times. All GS and MAS models and scenarios were analyzed using WSU's Kamiak high-performance computing cluster (Kamiak, 2021). Model, scenario, and training population comparisons were evaluated by using a Tukey's honestly significant difference (HSD) test implemented in the "agricolae" package in R (de Mendiburu and de Mendiburu, 2019). The comparison of models was then plotted for visual comparison using "ggplot2" in R (Wickham, 2011).

3 RESULTS

3.1 Phenotypic Data

The stripe rust phenotypes for both IT and SEV demonstrated variability for each scale (**Table 3**). For the DP, the IT and SEV values ranged the entire scale of each trait for the majority of the trials. Additionally, the means of the DP were higher than the BL trials, with lower coefficients of variation (CV). Furthermore, the BL trials ranged the entire scale for IT, but had lower means. The SEV in the BL trials did not reach the maximum value of SEV. Overall, the BL displayed a higher proportion of resistance than the DP trials. Every trial and trait displayed a positively skewed distribution, with the exception of SEV in the DP in Pullman in 2015. SEV for the majority of trials was extremely skewed for the BL, with Lind in 2018 displaying the highest skew of any trial and trait. Skewness decreased for combined analysis across environments. Positive values above three display long skinny tails as in the case for SEV for the BL population in Lind in 2018 at 19.77. The majority of distributions are skinny tailed, demonstrating the large amount of similar disease resistance around 0 and the large amount of resistance in the BL and DP populations.

The skewness and kurtosis of the distributions were further visualized (**Figure 1**). The DP is less skewed than the BL. For both IT and SEV, the DP displayed more variation than the BL, except for SEV in Central Ferry. Furthermore, there were significant differences between most years for each population and location (**Figure 1**). Heritability of the BL trials was moderately high for both IT and SEV, with values ranging from 0.76 to 0.97 and 0.52–0.63, respectively. For the DP, heritability ranged from 0.65 to 1.00 for IT and 0.71–1.00 for SEV (**Table 1**).

3.2 Analysis of Principal Components

After filtering and imputation, a total of 19,861 SNP markers for the 475 unique DP lines and the 2,630 BL lines were obtained from GBS. Principal component analysis using SNP markers for the DP and BL populations resulted in four clusters with Cluster 2 (green) overlapping with the other clusters (**Figure 2**). PC1 explained 5.8% of the variation whereas PC2 explained 3.4% of the variation. The biplot displayed four main clusters over the combined populations using *k*-means clustering. Cluster 1 consisted of lines common in both the BL and DP. Majority of lines in both the DP and BL were included in Cluster 3, which is composed of BL in Lind and Pullman and lines from the DP. Cluster 4 consisted mainly of lines from the BL in Lind, whereas majority of lines from the BL in Pullman comprised Cluster 2.

3.3 Cross-Validations for Regression Models

Multiple comparisons using HSD for RMSE and Pearson correlations for accuracy were conducted for the regression models in individual populations and years for IT and SEV. The SVMR model resulted in the highest accuracy ($r = 0.73$) in the 2018 Pullman BL trial for IT (**Figure 3**). Accuracy for the GLM model in 2018 Pullman BL was 0.72. The GLM displayed consistent high accuracies in the more skewed BL population than the less skewed DP but displayed the lowest accuracy for the most skewed trial in the BL in Lind in 2018 (0.23). Overall, there were no significant differences for the BL, whereas the LOG rrBLUP and the GLM model showed significant differences (HSD test, $p < 0.05$) in the DP. Additionally, the BL trials had higher mean accuracies than the DP trials with an increase in accuracy with the combination of years. Altogether, the rrBLUP had the highest accuracy over the transformed phenotypes (0.53). The rrBLUP model had similar RMSE than the SVMR and GLM

TABLE 3 | Stripe rust infection type (IT) and disease severity (SEV) heritability (H^2) and trial statistics for unadjusted phenotypes in the diversity panel (DP) and breeding line (BL) training population phenotypes from 2013 to 2016 and 2016 to 2020 growing seasons.

Population	Location	Trait	Year	H^2	CV[a]	Max[b]	Mean	Min[c]	SD[d]	Kurtosis	Skew
BL	Lind	IT	2017	0.82	87.97	8	2.91	0	2.56	−1.21	0.27
			2018	0.97	260.58	8	0.66	0	1.73	7.08	2.78
			2020	0.96	93.03	8	3.42	0	3.18	−1.60	0.20
			2017–2018	0.79	124.19	8	2.03	0	2.52	−0.67	0.84
			2017–2020	0.85	116.24	8	2.38	0	2.77	−0.97	0.70
		SEV	2017	0.82	125.51	70	13.83	0	17.36	1.33	1.45
			2018	0.76	304.83	30	1.29	0	3.93	19.77	4.15
			2020	0.97	125.06	80	18.04	0	22.56	0.42	1.25
			2017–2018	0.81	168.85	70	8.92	0	15.06	3.97	2.10
			2017–2020	0.83	157.72	80	11.23	0	17.71	2.90	1.89
	Pullman	IT	2016	0.53	87.56	8	2.76	0	2.41	−1.05	0.33
			2017	0.56	78.53	9	2.56	0	2.01	0.30	0.87
			2018	0.54	150.59	8	2.20	0	3.31	−0.93	0.96
			2016–2017	0.56	81.12	9	2.60	0	2.11	−0.17	0.70
			2016–2018	0.57	111.41	9	2.43	0	2.70	−0.57	0.85
		SEV	2016	0.63	152.31	80	8.49	0	12.94	6.06	2.36
			2017	0.54	133.03	90	17.00	0	22.62	2.00	1.70
			2018	0.53	177.64	80	15.58	0	27.68	0.76	1.55
			2016–2017	0.52	140.51	90	14.96	0	21.03	2.96	1.90
			2016–2018	0.53	158.47	90	15.23	0	24.14	1.78	1.73
DP	Central Ferry	IT	2013	—	55.97	8	3.14	1	1.76	0.55	1.03
			2014	0.93	61.93	9	3.06	1	1.90	1.97	1.30
			2015	1.00	42.38	9	4.57	1	1.94	−0.79	0.28
			2016	0.96	46.23	9	4.19	0	1.94	−0.02	0.57
			2013–2014	0.65	58.99	9	3.10	1	1.83	1.37	1.18
			2013–2015	0.85	55.46	9	3.59	1	1.99	0.00	0.78
			2013–2016	0.75	53.33	9	3.74	0	1.99	−0.07	0.71
		SEV	2013	—	99.64	90	24.08	2	24.00	−0.89	0.70
			2014	0.89	152.84	90	12.08	2	18.47	6.98	2.63
			2015	1.00	73.44	90	36.36	2	26.70	−1.25	0.19
			2016	0.97	70.57	100	36.15	0	25.51	−0.85	0.22
			2013–2014	0.71	123.34	90	17.98	2	22.18	0.91	1.40
			2013–2015	0.78	105.10	90	24.06	2	25.29	−0.47	0.90
			2013–2016	0.85	95.59	100	27.06	0	25.87	−0.74	0.70
	Pullman	IT	2013	1.00	47.89	9	3.78	1	1.81	0.66	0.62
			2014	1.00	46.38	9	4.83	1	2.24	−0.77	−0.02
			2015	1.00	44.17	9	5.13	1	2.27	−0.82	−0.07
			2013–2014	0.75	48.81	9	4.31	1	2.10	−0.42	0.33
			2013–2015	0.86	47.89	9	4.58	1	2.19	−0.65	0.21
		SEV	2013	0.97	117.71	100	20.33	2	23.93	2.46	1.67
			2014	0.89	84.60	90	36.28	2	30.69	−1.41	0.32
			2015	0.94	68.17	100	50.56	2	34.47	−1.34	−0.12
			2013–2014	0.71	101.13	100	28.33	2	28.65	−0.52	0.87
			2013–2015	0.85	91.07	100	35.54	2	32.37	−1.08	0.54

[a]*CV: coefficient of variation.*
[b]*Max: maximum.*
[c]*Min: minimum.*
[d]*SD: standard deviation.*

models with 2.15, 2.18, and 2.28, respectively (**Supplementary Figure S1**). The SQRT rrBLUP model had the lowest RMSE (0.51), and the BC and LOG rrBLUP models had the highest RMSE (5.67 and 5.93, respectively). Using SQRT transformation on the phenotypes reduced the error of the predictions compared to the other transformations.

