RESEARCH

# Predictive Association between Biotic Stress Traits and Eco-Geographic Data for Wheat and Barley Landraces

Dag Terje Filip Endresen,* Kenneth Street, Michael Mackay, Abdallah Bari, and Eddy De Pauw

**ABSTRACT**

Collections of crop genetic resources are a valuable source of new genetic variation for economically important traits, including resistance to crop diseases. New sources of useful crop traits are often identified through evaluation in field trials. The number of relevant accessions in genebank collections available to be evaluated is often substantially larger than the capacity of the evaluation project. The focused identification of germplasm strategy (FIGS) is an approach used to select subsets of germplasm from genetic resource collections in such a way as to maximize the likelihood of capturing a specific trait. This strategy uses a range of methods to link the expression of a specific trait (of a target crop) with the eco-geographic parameters of the original collection site. This study contributes to the development of the approach by which a FIGS subset could be assembled for biotic traits. We have evaluated trait-specific subset selection methods for two fungal crop diseases, namely stem rust (*Puccinia graminis* Pers.) in wheat (*Triticum aestivum* L. and *Triticum turgidum* L.) and net blotch (*Pyrenophora teres* Drechs.) in barley (*Hordeum vulgare* L.). The results indicate that the climate layers from freely available eco-geographic databases are well suited to model and predict the reaction in these crops to biotic stress traits. This result has the potential to improve the efficiency of field screening trials to find novel sources of economically valuable crop traits.

D.T.F. Endresen, Nordic Genetic Resources Center (NordGen), Smedjevägen 3, 230 53 Alnarp, Sweden; K. Street, A. Bari, and E. De Pauw, International Center for Agricultural Research in the Dry Areas (ICARDA), P.O. Box 5466, Aleppo, Syria; M. Mackay, Bioversity International, Via dei Tre Denari 472/a, 00057 Maccarese (Fiumicino) Rome, Italy. Received 21 Dec. 2010. *Corresponding author (dag.endresen@nordgen.org).

**Abbreviations:** BIOCLIM, bio-climatic; CI, confidence interval; FIGS, focused identification of germplasm strategy; FN, false negatives; FP, false positives; GRIN, Germplasm Resources Information Network; kNN, k-nearest neighbor; LDA, linear discriminant analysis; LR+, positive diagnostic likelihood ratio; NPGS, National Plant Germplasm System; PCA, principal component analysis; PET, potential evapotranspiration; PLS-DA, partial least squares discriminant analysis; PPV, positive predictive value; SIMCA, soft independent modeling by class analogy; TN, true negatives; TP, true positives.

The genetic resources conserved by ex situ genebanks around the world cover a vast range of genetic diversity underexploited in present day cultivars. The main objective of public genebanks is to conserve crop genetic diversity to sustain agricultural production systems by providing ready access to samples for research and plant breeding activities. A bottleneck for rational utilization is the availability and access to passport, characterization, and evaluation data. Obtaining good quality phenotypic trait data for genebank accessions requires large field or greenhouse experiments at great cost. The lack of evaluation data for useful traits is one of the major, current problems hindering the efficient use of plant genetic resources (FAO, 2010).

Further, the growing size of the genebank collections has been mentioned as a problem for the efficient use of genebank

Published in Crop Sci. 51:2036–2055 (2011).
doi: 10.2135/cropsci2010.12.0717
Published online 21 June 2011.
© Crop Science Society of America | 5585 Guilford Rd., Madison, WI 53711 USA

All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher.

collections (see for example Mackay, 1990, 1995) because the number of genebank accessions available to be evaluated for a specific trait is often substantially larger than the resources available to evaluate the material. Thus, finding the genebank accessions most likely to possess the desired trait can be compared to searching for a needle in a haystack. Clearly a rational and efficient strategy to mine genebanks for useful traits is required.

## Focused Identification of Germplasm Strategy

The challenges to using genetic resource collections, as detailed above, was one of the reasons for the introduction of the core collection concept (Frankel, 1984; van Hintum et al., 2000). A core collection seeks to represent most of the genetic variation present in the original collection in a core subset of 5 to 10% the size of the original. Core collection methods use statistical approaches to maximize diversity using a variety of input data including collection site descriptors, agro-morphological traits, and molecular marker data.

However, the core collection approach may not lead to the identification of rare useful traits in germplasm collections (Holbrook and Dong, 2005). Such concerns to capture rare traits and adaptive trait variation, much of which is thought to reflect plant functional variation (Wright and Gaut, 2005), have lead some workers to construct specific or thematic collections or use of other approaches (Brown and Spillane, 1999; Gepts, 2006; Dwivedi et al., 2007; Pessoa-Filho et al., 2010; Xu, 2010).

The focused identification of germplasm strategy (FIGS) strategy introduces a novel approach for constructing small subsets of accessions in that it selects genetic variation for just a single trait at a time. The FIGS strategy endeavors to maximize the likelihood of encountering specific adaptive traits in subsets by choosing accessions from collection sites that are most likely to impose a selection pressure for the trait being sought (Mackay and Street, 2004).

Nikolai Ivanovich Vavilov (1887–1943) was one of the first pioneers to recognize the importance of the eco-climatic conditions when searching for source material to include in plant breeding (Vavilov, 1992a, b, 1957; Kurlovich et al., 2000). Vavilov used the term "climatic analogy" for the selection of suitable strains guided by climate and soil data. His "differential phyto-geographical method" also has elements that link the morpho-physiological trait characters of species and strains to a definite environment and area (Vavilov, 1920, 1922, 1992c).

> "It is evident that when selecting species and strains for the U.S.S.R. it is necessary to take the climate and the soil conditions at their origin into consideration to introduce strains from areas that are more or less similar to those in our own country. Knowledge of the climate of our own country and that of the areas from where we collected the seeds is of great importance." (Vavilov, 1992b; see also Vavilov, 1992a:266).

This link between environment and phenotype was recently demonstrated by Endresen (2010), who successfully used the FIGS strategy to link morphological traits in barley (*Hordeum vulgare* L.) to the eco-climatic pattern from the original collecting sites for Nordic barley landraces. Put into practice, a simple example of applying the FIGS approach to selection of germplasm from a genebank could be if salinity tolerance is the target trait then accessions would be chosen from collection sites that have saline soils (Peeters et al., 1990). Hijmans et al. (2003) explored, based on a similar hypothesis, the link between frost tolerance and eco-climate with focus on temperature at the original collecting site for genebank accessions.

However, the problem becomes more complex if one is looking for tolerances to biotic constraints. The approach used by El Bouhssini et al. (2009) to identify bread wheat (*Triticum aestivum* L.) and durum wheat (*Triticum turgidum* L.) resistance to Sunn pest (*Eurygaster integriceps* Puton) and to the virulent Syrian Russian wheat aphid biotype (*Diuraphis noxia* Kurdjumov) (El Bouhssini et al., 2010) was to select accessions from agro-climatic environments that are likely to favor high pest populations during the growing season. Thousands of accessions from the ICARDA genebank had previously been chosen, largely at random, and screened for the two pests without success (El Bouhssini, personal communication, 2008). By contrast, the FIGS approach chose relatively small subsets (500 accessions) that contained multiple sources of resistance.

The first step when using the FIGS approach is to identify a group of geo-referenced landraces with known resistance to a given pest or disease. An eco-geographic profile of the collection sites of this "training set" is ascertained and statistical methods or models developed to select untested accessions from environments that are statistically similar to the trainer set environments. This was the approach successfully used by Bhullar et al. (2009, 2010) to identify a range of bread wheat accessions with resistance to various powdery mildew [*Blumeria graminis* (DC.) E.O. Speer f. sp. *tritici* Em. Marchal] isolates.

The above examples demonstrate the utility of FIGS as a means to choose germplasm with variation for specific adaptive traits. However, they cannot be used as a proof of concept because the frequencies of the resistant material in the collection from which the subset was chosen were not known. Further, the possibility exists that the resistant material occurred in the sets merely by chance. The aim of this present study was to use geospatial statistical analysis to predict the presence of resistance to stem rust (*Puccinia graminis* Pers.) in bread wheat and net blotch (*Pyrenophora teres* Drechs.) in barley in a set of landrace accessions that have been previously screened for the diseases, thus allowing an evaluation of the approach by comparing the predictions to a random selection.

### *Stem Rust*

The stem rusts are caused by the fungus *Puccinia graminis* Pers. 1794 and are a significant disease affecting cereal crops. The *formae speciales* of *Puccinia graminis* f.sp. *tritici* is responsible for stem rust in wheat (McIntosh et al., 1995). E. C. Stakman provided early pioneering work on stem rust and identified the first unique races of this pathogen (Stakman, 1915; Stakman and Piemeisel, 1917). After a number of devastating rust epidemics in important wheat producing areas of countries such as Australia, Canada, and the United States, a long-term global collaboration to combat wheat rust was so successful that the stem rust reached almost nonsignificant levels by the 1990s (Singh et al., 2008). In 1998 a new isolate of stem rust designated Ug99 and typed to race TTKS caused severe damage to wheat in Uganda and Kenya. Pretorius et al. (2000) discovered that this new stem rust race Ug99 showed virulence against the widely used Sr31 stem rust resistance gene in wheat. This grasped again the full attention of crop scientists and revived targeted international crop research collaboration against stem rust in 2005 with the Borlaug Global Rust Initiative (http://www.globalrust.org [verified 1 June 2011]). When Ug99 continued to spread north through the eastern African highlands and across the Red Sea into Yemen, it also reached the wider public through the media (see for example Koerner, 2010). Flood (2010) is warning that action to combat plant health problems in general, with a specific and concerted action to combat the Ug99 stem rust epidemic, is of vital importance to ensure food security. Identification of novel sources of resistance to stem rust is urgent and a current priority of a number of crop research groups around the world.

### *Net Blotch*

Net blotch is caused by a fungal pathogen (*Pyrenophora teres* Drechs.). It is known to cause serious harm to barley and typically reduce yields by 10 to 40% (Steffenson, 1997). The disease thrives most under wet (high relative humidity) and warm conditions with temperature optima between 15 and 25°C depending on the region (Krupinsky et al., 2002). There are two common forms of the net blotch fungus: *Pyrenophora teres* f. *teres* produces a net-like pattern, while the *Pyrenophora teres* f. *maculata* produces more spot-like lesions on the leaves of the crop plants (Liu and Friesen, 2010). Afanasenko et al. (1995) reported that the resistance against *Pyrenophora teres* f. *teres* (net type) and the *Pyrenophora teres* f. *maculata* (spot type) of net blotch are independently inherited. New epidemics of the net blotch recently have been reported around the world (McLean et al., 2009; Liu and Friesen, 2010). Development of resistant varieties is the most cost-efficient method for control of net blotch (Jalli, 2010a, b). Silvar et al. (2010) reported finding few examples of resistance to net blotch in the Spanish barley core collection.

## MATERIALS AND METHODS

This study used the results of disease screenings for landrace accessions maintained by the USDA National Plant Germplasm System (NPGS) Germplasm Resources Information Network (GRIN) (USDA-ARS, 2010a). Only those accessions with geo-referenced collection sites were used in the study. Agro-climatic data, obtained from ICARDA (De Pauw, 2008) and WorldClim (Hijmans et al., 2005a; WorldClim, 2010), describing the collection sites were used to develop models to predict the presence of resistant phenotypes.

