RESEARCH

Sources of Resistance to Stem Rust (Ug99) in Bread Wheat and Durum Wheat Identified Using Focused Identification of Germplasm Strategy

Dag Terje Filip Endresen,* Kenneth Street, Michael Mackay, Abdallah Bari, Ahmed Amri, Eddy De Pauw, Kumarse Nazari, and Amor Yahyaoui

ABSTRACT

The focused identification of germplasm strategy (FIGS) has been validated using predictive computer models in simulation studies to predict a priori known trait scores. This study was designed as a "blind" study where the person calculating the computer model did not know the actual trait scores. This study design provides a more realistic test of the predictive capacity of the FIGS approach compared to previous studies. Furthermore this study also explored the suitability of FIGS for the identification of resistance in bread wheat (*Triticum aestivum* L. subsp. *aestivum*) and durum wheat [*Triticum turgidum* L. subsp. *durum* (Desf.) Husn.] to Ug99—a strain of stem rust (*Puccinia graminis* Pers. f. sp. *tritici* Eriks. & Henn.) and typified to race TTKSK. The predictions were validated against a dataset with the screening of wheat accessions conducted in Yemen in 2008. Only a small training set representing 20% of the trait screening results was disclosed to the person conducting the data analysis for the calibration of the prediction model. The hit rate for identification of Ug99-resistant accessions was more than two times higher when using the FIGS approach compared to a random selection of accessions. These results suggested that FIGS was well suited for the identification of samples with resistance to fungal pathogens. It is therefore recommended that FIGS approach be used as a complement to expert knowledge and experience when selecting accessions for plant breeding and crop research activities.

D.T.F. Endresen, Global Biodiversity Information Facility (GBIF), Universitetsparken 15, DK-2100 Copenhagen, Denmark; K. Street, A. Bari, A. Amri, E. De Pauw, K. Nazari, and A. Yahyaoui, International Center for Agricultural Research in the Dry Areas (ICARDA), P.O. Box 5466, Aleppo, Syrian Arab Republic; M. Mackay, Bioversity International, Via dei Tre Denari 472/a, 00057 Maccarese (Fiumicino) Rome, Italy. Received 13 Aug. 2011. *Corresponding author (dag.endresen@gmail.com).

Abbreviations: FIGS, focused identification of germplasm strategy; kNN, k-nearest neighbor; LR+, positive diagnostic likelihood ratio; MR, moderately resistant; MS, moderately susceptible; NPGS, National Plant Germplasm System; PC, principal component; PCA, principal component analysis; PPV, positive predictive value; PRESS, predicted residual sum of squares; R, resistant; S, susceptible; SIMCA, soft independent modeling of class analogy.

The focused identification of germplasm strategy (FIGS) provides a sampling strategy to identify accessions from gene-bank collections for a target trait property (Mackay, 1986, 1990, 1995; Mackay and Street, 2004). The FIGS approach assumes a predictive link between some eco-geographic parameters of the original collecting site of germplasm (with focus on landraces and crop wild relatives) and a target adaptive trait, such as disease resistance. Recent studies have proposed the algorithms and methods for implementation of trait mining using FIGS (El Bouhssini et al., 2009, 2011; Endresen, 2010; Endresen et al., 2011; Bari et al., 2011). Zeven (1998) presented a review of definitions and classifications for landraces and suggested the definition provided by Mansholt (1909), which reads, "an autochthonous landrace is a variety with a high capacity to tolerate biotic and abiotic stress, resulting in a high

Published in Crop Sci. 52:764–773 (2012).
doi: 10.2135/cropsci2011.08.0427

yield stability and an intermediate yield level under a low input agricultural system" (Zeven 1998, 137).

The experiment followed the same methodology as described with a recent FIGS trait-mining study for stem rust (Endresen et al., 2011) and was designed to validate this approach in a "blind" study. The approach of using a blind study is important because the previous studies using FIGS to calibrate predictive computer models have been simulation studies, with crop trait scores known a priori. Even if a subset of the accessions is hidden to the model-calibration procedure, the person conducting the data analysis knows the trait scores for these accessions. This knowledge could influence modeling decisions made during the data analysis. The FIGS approach could be useful to plant breeders and crop scientists when selecting accessions of landraces and crop wild relatives for experiments to identify new sources of target genetic diversity. It is therefore important to verify the predictive performance of the FIGS approach for a dataset with unknown trait scores. This study provided a more realistic test because the trait scores to be predicted were not disclosed to the person conducting the data analysis.

Stem rust on wheat (*Triticum* spp.) is caused by the fungus *Puccinia graminis* Pers. f. sp. *tritici* Eriks. & Henn. and has a long history as one of the most destructive diseases of cultivated wheat (McIntosh et al., 1995). A recent outbreak of stem rust was caused by a new and exceptionally virulent strain designated Ug99, and according to the North American race analysis system it was typified to race TTKSK (Jin et al., 2008). A report from CIMMYT (2005) estimated yield loss of up to 71% in experimental fields. Yield losses experienced in Kenya were reported to reach 80% (KARI, 2005; CIMMYT, 2005). Ug99 was discovered in Kalengyere, Uganda, in February 1999 (Pretorius et al., 2000) and it immediately raised concerns because of susceptibility of a large number of wheat germplasms from CIMMYT and the striking virulence toward the most important stem rust resistance genes including *Sr31*. This gene is one of the most widely used to provide protection against stem rust in modern cultivars (Wanyera et al., 2006). During the last decade, Ug99 caused an epidemic, spreading through Eastern Africa before turning northward to enter Yemen in 2006 and more recently in Iran in 2007 (Nazari et al., 2009). Ug99 is likely to continue to spread and the identification of new sources of resistance is important to sustain the use of wheat for food production (Njau et al., 2010). A global collaborative initiative to fight this new epidemic of stem rust, called the Borlaug Global Rust Initiative (2012), was established in 2008 (replacing the Global Rust Initiative launched in 2005). Recent work at CIMMYT reported the successful development of "new Ug99-resistant varieties of wheat that yield more than current popular varieties" (Singh et al., 2011). Several Ug99-resistant promising lines were also selected from ICARDA-supplied international nurseries. Of these, two bread wheat (*Triticum aestivum* L. subsp. *aestivum*) and three durum wheat [*Triticum turgidum* L. subsp. *durum* (Desf.) Husn.] lines have already been released in Ethiopia (ICARDA, 2010).