Similar to IT, the highest accuracies for SEV were obtained in the 2018 Pullman BL trial, with the GLM reaching the highest accuracy (0.76), followed by the SQRT rrBLUP (0.74) and SVMR (0.73) models (**Supplementary Figure S2**). The lowest accuracies were also achieved with the GLM model in the 2018 Lind BL trial (0.18). The 2018 Lind BL trial had the lowest accuracies for the majority of models. Similar to IT, there were no statistical differences between models overall in the BL, and the SQRT rrBLUP, rrBLUP, and SVMR reached the highest accuracies in the DP. The SVMR and SQRT rrBLUP reached the highest

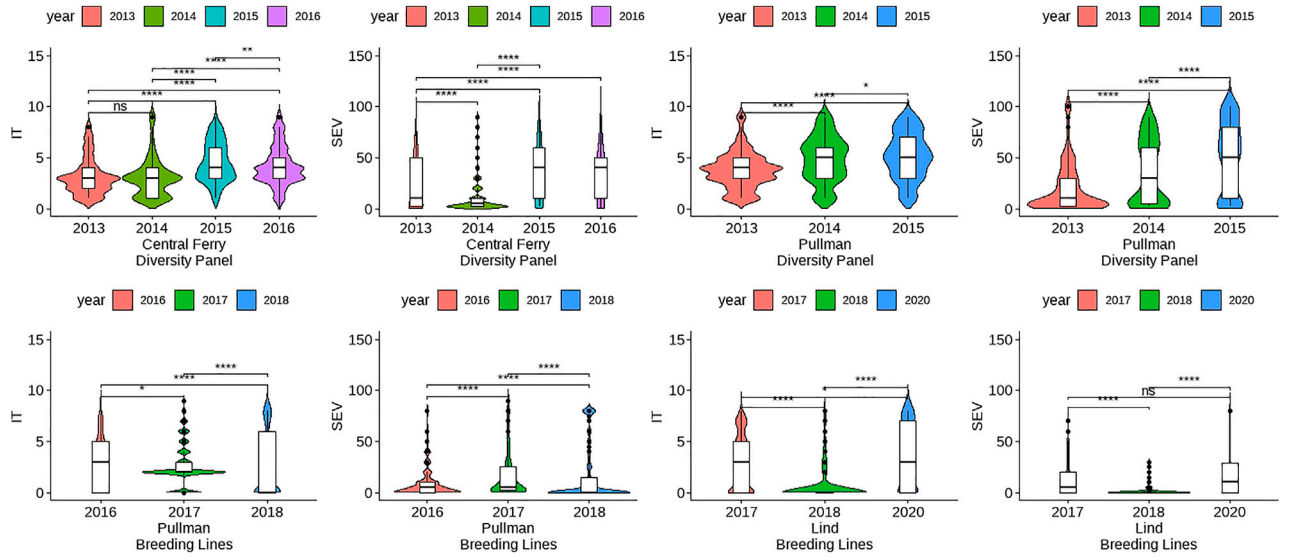

FIGURE 1 | Comparison of unadjusted phenotypes for infection type (IT) and disease severity (SEV) over years and locations in the diversity panel and breeding line training populations using Kruskal–Wallis test. Significant differences were based on *p*-values "*" < 0.05, "***" < 0.01, and "****" < 0.001.

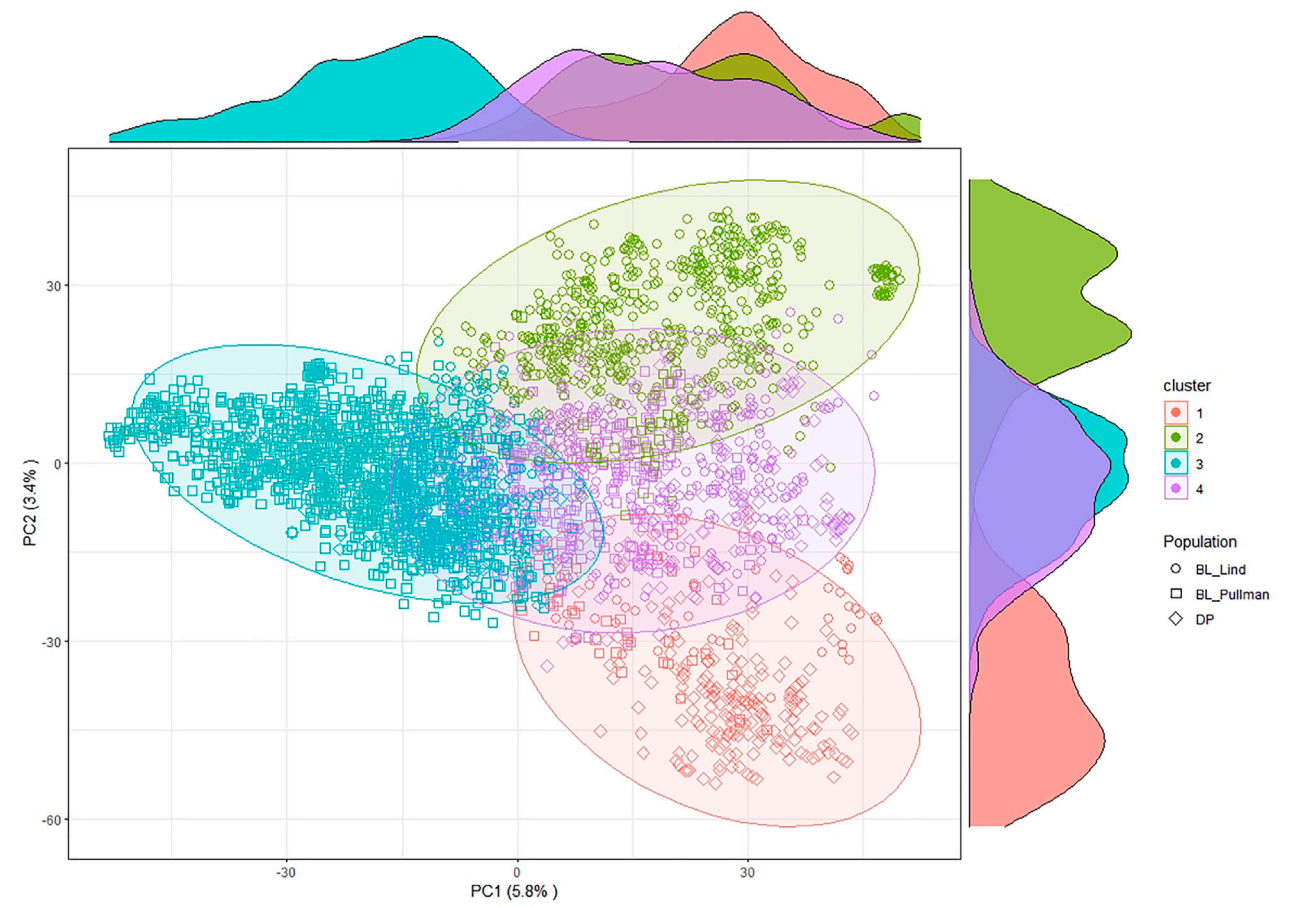

FIGURE 2 | Principal component (PC) biplot and k-means clustering of SNP GBS markers from the diversity panel (DP) and breeding line (BL) training populations.

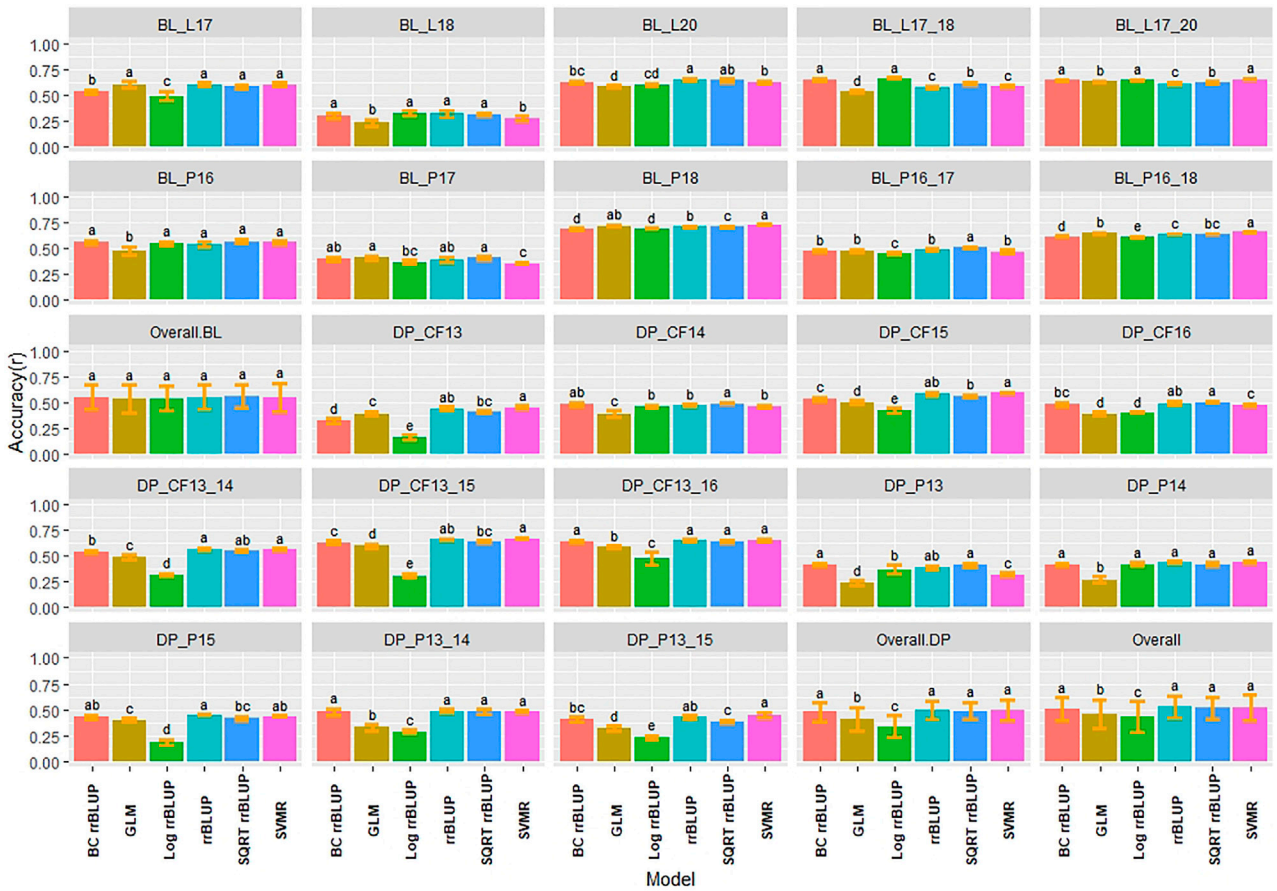

FIGURE 3 | Pairwise comparisons of genomic selection regression model accuracy (*r*) using cross-validations for stripe rust infection type. Pacific Northwest winter wheat diversity panel (DP) lines phenotyped from 2013 to 2016 in Central Ferry (DP_CF) and Pullman (DP_P), WA. Washington State University breeding lines phenotyped from 2016 to 2020 in Lind (BL_L) and Pullman (BL_P), WA. Model comparison across the DP (Overall.DP), BL (Overall.BL), and Overall scenarios. Models labeled with the same letter are not significantly different using Tukey's honestly significant difference. Bars indicate standard errors.

accuracies of 0.60 (**Supplementary Figure S2**). For SEV, the RMSE for the transformed rrBLUP models displayed much lower RMSE values than the rrBLUP, GLM, and SVMR models (**Supplementary Figure S3**). However, this discrepancy is presumably due to the phenotypic range of the transformations compared to the untransformed range for SEV, which is 0–100. The BC rrBLUP model displayed an extremely large RMSE in the DP in Central Ferry in 2015 (57.11). Overall, the rrBLUP models displayed statistically similar RMSE values with the transformed rrBLUP models.