### The Stem Rust Trait Dataset

The stem rust data is available online from the USDA NPGS GRIN database (USDA-ARS, 2010c). Bonman et al. (2007) described the experimental design for the field trials. Susceptibility to stem rust (*Puccinia graminis* Pers. f.sp. *tritici*) was measured for six different years (during 1988–1994) at the agricultural research stations at St Paul (44°59′17″ N, 93°10′48″ W) and Rosemount (44°43′01″ N, 93°05′56″ W) located in Minnesota in the northern United States. Dr. Don V. McVey made all of the trait observations for both locations. The trial experiments at Rosemont were inoculated by race TNMK. The trial experiments at St Paul were inoculated by race QFBS, RKQS, and RTQQ. The trials at St Paul in 1988 and 1989 were also inoculated by race QSHS and RHRS, since 1991 also by race HNLQ, and since 1992 also inoculated by race TNMK. The dataset contains observations for bread wheat (*Triticum aestivum* L.) and durum wheat (*Triticum turgidum* L.) and a total of 10 different subspecies. The original source locations for the wheat landraces are widely distributed across countries in Europe, Asia, and northern Africa (Fig. 1). Ethiopia and Turkey are the best represented countries with each having more than 20% of the landraces from the stem rust dataset. Complementary geo-referencing was made at ICARDA for landraces with missing geographic coordinates based on the description of the original collecting site. Only the 4932 landraces successfully geo-referenced were included in this study. The original stem rust ratings (6889 trait observations) were reported as classified into 10 classes according to the degree of reaction to the disease. The stem rust trait ratings 0 through 3 (1915 landraces or 28% of the total) were considered as resistant to stem rust, ratings 4 through 6 (2729 landraces or 40%) as intermediate, and ratings 7 through 9 (2245 landraces or 32%) as susceptible. The complete stem rust dataset for this study including the eco-climatic data is included as supplemental material (Supplemental File S1).

### The Net Blotch Trait Dataset

The net blotch trait dataset is available online from the USDA NPGS GRIN database (USDA-ARS, 2010b). The FIGS net blotch set was extracted from the USDA GRIN NPGS database by Dr. Harold Bockelman and includes trait observations for the reaction to net blotch (*Pyrenophora teres* Drechs.) for a total of 4645 barley (*Hordeum vulgare* L.) landraces (including greenhouse observations). Although net blotch seedling data from the greenhouse were available, they did not differentiate the landraces to the degree of the field trials. Thus, for this study we decided to focus on the analysis of the trait ratings from the field trials. From the net blotch dataset a total of 2786 geo-referenced accessions

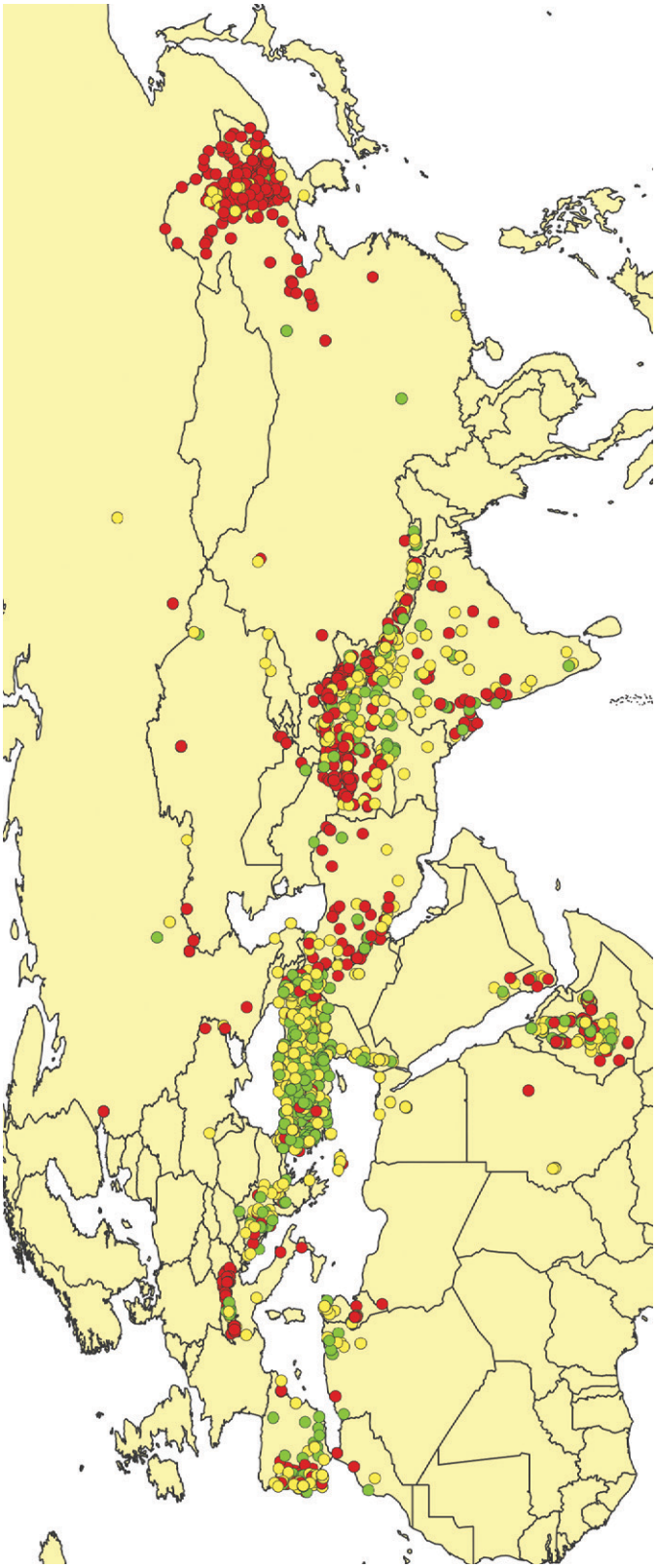

Figure 1. Distribution of the original source locations for the wheat landraces (stem rust dataset; 4932 accessions from 2013 collecting sites). Green circles indicate the collecting site for landraces resistant to stem rust, yellow circles show the medium susceptible, and red circles show where the susceptible landraces were collected.

were tested under field conditions during eight different years (1988–2004) at four different agricultural research stations: Langdon, ND (48°45′43″ N, 98°22′20″ W), Stephen, MN (48°27′03″ N, 96°52′30″ W), Fargo, ND (46°52′37″ N, 96°47′20″ W), and Athens, GA (33°57′18″ N, 83°22′59″ W). The field trial experiments were inoculated by isolate ND89-19 of net blotch (*Pyrenophora teres* f. *teres*) using infected barley straws from the previous season (Bonman et al., 2005). The original net blotch trait ratings (2786 trait observations) were reported as classified into nine classes with ratings 1 through 3 (1115 landraces or 40% of the total records) considered as resistant to net blotch, ratings 4 through 6 (1367 landraces or 49%) as intermediate, and ratings 7 through 9 (304 landraces or 11%) as susceptible. The original source locations for the barley landraces are widely distributed across 51 countries in Asia, Europe, and northern Africa (Fig. 2). A total of 1025 (36.8%) of the landraces originate from Ethiopia. The next country ranked by total number of records in the dataset was China with 365 (13.1%) landraces. The complete net blotch dataset for this study, including climate data, is included as supplemental material (Supplemental File S2).

## ICARDA Eco-Climatic Database

The ICARDA eco-climatic information system was created in 2003 covering the Central and West Asia and North Africa (CWANA) region (De Pauw, 2008). In 2005 the dataset was further extended to cover Europe and most of Asia. The "thin-plate smoothing spline" method of Hutchinson (1995), as implemented in the ANUSPLIN software (Hutchinson, 2000), was used to convert the station based climatic database into "climate surfaces" with a 30 arc-second (approximately 1 km) resolution grid. The dataset includes monthly mean values for minimum temperature (tmin), maximum temperature (tmax), precipitation (prec), and potential evapotranspiration (PET) as well as a wide range of derived climatic variables such as agroclimatic zone, aridity index, length of growing period, and others. The soil layers from the ICARDA eco-climatic information system are originally derived from the FAO Soil Map of the World (FAO, 1974, 2007; FAO-UNESCO, 1995). The climate data for this study was extracted for each accession using the longitude and latitude coordinates of the original collecting site.

## WorldClim Eco-Climatic Database

Climate data for the net blotch dataset was extracted from the WorldClim dataset (Hijmans et al., 2005a; WorldClim, 2010) with DIVA GIS (Hijmans et al., 2001, 2005b). The WorldClim dataset was developed (with the ANUSPLINE software [Hutchinson, 2000]) following a similar method as described above for the ICARDA eco-climatic information system. The climatic layers from the WorldClim dataset include monthly mean values for temperature (temp), minimum temperature (tmin), maximum temperature (tmax), precipitation (prec), and the derived bio-climatic (BIOCLIM) layers (Busby, 1991). The WorldClim dataset is available in different spatial resolutions: 30 arc-seconds (approximately 1 km), 2.5 min (approximately 4.5 km = 20 km$^2$), 5 min (approximately 9.3 km = 86 km$^2$), and 10 min (approximately 18.5 km = 342 km$^2$).

## Trait Mining Models

The underlying hypothesis for this study was that certain types of environments would favor the emergence of disease resistance within in situ populations of landraces. Negri et al. (2009:9) proposed to define a landrace as a cultivated plant with the "lack of formal crop improvement" and "characterized by a specific adaptation to the environmental conditions of the area of cultivation." The general approach was to classify collection site environments into those that are likely to yield a certain category of disease reaction. This was achieved by applying classification models to one set of data in which the disease score is used to "train" the model so that it correctly classifies environments according to disease type. Another "test" subset was used to proof the classification model using the disease scores for the selected accessions.

### *Validation Subsets (Training Set and Test Set)*

The training set used to calibrate the prediction models comprised 67% of the records, and the test set made up 33% of the records. For each subset the allocation of samples to the training set and the test set was a random process. The calibration of the model parameters was made with cross-validation (Hawkins et al., 2003). The samples from each of the test sets were only used to validate the predictive performance and were not included in any calibration steps (Hawkins, 2004).

## Classification Algorithms

The predictive performance for four different classification methods was compared.

(1) Linear discriminant analysis (LDA). The LDA (Fisher, 1936) classifier assumes a normal (Gaussian) distribution for the predictor variables. Also, the residuals are assumed to show the normal distribution. The normality assumption can be tested for example with a Q-Q plot or a Wilk-Shapiro test (Shapiro and Wilk, 1965). The LDA method requires more assumptions to be met than the other classification methods of this study. However, parametric methods usually outperform nonparametric methods when the assumptions to the underlying data distribution are met. It would thus always be wise to include a parametric classification method when the assumption of normality is evaluated to be reasonable.