The rationale or fundamental theory for trait mining using FIGS is that crop traits are linked to the eco-climatic environment associated with the collecting site of accessions and that this link can be exploited to build predictive models to identify a subset of accessions with a higher likelihood of holding a target trait property (Mackay and Street, 2004). The first studies to validate the FIGS approach were conducted using a heuristic approach in which expert knowledge from the scientific literature was used to identify the boundaries for each of the eco-climatic variables (Street et al., 2008; Bhullar et al., 2009; El Bouhssini et al., 2009, 2011). Other studies were designed to calibrate a predictive computer model to validate the FIGS approach (Endresen, 2010; Endresen et al., 2011). These FIGS studies were conducted as simulation studies to predict a priori known trait scores. The objective of this experiment was to validate the FIGS approach using a blind study in which the person calculating the computer model did not know the actual trait scores. This study design provides a more realistic test of the predictive capacity of the FIGS approach compared to previous studies.

MATERIALS AND METHODS

A set of 4563 genebank accessions with landraces of bread wheat (*Triticum aestivum* L. subsp. *aestivum*) and durum wheat [*Triticum turgidum* L. subsp. *durum* (Desf.) Husn.] were screened for resistance to the new isolate designated as Ug99 and typified to race TTKSK of *P. graminis* Pers f. sp. *tritici*. The accessions were screened under natural infection at Tehama experimental station in Yemen during the 2008 cropping season. The field responses of stem rust differentials for accessions planted as a trap nursery experiment and using inoculum designed to identify the most important stem rust genes (including *Sr31*) indicated a similar virulence pattern as is characteristic for Ug99. Confirmation of existence of Ug99 (race TTKSK) was obtained from rust samples collected in experimental fields in Tehama and analyzed by Dr. Tom Fetch at Winnipeg, MB (A. Yahyaoui, personal communication, 2011). The results are not yet published in the scientific press but are already mentioned in the Genetic Resources Section database (ICARDA, 2011).

A recent study conducted by Endresen et al. (2011) explored prediction of stem rust resistance for landraces of bread wheat and durum wheat using trait mining with FIGS. The trait dataset for stem rust resistance used in the study was obtained from the USDA-National Plant Germplasm System (NPGS) database (USDA-ARS, 2011b). This study was conducted using the methodology that was developed by Endresen et al. (2011).

When developing a predictive classification model, the fit of the model to the calibration dataset (training set) does not provide a good indicator for how suitable the model is for predicting new samples not used for calibration (Hawkins, 2004). The trait dataset was therefore split into two parts. The first part

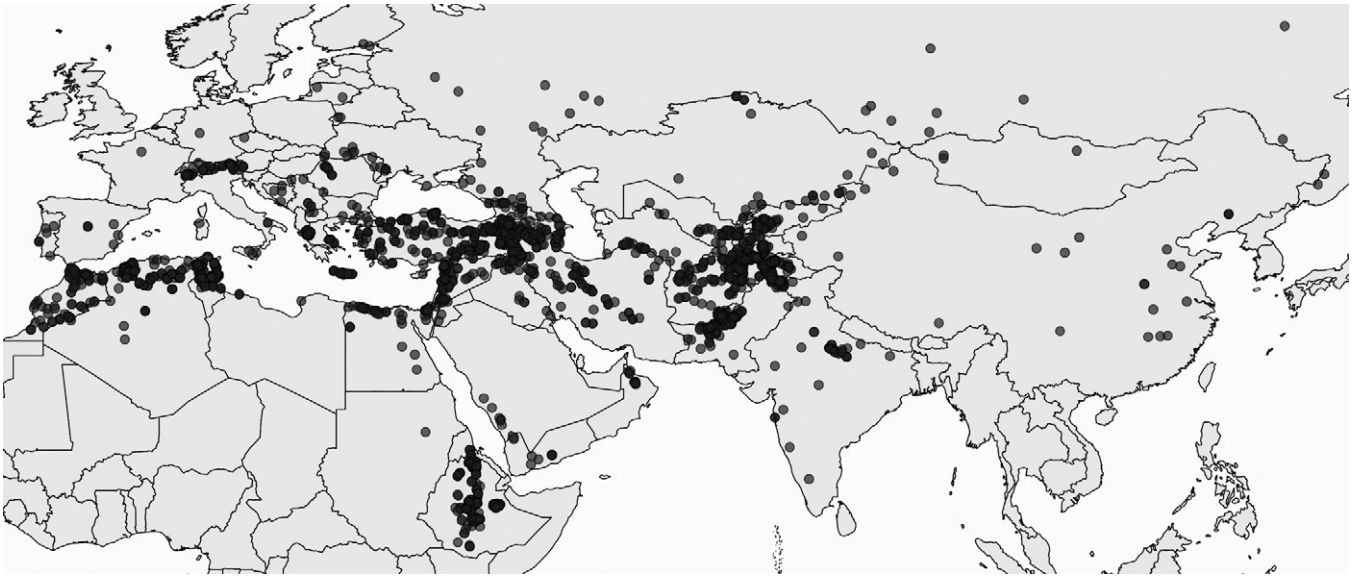

Figure 1. Original collecting sites for the wheat landraces (4563 accessions from 1928 collecting sites). Latitudes span from 4°48′0″ N to 62°43′12″ N and longitudes span from 10°4′12″ W to 134°4′12″ E.

(training set) was used to calibrate the classification model, and the second part (test set) was used to evaluate the performance of the model when predicting trait scores for samples not yet exposed to the model. The recommended size of the test set is normally given in the range of 25% of the samples (Myatt, 2007) to 33% (Brereton, 2006). Here we used a test set including 33% of the accessions randomly sampled from the full dataset. The accessions in the test set, however, were known to the modeler and could unintentionally have influenced choices made during the preparation of the adaptive trait-mining model.

A new experiment for blind predictions was designed to ensure that absolutely no test-set knowledge could influence the trait-mining model used for prediction of the trait scores. Dr. Kenneth Street representing ICARDA coordinated a follow-up experiment using the dataset with measurements of Ug99 resistance (K. Street, unpublished data, 2010). This experiment explored the performances of the trait-mining models in a simulation of a near real-life scenario in which the trait scores predicted by the model were not known to the modeler. A dataset including the accessions from the Ug99 trait dataset was prepared. This Ug99 set was randomly divided into two subsets with approximately 20% of the samples being used as a training set with trait scores included and the remainder 80% as the test set without any of the trait scores included. The next step was to predict the trait scores for the accessions in the test set and report the predictions back to the project coordinator (Dr K. Street), who had access to the actual trait scores of the test set.

For real-life screening experiments to search for useful genetic diversity, such as resistance to fungal pathogens, the final trait scores are not known when selecting accessions to include in the experiment. However, the passport data, including the geographic coordinates where the original material was collected, is available. This study is therefore a more realistic simulation of the conditions and information available for the planning of germplasm-evaluation experiments.