3.4 Cross-Validations for Classification Models

Due to the difference between regression and classification models, multiple comparisons using HSD for the kappa coefficient and overall class accuracy were conducted for the classification models in individual populations and years for IT and SEV. In contrast to the regression models where the 2018 Lind BL trial had the lowest regression accuracies, the classification models displayed the highest R^2 values with the 2-Class and 3-Class BOR models reaching an overall class accuracy of 0.88 for IT (**Figure 4**). Additionally, the SVM models displayed much higher accuracies than the BOR models overall. The full scale BOR model had very low accuracy for the majority of trials with the BL in 2018 in Pullman. The reduced class sizes, 2 and 3, displayed higher accuracy than the full IT scales. Overall, the selected BL displayed higher accuracies than the unselected DP. The 2-Class SVM reached the highest overall class accuracy with 0.76 in the BL and 0.69 in the DP. The 2-Class SVM reached the highest overall class accuracy of 0.72 in the overall comparison. The high-class accuracies in the BL in Lind in 2018 can be explained by the kappa values of 0 (**Supplementary Figure S4**), displaying the highly skewed data and the inability for the models to account for phenotypes of mostly zeros. The SVM displayed lower kappa values in the DP than in the BL, but the BOR models had the opposite trend. The BOR models displayed higher kappa values than the SVM models, but the SVM models showed higher accuracy.

The classification models for SEV had very similar results to IT, with the BOR and SVM 2-Class models reaching an accuracy

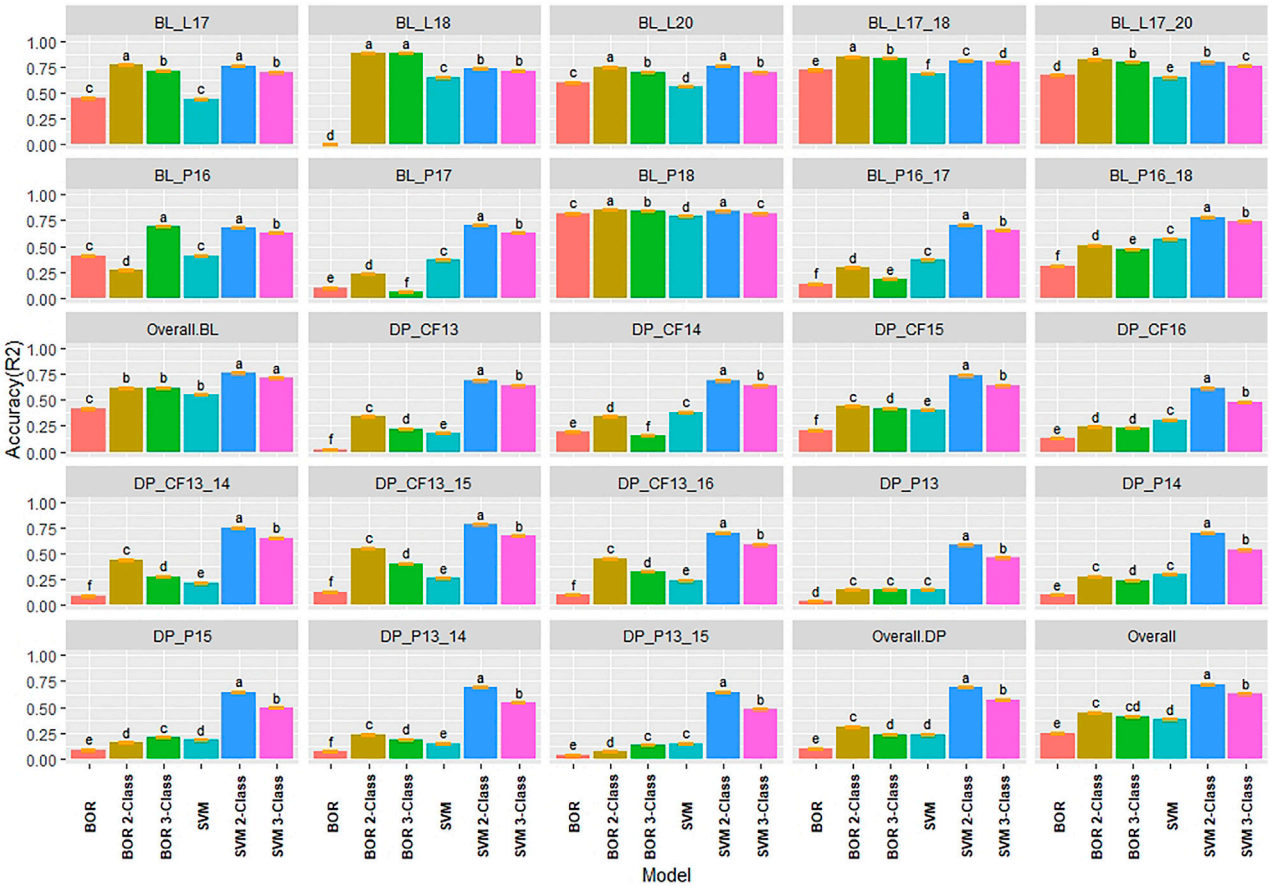

FIGURE 4 | Pairwise comparisons of genomic selection classification model overall class accuracy (R^2) using cross-validations for stripe rust infection type. Pacific Northwest winter wheat diversity panel (DP) lines phenotyped from 2013 to 2016 in Central Ferry (DP_CF) and Pullman (DP_P), WA. Washington State University breeding lines phenotyped from 2016 to 2020 in Lind (BL_L) and Pullman (BL_P), WA. Model comparison across the DP (Overall.DP), BL (Overall.BL), and Overall scenarios. Models labeled with the same letter are not significantly different using Tukey's honestly significant difference. Bars indicate standard errors.

of 0.99 and 0.98, respectively (**Supplementary Figure S5**). This was due to the very skewed and high levels of zeros in the data in the BL in Lind in 2018. Additionally, in the DP that had less skewed phenotypes, the BOR models showed very poor overall class accuracy with the majority of trials having R^2 values of 0.20, with moderate accuracies for the 2-Class BOR. The 2-Class SVM displayed the highest statistically significant class accuracy in scenarios with R^2 values of 0.86, 0.78, and 0.81 within the BL, DP, and overall comparisons, respectively. The kappa values were higher in the DP trials due to less skewed phenotypes and displayed low values in the high accuracy trial of the BL in 2018 Lind. Overall, the 2-Class SVM had the highest kappa value for SEV with 0.46 (**Supplementary Figure S6**).

3.5 Cross-Validation Relative Efficiency

RE was used to compare the selection differential between the GS models and phenotypic selection for the phenotypes. Overall, the highest relative efficiencies for IT were the regression models with the majority of models having statistically similar relative efficiencies. The regression models had very high RE values with the rrBLUP models reaching a maximum value of 0.94 in the 2018 Pullman BL trial (**Figure 5**). The SVMR model had statistically similar RE values to the rrrBLUP models in the overall comparisons. In contrast, the classification models had relatively low RE in the majority of trials with the three-class BOR model (−0.38) in the combined 2017 to 2018 Lind BL trials. This confirmed the bias seen in the kappa results with the majority of lines being predicted as zeros. Interestingly, the two- and three-class BOR and SVM displayed lower RE values overall than the full-scale models. Overall, the rrBLUP and SQRT rrBLUP reached RE values of 0.62.

Similar to IT, the regression models had very high RE for SEV, with the classification models reaching low to moderate values ranging between −0.58 and 0.89 (**Supplementary Figure S7**). The rrBLUP models had very high RE (0.98; BL Pullman 2018) compared to phenotypic selection. The rrBLUP models showed consistently higher REs than the GLM and SVMR models. The GLM displayed similar RE values in the BL, but lower in the DP. The SQRT rrBLUP model had the highest RE overall (0.81). The classification models had very low RE except the BL trials in 2018 Pullman and 2017 Lind, which showed very high RE compared to the other years and populations. Additionally, the combined trials for both the BL and DP

FIGURE 5 | Pairwise comparisons of genomic selection regression and classification model relative efficiency (RE) using cross-validations for stripe rust infection type. Pacific Northwest winter wheat diversity panel (DP) lines phenotyped from 2013 to 2016 in Central Ferry (DP_CF) and Pullman (DP_P), WA. Washington State University breeding lines phenotyped from 2016 to 2020 in Lind (BL_L) and Pullman (BL_P), WA. Model comparison across the DP (Overall.DP), BL (Overall.BL), and Overall scenarios. Models labeled with the same letter are not significantly different using Tukey's honestly significant difference. Bars indicate standard errors.

displayed higher RE than some of the individual years indicating an advantage of combining trials.