(2) Partial least squares discriminant analysis (PLS-DA). Partial least squares discriminant analysis (Barker and Rayens, 2003) is based on the partial least squares (Wold, 1966; Wold et al., 1984). With PLS-DA the training set is used to calibrate new latent variables that can be seen as a linear combination of the previous multivariate variables. The unknown samples are projected into this new multivariate space defined by the latent variables. A separate submodel is calibrated for each class. All of these submodels are calibrated together in an iterative process to simultaneously fit the independent predictor variables (climate data)

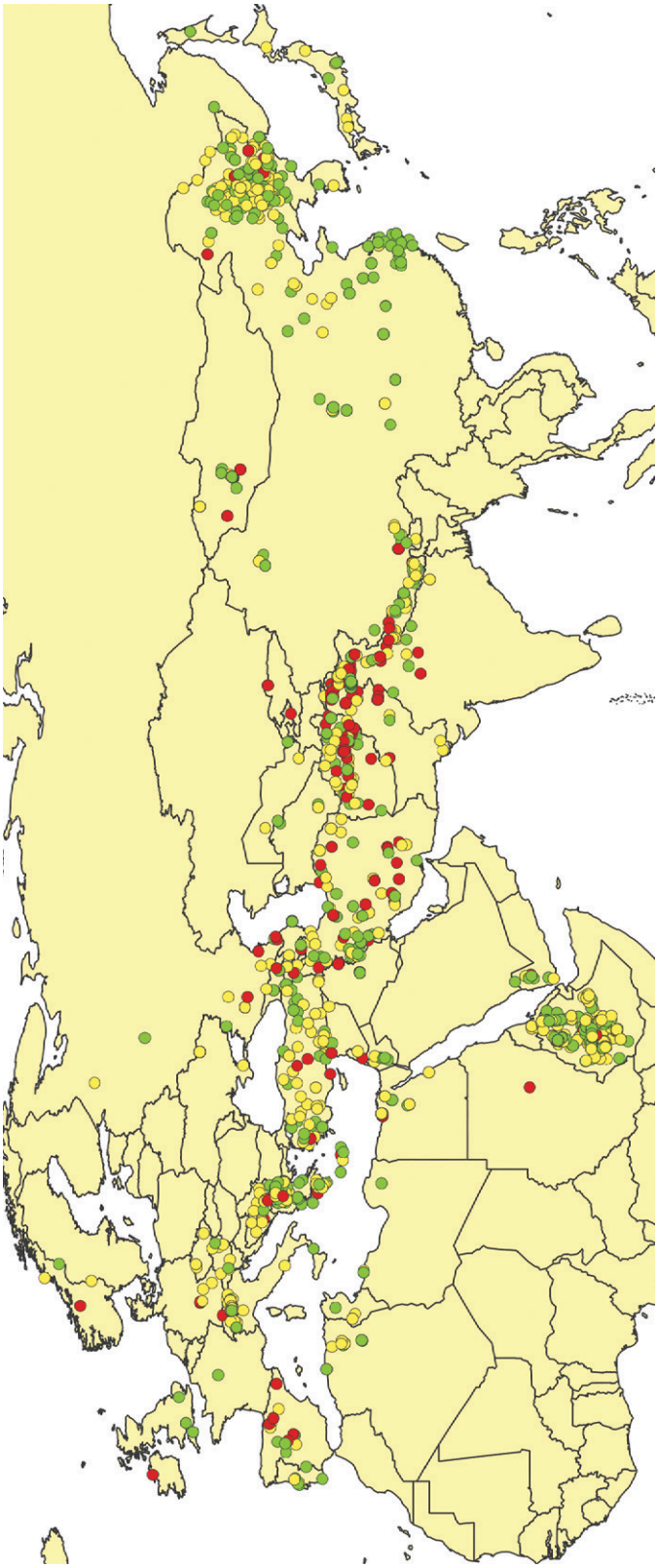

Figure 2. Distribution of the original source locations for the barley landraces (net blotch dataset; 2786 accessions). Green circles indicate the collecting site for landraces resistant to net blotch, yellow circles show the medium susceptible, and red circles show the locations where susceptible landraces were collected.

and the dependent response variables (trait data). Compression of the original multivariate variables to latent factors or principal components will often provide a solution to the common problem of colinearity between the predictor variables.

(3) k-nearest neighbor (kNN). k-nearest neighbor is a pattern recognition method for classification of objects based on the majority vote of the closest neighbors (Cover and Hart, 1967). This is one of the simplest classification algorithms. The kNN algorithm is a nonparametric classification method and thus makes no assumption on the underlying data distribution (Duda et al., 2001). For all the kNN classifiers in this study we used k = 1, in which the observation is assigned to the class of its nearest neighbor.

(4) Soft independent method of class analogy (SIMCA). Soft independent modeling by class analogy is a method of classification in which separate principal component analysis (PCA) models are calibrated for each class in the dataset (Wold and Sjostrom, 1977; Wold, 1976) (similar as for PLS-DA). These submodels are calibrated independently and only fitted to the independent predictor variables (different from the PLS-DA algorithm). The unknown samples are fitted to each of the submodels and assigned to the submodel of the closest fit. An unknown sample can in practice be assigned to multiple classes or to no class. The SIMCA method is a projection method in which the multivariate dataset under study is condensed to a set of lower-dimensional subspaces (PCA models). The SIMCA method requires few other assumptions than that the condensed subspaces provide a meaningful representation of the original dataset.

For each test set the classification results were compared to the results from an actual random selection. For the random selection, the trait scores for each test set where reassigned using random permutation of accession numbers. Using this approach the number of examples for each trait category remained unchanged, but any link between the trait and the eco-climatic description was broken. With the random selection experiments we can directly compare the performance of the classification methods to the effect of a random selection in practice. The behavior and stability of the performance indicators compared to a random selection are illustrated here with a practical test.

All of the classification tests for each subset were repeated three to five times to provide a series of replicated classification experiments. For each of these series the subset randomly was split into a new training set and a new test set. The classification indicators reported below are the average from these replicated classification experiments.

## Trait Mining Prestudy

Before starting the trait mining experiments a series of prestudy tests were made to identify the most appropriate classification algorithm (choosing from kNN, SIMCA, PLS-DA, or LDA), and the most appropriate number of levels for the trait score measurement scale (choosing from 2, 3, or 9 category levels). The most appropriate classification method and number of levels for the trait scale was next used for the corresponding series of trait mining experiments described below. The results from the prestudy tests are reported in Tables 1, 2, and 3.

### *Performance for the Models When the Stem Rust Disease Score Categories Were Rescaled*

This study was designed to explore if reclassification to fewer trait categories might contribute to improved predictions for the resistant landraces. The objective of this study was not to predict accurate levels of disease susceptibility but to identify the resistant landraces. The degree of disease susceptibility among the susceptible landraces was not the aim of this trait mining experiment. The disease score categories for stem rust were rescaled from a 0 to 9 scale as follows: S2 included disease scores reclassified to two classes to represent resistance (0–3) and susceptibility (4–9), S3 included disease scores reclassified to three classes (0–3), (4–6), and (7–9), and S9 included disease scores reclassified to nine classes (0–1), 2, 3, 4, 5, 6, 7, 8, and 9. The net blotch dataset did not include the original trait score zero (0), but the reclassification followed in every other respect exactly the same schema as for the stem rust set.

## Trait Mining Experiments

The modeling experiments described here explore the predictive performance when using different data stratification strategies and different eco-climatic data sets.

### *Experiment 1: Performance of Models When Different Eco-Climatic Data is Used*

In this experiment the predictive power of the models to identify stem rust resistance in wheat were compared using the three different eco-climatic data sets: the ICARDA eco-climatic dataset respectively with and without the layers for PET included (ICARDA 30″ and ICARDA 30″ with PET [30 s = 1 km spatial resolution]) (De Pauw, 2008), the WorldClim dataset (WorldClim 2.5′ [2.5 min = 4.5 km resolution]) (Hijmans et al., 2005a; WorldClim, 2010), and the BIOCLIM layers (Busby, 1991) derived from the WorldClim layers (BIOCLIM 2.5′ [4.5 km resolution]). The results from Exp. 1 are included in Table 4.

### *Experiment 2: Performance of Models After the Data is Stratified According to Genetic Background*

In this experiment the predictive power of the SIMCA models to identify stem rust resistance in wheat landraces were compared for two subsets containing data for *Triticum aestivum* L. and *Triticum turgidum* L. The hypothesis was that the different taxa (genetic background) of wheat might have slightly different mechanisms of resistance against stem rust and thus a different predictive association between the eco-geographic parameters and the trait score. The results from Exp. 2 are included in Table 5.

### *Experiment 3: Performance of Models After the Data is Stratified According to Test Site*

In this experiment the predictive power of the SIMCA models to identify stem rust resistance in wheat were compared for two subsets. One subset containing data for scores obtained in St. Paul, MN, and the other containing scores obtained in Rosemount, MN. The results from Exp. 3 are included in Table 5.

Table 1. Comparison of the performance of (i) different classification models and (ii) the performance of the soft independent modeling by class analogy (SIMCA) model using different levels of the trait measurement scale to predict the occurrence of stem rust resistance in wheat (prestudy 1).†

| Dataset | Model‡ | Scale | PPV§ | LR+¶ | Estimated gain |
|---|---|---|---|---|---|
| Stem rust | LDA | S3 | 0.48 (0.45–0.52)# | 2.40 (2.18–2.64) | 1.73 (1.61–1.85) |
| Stem rust | PLS-DA | S3 | 0.44 (0.41–0.47) | 2.01 (1.86–2.17) | 1.57 (1.46–1.68) |
| Stem rust | kNN | S3 | 0.49 (0.45–0.53) | 2.46 (2.17–2.77) | 1.75 (1.61–1.88) |
| Stem rust | SIMCA | S3 | 0.54 (0.50–0.59) | 3.07 (2.66–3.54) | 1.95 (1.79–2.09) |
| Stem rust | SIMCA | S9 | 0.53 (0.48–0.57) | 2.86 (2.47–3.32) | 1.88 (1.73–2.04) |
| Stem rust | SIMCA | S2 | 0.51 (0.48–0.55) | 2.72 (2.42–3.07) | 1.84 (1.70–1.97) |
| *Random* | | | *0.29 (0.26–0.33)* | *1.04 (0.90–1.20)* | *1.03 (0.91–1.16)* |

†The reported performance indicators are the average from three to five replications (with different random split to training set and test set) for the same full stem rust dataset. The record level data unit is here the trait measurement from the USDA net blotch set (6889 observations). The proportion true resistant samples in this dataset were 28% (prevalence = 0.28).

‡kNN, k-nearest neighbor; LDA, linear discriminant analysis; PLS-DA, partial least squares discriminant analysis.

§The positive predictive value (PPV) provides an indicator for classification performance of resistant samples (positives).

¶The positive diagnostic likelihood ratio (LR+) provides a similar indicator that is less sensitive to the prevalence or proportion of resistant samples (positives) in the dataset.

#The 95% confidence interval is included inside the parentheses.

Table 2. Comparison of the ability of different classification models and different number of levels of the trait measurement scale to predict the occurrence of stem rust resistance in wheat (prestudy 2).†

| Dataset | Model‡ | Scale | PPV§ | LR+¶ | Estimated gain |
|---|---|---|---|---|---|
| Stem rust (site) | LDA | S3 | 0.39 (0.33–0.46) # | 2.60 (2.12–3.23) | 1.97 (1.65–2.32) |
| Stem rust (site) | PLS-DA | S3 | 0.37 (0.31–0.44) | 2.42 (2.00–2.93) | 1.89 (1.57–2.20) |
| Stem rust (site) | kNN | S3 | 0.44 (0.36–0.53) | 3.21 (2.44–4.28) | 2.23 (1.82–2.65) |
| Stem rust (site) | SIMCA | S3 | 0.50 (0.40–0.60 | 4.00 (2.85–5.66) | 2.51 (2.02–2.98) |
| Stem rust (site) | SIMCA | S9 | 0.49 (0.41–0.59) | 3.96 (2.91–5.43) | 2.50 (2.03–2.93) |
| Stem rust (site) | SIMCA | S2 | 0.47 (0.39–0.55) | 3.58 (2.75–4.69) | 2.37 (1.96–2.77) |
| *Random* | | | *0.19 (0.13–0.26)* | *0.94 (0.63–1.39)* | *0.95 (0.66–1.33)* |

†The record level data unit is here the collecting site from the stem rust dataset (total 2013 sites; prevalence = 0.20). The disease trait scores for all wheat landraces collected at the same site were averaged to make one data record for each distinct collecting site.