The dataset with accessions screened for resistance to Ug99 in Yemen in 2008 was matched with the accessions from the online System-wide Information Network for Genetic Resources (SINGER) (SGRP, 2011) and the USDA-NPGS (USDA-ARS, 2011a) databases to find more complete germplasm passport data. The dataset included a total of 4563 accessions from a total of 1928 collecting sites. Most of the accessions (4438 samples) were of bread wheat, but the dataset also included 114 accessions of durum wheat and 11 wheat accessions of unidentified species (genus *Triticum*). Figure 1 provides the location of the original collecting sites for all of the 4563 landraces included in the Ug99 dataset. The trait observations for the training set with disclosed trait scores for 825 accessions were reclassified to include three measurement levels. Trait scores reported as resistant (R) and moderately resistant (MR) were classified as resistant samples (class 1). Trait scores reported as moderately susceptible (MS) were assigned to class 2, and those reported as susceptible (S) to class 3. For the spring type accessions, the Cobb's scale for disease response and severity (Peterson et al., 1948) was used and for the winter type and the day-length-sensitive accessions, only disease response of R, MR, MS, and S was scored during the vegetative stage. The explorative principal component analysis (PCA) indicated that factorial analysis with decomposition of the climate data into principal components (PCs) was a suitable approach for this dataset; the accessions were well separated in the score plots from this PCA. This initial analysis did not indicate any outliers.

ICARDA Eco-Climatic Database

The climate data for this study were extracted from the eco-climatic information system maintained at the Geographic Information Systems Unit at ICARDA (De Pauw, 2008). The climate data were extracted using the latitude and longitude coordinates of the original landrace collection sites.

Prediction Based on a Model Calibrated from USDA Stem Rust Data

An initial prediction experiment was conducted using the same classification models as developed for a previous study (Endresen et al., 2011). These models were calibrated with stem rust scores from field trials in Minnesota during 1988 to 1994

(Bonman et al., 2007; USDA-ARS, 2011b). However, using this model, prediction of the Ug99 scores recorded in Yemen was surprisingly low. Ranked by the predicted resistance, a subset of 500 accessions was selected. From these accessions 9.55% were scored as resistant to Ug99 in the Yemen trials. The overall ratio of resistant accessions in the Ug99 set was reported to be 10.2% (K. Street, unpublished data, 2011). A second prediction with accessions from the same collecting sites, grouped together using the mean trait score for accessions from the same site, resulted in an even lower hit rate.

Prestudy: Ug99 Set (550 and 275 Accessions)

The next experiment was conducted in which only the accessions from the Ug99 set and only the 825 accessions with the trait scores reported were included. These accessions were split into a training set with 550 samples (67% of the accessions) and a test set with 275 samples (33%). For this experiment, the reported performance indicators were the averages from 15 repeated prestudy experiments. Each of the replicated prestudy experiments was made from a new and different allocation of accessions as training set and test set. The desktop study was conducted as described by Endresen et al. (2011), with the comparison of confidence intervals for the performance indicators calculated, respectively, for the accessions sampled by soft independent method of class analogy (SIMCA) (Wold, 1976; Wold and Sjöström, 1977) and k-nearest neighbor (kNN) (Cover and Hart, 1967). A "classifier ensemble" (Kuncheva, 2004; Rokach, 2010) was calculated as the mean of the predicted resistance class for each individual accession from the kNN and the SIMCA classifiers. Predictive performance was measured using the positive predictive value (PPV) and positive diagnostic likelihood ratio (LR+) (Altman and Bland, 1994).

Blind Prediction of Ug99 Resistance (825 and 3738 Accessions)

The final experiment was the predictive sampling of resistant accessions from the 3738 accessions of the Ug99 set with no trait scores disclosed to the modeler (test set). The classification model was here calibrated using all of the 825 accessions from the Ug99 set with trait scores disclosed (as the training set including 18% of the accessions in the full dataset). This trait-mining experiment was designed to simulate the sampling of accessions for a germplasm evaluation project with the (imaginary) capacity to screen 500 accessions. The task was therefore to select 500 accessions from the Ug99 test set predicted to have a higher likelihood to be resistant to the Ug99 stem rust pathogen.

The predictions from the SIMCA model and the kNN model were combined with equal weight to form a so-called classifier ensemble (Kuncheva, 2004; Rokach, 2010) following the same method as described for the prestudy. The top 500 accessions ranked by the predicted resistance to Ug99 were selected as the sampled subset with resistant samples. These 500 accessions corresponded to 13.4% of the samples in the test set with a total of 3738 accessions.

Classification Methods (kNN and SIMCA)

The previous study by Endresen et al. (2011) indicated that the SIMCA (Wold, 1976; Wold and Sjöström, 1977) and the kNN (Cover and Hart, 1967) were well suited for the classification of stem rust trait scores. The results reported here were calculated using these two classification methods. Each of the predictions was also compared with a corresponding random sampling approach. The random sampling was conducted by permutation of the true trait scores across the accessions in the test set and simply observing the coincidental hit rate against the true trait scores for these accessions.

Software

The classification models were calculated using the PLS Toolbox (Eigenvector Research Inc., 2010) for MATLAB (MathWorks Inc., 2009). The confidence interval for the performance indicators was calculated using the online Statistics Calculator from the Centre for Evidence-Based Medicine (CEBM, 2011).

RESULTS

Prestudy: Ug99 Set (555 and 275 Accessions)

The plots in Fig. 2 provide the cross-validation results for the SIMCA model from the prestudy using a training set with 550 accessions. The predicted residual sum of squares (PRESS) decreases toward a minimum around seven PCs. The increase of PRESS after this level indicates that the SIMCA classifier starts to overfit the model to the data when more than seven PCs are included in the model. Cross-validation when using all of the 825 accessions with disclosed trait scores resulted in similar plots with the minimum PRESS around seven PCs. The optimal model complexity of seven PCs found here was chosen as the model complexity for the final SIMCA model for prediction of the 3738 blind samples.

Visual Inspection of the Characteristic Plots for the SIMCA Model

The score plots indicated that the first PC focused on landrace accessions collected in Ethiopia. The second PC did not express focus on accessions from any individual country. The third PC focused on accessions collected in Greece. Ethiopia is the country of origin of most samples (1260 accessions) and Turkey the second ranked country (725 accessions) whereas Greece is ranked seventh (133 accessions) using this approach. The influence plot for the SIMCA model of class 1 indicated that the samples from Ethiopia were split into one group well described by the model (low error indicated by small residuals [Q] and low leverage [T^2]) and another group less well described by the model (with large residuals and high leverage). Most of the accessions from Greece have high values for both error and leverage. The accessions originating from Turkey have generally low influence and error and are therefore well described by the model.