3.6 Validation Sets for Regression Models

The training populations were evaluated for validation sets on a yearly basis and over combined years and trials. We used the earliest trial to predict the following year and then a new model with the addition of each subsequent trial to evaluate genotype-by-environment interaction of a prediction model. We then compared the combination of all trials for one population to predict the combination of all trials in the other population. The highest accuracy for IT was in the continuous training scenario of the DP combined 2013–2015 to predict the DP 2016 with SQRT rrBLUP reaching 0.65 (**Figure 6**). There were only a few significant differences, with none in the overall BL or DP. Overall, the SQRT rrBLUP displayed the highest accuracy (0.46). Furthermore, there was an increase in accuracy as the years were combined within the same population. However, the accuracy was much lower when predicting into the combined trials of the other population. Similar RMSE values to the cross-validations were displayed with SQRT rrBLUP having the lowest RMSE (1.31; **Supplementary Figure S8**).

The validation accuracy for SEV displayed similar trends to IT, with the highest accuracy of 0.72 for the SQRT rrBLUP and rrBLUP (**Figure 6**). Interestingly, the combined BL trials predicting into the combined DP displayed the highest accuracy in the BL prediction scenarios with BC rrBLUP reaching 0.53. This trend was in contrast to IT. However, the opposite was seen in the DP. The validation set accuracy for the DP was higher than the validation sets for BL. In the overall comparison, there were no statistical differences between the models. The BC and Log rrBLUP displayed the highest accuracies in some scenarios, which was not seen in cross-validations and was only observed in the BL. The RMSE values were much higher for SEV with the SQRT rrBLUP displaying similar RMSE to IT (**Supplementary Figure S9**).

3.7 Validation Sets for Classification Models

The classification models had contrasting results for the validation sets compared to the regression models. The

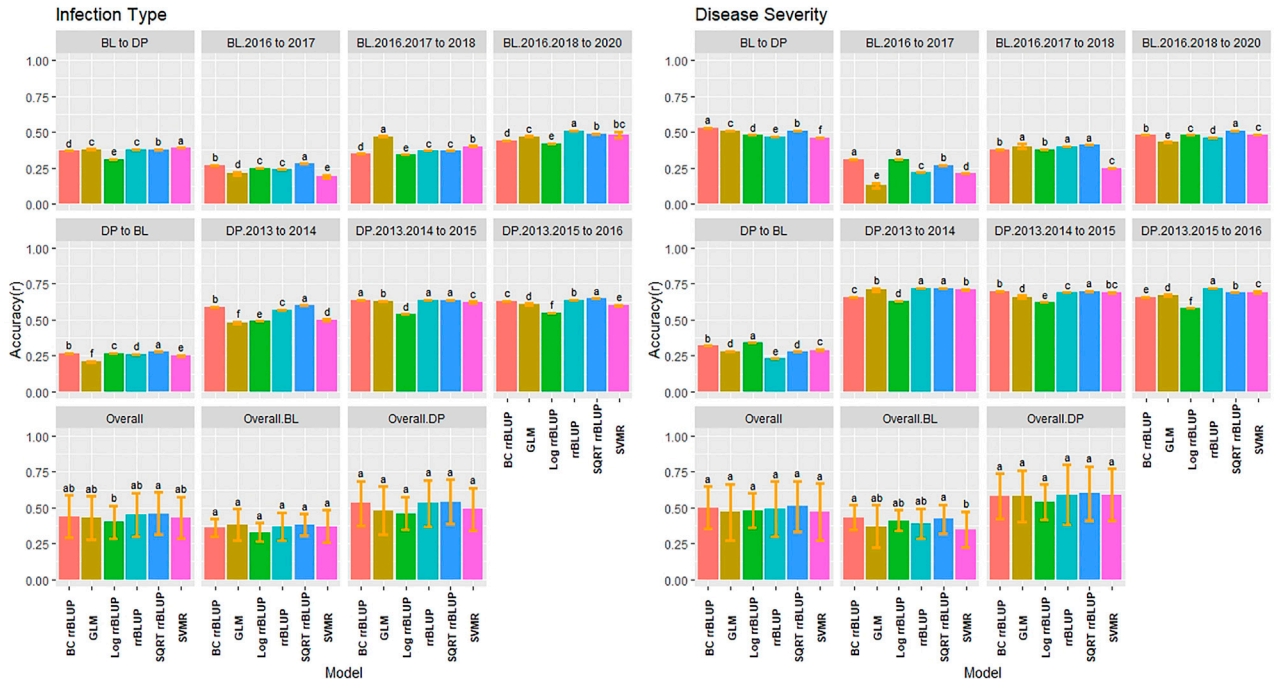

FIGURE 6 | Pairwise comparisons of genomic selection regression model accuracy (*r*) using validation sets for stripe rust infection type and disease severity. Pacific Northwest winter wheat diversity panel (DP) lines phenotyped from 2013 to 2016 in Central Ferry (DP_CF) and Pullman (DP_P), WA. Washington State University breeding lines phenotyped from 2016 to 2020 in Lind (BL_L) and Pullman (BL_P), WA. Model comparison across the DP (Overall.DP), BL (Overall.BL), and Overall scenarios. Models labeled with the same letter are not significantly different using Tukey's honestly significant difference. Bars indicate standard errors.

validation set class accuracy for the classification models were all relatively low except for the two- and three-class SVM model. Furthermore, there was no trend in increasing overall class accuracy by combining trials. The BL trials displayed the highest overall class accuracy with R^2 values reaching 0.78 for the two class SVM model (**Figure 7**). The low accuracies were presumably due to the increase in resistance and the models predicting zeros in the IT scale. Similar to the cross-validation scenarios, the reduced two class models reached a much higher accuracy across the majority of trials. Furthermore, the prediction accuracy can be accounted for by the low kappa values in the majority of models except the 2-Class SVM model reaching 0.40 (**Supplementary Figure S10**).

SEV displayed similar results with IT, but the BOR model had zero *r* for all scenarios. However, the accuracies increased in the BOR with the reduced class scales (**Figure 7**). The two- and three-class SVM models displayed very high accuracy with two-Class SVM reaching an overall class accuracy of 0.83 and maintained the high accuracy predicting the other population for both the BL and DP validation scenarios. Combining years did not result in improved accuracy. Furthermore, the kappa values were very low except for the two- and three-Class SVM models reaching kappa values of 0.63 in the DP (**Supplementary Figure S11**).

3.8 Validation Set Relative Efficiency

The RE of the regression models were high in the validation scenarios reaching RE values of 0.85 using the SQRT rrBLUP model (**Figure 8**). The BOR and SVM models displayed relatively low RE values compared to the regression models. This was presumably due to the BOR not being able to predict the phenotypic values of the majority of the lines; however, the RE was higher for the classification models than the cross-validations with only one scenario having a negative value (−0.20). For overall comparisons, there were significant differences compared to the cross-validation scenarios. The SQRT rrBLUP reached the highest overall RE with 0.60. Furthermore, the RE values were higher in the DP than the BL. Combining years was related to an increased RE for the regression models.

Consistent trends for SEV were observed with the transformed rrBLUP model RE values of 0.97, displaying very high RE compared to phenotypic selection (**Supplementary Figure S12**). The RE for SEV was relatively high for the rrBLUP and SVMR models predicting into the other population using the DP as the training population ranging from 0.58 to 0.86, further displaying the ability for the regression models to accurately predict across years and populations while dealing with skewed phenotypes.

4 DISCUSSION

GS has many advantages over traditional phenotypic selection and marker-assisted selection. Increased genetic gain and improved trait selection can be achieved by using GS (Heffner et al., 2010; Rutkoski et al., 2015; Michel et al., 2017).

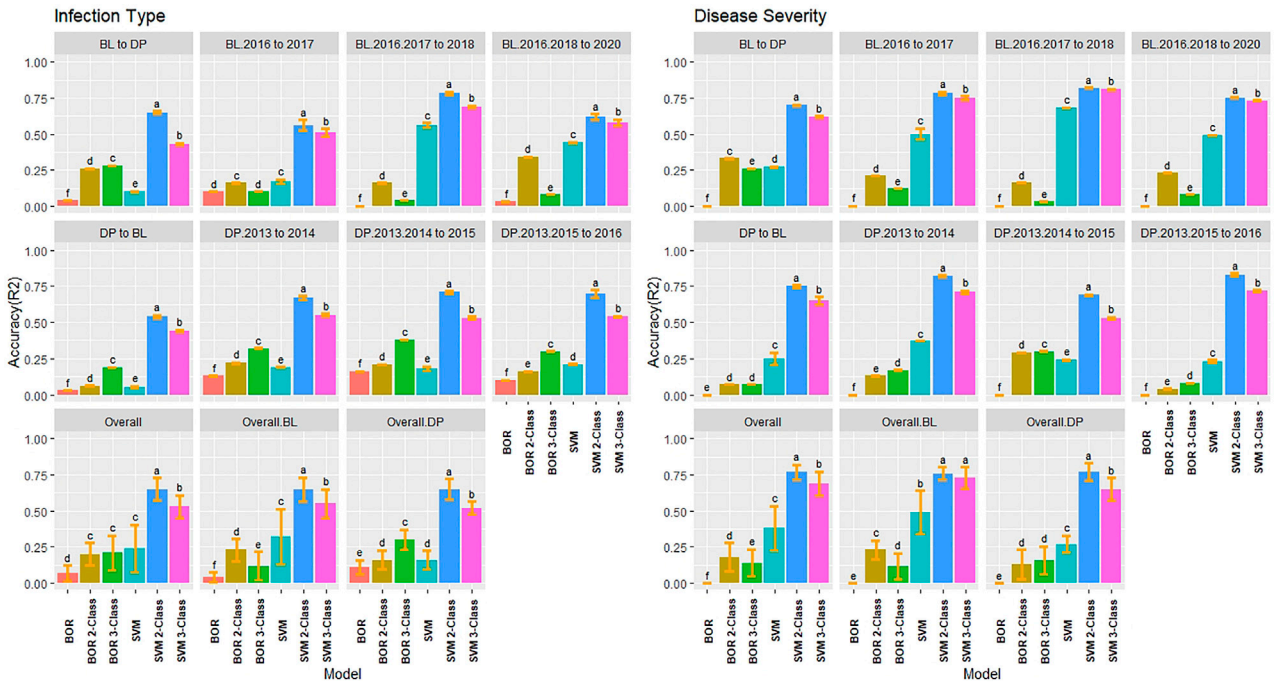

FIGURE 7 | Pairwise comparisons of genomic selection classification model accuracy (R^2) using validation sets for stripe rust infection type and disease severity. Pacific Northwest winter wheat diversity panel (DP) lines phenotyped from 2013 to 2016 in Central Ferry (DP_CF) and Pullman (DP_P), WA. Washington State University breeding lines phenotyped from 2016 to 2020 in Lind (BL_L) and Pullman (BP_L), WA. Model comparison across the DP (Overall.DP), BL (Overall.BL), and Overall scenarios. Models labeled with the same letter are not significantly different using Tukey's honestly significant difference. Bars indicate standard errors.