‡kNN, k-nearest neighbor; LDA, linear discriminant analysis; PLS-DA, partial least squares discriminant analysis; SIMCA, soft independent modeling by class analogy.

§PPV, positive predictive value.

¶LR+, positive diagnostic likelihood ratio.

#The 95% confidence interval is included inside the parentheses.

Table 3. Comparison of the ability of different classification models to predict the occurrence of net blotch resistance in barley (prestudy 3).†

| Dataset | Model‡ | Scale | PPV§ | LR+¶ | Estimated gain |
|---|---|---|---|---|---|
| Net blotch | LDA | S3 | 0.50 (0.45–0.56)# | 1.52 (1.29–1.79) | 1.26 (1.13–1.39) |
| Net blotch | PLS-DA | S3 | 0.48 (0.42–0.55) | 1.41 (1.15–1.77) | 1.21 (1.06–1.37) |
| Net blotch | kNN | S3 | 0.51 (0.46–0.56) | 1.56 (1.33–1.82) | 1.27 (1.14–1.40) |
| Net blotch | SIMCA | S3 | 0.54 (0.48–0.60) | 1.75 (1.42–2.17) | 1.35 (1.19–1.50) |
| *Random* | | | *0.40 (0.35–0.45)* | *0.99 (0.84–1.17)* | *0.99 (0.87–1.12)* |

†The record level data unit is here the trait measurement from the USDA net blotch set (2786 observations; prevalence = 0.40).

‡kNN, k-nearest neighbor; LDA, linear discriminant analysis; PLS-DA, partial least squares discriminant analysis; SIMCA, soft independent modeling by class analogy.

§PPV, positive predictive value.

¶LR+, positive diagnostic likelihood ratio.

#The 95% confidence interval is included inside the parentheses.

### *Experiment 4: Performance of Models After the Data was Stratified According to Year of Screening*

In this experiment the predictive power of the SIMCA models to identify stem rust resistance in wheat was compared for six subsets corresponding to what year the disease scoring was undertaken: 1988, 1989, 1991, 1992, 1993, or 1994. The results from Exp. 4 are included in Table 5.

Note that, for Exp. 1, 2, 3, and 4, collection site eco-climatic data were included for all accessions scored. A total of 6889 records with stem rust ratings (corresponding to a total of 4932 genebank accessions) were processed, meaning that some of the sites (2013 sites in total) were represented in the analysis multiple times.

### *Experiment 5: Performance of the Models When Only One Accession per Site is Included in the Analysis*

The complete trait datasets from the USDA includes only very few multiple measurements for the same landrace (replications), but there are often multiple landraces in the dataset that originate from the same source location. Thus, the eco-geographic

Table 4. Performance of the soft independent modeling by class analogy (SIMCA) model to select stem rust resistant accessions using different eco-climatic layers (ICARDA, WorldClim, and bio-climatic [BIOCLIM]) compared to a random selection (Exp. 1).†

| Eco-climate‡ | PPV§ | LR+¶ | Estimated gain |
|---|---|---|---|
| ICARDA 30″ with PET | 0.55 (0.51–0.59)# | 3.08 (2.67–3.54) | 1.94 (1.81–2.11) |
| ICARDA 30″ | 0.57 (0.52–0.61) | 3.31 (2.87–3.82) | 2.00 (1.87–2.17) |
| WorldClim 2.5′ | 0.57 (0.52–0.61) | 3.54 (3.05–4.10) | 2.10 (1.94–2.25) |
| BIOCLIM 2.5′ | 0.49 (0.45–0.53) | 2.48 (2.16–2.86) | 1.76 (1.59–1.89) |
| *Random selection* | *0.29 (0.26–0.33)* | *1.04 (0.90–1.20)* | *1.03 (0.91–1.16)* |

†The record level data unit is here the trait measurement from the USDA stem rust set (6889 observations; prevalence = 0.28).

‡ICARDA 30 sec with potential evapotranspiration (PET), ICARDA 30 sec, ICARDA eco-climatic dataset (De Pauw, 2008) with the layer for potential evapotranspiration respectively included and omitted and 30 sec spatial resolution; WorldClim 2.5 min (Hijmans et al., 2005a; WorldClim, 2010); BIOCLIM 2.5 min (Busby, 1991).

§PPV, positive predictive value.

¶LR+, positive diagnostic likelihood ratio.

#The 95% confidence interval is included inside the parentheses.

Table 5. Performance of the soft independent modeling by class analogy (SIMCA) model to select stem rust resistant accessions, after stratifying the trait data according to (i) species, (ii) location of trial, and (iii) year of trail compared to a random selection (Exp. 2, 3, and 4).†

| Stratified subset | LR+‡ | Estimated gain | Prevalence§ |
|---|---|---|---|
| Bread wheat | 4.27 (3.42–5.28)¶ | 2.57 (2.28–2.90) | 0.20 |
| Durum wheat | 1.76 (1.43–2.17) | 1.32 (1.17–1.45) | 0.44 |
| St. Paul, MN | 3.07 (2.62–3.59) | 2.13 (1.93–2.40) | 0.21 |
| Rosemount, MN | 2.54 (2.01–3.21) | 1.55 (1.40–1.71) | 0.41 |
| Trial year 1988 | 4.44 (3.23–6.12) | 3.70 (2.37–5.28) | 0.06 |
| Trial year 1989 | 2.16 (1.82–2.55) | 1.38 (1.28–1.49) | 0.48 |
| Trial year 1991 | 2.53 (2.04–3.12) | 1.92 (1.58–2.21) | 0.21 |
| Trial year 1992 | 4.67 (3.19–6.82) | 2.28 (1.87–2.57) | 0.29 |
| Trial year 1993 | 6.18 (3.35–11.40) | 2.17 (1.74–2.44) | 0.36 |
| Trial year 1994 | 2.86 (1.09–7.51) | 2.18 (0.89–3.80) | 0.17 |
| *Random selection* | *1.30 (1.02–1.68)* | *1.21 (0.98–1.47)* | *0.26* |

†This stem rust dataset is stratified by the St. Paul, MN, Agricultural Research Station and the Rosemount, MN, Agricultural Research Station, both located in the northern United States; by six distinct trial years, 1988, 1989, 1991, 1992, 1993, and 1994; and by the two wheat subspecies with the most data records (landraces). The record level data unit is here the original observation measurement from the USDA stem rust set.

‡LR+, positive diagnostic likelihood ratio.

§The positive predictive value (PPV) indicator is not included to this table because the PPV is sensitive to the prevalence, which is very variable for these stratified subsets.

¶The 95% confidence interval is included inside the parentheses.

parameters for the landraces collected at the same source location are identical. A total of 1124 landraces from the Ethiopian stem rust set originated from only 191 different source locations. For India, Afghanistan, and Turkey there were also more than twice as many landraces as source locations. This trait mining experiment in the stem rust set was designed to explore the effect of pooling observations by distinct source locations. The hypothesis was that source locations with high sample number would tend to dominate the models, and that reducing the number of landraces to one per collecting site would give the eco-geography of each site more equal influence on the prediction. In Exp. 5 the modeling design for Exp. 1 above was repeated using only one accession per site. Thus only 2013 records were processed in total for each experiment. The disease scores were averaged across accessions for sites containing multiple entries; thus only one (average) disease rating represented sites with multiple accessions. The results from Exp. 5 are included in Table 6.

### *Experiment 6: Performance of the Models to Select Net Blotch Resistant Barley Accessions When using Different Eco-Climatic Datasets*

In this experiment the predictive power of the models to identify net blotch resistance in barley were compared using the three different eco-climatic datasets: the ICARDA eco-climatic dataset with and without the layers for PET (ICARDA 30″ and ICARDA 30″ with PET [30 s = 1 km spatial resolution]) (De Pauw, 2008), the WorldClim dataset (WorldClim 2.5′ [4.5 km resolution], WorldClim 5′ [9.3 km resolution], and WorldClim 10′ [18.5 km resolution]) (Hijmans et al., 2005a; WorldClim, 2010), and the BIOCLIM layers (Busby, 1991) derived from the WorldClim layers (BIOCLIM 2.5′, BIOCLIM 5′, and BIOCLIM 10′). The results from Exp. 6 are included in Table 7.

## Evaluation of the Trait Mining Results

The target of this trait mining study was concerned with the identification of resistant landraces rather than with the accurate classification of the samples to the different stem rust score categories. Thus, to assess the positive predictive performance for the identification of resistant landraces, the so-called confusion matrices (Kohavi and Provost, 1998) for each trait mining experiment was first collapsed to a two by two table. The collapsed confusion matrix tabulated the predicted resistant and susceptible landraces against the actual observed number of resistant and susceptible samples from the test sets. The samples, scored as 0 through 3 for the stem rust set and 1 through 3 for net blotch (on the original trait scale) were pooled together and classified as resistant samples. The samples scored as 4 through 9 on the original trait scale were classified as susceptible samples. The landraces predicted to be resistant were either true positive (TP), when a landrace observed to be resistant was also predicted to be resistant (positive), or they were false positives (FP). Likewise the landraces predicted to be susceptible were either true negatives (TN) or false negatives (FN).

To establish and compare the predictive performance for the different test models, we need metrics to compare the models. A number of different indicators of inter-rater agreement have been developed for different purposes (Gwet, 2010). When choosing the indicator to use it is important to remember the aim of the current study was to identify a smaller subset of landraces more likely to be resistant to stem rust than what would be expected when selecting landraces by chance. We are thus more concerned with the specificity of the model to identify

**Table 6. Comparison of the performance of the soft independent modeling by class analogy (SIMCA) model to select wheat accessions resistant to stem rust using different eco-climatic data (ICARDA, WorldClim, and bio-climatic [BIOCLIM]) and with distinct collecting sites as the record level data unit (Exp. 5).†**

| Eco-climate‡ | PPV§ | LR+¶ | Estimated gain |
|---|---|---|---|
| ICARDA 30″ with PET | 0.53 (0.44–0.65)# | 4.70 (3.41–7.47) | 2.76 (2.29–3.39) |
| ICARDA 30″ | 0.43 (0.33–0.53) | 3.16 (2.13–4.51) | 2.23 (1.73–2.74) |
| WorldClim 2.5′ | 0.55 (0.46–0.67) | 4.46 (3.20–7.02) | 2.55 (2.09–3.09) |
| BIOCLIM 2.5′ | 0.46 (0.40–0.56) | 3.55 (2.93–4.96) | 2.38 (2.08–2.90) |
| *Random selection* | *0.19 (0.13–0.26)* | *0.94 (0.63–1.39)* | *0.95 (0.66–1.33)* |

†The record level data unit is here the collecting site from the stem rust dataset (total 2013 sites; prevalence = 0.20). The disease trait scores for all wheat landraces collected at the same site were averaged to make one data record for each distinct collecting site.

‡ICARDA 30 sec with potential evapotranspiration (PET), ICARDA 30 sec, ICARDA eco-climatic dataset (De Pauw, 2008) with the layer for potential evapotranspiration respectively included and omitted and 30 sec spatial resolution; WorldClim 2.5 min (Hijmans et al., 2005a; WorldClim, 2010); BIOCLIM 2.5 min (Busby, 1991).

§PPV, positive predictive value.

¶LR+, positive diagnostic likelihood ratio.

#The 95% confidence interval is included inside the parentheses.