The performance indicators for these classification models are reported in Table 1. The hit rate for this trait-mining experiment is notably higher than the predictive performance reported by Endresen et al. (2011) for a

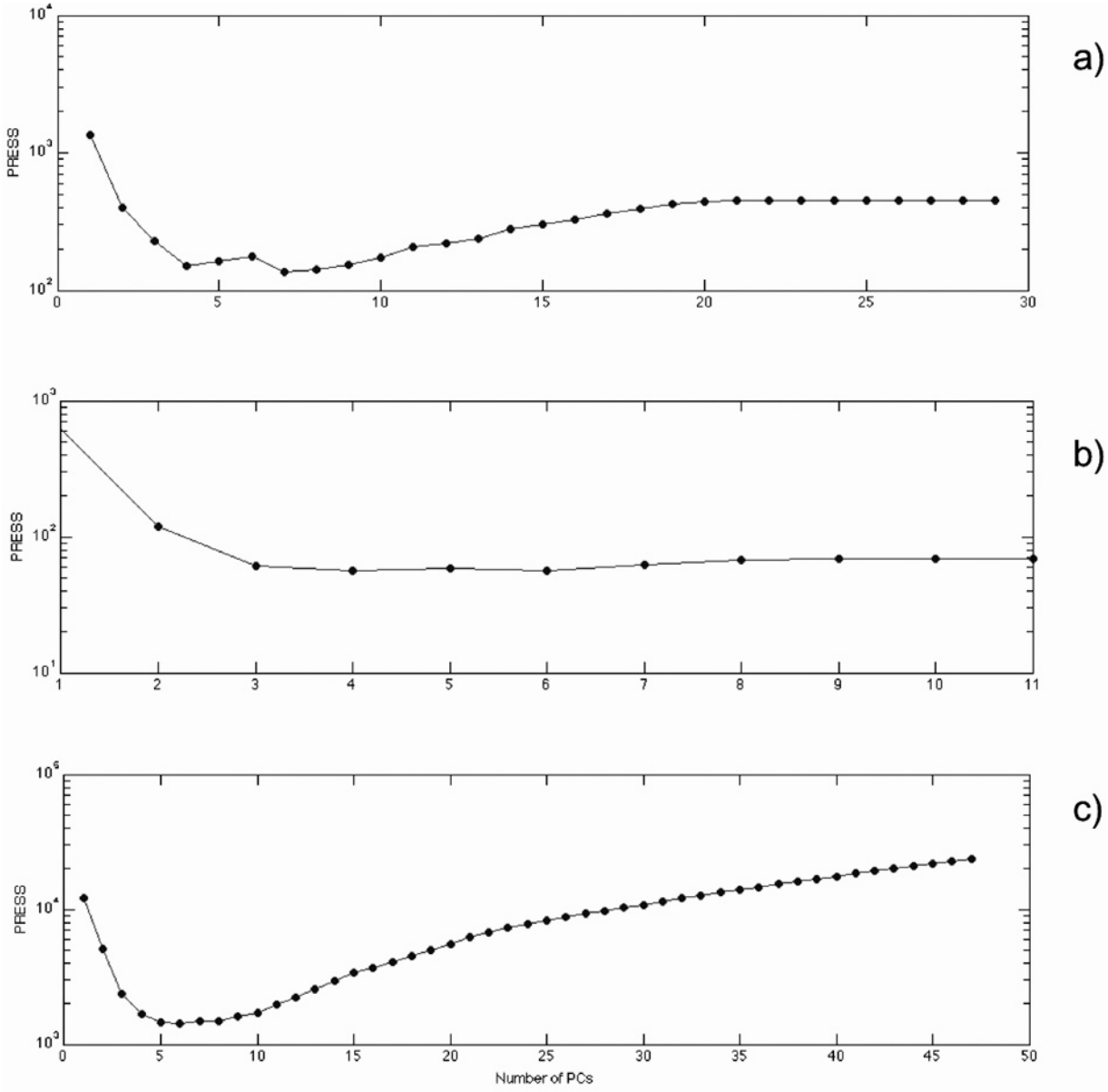

Figure 2. Cross-validation results for the soft independent modeling of class analogy (SIMCA) model of the Yemen Ug99 set including 550 training set accessions and 275 test accessions. The predicted residual sum of squares (PRESS) is plotted against the number of principal components (PCs) for each of the principal component analysis class models of the SIMCA model. (a) Principal component analysis model for class 1 with the resistant accessions, (b) class 2 with intermediate resistance, and (c) class 3 with the susceptible accessions.

similar study using a stem rust dataset from the USDA (USDA-ARS, 2011b). The predictive performance for the trait-mining models in this experiment is measured as four times higher than a random selection (Table 1) compared with an improved hit rate of two times reported by Endresen et al. (2011). The LR+ indicator should be used when comparing these results with the USDA stem rust results because of the different ratios of resistant accessions (prevalence) between these two datasets.

The predictions from the classifier ensemble suggest a gain in predictive performance of more than six times (Table 1). This is significantly higher than for each the kNN and SIMCA classifiers alone and suggests that the classifier ensemble approach is suitable for this dataset. Notice, however, the overlapping 95% confidence interval for the classifier ensemble compared with the kNN and SIMCA classifiers.

It is also important to note that the 95% confidence interval for the SIMCA, kNN, and ensemble performance indicators overlaps with the confidence interval for the randomly sampled accessions. Therefore, there is no statistically significant support (at the 95% confidence level) for a claim that these classification models perform better than the random sampling method. It is possible that the relatively broad confidence intervals are a result of too few samples in the test set. The correctly predicted resistant accessions (true positives) ranged from one to ten for the prediction models and from zero to three for the random sampling method.

Table 1. Results from the evaluation of predictive performance in the 825 accessions with trait score included (training set 550 accessions and test set 275 accessions).

Model†	PPV‡	LR+§	Estimated gain¶
kNN	0.29 (0.13–0.53)#	5.61 (2.21–14.28)	4.14 (1.86–7.57)
SIMCA	0.28 (0.14–0.48)	5.26 (2.51–11.01)	4.00 (2.00–6.86)
Ensemble classifier††	0.33 (0.12–0.65)	8.09 (2.23–29.42)	6.47 (2.05–11.06)
Random selection	0.06 (0.01–0.27)	0.95 (0.13–6.73)	0.97 (0.16–4.35)

†kNN, k-nearest neighbor (Cover and Hart, 1967); SIMCA, soft independent modeling of class analogy (Wold, 1976; Wold and Sjöström, 1977).

‡The positive predictive value (PPV) provides an indicator for classification performance of resistant accessions (positives) (Altman and Bland, 1994).

§The positive diagnostic likelihood ratio (LR+) provides a similar indicator that is less sensitive to the prevalence or proportion of resistant accessions (positives) in the dataset (Altman and Bland, 1994).

¶The estimated gain was calculated as the PPV divided by the proportion resistant samples (prevalence) for each subset.

#The 95% confidence interval for each of the performance indicators is included inside the parentheses.

††The mean of the predicted class for each record (accession) from the kNN and the SIMCA classifier was calculated to form the predicted class for the ensemble classifier.