Furthermore, GS can aid in selection for traits dependent on the environment to display variation especially in years with little to no phenotypic variation for phenotypic selection. Plant breeding programs continually select and improve disease resistance due to the evolving race and pathogen changes along with the breakdown of resistance genes. Due to the high levels of resistance targeted within most plant breeding programs, positively skewed phenotypes generally result when selecting for disease resistance. Furthermore, disease resistance is commonly phenotyped in ordinal scales and percentages. The skewed and ordinal phenotypes pose challenges to utilizing regression models for GS (Montesinos-López et al., 2015a). However, most GS studies treat disease resistance as continuous values and utilize regression models and transformations for prediction, while only a few studies have used classification methods (Ornella et al., 2012, 2014; Rutkoski et al., 2014; Arruda et al., 2016; Muleta et al., 2017; González-Camacho et al., 2018; Merrick et al., 2021). In the current study, we compared several regression and classification methods for genomic prediction for skewed phenotypes in the context of stripe rust resistance in winter wheat and identified the best approaches to use for predicting traits with skewed distributions.

When utilizing GS for resistance to diseases such as stripe rust, GS approaches can capture the additive effects of APR and are therefore relevant for accumulating favorable alleles for rust resistance. GS can reach high levels of accuracy for stripe rust and other rust diseases (Ornella et al., 2012; Rutkoski et al., 2014, 2015; Muleta et al., 2017; Merrick et al., 2021). Because of the high levels of resistance and high heritability of disease resistance in most breeding programs, phenotypic selection and marker-assisted selection have been shown to be successful (Lande and Thompson, 1990). Even so, GS has been shown to be superior to marker-assisted selection in selecting for APR in the presence of major resistance genes (Merrick et al., 2021).

4.1 Accuracy of Regression Models

Regression models assume continuous and normally distributed phenotypes (Montesinos-López et al., 2015c). In the current study, the BL and DP populations displayed skewed distributions for both IT and SEV with inflations of zero due to the high levels of disease resistance. Among the primary approaches used for phenotypes that do not follow a normal distribution are disregarding the lack of normality or transforming the phenotypes to a normal distribution (Montesinos-López et al., 2015b). In the current study, we observed that even with the skewed distributions, the rrBLUP model without transformed phenotypes still displayed high accuracies and performed similarly to the highest-performing SQRT rrBLUP model in many scenarios. For example, there were no significant differences between SQRT rrBLUP and rrBLUP in the overall comparisons in the cross-validation (**Figure 3**) or validation set scenarios (**Figures 6**, **7**). These results support previous studies that utilized rrBLUP models for disease resistance (Rutkoski et al., 2014; Rutkoski et al., 2015; Juliana et al., 2017; Muleta et al., 2017; Merrick et al., 2021). The performance of the untransformed rrBLUP model may be due

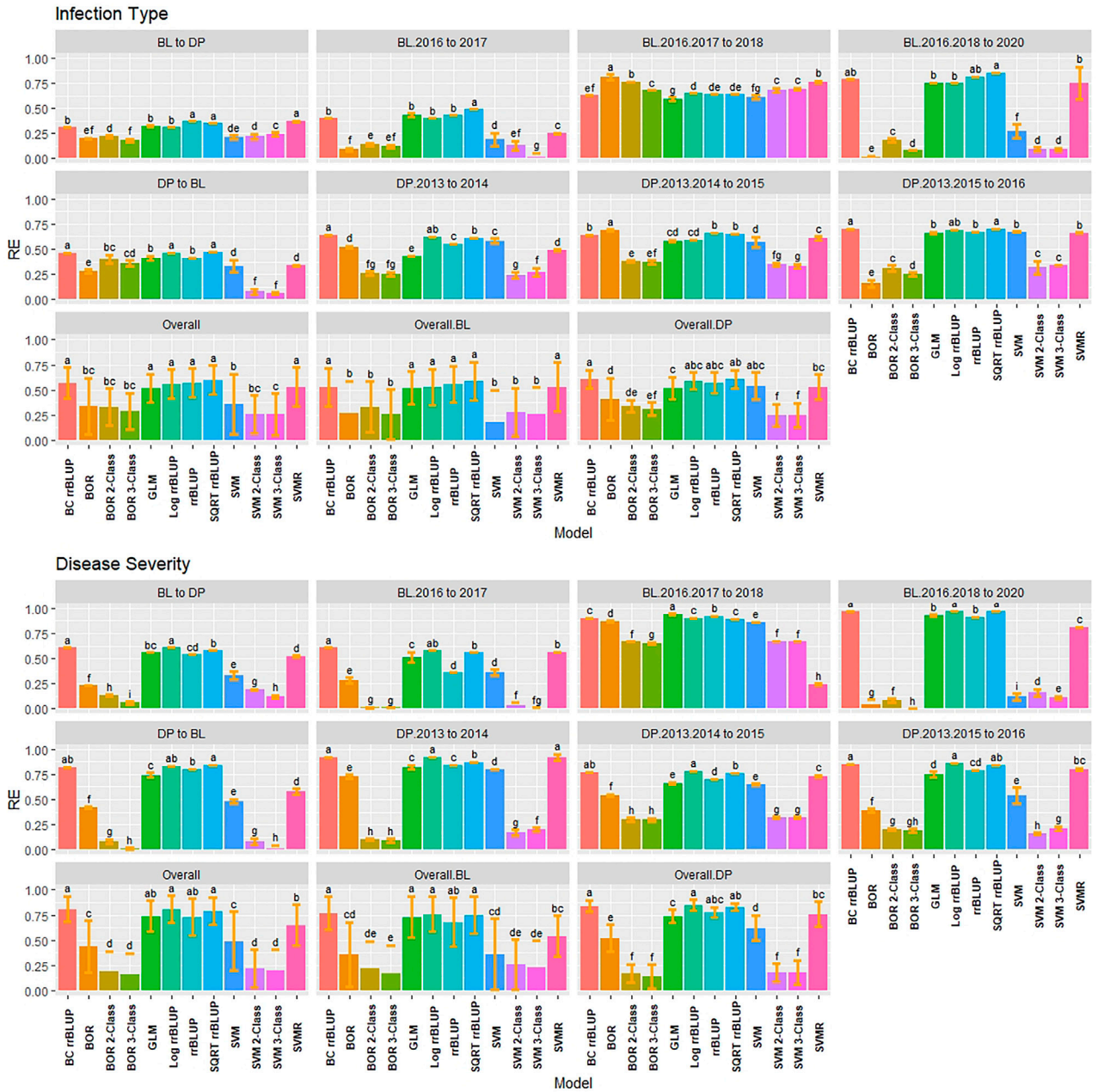

FIGURE 8 | Pairwise comparisons of genomic selection regression and classification model relative efficiency (RE) using validation sets for stripe rust infection type. Pacific Northwest winter wheat diversity panel (DP) lines phenotyped from 2013 to 2016 in Central Ferry (DP_CF) and Pullman (DP_P), WA. Washington State University breeding lines phenotyped from 2016 to 2020 in Lind (BL_L) and Pullman (BL_P), WA. Model comparison across the DP (Overall.DP), BL (Overall.BL), and Overall scenarios. Models labeled with the same letter are not significantly different using Tukey's honestly significant difference. Bars indicate standard errors.

to the central limit theorem, which argues that given a sufficient number of observations, the sampling distribution of the means can be assumed to be approximately normal (Stroup, 2015).

Transformations were introduced to stabilize variance and fulfill the homogenous variance assumption of linear regression models (Bartlett, 1947). However, transformations have shown to produce a loss of accuracy and power in small sample size (Stroup, 2015). Furthermore, in our study, the log and BC transformations displayed lower accuracy than the SQRT transformation. One of the problems with log transformations is the large number of zeros due to the presence of highly resistant lines in both the BL and DP populations. This occurrence constrains the transformation to stabilize variance and transform the phenotypes to follow a normal distribution (O'Hara and Kotze, 2010). Furthermore, log transformations yield downwardly biased estimates, whereas SQRT does not (Stroup, 2015). The BC

transformations is a powerful transformation that raise numbers to an exponent; nonetheless, BC requires lambda estimation and can theoretically be the same as the SQRT transformation at $\lambda = 0.50$ (Osborne, 2010). Therefore, if the optimal λ is not chosen correctly, the BC may not appropriately stabilize the variance of the data.