**Table 7. Comparison of the performance of the soft independent modeling by class analogy (SIMCA) model to select barley accessions resistant to net blotch when using different eco-climatic data (and different spatial resolutions) (Exp. 6).†**

| Eco-climate‡ | Spatial resolution | PPV§ | LR+¶ | Estimated gain |
|---|---|---|---|---|
| ICARDA 30″ with PET | 1 km | 0.52 (0.47–0.57)# | 1.67 (1.43–1.95) | 1.32 (1.19–1.45) |
| ICARDA 30″ | 1 km | 0.53 (0.47–0.58) | 1.69 (1.44–1.99) | 1.33 (1.19–1.46) |
| WorldClim 2.5′ | 4.5 km | 0.56 (0.50–0.61) | 1.81 (1.52–2.17) | 1.36 (1.23–1.49) |
| WorldClim 5′ | 9.3 km | 0.53 (0.48–0.59) | 1.71 (1.44–2.04) | 1.33 (1.20–1.47) |
| WorldClim 10′ | 18.5 km | 0.54 (0.47–0.60) | 1.71 (1.37–2.14) | 1.33 (1.18–1.50) |
| BIOCLIM 2.5′ | 4.5 km | 0.55 (0.46–0.64) | 1.90 (1.34–2.72) | 1.40 (1.17–1.63) |
| BIOCLIM 5′ | 9.3 km | 0.55 (0.46–0.63) | 1.86 (1.34–2.56) | 1.39 (1.16–1.58) |
| BIOCLIM 10′ | 18.5 km | 0.56 (0.48–0.63) | 1.89 (1.43–2.53) | 1.40 (1.20–1.59) |
| *Random selection* | | *0.40 (0.35–0.45)* | *0.99 (0.84–1.17)* | *0.99 (0.87–1.12)* |

†The bio-climatic (BIOCLIM) eco-climatic layers are included with the same resolutions. The record level data unit is here the original observation measurement from the USDA net blotch dataset (total 2786 accessions; prevalence = 0.40).

‡ICARDA 30 sec with potential evapotranspiration (PET), ICARDA 30 sec, ICARDA eco-climatic dataset (De Pauw, 2008) with the layer for potential evapotranspiration respectively included and omitted and 30 sec spatial resolution; WorldClim 2.5 min, WorldClim 5 min, and WorldClim 10 min (Hijmans et al., 2005a; WorldClim, 2010); BIOCLIM 2.5 min, BIOCLIM 5 min, and BIOCLIM 10 min (Busby, 1991).

§PPV, positive predictive value.

¶LR+, positive diagnostic likelihood ratio.

#The 95% confidence interval is included inside the parentheses.

the resistant samples than on the overall agreement related to correctly predicted samples.

Cohen's Kappa (Kappa) is a popular measure of inter-rater agreement for categorical (qualitative) items (Cohen, 1960). Cohen's Kappa is applied here as an indicator for the agreement between the actual observed trait scores and the predicted trait category as calculated by the models. The Kappa inter-rater agreement indicator aims to calculate the observed agreement adjusted for the level of agreement that would be expected by chance. The value of Cohen's Kappa ranges from –1 to +1, in which +1 implies perfect agreement, 0 implies no relationship, and –1 implies perfect disagreement (Landis and Koch, 1977). There remains some controversy in particular related to the application and calculation of agreement expected by chance. Cohen's Kappa, however, is widely used in the absence of standard, generally accepted alternatives (Gwet, 2010).

Among the other alternative indicators for predictive performance the raw proportion observed agreement (PO) is perhaps the most intuitive (Altman and Bland, 1994a). The total number of overall observed agreement (TP + TN) is simply divided by the total number of samples (N). The proportion positive agreement [PA = 2 × TP/(2 × TP + FP + FN)] might be a more appropriate indicator for our aim to identify the resistant (positive) samples. The positive predictive value (PPV = TP/(TP + FP)) is yet another suitable indicator to measure the predictive performance for identification of resistant landraces (Altman and Bland, 1994b). The PPV measures the probability that a landrace predicted by the model to be resistant is truly resistant (observed as resistant in the test set). The PPV is inherently dependent on the prevalence of resistant samples. However, most of the trait mining test sets for this study have the same proportion of resistant samples as the overall trait dataset, making the PPV a suitable indicator to compare the performance of these modeling experiments. Some subsets (e.g., split by experiment year or location) show very different prevalence for the proportion resistant samples. The prevalence was evaluated as the number of resistant samples in relation to all samples for each test set. The positive diagnostic likelihood ratio (LR+ = [TP/(TP + FN)]/[FP/(FP + TN)] = sensitivity/[1 – specificity]) is more appropriate when comparing test sets with very different prevalences (Altman and Bland, 1994b). The positive diagnostic likelihood ratio (LR+) measures how much more likely it is for the model to predict a landrace to be resistant (positive) in the group of landraces observed to be resistant compared to making this prediction in the group of landraces observed to be susceptible. For each of the modeling tests we have also calculated the sensitivity [sensitivity = TP/(TP + FN)] and the specificity [specificity = TN/

(TP + FP)] indicators (Altman and Bland, 1994a). The sensitivity is the proportion of actual resistant landraces that are correctly identified (predicted by the model) as such. The specificity is the proportion of susceptible landraces that are correctly identified. The last performance indicator reported here is the Yule's Q (Yule's Q = [OR − 1]/[OR + 1], in which the odds ratio [OR] = [TP × TN]/[FN × FP]). The odds ratio can be interpreted as the magnitude of association between the model predictions and the actual observed trait values. The Yule's Q is only a transformation of the odds ratio so that the indicator takes values in the range between −1 and 1. For more information and additional other alternative indicators for the predictive performance for classification tests see the text book by Agresti (2002) or the more recent text book by Gwet (2010).

All of the reported performance metrics are the average from a series of replicated trait mining tests for each subset. The estimate for the gain or improved predictive performance compared to a random selection for each trait mining test set is included in the right-side column of the results tables as the estimated gain. The gain is calculated as the PPV divided by the proportion resistant samples (prevalence) for each subset.

To further illustrate in practice the performance for a complete random selection, a series of actual random selections were made for each test set (with the same number of replications as for the trait mining models). The average of the tests with the random selection is reported at the bottom of the tables (rows marked "random selection").

For this study all performance indicators are reported in the supplemental material (Supplemental Table S3). For clarity, however, in the results section only the PPV, LR+, and the estimated gain indicators are reported.

The 95% confidence interval is reported for the performance indicators. The confidence intervals are included in a parenthesis after each indicator. The reported confidence interval upper and lower boundaries were calculated with the online Statistics Calculator from the Centre for Evidence-Based Medicine (CEBM, 2010).

### Software

The classification models were calculated with the MATLAB software (MathWorks, 2009) and the PLS Toolbox (version 5.8; Eigenvector, 2010). The Cohen's Kappa indicator was calculated with a MATLAB script by Cardillo (2007).

## RESULTS

The initial explorative PCA showed no clear grouping for the trait score categories (for either the stem rust or the net blotch sets). Only one sample was identified as a very strong outlier. This was landrace PI 212925 from the stem rust set. We could not identify why this landrace behaves as an outlier, yet the sample was very atypical of the other samples. We decided to remove this sample from the further analysis as reported below (6889 records with trait observations for 4932 wheat accessions collected from 2013 distinct sites). The initial PCA did not identify any notable outliers in the net blotch set (2786 trait observations from the same number of barley accessions).

### Prestudy 1

The purpose of the first prestudy (prestudy 1), reported in Table 1, was to identify the most appropriate classification algorithm to use in the experiments reported below as well as the most appropriate score reclassification scheme to use. All of the models used were able to significantly improve on a random selection to capture stem rust resistance. For example, even the lowest performing classification model (PLS-DA) was 1.57 (estimated gain in Table 1) times more effective when compared to a random selection. However, there were significant differences in the performance of the classification models. The PPV, LR+, and estimated gain performance indicators reported in Table 1 demonstrate that the SIMCA model had the most effective predictive power; followed by the kNN, LDA, and PLS-DA models, the latter of which yielded performance indicators that were significantly lower than the others. All the classification algorithms have significant higher predictive performance indicators than the random selection (not overlapping confidence interval).

The degree to which the score data was recategorized had an effect on the predictive power of all of the models. The results for the SIMCA model, detailed in Tables 1, 2, and 3, show that when the original 0 through 9 scoring categories are reclassified to three categories the models perform better than if asked to predict membership in two categories or nine categories.

### Prestudy 2

The second prestudy (prestudy 2) also examined the most appropriate classification algorithm to use as well as the most appropriate score reclassification scheme. This study differs from prestudy 1 in that the disease scores for sites represented by multiple accessions were averaged. Thus only one record per site was included in the analysis.

The trends demonstrated in Table 1 are corroborated by the results in Table 2. That is, the most effective classification algorithm to use is SIMCA, and three trait categories is the most appropriate trait scale to maximize the predictive performance of the models. However, processing the USDA stem rust dataset to include unique records for each distinct collecting site made a substantial improvement to the predictive performance. Note that the proportion of resistant cases (prevalence) is lower in the set with distinct sites (20% compared to 28% for the full set; see Supplemental Table S3.1). The PPV indicator is sensitive to the prevalence, so here we should focus on the more robust LR+ (and the estimated gain) than on the PPV indicator when comparing the performance for this dataset (prestudy 2) with the full set (prestudy 1). The estimated gain for the SIMCA model when using unique collection sites is 2.51 (150% higher hit rate compared to a random selection) compared to the gain of 1.95 (95%) when multiple accessions per site are processed. This is a substantial improvement for this indicator (even if the confidence intervals for the same classification algorithms in prestudy 1 and 2

are marginally overlapping). The LR+ indicator also shows a substantial improvement for the experiment with distinct collecting sites. In other words, these results indicate that it is better to average the scores across accessions when there are multiple accessions per site than to have a particular site represented multiple times in the analysis. In prestudy 2 (same as for prestudy 1) all of the classification algorithms have significant higher predictive performance indicators than the random selection (not overlapping confidence interval).

## Prestudy 3

The aim of the third prestudy (prestudy 3) was to determine the most effective classification model to predict the occurrence of net blotch resistance in barley, the results of which are detailed in Table 3. As for the prestudies in the stem rust set (prestudy 1 and 2), the SIMCA model outperformed the other models. Similar trends in relative predictive powers of the models were also demonstrated (SIMCA > kNN > LDA > PLS-DA). In this study the performance of the SIMCA model was 35% higher than for an expected random selection.

What is notable though is that while the LR+ and estimated gains demonstrate that net blotch resistance can be effectively predicted using eco-climatic parameters, the predicative power of the models is significantly lower than that demonstrated for the predictions in the stem rust set. For both LR+ and the estimated gain, the confidence intervals (95% CI) in the stem rust set are above and not overlapping with the confidence interval for these indicators in net blotch set. For the stem rust set, when using the SIMCA method, the 95% confidence interval estimated the value for LR+ to be between 2.66 and 3.54 (and for the stem rust set with distinct sites to be between 2.85 and 5.66). For the net blotch set the respective confidence interval was estimated to be between 1.42 and 2.17. For prestudy 3 the PLS-DA algorithm has marginally overlapping confidence interval for the performance indicators when compared to the random selection. The other classification algorithms have significant higher predictive performance than the random selection, the same as for prestudy 1 and 2.

Based on the above prestudies (Tables 1, 2, and 3), the models for the trait mining experiments reported with Tables 4, 5, 6, and 7 were calibrated with the SIMCA classification algorithm using a trait measurement scale with three categories.

## Experiment 1

The aim of this experiment was to see if there were significant differences between how the different eco-climatic datasets affect the predictive power of the models. The results for the SIMCA model are reported in Table 4.