Blind Predictions for the Ug99 Set (825 and 3728 Accessions)

The classification model for prediction of the 3728 unknown accessions from the Ug99 set was calibrated using all of the 825 accessions with revealed trait scores. The complete Ug99 set (4563 accessions) was reported to have 10.2% resistant accessions. The subset with the 3738 blind accessions included 11.1% resistant samples. The 500 accessions selected by the kNN and SIMCA classifier ensemble were reported to the project coordinator who found that 129 accessions (25.8%) were correctly predicted as resistant to Ug99. The proportion of resistant accessions in the set selected by the trait-mining models were therefore 2.3 times higher than the ratio of resistant samples in the Ug99 subset with hidden trait scores. The collection site locations of the 500 selected accessions predicted as resistant to Ug99 are indicated geographically in Fig. 3. The predictive performance is reported in Table 2. The predictive performance was at the same level as that reported by Endresen et al. (2011). Note that the 95% confidence intervals for the performance indicators are well above the corresponding confidence intervals for the random selection. There is therefore good statistically significant support for a claim that these classification models perform better than the random sampling method.

DISCUSSION

From Minnesota to Yemen

The first predictions using the USDA stem rust set to calibrate the model and the Ug99 set (tested in Yemen) as the test set did not provide any advantage over a randomly selected set. The USDA set and the Yemen set were both screened in open-field experiments at the adult growth stage of the wheat plants. The Ug99 set included data from the screening of the Ug99 strain of stem rust. This is a different (and much more virulent) type of stem rust than the type that was screened for the USDA set. The USDA set was screened in Minnesota (Rosemount and St. Paul) whereas the Ug99 set was screened in Yemen. The environmental conditions are very different in Tehama, Yemen, (dry and warm) compared with Minnesota (wet and cold). These differences in trial site could perhaps cause a difference in the expression of stem rust resistance for these landrace genotypes. We also note that the environmental conditions in Tehama are more conductive to stem rust development and that the race composition of natural infection is more aggressive than in Minnesota in particular with the presence of Ug99.

This result indicates that the calibration could be specific to the *P. graminis* population and could not find a general model to explain the resistance to other populations

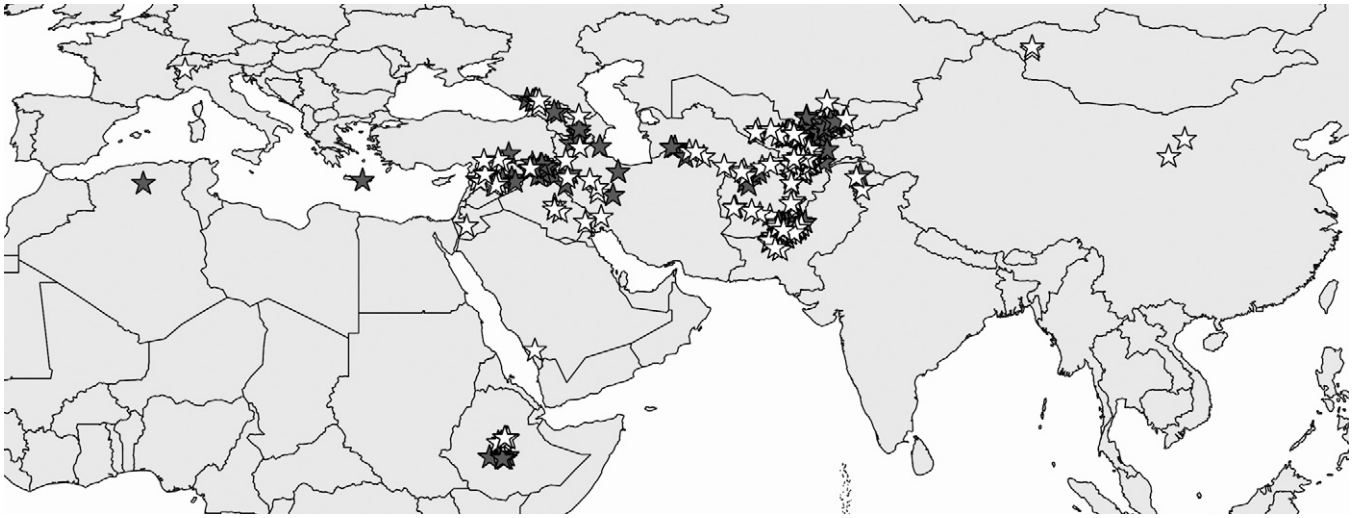

Figure 3. Map with the 500 accessions predicted to be resistant to stem rust. Gray stars indicate accessions predicted as resistant by all of the trait mining models (soft independent modeling of class analogy [SIMCA] and k-nearest neighbor [kNN]). White stars indicate accessions predicted to be resistant only by one of these trait mining models (SIMCA or kNN).

(races). It is also possible that the virulence of the race is specific to the environment or that the functional response of the crop to the disease is specific to the environment (Bolnick et al., 2011; Fynn et al., 2011). This problem can perhaps be compared with the problem of overfitting of the model to the training set (Hawkins, 2004). It is possible that a reduction of the model complexity resulting in lower precision would improve the generality of the model.

The prestudy tests including the 825 accessions with known trait scores indicated four to six times improved hit rate compared with random sampling (Table 1). This is higher than the improved hit rate of 2.5 times reported by Endresen et al. (2011) for a similar study. The Ug99 set includes trait scores from one single trial season and at only one site whereas the USDA set explored by Endresen et al. (2011) is a compilation of trait scores from six different seasons and two different experiment stations. The different screening conditions for the USDA set could therefore introduce variation (genotype × environment interaction) in the trait scores that would be interpreted as noise by the classification algorithm. These observations suggest a best practice of choosing the training set for future FIGS studies should include trait scores with the same or at least similar field trial conditions.

Model Structure

It is possible that a predictive signal exists between the USDA stem rust set and the Ug99 set that the classification models used in this study were unable to find. The classification model might be "overfitted" with respect to the training data (USDA stem rust set). It is relatively easy to make the model fit the training set but much more challenging to make the model describe the test set (samples not "seen" by the model calibration algorithm). The classifier could also be too simple or have an inappropriate structure. Fuzzy samples are known to disturb the central covariance matrix of discriminant analysis models (Fielding, 2007). There might be a similar effect disturbing the SIMCA and kNN models used here. Other methods such as artificial neural networks (ANN) (Bishop, 1996) and random forest (Breiman, 2001) are less sensitive to such problems and might give a stronger predictive signal where other methods fail.

These stem rust datasets hold unequal proportions of resistant accessions compared to the much larger number of susceptible accessions. The stem rust datasets have an asymmetric internal class data structure, also called "unbalanced class." Different classifiers will be affected differently by the challenge of modeling very different class sizes. Artificial neural network has been reported to make better provision for asymmetrical internal class structures (Davies and Silverstein, 1995).