The SQRT transformation proved to perform very well for both accuracy and RE across populations, cross-validations, and validation scenarios in the current study. The SQRT transformations showed the ability to have higher accuracy and reduced RMSE compared with the untransformed data for the rrBLUP model. In Poisson distributions similar to the skewed phenotypes of our study, the variance is equal to the mean, and the SQRT is recommended to stabilize variance in those scenarios (Bartlett 1947); this could have resulted in increased performance for the SQRT transformation. Overall, the appropriate method must be chosen carefully when implementing data transformation on breeding programs.

Using the GLM model, high accuracy (0.66 and 0.76) in both the DP and BL training populations were observed. The performance of GLM was noted to be dependent on the distribution of the phenotypes. The GLM performed similarly to the rrBLUP model in the highly skewed selected BL population, but displayed statistically significant lower accuracies in the less skewed unselected DP population. Poisson GLMs, which were implemented in the present work, have been shown to display superior accuracy while correctly fitting the data (O'Hara and Kotze, 2010; Montesinos-López et al., 2015b; Montesinos-López et al., 2016; Montesinos-López et al., 2020; Stroup, 2015). The Poisson GLM accurately models count and ordinal data and is therefore suited for skewed phenotypes such as disease resistance (Ornella et al., 2014; Montesinos-López et al., 2015a; Montesinos-López et al., 2016). Furthermore, the GLM models outperformed deep learning models in a previous study (Montesinos-López et al., 2020). The utilization of GLMs should be implemented in scenarios with the appropriate distribution of phenotypes.

Non-parametric models such as SVMR, which has no underlying assumption on the distribution of the phenotypes, performed better than the LOG and BC transformations, and similar to the GLM model in the current study. Previously, the SVMR model has been shown to have superior prediction and RE values over parametric and semi-parametric models for predicting disease resistance due to the skewed phenotypes (González-Camacho et al., 2018). This demonstrates that the SVMR can accurately predict skewed phenotypes without the need to transform the data. SVM regression maps samples from a predictor space to a high-dimensional feature space using a non-linear kernel function and then completes linear regression in the feature space (Jannink et al., 2010). Consequently, this creates the ability for the SVMR to predict skewed phenotypes and allows the model to learn the complexity of the training population without imposing structure on the data (González-Camacho et al., 2018).

The SQRT rrBLUP models performed better than the SVMR model in overall prediction accuracy across many scenarios. The lack of advantage in regression scenarios was also observed by Ornella et al. (2014), where reproducing kernel Hilbert Space models were observed to be statistically significant for all yield datasets over SVM and random forest models. In the current study, the subordinate performance of the SVMR models is presumably due to the mostly additive effect of stripe rust resistance. Once the skewed phenotypes are properly modeled, the advantage of non-parametric models that also model non-additive effects disappears (Ornella et al., 2014; Poland and Rutkoski, 2016).

4.2 Accuracy of Classification Models

In the present study, BOR models displayed the lowest accuracies and RE across all the classification and regression models, particularly in the DP population. Conversely, the BOR models using reduced classes reached the highest overall class accuracy over all models with $r = 0.99$ for the BL. In contrast, when the accuracy was high in the BL, the kappa values were low. The opposite was shown in the DP with low overall class accuracies and moderate kappa values indicating that the high overall class accuracy and low kappa values were a result of the BOR model consistently predicting zeros and the inability to predict the other classes. Furthermore, in the validation sets, the BOR performed very poorly, and resulted in near-zero overall class accuracy and kappa values for both IT and SEV. The BOR model uses ordinal regression that is suitable for count and censored data; nevertheless, the BOR model uses the probit link function that does not explicitly model non-normal distribution such as the Poisson distribution model by the GLM model in our study (Montesinos-López et al., 2015a). Altogether, our results showed that the BOR model is not appropriate for the highly skewed phenotypes in our study.

We also used SVMs as the non-parametric machine learning model for both regression and classification. The advantage in classification and regression using SVM models for disease resistance has been previously demonstrated (Ornella et al., 2014; González-Camacho et al., 2018). Contrary to the BOR results, the SVMs consistently displayed high accuracies throughout the locations and years for both DP and BL training populations. However, the SVM showed lower kappa values than the BOR in many scenarios. This was not the trend in the validation sets, where the full-scale BOR and SVM displayed poor accuracy and kappa values. The consistent accuracy of SVM over BOR may be due to the non-parametric nature of the SVM models. The SVM model is implemented similar to the SVMR model and uses soft classifiers to calculate the probability of the class rather than hard classifiers that directly target the decision boundary and allow the model to be flexible (Ornella et al., 2014). Based on the results for BOR and SVM, classification models need to be compared by both overall class accuracy as well as a metric such as kappa that accounts for individual class accuracy.

The precision of the classification models depends on the number of individuals in a given class. In our study, we implemented up-sampling (i.e., random sampling with replacement) to increase the minority class to the same size of the majority class and reduce the effect of class imbalance (Kuhn, 2008). However, our results showed that with imbalanced class frequency due to skewed phenotypes, even resampling techniques such as up-sampling failed to accurately predict disease resistance. Another approach to deal with class frequency is to reduce the number of overall classes. We then binned classes to create 2- and 3-Class prediction scenarios. Reducing the class

scale to two creates a binary classification model that has been shown to outperform other regression and classification models (Ornella et al., 2014). By reducing the number of classes, we also decreased the effect of class imbalances. Accuracy as well as kappa increased specifically for the SVM by reducing the class scales. This observation was seen even in the validation sets, which resulted in the SVM 2-Class models achieving both high accuracy and kappa values, consistent with previous studies on the effects of reduced classes (Ornella et al., 2014; González-Camacho et al., 2016). Therefore, by reducing the class scale, classification models such as SVM can accurately predict skewed phenotypes such as disease resistance.

4.3 Relative Efficiency

RE compares the expected genetic gain when selecting based on GEBVs compared to phenotypic selection. The RE can be used as an indicator of the performance of a model when used for truncation selection and expected genetic gain (Ornella et al., 2014; González-Camacho et al., 2018). Since classification and regression do not use the same metrics for performance, simply comparing accuracies is not possible; hence, we used RE for comparisons. A selection intensity of 15% was used based on a previous study (Ornella et al., 2014). In the current work, the rrBLUP models and SMVR displayed high RE values across both cross-validation and validation sets for IT and SEV with values above 0.90. SVMR models have been shown to have superior RE values for classification in disease resistance (Ornella et al., 2014). The high RE values indicated that accuracy is linear in the regression models, but this was not the case for the classification models. The classification models displayed relatively low RE values and, in some cases, negative values. Both the SVM and BOR models displayed the inability to select the top 15% performers for stripe rust resistance. The large amounts of zeros (i.e., disease resistant phenotypes) skew the prediction accuracy for the classification models to the very high, with low kappa and RE values. The classification models failed to overcome the skewed phenotypes even with up-sampling and reduction of classes. Therefore, similar to our results for prediction accuracy, regression models outperformed classification models and displayed their ability to predict and select skewed phenotypes.

4.4 Training Population Comparison

We compared the performance of GS models in different training populations, environments, and phenotypic distributions. The effect of environment was less apparent than the effect of distribution. The differences in distribution of phenotypes for disease resistance is readily apparent between populations. The two populations were used to compare the effects of a selected and unselected population with varying degrees and sources of resistance. The BL population, consisting of WSU breeding lines that were selected for disease resistance prior to field trials, is extremely skewed for both IT and SEV. Therefore, there is already a selection pressure for high levels of resistance to stripe rust in the current study. In contrast, the DP appears less skewed with more variation for disease resistance, a consequence of the population consisting of diverse varieties from multiple breeding programs in the Pacific Northwest region of the US. The DP included lines from the WSU breeding program, but the other varieties were not bred and selected specifically for resistance to the stripe rust races present in our study. Additionally, the sources of stripe rust resistance genes vary more in the DP compared to the BL. The frequency and type of stripe rust races along with major genes for stripe rust resistance for these two populations were compared in depth in Merrick et al. (2021).

The differences in skewness between the populations affected the performance of the GS models in each population. The GLM models accurately predicted the extremely skewed BL trials similar to the other regression models because the skewed phenotypes follow the Poisson distribution rather than the normal distribution. However, the GLM model displayed lower accuracies in the less skewed DP. In addition to the distribution that is modeled, the skewness affects the frequency of classes used in classification models. In the extremely skewed BL, the classification models have high accuracy and low kappa, displaying the prediction of mainly zeros. However, as mentioned previously, the reduction of classes helps decrease the effect of class imbalance and increased accuracy.

The differences in accuracies between populations can also be attributed to the genetic relatedness of the populations (Asoro et al., 2011). The effect of the population on accuracy is due to both population structure and genetic relatedness (Habier et al., 2007; Asoro et al., 2011; Mirdita et al., 2015). We used the elbow method to determine the number of clusters when examining PCs for our populations and resulted in four distinct clusters. Consequently, the prediction accuracy for the BL cross-validations was higher than the DP. When independently predicting other populations as seen in the validations sets, we generally observe a decrease in accuracy (Merrick and Carter, 2021; Merrick et al., 2021). Interestingly, though, there was an increase in accuracy when using the BL to predict the DP in the validation sets. However, a decrease in prediction accuracy was observed when the DP predicted the BL. However, this was only seen in the regression models for predicting SEV in the validation sets. Furthermore, this trend is not seen in the classification models that display consistent accuracy across validation scenarios. This may be due to the effect of predicting a less skewed population in which regression models generally have better performance compared to predicting more skewed distributions (Montesinos-López et al., 2015a).