While the performance indicators for the SIMCA model all score slightly higher for the WorldClim 2.5′ dataset than the two ICARDA sets, the confidence intervals indicate that there is little difference between these sets (Table 4). This result was expected because both the WorldClim and the ICARDA ecological database are constructed using the same spatial interpolation method and is to a large extent (although not fully) based on the same climatic data sources. However, the model's predictive power was significantly lowered when the BIOCLIM 2.5′ set (which are a different class of eco-climatic parameters) was used. Despite this, using the BIOCLIM 2.5′ set of eco-climatic parameters was still 73% more effective than making a random selection. The BIOCLIM set is derived from the WorldClim set, and these results thus indicate that using the "raw" WorldClim layers give a better predictive performance for this dataset than using the derived BIOCLIM layers.

The ICARDA eco-climate layers with PET included were used for the results reported from Exp. 2, 3, and 4 (Table 5). The choice of using these eco-climatic layers was made before analyzing the results from Exp. 1 (Table 4).

## Experiment 2

The aim of this study was to see if there was a difference in the predictive performance of the SIMCA model when applied to discrete datasets for *Triticum aestivum* and *Triticum turgidum*. When the dataset was stratified based on genetic background, the LR+ indicator showed an even higher response: 4.27 for bread wheat compared to 1.76 for durum wheat (Table 5). The bread wheat subset has substantially higher predictive performance compared to the random selection. The durum wheat subset has overlapping confidence interval (95% CI) with the random selection and thus we have no significant predictive effect from the models in this subset.

## Experiment 3

The predictive performance of the SIMCA model also differed when the stem rust data was split according to disease test site. The models performance using the St. Paul data was significantly higher than those obtained for Rosemount (LR+ = 3.07 and 2.54, respectively; Table 5). Both subsets (St. Paul and Rosemount), however, performed significantly better when compared to a random selection.

## Experiment 4

Stratifying the data according to year of trial also significantly affected the performance of the models. Again, notice that the stratified subsets often have very different proportions of resistant samples (prevalence), so the LR+ should be compared rather than the PPVs. Despite the differences in the models' performance when the data were split into years of experiment each result was significantly higher than a random selection process (LR+; Table 5).

## Experiment 5

Processing the USDA stem rust dataset to include unique records for each distinct collecting site made a substantial

improvement to the predictive performance in that the magnitude of the LR+ and estimated gain was greater in this experiment (Table 6) than those reported for Exp. 1 (Table 4). Note that the proportion resistant cases are lower in the set with distinct sites (20% compared to 28% for the full set).

In Exp. 1 (Table 4) the WorldClim-2.5′ eco-climatic layers gave the most effective identification of samples resistant to stem rust. By contrast, in Exp. 5 (Table 6) the ICARDA-30″ with PET eco-climatic layers showed the highest PPV, LR+, and estimated gain when compared to the other eco-climatic sets. However there was very little difference in predictive performance between the eco-climatic layers. None of the eco-climatic layers in Exp. 5 produced significant higher performance than any of the other layers (overlapping 95% confidence intervals; LR+; Table 6).

All of the models (eco-climatic sets) in Exp. 5 have substantially higher predictive performance compared to the random selection (by a very good margin not overlapping 95% confidence intervals).

### Experiment 6

Since the predictive performance in prestudy 3 for net blotch was significantly lower than that observed for stem rust (compare Tables 1 and 3) we explored the predictive performance for different climate data sets in more detail in this experiment than we did for the stem rust set to see if the performance could be improved.

All the eco-climatic sets yielded hit rates, for identification of samples resistant to net blotch, 32 to 40% higher than if selections had been made at random (Table 7). However, while the magnitudes of the LR+ and estimated gain for the BIOCLIM data show a marginally better performance than other classes of eco-climatic data, the confidence intervals indicate there was no appreciable difference between using the different eco-climatic sets or by using different degrees of resolution for the climatic surfaces (Table 7). That is, the finer resolutions did not improve the predictive performance of the model. Further, the predictive performance for the net blotch dataset still remained significantly poorer than those achieved for the stem rust dataset.

## DISCUSSION

This study clearly shows that the eco-geographic distribution of both stem rust resistance in wheat and net blotch in barley is not random but rather is linked to climatic factors. This supports the findings of Bonman et al. (2005, 2007) for rust and blotch diseases. Further, if this holds true for stem rust and net blotch then it is reasonable to assume it would hold true for other pests and diseases. As such, we can conclude that variables describing collection site environments can be used to identify disease or pest resistant landraces or wild relative accessions conserved in genetic resource collections at a better frequency than if material is selected at random or with a core collection. Trait mining using FIGS requires a small training set with trait scores to be available for the calibration of the model. While the work of El Bouhssini et al. (2009, 2010) and Bhullar et al. (2009, 2010) showed how this can be done to good effect for both pests and diseases, their studies did not include a comparison to a random selection process. By contrast this study did and thus can be considered as the first definitive proof of concept for the FIGS strategy applied to genebank mining for useful traits.

An important limitation for the exploitation of the link between the eco-climatic data and the trait property is the requirement of a small set a priori trait data to train the trait mining model. A heuristic approach to incorporate expert knowledge to select samples for the initial training set will help to reduce this dependence of a priori trait data. However, this training set needs to be screened before the first trait mining model can be calibrated. The further FIGS sets can be developed in a stepwise manner to incorporate all samples screened each trial season in the trait mining model to select the samples for the next trial season. Another limitation when using the FIGS strategy is the requirement of geo-referenced collecting sites. The eco-climatic dataset is extracted based on the geographic coordinate. Some of the genebank accessions lack appropriate information required to identify the collecting site.

The indications are that the modeling approaches used in this study could be useful to predict disease resistances in untested germplasm, provided there are data that can be used as a trainer set. However, this study also indicates that the models used here are sensitive to differences in where and when screenings take place (Exp. 3 and 4), the pathogen being tested (Exp. 1 compared to Exp. 5), and the host crop (Exp. 2). For the stem rust set the race used as inoculum was also different between trial years and trial season, which likely contribute to the difference in predictive performance observed for Exp. 2. The utility of these models in an applied context are still yet to be established. In a follow on study by the same authors the same models and trainer data deployed in this study were used to select small subsets of wheat landrace material with a higher than random frequency of Ug99 resistance. This study included 4563 wheat landraces screened for resistance to stem rust Ug99 in Yemen during 2007. The observed trait scores were not revealed to the person making the trait mining models except from a small training set of 825 samples (20%). The trait mining model developed using the same method as described here was used to select a subset of 500 samples predicted to have a higher likelihood of resistance for Ug99. The complete set included 10.2% resistant samples while the selected set of 500 samples was found to have 25.8% resistant samples; thus demonstrating that the models and approach used here can indeed be applied to a real life genebank situation (unpublished data, 2010). The challenge now is to improve the robustness and predictive power of the approaches used.

The initial prestudy tests revealed that for these trait datasets (stem rust and net blotch sets) the SIMCA classification algorithm produced the best models. If the datasets had fulfilled the parametric assumptions (normal distribution of variables and residuals) then we would have expected the LDA and PLS-DA algorithms to produce the best models. This is not the case for these datasets. The SIMCA algorithm can be seen as similar to the kNN approach in that it selects the class of PCA model to which the sample is most similar, while kNN selects the class that is most similar to the nearest sample. The prestudy tests revealed that the SIMCA models performed better than the kNN models (Tables 1, 2, and 3) and thus would be the model of choice to use in an applied context when the data are not normally distributed.

## Reclassification of the Measurement Scale

For the stem rust set the SIMCA models for different reclassification of the measurement scale indicated that the reclassification to use a scale with three levels improved the predictive performance. However, it must be noted that the reclassification must make sense in terms of what the measurement scale represents. That is, in this case disease scores from 0 through 3 can be considered as resistant, 4 through 6 as moderately resistant, and 7 through 9 as susceptible. The original trait measurement scale with 9 or 10 category levels caused problems for the calibration of some of the models (in particular for the stratified subsets with fewer number of samples) because of the lack of samples to represent some of the category levels. Even for the training sets with samples to represent all category levels a high number of levels may cause other issues. One such issue relates to how the classifier relates to the so-called level of measurement. Stevens (1946) suggested a formal taxonomy for different types of measurement scales (nominal, ordinal, interval, and ratio). All of the classifiers we used in this experiment (LDA, PLS-DA, kNN, and SIMCA) only make use of information from a nominal type measurement scale. This means that the classifiers do not assume any order in the category levels and does not "know" that the trait score 2 is between trait score 1 and trait score 3. This means that including more category levels in the dataset does not give the classifier more useful information to identify the resistant samples based on the order of the category levels. The reduction to a measurement scale with all the target samples (resistant landraces) grouped together is likely to give the classifier more information relevant to the task at hand—that is to discriminate the resistant samples from the susceptible samples. Following this argument one might expect that the trait scale with two category levels would be the best alternative. This was not the case in this study (prestudy 1 [Table 1] and prestudy 2 [Table 2]) in which the S3 scale showed a tendency to give the highest predictive performance (however, not statistically significant as evaluated by the overlapping 95% confidence intervals). It is possible that the samples from the original trait scores 7 through 9 provides the classifier with more coherent examples of the difference between the resistant and the susceptible samples than using original trait scores 4 through 9 as examples of susceptible samples. A combination of samples from the intermediate resistance and the susceptible groups would thus remove information that the model otherwise was able to exploit, while reducing the original measurement scale to three levels only removes information that the model was not able to exploit.

In the final step to evaluate the classification performance the predicted trait scores were reduced to only two levels with the confusion matrix. This last step of the trait mining experiments reported here, however, was made after the classifier has extracted information from the dataset and was motivated by the primary interest of this study to evaluate the performance of the classifier to identify resistant samples (positives) rather than to distinguish resistant samples from intermediate and susceptible samples.

## Stratification by Species and Screening Year and Location

For all of the stratification subsets the trait mining models perform better than a random selection. For some subsets the predictive performance is notably higher than for other subsets. But when we compare the predictive performance for different stratified subsets in the same group to the overall average predictive performance, we generally find that some subsets perform better while other subsets in the same groups of subsets perform lower than the overall average. For the stem rust set (Table 5) we found that some of the models limited to one single species, experiment site, or trial year perform better than the overall average performance, while others in the same group perform lower than the overall average (see Table 1, line 4). Similar tests in the net blotch set (not reported; see Supplemental Table S3.5) with stratification by experiment site and year show the same pattern. Thus we did not observe any clear evidence that splitting the full dataset into smaller stratified subsets containing similar samples provides any general improvement in predictive performance. In terms of creating data subsets across different models there is still the question: Does adding more predictive variables (more eco-geographical layers; see below) and splitting the data to different genetic backgrounds improve the prediction?

## The Bio-Climatic Layers

The BIOCLIM variables (Busby, 1991) are derived from the raw climate variables with the aim to better describe the eco-climatic environment with parameters that are more directly relevant for the description of the ecological niche. However, for the trait mining experiments with the stem rust set, the predictive performance is significantly higher when using the raw climate variables rather than the

derived BIOCLIM variables (not overlapping 95% CI for the WorldClim layers compared to the BIOCLIM layers in Exp. 1; Table 4). Note that in the net blotch set this result is reversed, and here the BIOCLIM eco-climatic set calibrates the models with the highest predictive performance (Exp. 6; Table 7). The original purpose of developing the BIOCLIM variables was for the calibration of envelope models in which the maximum and minimum value for each of the eco-climatic variables is assumed to define the boundaries of the species habitat (Busby, 1991; Franklin, 2010). For this study we have not explored the envelope modeling principle but rather used standard multivariate classification methods. Our study indicates that for the stem rust set the raw eco-climatic variables perform better than the derived BIOCLIM variables when using the SIMCA classifier. It is possible that the processing of the raw climate variables into the BIOCLIM variables does not always preserve all the predictive information content for our approach.