Predictive Performance and the Size of the Training Set

The predictive performance for the prestudy experiment was high, but the 95% confidence interval, when compared with the random sampling approach, overlapped. The reason for the wide confidence interval in the prestudy was the low number of samples (275 accessions) in the test set. Allocating fewer accessions to the training set and more accessions to the test set is likely to give lower actual predictive performance (fewer training samples for the model to learn from). More accessions in the test set are also likely to give narrower confidence intervals. This suggests that a training set of 550 accessions is sufficient to calibrate the trait-mining model, but a test set of 275 accessions is too small to validate the model.

Table 2: Predictive performance in the 3738 "blind" samples with trait score hidden from the modeler (training set 825 accessions and test set 3738 accessions).

Model	PPV†	LR+‡	Estimated gain§
Ensemble classifier¶	0.26 (0.22–0.30)#	2.78 (2.34–3.31)	2.32 (2.00–2.68)
Random selection	0.11 (0.09–0.15)	1.02 (0.77–1.36)	1.02 (0.77–1.32)

†The positive predictive value (PPV) provides an indicator for classification performance of resistant accessions (positives) (Altman and Bland, 1994).

‡The positive diagnostic likelihood ratio (LR+) provides a similar indicator that is less sensitive to the prevalence or proportion of resistant accessions (positives) in the dataset (Altman and Bland, 1994).

§The estimated gain was calculated as the PPV divided by the proportion resistant samples (prevalence) for each subset.

¶The mean of the predicted class for each record (accession) from the k-nearest neighbor (kNN; Cover and Hart, 1967) and the soft independent modeling of class analogy (SIMCA; Wold, 1976; Wold and Sjöström, 1977) classifier was calculated to form the predicted class for the ensemble classifier.

#The 95% confidence interval for each of the performance indicators is included inside the parentheses.

In the case of the blind study to predict resistant accessions within the test set, a training set size of 825 accessions proved to be sufficient to calibrate the trait-mining models with good predictive performance. Further work is recommended to investigate the number of accessions required for training sets for a range of adaptive traits. In the practical application of FIGS, information about the minimum number of positives and negatives would be helpful in reducing the cost of the initial field experiment while avoiding the risk of generating too small a training set. Notice that the model calibration procedure requires inclusion of a minimum number of both resistant and susceptible accessions in the training set. In this study, the results indicated that it was the number of resistant samples in the dataset that provided a limitation and not the number of susceptible samples.

The prestudy indicated a gain in predictive performance of four to six times when compared with a random selection (Table 1). The blind study provided a gain in prediction quality of 2.3 times (Table 2). The relative size of the training set was different for these experiments. For the prestudy, the training set included 67% of the total number of accessions whereas the training set for the blind study included only 18% of the accessions. The substantially lower predictive performance for the blind study compared with the prestudy experiments could be an effect of the lower ratio accessions allocated

to the training set. Roy et al. (2008) explored the impact of the size of the training set relative to the test set on the predictive performance for classification studies. Their study included three different datasets and found a reduction of predictive performance with fewer samples allocated to the training set for two of the datasets. They concluded that the optimum size of the training set varied with different datasets. If the pattern the classification model should find is complex and involves a large number of interactions between many different features of the eco-climatic data, then the training set needs to be larger to provide enough examples for the classifier to learn from. This effect has important implications for trait mining using FIGS. For complex trait–environment relations, the quality of the predictions is dependent on a relatively larger training set. Using different sizes of the training set for the prestudy experiments could give a useful indication of this effect and is recommended for future FIGS studies.

Limitations of the Focused Identification of Germplasm Strategy Approach

The methodology followed here requires an initial training set to calibrate the prediction model. This requirement would often demand a smaller set to be screened in a field experiment a priori to data analysis. We also observed that trait data for stem rust from previous screening experiments failed to produce good models for this Ug99 set. It is therefore suggested that the initial training set be sensitive to the experimental conditions. Additionally, the FIGS approach requires accessions to be geo-referenced with reasonably accurate spatial coordinates to enable the extraction of eco-climatic data for the source environments.

The FIGS approach will identify a subset of accessions predicted to hold a desired trait property by searching for accessions from similar eco-geographic environments as accessions already identified to hold the trait property of interest. It is therefore possible that the FIGS approach will show less merit to identify new sources of genetic diversity but rather identify more accessions with the same genetic diversity as is already known. Bhullar et al. (2009) reported use of the FIGS approach together with allele mining to identify new sources of resistance to powdery mildew [*Blumeria graminis* (DC.) Spear f. sp. *tritici*] at a known locus in wheat (*Triticum* spp.). A subset of 1320 accessions was derived using FIGS. Field trials including these 1320 accessions resulted in the identification of 211 resistant accessions (16% hit rate). Molecular analysis of these 211 accessions identified seven new alleles for powdery mildew resistance compared with the previously known seven functionally distinct alleles (*Pm3a* to *Pm3g*). This result clearly indicates the suitability of the FIGS approach to identify new sources of disease resistance for wheat. The hypothesis is that functionally distinct resistant alleles are more likely to be found at similar eco-geographic environments.

Further studies are recommended to explore the suitability of the FIGS approach to identify new sources of useful genetic diversity. This study indicates that the FIGS approach is efficient in the identification of accessions with a target trait property. If further studies confirm that this approach is also efficient in the identification of new sources of target genetic diversity, then FIGS could provide an important tool for improving future plant breeding efforts.

CONCLUSION

This study highlights a number of issues regarding how the selection of germplasm for evaluation to discover new genetic variation for specific adaptive traits might become more effective. Two sets of a priori data were used to train models to predict which accessions might possess resistance to stem rust based on the classification of the environment from which they were collected.

The location and year of the experiment in which the a priori evaluation data were obtained were seen to influence the predictive power of the models. Another factor requiring consideration was the size and perhaps the ratio of resistant to susceptible expression of the training set. This factor might also be influenced by the specific adaptive trait under consideration. It was also noted that the FIGS approach requires the accessions to be geo-referenced before an association between the germplasm and the environment (in which it evolved and was subject to selection) can be established.

The blind prediction example, however, suggests that trait mining using the FIGS approach to select accessions for a resistance to Ug99 can be as much as 2.3 times more effective than a random sample if the training set is of sufficient size and scope. This difference indicates that the FIGS approach can contribute significant efficiency to the selection of germplasm for evaluation and suggests that further research into refinement is justified.

Acknowledgments

Associate professor Dvora-Laiô Wulfsohn (Copenhagen University) provided feedback and suggestions on the draft manuscript. This research project was supported by a grant from the Nordic Genetic Resources Center (NordGen; http://www.nordgen.org). All authors helped assemble the data and to develop the experimental design for the modeling experiments. Kumarse Nazari, and Amor Yahyaoui conducted the disease screening experiments. Dag Endresen made the data analysis and wrote the first version of the manuscript. All authors contributed to the final manuscript.