The increase in prediction accuracy with the increased combination of years in both our cross-validation and validation sets can be attributed to the increase of phenotypic data points and decrease in skewness and accounting for the genotype-by-environment interaction (GEI). The trials in our study were dependent on the natural occurrence and pressure of stripe rust. Therefore, the skewness of the populations, individual years, and locations may be due to not only the levels of resistance within the populations, but also the general disease pressure for stripe rust. By combining environments, we can account for the GEI in our phenotypic adjustments and increase our prediction accuracy (Crossa et al., 2014; Jarquín et al., 2014; Haile et al., 2020; Merrick and Carter 2021; Merrick et al., 2021). The increased accuracy by accounting for GEI can be seen in the validation sets. The DP displayed higher accuracies in the validation sets as the

DP consisted of the same lines each year, whereas the BL consists of different lines in both years and locations. By screening the same lines each year, the environmental effect can be effectively accounted for. However, the trend for increasing accuracy and RE values by combining years was not seen in the classification models. This was due to the continued large class imbalances even when combining years. Therefore, there is a need to develop training populations carefully to balance class frequencies for the classification models. Even so, the reduced class SVM models displayed the ability to overcome the class frequencies regardless of year combinations. Overall, the rrBLUP and reduced class classification models displayed the ability to accurately predict populations and environments with skewed phenotypes.

4.5 Applications in Breeding

GS is becoming more cost-effective due to the decreasing costs of high-throughput genotyping. With the increased use of GS comes its utilization for the prediction of complex traits (e.g., disease resistance), which do not always follow the assumptions of the commonly used models (Montesinos-López et al., 2015a). Instead of applying the same approach to every trait, breeders will need to customize their GS models to achieve accurate GEBVs for selection. With the integration of data science and plant breeding, the availability of different prediction models has resulted in an increased efficiency of implementing GS for a wide range of traits. This study showed that with the appropriate choice of model and transformation, even the commonly used GS regression model, rrBLUP, can be utilized for predicting complex traits, such as stripe rust resistance, that do not follow a normal distribution. Furthermore, this study demonstrated the ability to integrate selection decisions and GS by utilizing classification models. Reducing classes resulted in higher predictions due to decreasing the number of outcomes the models need to account for, especially for classes with only a few observations. Moreover, by reducing the number of classes, we not only predict resistance more accurately, but also couple in selection decisions. By reducing the number of classes for IT from ten to two, we can either keep or discard lines. Ultimately, by using various GS schemes with regression and classification models, breeders can reduce the number of selection decisions made for disease resistance and focus on selecting other important traits such as grain yield.

5 CONCLUSION

This study compared GS regression and classification models' ability to accurately predict populations with different levels of disease resistance and distributions. The varying results for the classification and transformation methods displayed the need to choose the prediction model carefully based on the phenotype distribution. For trials that display a Poisson distribution that is skewed to lower ordinal values, a GLM or reduced class binomial classification model can be implemented. However, the SQRT and SVMR models displayed the flexibility across varying distributions, and consistently predicted stripe rust with high accuracies. Moreover, combining years increased the prediction accuracies for regression models, but failed to increase the overall class accuracy for classification models due to imbalance class frequencies. Additionally, regression models displayed high RE, indicating their ability to select accurately like phenotypic selection. Overall, SQRT transformation using rrBLUP and SVM regression models displayed the highest combination of accuracy and RE across the regression and classification models. Furthermore, a classification system based on SVM with a 2-Class scale can be implemented not only to predict resistance more accurately, but also to couple in selection decisions. This study showed that breeders can use linear and non-parametric regression models using their own breeding lines over combined years to accurately predict skewed phenotypes.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://github.com/lfmerrick21/Regression-vs-Classification.

AUTHOR CONTRIBUTIONS

LM: conceptualized the idea, analyzed data, and drafted the manuscript; DL: reviewed and edited the manuscript; XC: reviewed and edited the manuscript; AC: supervised the study, conducted field trials, edited the manuscript, and obtained the funding for the project.

FUNDING

This research was partially funded by the National Institute of Food and Agriculture (NIFA) of the U.S. Department of Agriculture (Award Number 2016-68004-24770), Hatch project 1014919, and the O.A. Vogel Research Foundation at Washington State University.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Washington State University Winter Wheat Breeding Program personnel Gary Shelton and Kyall Hagemeyer for plot maintenance and data collection under field conditions. We would also like to thank Adrienne Burke, Gina Brown-Guedira, Jared Smith, Brian Ward, and staff at the Eastern Regional Small Grains Genotyping Laboratory for their assistance with DNA library prep and GBS sequencing and analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.835781/full#supplementary-material

REFERENCES

Appels, R., Eversole, K., Appels, R., Eversole, K., Feuillet, C., Keller, B., et al. (2018). Shifting the Limits in Wheat Research and Breeding Using a Fully Annotated Reference Genome. *Science* 361, eaar7191. doi:10.1126/science.aar7191

Arruda, M. P., Lipka, A. E., Brown, P. J., Krill, A. M., Thurber, C., Brown-Guedira, G., et al. (2016). Comparing Genomic Selection and Marker-Assisted Selection for Fusarium Head Blight Resistance in Wheat (*Triticum aestivum* L.). *Mol. Breed.* 36, 84. doi:10.1007/s11032-016-0508-5

Asoro, F. G., Newell, M. A., Beavis, W. D., Scott, M. P., and Jannink, J. L. (2011). Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats. *Plant Genome* 4, 132. doi:10.3835/plantgenome2011.02.0007

Bartlett, M. S. (1947). The Use of Transformations. *Biometrics* 3, 39–52. doi:10.2307/3001536

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: Software for Association Mapping of Complex Traits in Diverse Samples. *Bioinformatics* 23, 2633–2635. doi:10.1093/bioinformatics/btm308

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* 103, 338–348. doi:10.1016/j.ajhg.2018.07.015

Chen, X. (2013). High-temperature Adult-Plant Resistance, Key for Sustainable Control of Stripe Rust. *Am. J. Plant Sci.* 04 (03), 608–627. doi:10.4236/ajps.2013.43080

Chen, X. (2020). Pathogens Which Threaten Food Security: Puccinia Striiformis, the Wheat Stripe Rust Pathogen. *Food Sec* 12, 239–251. doi:10.1007/s12571-020-01016-z

Chen, X., Washington, S. U., and Line, R. F. (1995a). Gene Action in Wheat Cultivars for Durable, High-Temperature, Adult-Plant Resistance and Interaction with Race-specific , Seedling Resistance to Puccinia Striiformis. *Phytopathol. USA* 85 (5), 567. doi:10.1094/phyto-85-567

Chen, X., Washington, S. U., and Line, R. F. (1995b). Gene Number and Heritability of Wheat Cultivars with Durable, High-Temperature, Adult-Plant (HTAP) Resistance and Interaction of HTAP and Race-specific Seedling Resistance to Puccinia Striiformis. *Phytopathol. USA* 85 (5), 573. doi:10.1094/phyto-85-573

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., et al. (2014). Genomic Prediction in CIMMYT maize and Wheat Breeding Programs. *Heredity* 112, 48–60. doi:10.1038/hdy.2013.16

Cullis, B. R., Smith, A. B., and Coombes, N. E. (2006). On the Design of Early Generation Variety Trials with Correlated Data. *Jabes* 11, 381–393. doi:10.1198/108571106X154443

de Mendiburu, F., and de Mendiburu, M. F. (2019). *Package 'agricolae.' R Package Version*, 2–8.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A Robust, Simple Genotyping-By-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* 6, e19379. doi:10.1371/journal.pone.0019379

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome* 4, 250–255. doi:10.3835/plantgenome2011.08.0024

Federer, W. F. (1956). *Experimental Design, Theory and Application*. New York: Macmillan.

Gareth, J., Daniela, W., Trevor, H., and Robert, T. (2013). *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *PLOS ONE* 9, e90346. doi:10.1371/journal.pone.0090346

Goldman, I. (2019). *Plant Breeding Reviews*. Chichester, UK: John Wiley & Sons.

González-Camacho, J. M., Ornella, L., Pérez-Rodríguez, P., Gianola, D., Dreisigacker, S., and Crossa, J. (2018). Applications of Machine Learning Methods to Genomic Selection in Breeding Wheat for Rust Resistance. *Plant Genome* 11, 170104. doi:10.3835/plantgenome2017.11.0104

González-Camacho, J. M., Crossa, J., Pérez-Rodríguez, P., Ornella, L., and Gianola, D. (2016). Genome-enabled Prediction Using Probabilistic Neural Network Classifiers. *BMC Genomics* 17, 1–16.

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. *Genetics* 177, 2389–2397. doi:10.1534/genetics.107.081190

Haile, T. A., Walkowiak, S., N'Diaye, A., Clarke, J. M., Hucl, P. J., Cuthbert, R. D., et al. (2020). Genomic Prediction of Agronomic Traits in Wheat Using Different Models and Cross-Validation Designs. *Theor. Appl. Genet.* 134, 381–398. doi:10.1007/s00122-020-03703-z

Hastie, T., Qian, J., and Tay, K. (2016). *An Introduction to Glmnet*.

Heffner, E. L., Lorenz, A. J., Jannink, J. L., and Sorrells, M. E. (2010). Plant Breeding with Genomic Selection: Gain Per Unit Time and Cost. *Crop Sci.* 50, 1681–1690. doi:10.2135/cropsci2009.11.0662

Hyndman, R. J., and Khandakar, Y. (2008). Automatic Time Series Forecasting: the Forecast Package for R. *J. Stat. Softw.* 27, 1–22. doi:10.18637/jss.v027.i03

Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). Genomic Selection in Plant Breeding: from Theory to Practice. *Brief. Funct. Genomics* 9, 166–177. doi:10.1093/bfgp/elq001

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A Reaction Norm Model for Genomic Selection Using High-Dimensional Genomic and Environmental Data. *Theor. Appl. Genet.* 127, 595–607. doi:10.1007/s00122-013-2243-1

Juliana, P., Singh, R. P., Singh, P. K., Crossa, J., Huerta-Espino, J., Lan, C., et al. (2017). Genomic and Pedigree-Based Prediction for Leaf, Stem, and Stripe Rust Resistance in Wheat. *Theor. Appl. Genet.* 130, 1415–1430. doi:10.1007/s00122-017-2897-1

Kamiak (2021). *High Performance Computing*. Pullman, WA: Washington State University. Available at: https://hpc.wsu.edu/ [Accessed January 21, 2021].