## The Potential Value of Additional Eco-Geographic Layers

The second observation related to the eco-climatic variables is that when the PET, from the ICARDA eco-climatic database, is added to the eco-climatic variables, the predictive performance is sometimes slightly higher (Table 6). This climate variable is only available for the ICARDA eco-climatic dataset. When we repeat the same trait mining experiments with PET included or excluded, we find for Exp. 5 (Table 6) (but not for Exp. 1 [Table 4] and Exp. 6 [Table 7]) that including this eco-climatic property improves the predictive performance. This indicates that the PET climatic layer could in some contexts carry independent predictive information that is useful to the trait mining models. This is hardly surprising in that atmospheric humidity around the plant directly impacts the PET and it is widely accepted that high humidity is associated with infection by fungal type plant pathogens (e.g., Hoffmann and Schmutterer, 1983).

Whatever the underlying reason, this result leads us to suggest that choosing appropriate eco-climatic variables will be crucial to improving the predictive performance of FIGS. For example, in this study monthly variables were used that described the entire year from January to December. However, for a given collection site the growing season usually does not start in January and does not last the whole year thus making some monthly variables not as relevant as others. A suggestion for further FIGS studies of this kind would thus be to explore the effects on the predicative power by (i) aligning monthly variable according to onset of growing period and (ii) only including monthly variables that are within the growing period for a given site. This, however, would necessitate accurate estimations of the onset of growing period. Continuous surfaces for this variable have been developed at ICARDA by De Pauw et al. (personal communication, 2010) and are currently being used in a study as suggested here.

## Predictive Performance is Lower for Net Blotch than for Stem Rust

Although we see a similar pattern of the performance indicators when compared to the random selection, the predictive performance indicators are substantially lower for the net blotch set than they are for the stem rust set. Afanasenko et al. (1995) discovered that the resistance in barley against net blotch caused by *Pyrenophora teres* f. *teres* (net form) and *Pyrenophora teres* f. *maculata* (spot form) are inherited independently. Perhaps resistance to net blotch in barley is more complex than the resistance to stem rust in wheat and thus more difficult to capture using the models developed in this study. Bonman et al. (2005) found that the response to net blotch in this dataset is correlated to the winter habit of the germplasm samples. It is possible that a trait mining study on subsets for each winter habit separate would give a higher predictive performance. However, overall both datasets show a very satisfactory predictive performance for the FIGS strategy in this study.

## Distinct Collecting Sites

Many genebanks contain multiple accessions from the same collection site. This could be due to a variety of reasons including multiple accessions being collected from the same site, one accession being split into different genotypes, non-geo-referenced accessions from a given province being assigned a collection site geo-coordinate that corresponds with the central point of the province, or material being stored in a collection are assigned collection site geo-coordinates corresponding to the physical location of the collection. Clearly, to use accessions in which the latter example is the case in a FIGS analysis would not be appropriate. However, the FIGS approach is relevant if the collection site geo-coordinates are reasonably accurate; thus, the question then becomes how one treats multiple accessions per collection site when using a FIGS approach.

The results of Exp. 5 (Table 6) suggest that, for analyses such as those reported, it is better to use an average score across accessions from the same site so that the site representation in the analysis is kept to a single entry. Many of the collecting sites in the stem rust set have a very high number of accessions collected at the same site. During the calibration of the classification models these collecting sites provide a very high number of examples from which the models could learn. The models could thus be biased, focusing too much on these collecting sites and neglecting useful information from the collecting sites with fewer accessions. Another contributing explanation could perhaps be that some of the accessions from the same collecting site have very different trait score values. The calibration of classification models would thus receive

a number of conflicting examples linking the same eco-geographic pattern to both high trait scores and to low trait scores. For the dataset with distinct sites, the average of the different trait scores for each site gave the calibration routine only one example of the link between the eco-geography and trait score for each site.

## Different Resolutions for the Eco-Climatic Layers

A somewhat unexpected result was that the finer spatial resolution for the eco-climatic layers did not improve the predictive performance (Exp. 6; Table 7). This experiment was only made for the net blotch set. Even if the predictive performance was slightly higher for the finer resolutions, the 95% confidence intervals clearly show that the observed differences between the different resolutions are insignificant. In mountainous areas the differences in particular for temperature, but also for precipitation can be substantial within the different spatial resolutions explored. However the area of (adaptive) cultivation for a landrace will sometimes be larger than even the largest grid cell we explored (10 arc-minutes = 18.5 km = 342 $km^2$). It is further possible that the geo-referenced coordinates for the reported collecting site is not the center point for the area of cultivation for the landraces. In some cases the collecting site might even be a farmers market in close proximity to the typical cultivation area. For these examples the eco-climate of the coarser spatial resolutions might be a better representative of the typical eco-climate of the landrace than the eco-climate of the smaller resolution pixel centered at the collecting site. Further experiments to explore the effect of the spatial resolution for the eco-climatic layers would be useful.

## Assumptions for the Focused Identification of Germplasm Strategy Approach

The FIGS strategy is based on the assumption that the expression of a useful trait, for example pest resistance, in landraces (and crop wild relatives) is linked to the environmental parameters describing the collection site and that we can build a statistical model to define a signature for the eco-geography of these landraces. The model in this study is applied as a search pattern to identify other landraces originating from locations with similar eco-geography as the resistant landraces. In practice these landraces would be selected as candidate samples for a field trial to screen for the target trait. Trait mining with FIGS aims to identify a higher proportion of resistant landraces than would be expected without the application of this selection strategy (Mackay and Street, 2004).

When modeling the crop resistance against a pathogen it is important to remember that the distribution of the pathogen is directly linked to the eco-geography. For example, given that many pathogens are sensitive to humidity, it is possible that the improved performance of the SIMCA model demonstrated in Exp. 5 was due to the inclusion of the evapotranspiration parameters.

For pathogens such as stem rust the distribution of the alternative host, barberry (*Berberis* L.), is required for sexual reproduction of the pathogen. The virulence of the pathogen is thus expected to be higher in areas where barberry grows in the proximity of the cultivated crop plants. The predictive association between the resistance trait and the eco-climatic variables we have identified with this study is thus, at least partly, an indirect link. The models are likely to describe the suitable eco-geography where the pathogen thrives and thus are likely to impose a selection pressure for the emergence of resistance genes within in situ populations. This was illustrated by Paillard et al. (2000) who report that populations of winter wheat with the highest level of resistance to powdery mildew originated from sites where powdery mildew pressure was high, due to environmental factors, while the reverse was true of those populations where the pressure was low. On the other hand, the models reported here are less likely to describe the eco-climatic conditions favorable for the crop to develop traits that would protect it against the pathogen without the presence of the pathogen. The development of useful resistance in the landraces is an adaptive response to the biotic stress from the pathogen and not the environment per se. However, to complicate the picture further, Stukenbrock and McDonald (2008) pointed out that many crop pathogens have been domesticated together with their host crop and are thus linked back to the geographic distribution of the crop.

The most important aspect of FIGS is that it is predictive. Focused identification of germplasm strategy does not aim to describe the mechanism behind the crop traits. If the models used to develop FIGS sets are predictive then they could be used to develop smaller subsets with a higher hit rate for a targeted crop trait.

## Focused Identification of Germplasm Strategy Models Provide a Complement to Expert Knowledge

The FIGS approach is not intended to replace the valuable expert knowledge held by crop breeders and genebank curators. When planning a new field experiment, the predictions from FIGS will assist the crop expert to select the most appropriate genebank accessions to include. The size of the smaller subset could be limited by the capacity given by the size of the available field area, the laboratory capacity, or the project funding available for human resources.

## Possible Causes of (Eventual) Prediction Problems

The predictive performance for the experiments in the stem rust and net blotch set from this current study was good. However, if the predictive performance is low when

the approach described here is followed, the list below provides some suggestions on how to improve the hit rate.

- The algorithm of the classifier is not able to recognize and discriminate all the samples. Further additional classification methods and other preprocessing methods can be explored.
- The models explored and compared in this study used long-term monthly climatic data arranged from January to December. However, when refining these processes it will be interesting to test if the predictive power of the techniques is improved when start of growing seasons are aligned so that only those months in which the crop would normally develop in situ are used in the models. In other words, sites are agro-climatically compared for similarities based on conditions prevailing during the actual growing seasons.
- The eco-geographic data from the source location of the landrace does not contain enough relevant information that could be linked to the evolution of a given crop trait. Other eco-geographic datasets or other grid resolutions can be explored. For example, measures of long-term season-to-season variation for climatic parameters would be useful when considering crop adaptation strategies.
- Data quality, precision, or error issues of the germplasm passport data. Data quality is of course paramount in any data analysis study. Written logbooks, collection mission reports, and similar sources can be revisited to complete missing data and improve on data accuracy, particularly the precision of collection site geo-coordinates.
- Lack of replicated measurements. Many datasets with evaluation of genebank material includes only one single observation for each genebank accession. With the lack of replication across multiple experiment years and experiment locations (agricultural research stations) it is very difficult to assess the precision and to estimate the natural variance of the trait scores or observations. It is also difficult to estimate any genotype × environment interaction effects in the trait dataset. Care should thus be made whenever possible to include replicated measurements across both experiment site and year for future trait evaluations. It is also important to apply an appropriate sampling design to avoid systematic bias in the recorded data.
- Assessment of trait variation could also be a problem as trait observations might include unexpected bias and mistakes. Some of the individual observations from the crop trait training set could be the result of an unusual experimental condition, for example, as locally higher pest stress pressure in smaller parts of the field plot or unusual low or high pest activity during some of the trial seasons. With the absence of repetitions it is difficult to evaluate this aspect. The initial data analysis can be made to explore outliers and to identify the most important problem samples. However, outliers can be valid data points and should not always be removed.
- When working with cultivated material (such as landraces), the adaptive development of the crop trait might be more dominantly explained by the breeding decisions made by the farmer. For more modern cultivated material there is no appropriate location of origin, as the breeding lines are often the complex result of crossing between genetic resources from very many different source locations. For this problem FIGS strategy may not be the most appropriate approach.

## Future Work

The predictive performance from other different classification methods should be explored. In this study we found significant variation between the four different classification methods we used. The artificial neural networks (Bishop, 1996) is one particular interesting method to explore because the algorithm is so different from the algorithms of the methods used here and also because the failure of the LDA (Fisher, 1936) classifier indicates that the classification problem here is not typical for a parametric solution. Another classifier that could be explored is decision tree methods such as the random forest algorithm (Breiman, 2001; Stockwell, 2007). With multiple different classification methods the so-called ensemble classifier method (Kuncheva, 2004; Rokach, 2010) could be used to combine the predictive information from each classifier. The classifier ensemble will often provide a higher predictive performance than even the best individual classifier. The performance of the classifier ensemble is based on the assumption that each classifier describes the dataset independently and in a different way from the other classifiers.

A significant amount of work still needs to be done to identify or create environmental parameters that are more tightly linked to the evolution of traits so that the predictive power of models can be improved.