References

Altman, D.G., and J.M. Bland. 1994. Statistical notes: Diagnostic tests 2: Predictive values. BMJ 309(6947):102. doi:10.1136/bmj.309.6947.102

Bari, A., K. Street, M. Mackay, D.T.F. Endresen, E. De Pauw, and A. Amri. 2011. Focused identification of germplasm strategy (FIGS) detects wheat stem rust resistance linked

to environmental variables. Genet. Resour. Crop Evol. doi:10.1007/s10722-011-9775-5.

Bhullar, N.K., K. Street, M. Mackay, N. Yahiaoui, and B. Keller. 2009. Unlocking wheat genetic resources for the molecular identification of previously undescribed functional alleles at the *Pm3* resistance locus. PNAS 106(23):9519–9524. doi:10.1073/pnas.0904152106

Bishop, C. 1996. Neural networks for pattern recognition. Oxford Univ. Press, Oxford, UK.

Bolnick, D.I., P. Amarasekare, M.S. Araújo, R. Bürger, J.M. Levine, M. Novak, V.H.W. Rudolf, S.J. Schreiber, M.C. Urban, and D.A. Vasseur. 2011. Why intraspecific trait variation matters in community ecology. Trends Ecol. Evol. 26(4):183–192. doi:10.1016/j.tree.2011.01.009

Bonman, J.M., H.E. Bockelman, Y. Jin, R.J. Hijmans, and A.I.N. Gironella. 2007. Geographic distribution of stem rust resistance in wheat landraces. Crop Sci. 47:1955–1963. doi:10.2135/cropsci2007.01.0028

Borlaug Global Rust Initiative. 2012. Borlaug Global Rust Initiative – A resource for scientists and policymakers about the rusts of wheat. Available at http://www.globalrust.org (verified 5 Jan. 2012). Cornell University, Ithaca, NY.

Breiman, L. 2001. Random forests. Mach. Learn. 45(1):5–32. doi:10.1023/A:1010933404324

Brereton, R.G. 2006. Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. Trends Anal. Chem. 25(11):1103–1111. doi:10.1016/j.trac.2006.10.005

Centre for Evidence-Based Medicine (CEBM). 2011. Statistics calculator. Available at http://ktclearinghouse.ca/cebm/practise/ca/calculators/statscalc (verified 2 Jan. 2012). Centre for Evidence-Based Medicine, University Health Network, Toronto, ON, Canada.

CIMMYT. 2005. Sounding the alarm on global stem rust. An assessment of race Ug99 in Kenya and Ethiopia and the neighboring regions and beyond, by the expert panel on the stem rust outbreak in Eastern Africa. CIMMYT, El Batán, Edo Mex, Mexico. 29 May 2005 Available at http://www.globalrust.org/db/attachments/about/2/1/Sounding%20the%20Alarm%20on%20Global%20Stem%20Rust.pdf (verified 2 Jan. 2012). CIMMYT, El Batán, Edo Mex, Mexico.

Cover, T.M., and P.E. Hart. 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13(1):21–27. doi:10.1109/TIT.1967.1053964

Davies, P., and B.R. Silverstein. 1995. A comparison of neural nets to statistical stubborn classification problems. p. 3467–3470. *In* Int. Conf. Acoustics, Speech, Signal Processing. ICASSP-95, Vol. 5, Detroit, MI. 9–12 May 1995. IEEE Signal Processing Society, Piscataway, NJ. doi:10.1109/ICASSP.1995.479732

De Pauw, E. 2008. Climatic and soil datasets for the ICARDA wheat genetic resource collections of the Eurasia region. Explanatory notes. Available at http://geonet.icarda.cgiar.org/geonetwork/data/regional/GRU_NetBlotch/Doc/Report_NetBlotch.pdf (verified 2 Jan. 2012). ICARDA GIS Unit, Aleppo, Syria.

Eigenvector Research Inc. 2010. PLS toolbox 5.8 (R5.8.3). Available at http://software.eigenvector.com/toolbox (verified 2 Jan. 2012). Eigenvector Research Inc., Wenatchee, WA.

El Bouhssini, M., K. Street, A. Amri, M. Mackay, F.C. Ogbonnaya, A. Omran, O. Abdalla, M. Baum, A. Dabbous, and F. Rihawi. 2011. Sources of resistance in bread wheat to Russian wheat aphid (*Diuraphis noxia*) in Syria identified using the focused identification of germplasm strategy (FIGS). Plant Breed. 130(1):96–97. doi:10.1111/j.1439-0523.2010.01814.x

El Bouhssini, M., K. Street, A. Joubi, Z. Ibrahim, and F. Rihawi. 2009. Sources of wheat resistance to Sunn pest, *Eurygaster integriceps* Puton, in Syria. Genet. Resour. Crop Evol. 56:1065–1069. doi:10.1007/s10722-009-9427-1

Endresen, D.T.F. 2010. Predictive association between trait data and ecogeographic data for Nordic barley landraces. Crop Sci. 50(6):2418–2430. doi:10.2135/cropsci2010.03.0174

Endresen, D.T.F., K. Street, M. Mackay, A. Bari, and E. De Pauw. 2011. Predictive association between biotic stress traits and eco-geographic data for wheat and barley landraces. Crop Sci. 51(5):2036–2055. doi:10.2135/cropsci2010.12.0717

Fielding, A.H. 2007. Cluster and classification techniques for the biosciences. Cambridge Univ. Press, Cambridge, UK.

Fynn, R., C. Morris, D. Ward, and K. Kirkman. 2011. Trait-environment relations for dominant grasses in South African mesic grassland support a general leaf economic model. J. Veg. Sci. 22:528–540. doi:10.1111/j.1654-1103.2011.01268.x

Hawkins, D.M. 2004. The problem of overfitting. J. Chem. Inf. Comput. Sci. 44(1):1–12. doi:10.1021/ci0342472

ICARDA. 2010. ICARDA annual report 2010. ICARDA, Aleppo, Syria.

ICARDA. 2011. Genetic resources section database. Available at http://singer.cgiar.org (verified 12 Aug 2011). ICARDA, Aleppo, Syria.

Jin, Y., L.J. Szabo, Z.A. Pretorius, R.P. Singh, R. Ward, and T. Fetch. 2008. Detection of virulence to resistance gene *Sr24* within race TTKS of *Puccinia graminis* f. sp. *tritici*. Plant Dis. 92:923–926. doi:10.1094/PDIS-92-6-0923

Kenya Agricultural Research Institute (KARI). 2005. Effect of a new race on wheat production/use of fungicides and its cost in large vs. small scale farmers, situation of current cultivars. KARI, Njoro, Kenya.