Karatzoglou, A., Smola, A., Hornik, K., and Karatzoglou, M. A. (2019). *Package 'kernlab.'* CRAN R Proj.

Kassambara, A., and Kassambara, M. A. (2020). *Package 'ggpubr*.

Klarquist, E., Chen, X., and Carter, A. (2016). Novel QTL for Stripe Rust Resistance on Chromosomes 4A and 6B in Soft White Winter Wheat Cultivars. *Agronomy* 6, 4. doi:10.3390/agronomy6010004

Kuhn, M. (2008). Building Predictive Models inRUsing thecaretPackage. *J. Stat. Soft.* 28, 1–26. doi:10.18637/jss.v028.i05

Lande, R., and Thompson, R. (1990). Efficiency of Marker-Assisted Selection in the Improvement of Quantitative Traits. *GENETICS* 124, 743–756. doi:10.1093/genetics/124.3.743

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Line, R. F., and Qayoum, A. (1992). Virulence, Aggressiveness, Evolution and Distribution of Races of Puccinia Striiformis (The Cause of Stripe Rust of Wheat) in North America, 1968-87. *Tech. Bull. USA* 1788. https://handle.nal.usda.gov/10113/CAT92983836

Liu, L., Yuan, C. Y., Wang, M. N., See, D. R., Zemetra, R. S., and Chen, X. M. (2019). QTL Analysis of Durable Stripe Rust Resistance in the North American winter Wheat Cultivar Skiles. *Theor. Appl. Genet.* 132, 1677–1691. doi:10.1007/s00122-019-03307-2

Liu, Y., Qie, Y., Wang, M., and Chen, X. (2020). Genome-Wide Mapping of Quantitative Trait Loci Conferring All-Stage and High-Temperature Adult-Plant Resistance to Stripe Rust in Spring Wheat Landrace PI 181410. *Ijms* 21, 478. doi:10.3390/ijms21020478

Merrick, L. F., Burke, A. B., Chen, X., and Carter, A. H. (2021). Breeding with Major and Minor Genes: Genomic Selection for Quantitative Disease Resistance. *Front. Plant Sci.* 12, 1599. doi:10.3389/fpls.2021.713667

Merrick, L. F., and Carter, A. H. (2021). Comparison of Genomic Selection Models for Exploring Predictive Ability of Complex Traits in Breeding Programs. *Plant Genome* 14, e20158. doi:10.1002/tpg2.20158

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-wide Dense Marker Maps. *Genetics* 157, 1819–1829. doi:10.1093/genetics/157.4.1819

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., et al. (2019). Package 'e1071. *R. J.*

Michel, S., Ametz, C., Gungor, H., Akgöl, B., Epure, D., Grausgruber, H., et al. (2017). Genomic Assisted Selection for Enhancing Line Breeding: Merging Genomic and Phenotypic Selection in winter Wheat Breeding Programs with Preliminary Yield Trials. *Theor. Appl. Genet.* 130, 363–376. doi:10.1007/s00122-016-2818-8

Mirdita, V., He, S., Zhao, Y., Korzun, V., Bothe, R., Ebmeyer, E., et al. (2015). Potential and Limits of Whole Genome Prediction of Resistance to Fusarium Head Blight and Septoria Tritici Blotch in a Vast Central European Elite winter Wheat Population. *Theor. Appl. Genet.* 128, 2471–2481. doi:10.1007/s00122-015-2602-1

Montesinos-López, A., Montesinos-López, O. A., Crossa, J., Burgueño, J., Eskridge, K. M., Falconi-Castillo, E., et al. (2016). Genomic Bayesian Prediction Model for Count Data with Genotype × Environment Interaction. *G3amp58 GenesGenomesGenetics* 6, 1165–1177. doi:10.1534/g3.116.028118

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Burgueño, J., and Eskridge, K. (2015a). Genomic-Enabled Prediction of Ordinal Data with Bayesian Logistic Ordinal Regression. *G3amp58 GenesGenomesGenetics* 5, 2113–2126. doi:10.1534/g3.115.021154

Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., de los Campos, G., Eskridge, K., and Crossa, J. (2015b). Threshold Models for Genome-Enabled Prediction of Ordinal Categorical Traits in Plant Breeding. *G3amp58 GenesGenomesGenetics* 5, 291–300. doi:10.1534/g3.114.016188

Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., Eskridge, K., He, X., Juliana, P., et al. (2015c). Genomic Prediction Models for Count Data. *Jabes* 20, 533–554. doi:10.1007/s13253-015-0223-4

Montesinos-López, O. A., Montesinos-López, J. C., Singh, P., Lozano-Ramirez, N., Barrón-López, A., Montesinos-López, A., et al. (2020). A Multivariate Poisson Deep Learning Model for Genomic Prediction of Count Data. *G3 Genes Genomes Genet.* 10, 4177–4190. doi:10.1534/g3.120.401631

Muleta, K. T., Bulli, P., Zhang, Z., Chen, X., and Pumphrey, M. (2017). Unlocking Diversity in Germplasm Collections via Genomic Selection: A Case Study Based on Quantitative Adult Plant Resistance to Stripe Rust in Spring Wheat. *Plant Genome* 10, 0. doi:10.3835/plantgenome2016.12.0124

O'Hara, R. B., and Kotze, D. J. (2010). Do Not Log-Transform Count Data. *Methods Ecol. Evol.* 1, 118–122. doi:10.1111/j.2041-210X.2010.00021.x

Ornella, L., Pérez, P., Tapia, E., González-Camacho, J. M., Burgueño, J., Zhang, X., et al. (2014). Genomic-enabled Prediction with Classification Algorithms. *Heredity* 112, 616–626. doi:10.1038/hdy.2013.144

Ornella, L., Singh, S., Perez, P., Burgueño, J., Singh, R., Tapia, E., et al. (2012). Genomic Prediction of Genetic Values for Resistance to Wheat Rusts. *Plant Genome* 5, 136–148. doi:10.3835/plantgenome2012.07.0017

Osborne, J. (2010). Improving Your Data Transformations: Applying the Box-Cox Transformation. *Pract. Assess. Res. Eval.* 15. doi:10.7275/qbpc-gk17

Pérez, P., and de los Campos, G. (2014). Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics* 198, 483–495. doi:10.1534/genetics.114.164442

Peterson, R. F., Campbell, A. B., and Hannah, A. E. (1948). A DIAGRAMMATIC SCALE FOR ESTIMATING RUST INTENSITY ON LEAVES AND STEMS OF CEREALS. *Can. J. Res.* 26c, 496–500. doi:10.1139/cjr48c-033

Poland, J. A., and Rife, T. W. (2012). Genotyping-by-Sequencing for Plant Breeding and Genetics. *Plant Genome* 5, 92. doi:10.3835/plantgenome2012.05.0005

Poland, J., and Rutkoski, J. (2016). Advances and Challenges in Genomic Selection for Disease Resistance. *Annu. Rev. Phytopathol.* 54, 79–98. doi:10.1146/annurev-phyto-080615-100056

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/ (Accessed January 18, 2022).

Riedelsheimer, C., Technow, F., and Melchinger, A. E. (2012). Comparison of Whole-Genome Prediction Models for Traits with Contrasting Genetic Architecture in a Diversity Panel of maize Inbred Lines. *BMC Genomics* 13, 452. doi:10.1186/1471-2164-13-452

Rutkoski, J. E., Poland, J. A., Singh, R. P., Huerta-Espino, J., Bhavani, S., Barbier, H., et al. (2014). Genomic Selection for Quantitative Adult Plant Stem Rust Resistance in Wheat. *Plant Genome* 7, 0. doi:10.3835/plantgenome2014.02.0006

Rutkoski, J., Singh, R. P., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J. L., et al. (2015). Efficient Use of Historical Data for Genomic Selection: A Case Study of Stem Rust Resistance in Wheat. *Plant Genome* 8, 0. doi:10.3835/plantgenome2014.09.0046

SAS Institute, Inc (2011). *SAS® 9.3 System Options: Reference*. NC: SAS Institute Inc Cary.

Schmidt, P., Hartung, J., Bennewitz, J., and Piepho, H.-P. (2019). Heritability in Plant Breeding on a Genotype-Difference Basis. *Genetics* 212, 991–1008. doi:10.1534/genetics.119.302134

Stroup, W. W. (2015). Rethinking the Analysis of Non-Normal Data in Plant and Soil Science. *Agron.j.* 107, 811–827. doi:10.2134/agronj2013.0342

Wang, X., Xu, Y., Hu, Z., and Xu, C. (2018). Genomic Selection Methods for Crop Improvement: Current Status and Prospects. *Crop J.* 6, 330–340. doi:10.1016/j.cj.2018.03.001

Ward, B. P., Brown-Guedira, G., Tyagi, P., Kolb, F. L., Van Sanford, D. A., Sneller, C. H., et al. (2019). Multienvironment and Multitrait Genomic Selection Models in Unbalanced Early-Generation Wheat Yield Trials. *Crop Sci.* 59, 491–507. doi:10.2135/cropsci2018.03.0189

Wickham, H. (2011). ggplot2. *Wires Comp. Stat.* 3, 180–185. doi:10.1002/wics.147

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.