Kuncheva, L.I. 2004. Combining pattern classifiers. Methods and algorithms. John Wiley & Sons, Hoboken, NJ.

Mackay, M.C. 1986. Utilizing wheat genetic resources in Australia. p. 56–61. *In* R. McLean (ed.). Proc. 5th Assembly Wheat Breed. Soc., Merredin, WA, Australia. 18–22 Aug. 1986. Western Australian Department of Agriculture, Perth, WA, Australia.

Mackay, M.C. 1990. Strategic planning for effective evaluation of plant germplasm. p. 21–25. *In* J.P. Srivastava and A.B. Damania (ed.). Wheat genetic resources: Meeting diverse needs. John Wiley & Sons, Chichester, UK.

Mackay, M.C. 1995. One core collection or many? p. 199–210. *In* T. Hodgkin, A.H.D. Brown, Th.J.L. van Hintum, and A.A.V. Morales (ed.). Core collections of plant genetic resources. John Wiley & Sons, Chichester, UK.

Mackay, M.C., and K. Street. 2004. Focused identification of germplasm strategy – FIGS. p. 138–141. *In* C.K. Black, J.F. Panozzo, and G.J. Rebetzke (ed.). Proc. 54th Australian Cereal Chem. Conf. 11th Wheat Breed. Assembly, Canberra, ACT, Australia. 21–24 Sept. 2004. Cereal Chem. Div., Royal Australian Chemical Institute (RACI), Melbourne, VIC, Australia.

Mansholt, U.J. 1909. Van Pesch Plantenteelt, beknopte handleiding tot de kennis van den Nederlandschen landbouw. 3rd revised edition, pt 2. (In Dutch.) Plantenteelt. Zwolle, The Netherlands.

MathWorks Inc. 2009. MATLAB & Simulink student version, release 2009a – Mac. Available at http://www.mathworks.com/products/matlab/ (verified 12 Aug 2011). MathWorks Inc., Natick, MA.

McIntosh, R.A., C.R. Wellings, and R.F. Park. 1995. Wheat rusts: An atlas of resistance genes. CSIRO, Melbourne, VIC, Australia.

Myatt, G.J. 2007. Making sense of data: A practical guide to exploratory data analysis and data mining. John Wiley & Sons Inc., Hoboken, NJ.

Nazari, K., M. Mafi, A. Yahyaoui, R.P. Singh, and R.F. Park. 2009. Detection of wheat stem rust (*Puccinia graminis* f. sp. *tritici*) race TTKSK (Ug99) in Iran. Plant Dis. 93(3):317. doi:10.1094/PDIS-93-3-0317B

Njau, P.N., Y. Jin, J. Huerta-Espino, B. Keller, and R.P. Singh. 2010. Identification and evaluation of sources of resistance to stem rust race Ug99 in wheat. Plant Dis. 94(4):413–419. doi:10.1094/PDIS-94-4-0413

Peterson, R.F., A.B. Campbell, and A.E. Hannah. 1948. A diagrammatic scale for estimating rust intensity of leaves and stems of cereals. Can. J. of Res. 26c(5):496–500. doi:10.1139/cjr48c-033

Pretorius, Z.A., R.P. Singh, W.W. Wagoire, and T.S. Payne. 2000. Detection of virulence to wheat stem rust resistance gene *Sr31* in *Puccinia graminis* f. sp. *tritici* in Uganda. Plant Dis. 84(2):203. doi:10.1094/PDIS.2000.84.2.203B

Rokach, L. 2010. Pattern classification using ensemble methods. Series in machine perception and artificial intelligence – Vol. 75. World Scientific Publishing Co. Pte. Ltd., Singapore.

Roy, P.P., J.T. Leonard, and K. Roy. 2008. Exploring the impact of size of training sets for the development of predictive QSAR models. Chemometrics and Intelligent Laboratory Systems 90(1):31–42. doi:10.1016/j.chemolab.2007.07.004

Singh, R.P., D.P. Hodson, J. Huerta-Espino, Y. Jin, S. Bhavani, P. Njau, S. Herrera-Foessel, P.K. Singh, S. Singh, and V. Govindan. 2011. The emergence of Ug99 races of the stem rust fungus is a threat to world wheat production. Annu. Rev. Phytopathol. 49:465–481. doi:10.1146/annurev-phyto-072910-095423

Street, K., M. Mackay, E. Zuev, N. Kaul, M. El Bouhssini, J. Konopka, and O. Mitrofanova. 2008. Diving into the genepool: A rational system to access specific traits from large germplasm collections. p. 28–31. *In* R. Appels, R. Eastwood, E. Lagudah, P. Langridge, and M. Mackay (ed.) Proc. 11th Int. Wheat Genet. Symp., Brisbane, QLD, Australia. 24–29 Aug. 2008. Sydney Univ. Press, Sydney, NSW, Australia.

System-wide Genetic Resources Programme (SGRP). 2011. System-wide information network for genetic resources (SINGER). Available at http://singer.cgiar.org (updated 21 Dec. 2010, verified 2 Jan. 2012). SGRP Secretariat, Bioversity International, Maccarese, Rome, Italy.

USDA-Agricultural Research Service (USDA-ARS). 2011a. Germplasm resources information network (GRIN) national plant germplasm system (NPGS). Available at http://www.ars-grin.gov/npgs/ (last updated 25 Mar. 2010, verified 2 Jan. 2012). USDA-ARS, National Germplasm Resources Laboratory, Beltsville, MD.

USDA-Agricultural Research Service (USDA-ARS). 2011b. Trait: Stem rust adult (STEMRUSTAD). Available at http://www.ars-grin.gov/cgi-bin/npgs/html/desc.pl?65049 (verified 2 Jan. 2012). USDA-ARS, National Germplasm Resources Laboratory, Beltsville, MD.

Wanyera, R., M.G. Kinyua, Y. Jin, and R.P. Singh. 2006. The spread of stem rust caused by *Puccinia graminis* f. sp. *tritici* virulence on *Sr31* in wheat in Eastern Africa. Plant Dis. 90(1):113. doi:10.1094/PD-90-0113A

Wold, S. 1976. Pattern recognition by means of disjoint principal component models. Pattern Recog. 8:127–139. doi:10.1016/0031-3203(76)90014-5

Wold, S., and M. Sjöström. 1977. SIMCA: A method for analyzing chemical data in terms of similarity and analogy. p. 243–282. *In* B.R. Kowalski (ed.). Chemometrics Theory and Application, American Chemical Society Symposium Series 52. American Chemical Society Washington, DC. doi:10.1021/bk-1977-0052.ch012

Zeven, A.C. 1998. Landraces: A review of definitions and classifications. Euphytica 104(2):127–139. doi:10.1023/A:1018683119237