Another obvious use case would be to apply the FIGS approach to analyzing gaps in genebank collections. The eco-geographic signature for a particular crop trait can be applied to identify likely locations with specific genetic diversity not yet represented in the collection. Focused identification of germplasm strategy could thus guide new collecting expeditions to interesting new locations based on particular target crop traits. This can be compared to similar gap analysis studies with species distribution models in which the purpose is to complete the genebank collection with overall genetic diversity not yet represented in the collection (Jarvis et al., 2003, 2005, Upadhyaya et al., 2009; Ramírez-Villegas et al., 2010). This use case for the FIGS strategy can be seen as a natural extension of the ecological niche modeling methods to estimate species' distributions.

## CONCLUSIONS

This study contributes to the development of methods for identifying FIGS subsets of geo-referenced genebank accessions more likely to contain sought-after novel genetic variation for adaptive traits. The objective of the FIGS strategy is to more efficiently identify and utilize plant genetic resources, particularly the landrace and wild relatives of crop plants. The results support the assertion that trait mining using the FIGS approach can significantly improve the hit rate for identification of landrace samples with resistance to target crop pests. Focused identification of germplasm strategy subset selection is proposed as an alternative approach to the selection of a core collection to assess rare and useful traits such as resistance to diseases, pests, and abiotic constraints.

## AUTHOR CONTRIBUTIONS

All authors helped to assemble the data and to develop the experimental design for the modeling studies. Dag Terje Filip Endresen made the data analysis and wrote the first version of the manuscript. All authors contributed to the final manuscript.

### Supplemental Information Available

Supplemental material is available free of charge at http://www.crops.org/publications/cs. Supplemental File S1 includes the stem rust dataset. Supplemental S2 includes the net blotch dataset. Supplemental Files S1 and S2 include passport data and eco-climatic layers provided as tab-delimited text and as a spreadsheet.

Supplemental Files S1 and S2 also include a descriptor list to describe the data columns of the datasets. Supplemental Table S3 provides more details on the performance indicators and also additional experimental results not reported in the manuscript.

### Acknowledgments

Dr. Harold E. Bockelman, head of the USDA ARS National Small Grain Collection in Aberdeen, ID, extracted the stem rust and net blotch dataset from the USDA NPGS GRIN database. Associate professor Dvora-Laiô Wulfsohn (Copenhagen University) provided feedback and suggestions to the draft manuscript. Dr Axel Diederichsen and other colleagues at NordGen provided challenging feedback and discussions on the research topics of this manuscript. Many thanks also to the late Dr Bent Skovmand (NordGen) for help and advice when this research project was started. This research project is supported by a grant from the Nordic Genetic Resources Center (NordGen; www.nordgen.org).

### References

Afanasenko, O.S., H. Hartleb, N.N. Guseva, V. Minarikova, and M. Janosheva. 1995. A set of differentials to characterize populations of *Pyrenophora teres* Drechs. for international use. J. Phytopathol. 143(8):501–507. doi:10.1111/j.1439-0434.1995.tb04562.x

Agresti, A. 2002. Categorical data analysis. Second ed. John Wiley & Sons, Hoboken, NJ.

Altman, D.G., and J.M. Bland. 1994a. Statistical notes: Diagnostic tests 1: Sensitivity and specificity. BMJ 308(6943):1552.

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

Barker, M., and W. Rayens. 2003. Partial least squares for classification. J. Chemometr. 17(3):166–173. doi:10.1002/cem.785

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. Proc. Natl. Acad. Sci. USA 106:9519–9524. doi:10.1073/pnas.0904152106

Bhullar, N.K., Z. Zhang, T. Wicker, and B. Keller. 2010. Wheat gene bank accessions as a source of new alleles of the powdery mildew resistance gene *Pm3*: A large scale allele mining project. BMC Plant Biol. 10:88. doi:10.1186/1471-2229-10-88

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

Bonman, J.M., H.E. Bockelman, L.F. Jackson, and B.J. Steffenson. 2005. Disease and insect resistance in cultivated barley accessions from the USDA national small grains collection. Crop Sci. 45:1271–1280. doi:10.2135/cropsci2004.0546

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

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

Brown, A.H.D., and C. Spillane. 1999. Implementing core collections principles, procedures, progress, problems and promise. p. 1–9. *In* R.C. Johnson and T. Hodgkin (ed.) Core collections for today and tomorrow. International Plant Genetic Resources Institute, Rome.

Busby, J.R. 1991. BIOCLIM – A bioclimatic analysis and prediction system. p. 64–68. *In* C.R. Margules and M.P. Austin (ed.) Nature conservation: Cost effective biological surveys and data analysis. CSIRO, Canberra, Australia.

Cardillo, G. 2007. Cohen's kappa: Compute the Cohen's kappa ratio. Available at http://www.mathworks.com/matlabcentral/fileexchange/15365 (MATLAB script, downloaded on 27 July 2010, verified 31 May 2011). MathWorks, Natick, MA.

Centre for Evidence-Based Medicine (CEBM). 2010. Statistics calculator. Available at http://ktclearinghouse.ca/cebm/practise/ca/calculators/statscalc (verified 31 May 2011). Center for Evidence-Based Medicine, University Health Network, Toronto, ON, Canada.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20:37–46. doi:10.1177/001316446002000104

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

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 31 May 2011). ICARDA GIS Unit, Aleppo, Syria.

Duda, R.O., P.E. Hart, and D.G. Stork. 2001. Pattern classification, 2nd ed. Wiley, Hoboken, NJ.

Dwivedi, S.L., J.H. Crouch, D.J. Mackill, Y. Xu, M.W. Blair, M.

Ragot, H.D. Upadhyaya, and R. Ortiz. 2007. The molecularization of public sector crop breeding: Progress, problems, and prospects. Adv. Agron. 95:163–318. doi:10.1016/S0065-2113(07)95003-8

Eigenvector. 2010. PLS toolbox 5.8 (R5.8.3). Available at http://software.eigenvector.com/toolbox (verified 13 June 2011). 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. 2010. 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. 130:96–97. 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. 50:2418–2430. doi:10.2135/cropsci2010.03.0174

FAO. 1974. FAO-UNESCO soil map of the world. Vol. I: Legend. UNESCO, Paris, France.

FAO. 2007. Digital soil map of the world. Version 3.6. Available at http://www.fao.org/geonetwork/srv/en/metadata.show?id=14116 (verified 31 May 2011). FAO, Rome, Italy.

FAO. 2010. The second report on the state of the world's plant genetic resources for food and agriculture. FAO, Rome, Italy.

FAO-UNESCO. 1995. The digital soil map of the world and derived soil properties [CD-ROM]. Land and water digital media series 1. FAO, Rome, Italy.

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7:179–188. doi:10.1111/j.1469-1809.1936.tb02137.x

Flood, J. 2010. The importance of plant health to food security. Food Security 2:215–231. doi:10.1007/s12571-010-0072-5

Frankel, O. 1984. Genetic perspectives of germplasm conservation. p. 161–170. *In* W. Arber, K. Illmensee, W.J. Peacock, and P. Starlinger (ed.) Genetic manipulation: Impact on man and society. Cambridge Univ. Press, Cambridge, UK.

Franklin, J. 2010. Mapping species distributions. Spatial inference and prediction. Cambridge Univ. Press, Cambridge, UK.

Gepts, P. 2006. Plant genetic resources conservation and utilization: The accomplishments and future of a societal insurance policy. Crop Sci. 46:2278–2292. doi:10.2135/cropsci2006.03.0169gas

Gwet, K.L. 2010. Handbook of inter-rater reliability (second edition), The definitive guide to measuring the extent of agreement among multiple raters. Advanced Analytics, LLC, Gaithersburg, MD.

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

Hawkins, D.M., S.C. Basak, and D. Mills. 2003. Assessing model fit by cross-validation. J. Chem. Inf. Comput. Sci. 43:579–586. doi:10.1021/ci025626i

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones, and A. Jarvis. 2005a. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25:1965–1978. doi:10.1002/joc.1276

Hijmans, R.J., L. Guarino, M. Cruz, and E. Rojas. 2001. Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS. Plant Genet. Resour. Newsl. 127:15–19.

Hijmans, R.J., L. Guarino, A. Jarvis, R. O'Brien, P. Mathur, E. Rojas, M. Cruz, and I. Barrantes. 2005b. DIVA-GIS 5.2. Available at http://www.diva-gis.org/download (verified 21 Dec. 2010). Bioversity International, Rome, Italy; CIP, Lima, Peru; IRRI, Los Baños, Philippines; and the Museum of Vertebrate Zoology, University of California, Berkeley, CA.

Hijmans, R.J., M. Jacobs, J.B. Bamberg, and D.M. Spooner. 2003. Frost tolerance in wild potato species: Unraveling the predictivity of taxonomic, geographic, and ecological factors. Euphytica 130:47–59. doi:10.1023/A:1022344327669

Hoffmann, G.M., and H. Schmutterer. 1983. Parasitäre Krankheiten und Schädlinge an landwirtschaftlichen Kulturpflanzen. (In German.) Eugen Ulmer GmbH & Co., Stuttgart, Germany.

Holbrook, C.C., and W. Dong. 2005. Development and evaluation of a mini core collection for the U.S. peanut germplasm collection. Crop Sci. 45:1540–1544.

Hutchinson, M.F. 1995. Interpolating mean rainfall using thin plate smoothing splines. Int. J. Geogr. Inform. Syst. 9:385–403. doi:10.1080/02693799508902045

Hutchinson, M.F. 2000. ANUSPLIN version 4.1. User guide. Center for Resource and Environmental Studies, Australian National University, Canberra, Australia.

Jalli, M. 2010a. The virulence of Finnish *Pyrenophora teres* f. *teres* isolates and its implications for resistance breeding. Ph.D. diss. Available at http://www.mtt.fi/mtttiede/pdf/mtttiede9.pdf (verified 31 May 2011). MTT Agrifood Research Finland, Jokioinen, Finland.

Jalli, M. 2010b. Sexual reproduction and soil tillage effects on virulence of *Pyrenophora teres* in Finland. Ann. Appl. Biol. 158:95–105. doi:10.1111/j.1744–7348.2010.00445.x

Jarvis, A., M.E. Ferguson, D.E. Williams, L. Guarino, P.G. Jones, H.T. Stalker, J.F.M. Vallis, R.N. Pittman, C.E. Simpson, and P. Bramel. 2003. Biogeography of wild Arachis: Assessing conservation status and setting future priorities. Crop Sci. 43:1100–1108. doi:10.2135/cropsci2003.1100

Jarvis, A., K. Williams, D. Williams, L. Guarino, P.J. Caballero, and G. Mottram. 2005. Use of GIS for optimizing a collecting mission for a rare wild pepper (*Capsicum flexuosum* Sendtn.) in Paraguay. Genet. Resour. Crop Evol. 52:671–682. doi:10.1007/s10722-003-6020-x

Koerner, B.I. 2010. Red menace: Stop the Ug99 fungus before its spores bring starvation. Available at http://www.wired.com/magazine/2010/02/ff_ug99_fungus/all/1 (verified 31 May 2011). WIRED Magazine, March 2010.

Kohavi, R., and F. Provost. 1998. Glossary of terms. Mach. Learn. 30:271–274. doi:10.1023/A:1017181826899

Krupinsky, J.M., K.L. Bailey, M.P. McMullen, B.D. Gossen, and T.K. Turkington. 2002. Managing plant disease risk in diversified cropping systems. Agron. J. 94:198–209. doi:10.2134/agronj2002.0198

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

Kurlovich, B.S., S.I. Rep'ev, M.-V. Petrova, T.V. Buravtseva, L.T. Kartuzova, and T.A. Voluzneva. 2000. The significance of Vavilov's scientific expeditions and ideas for development and use of legume genetic resources. Plant Genet. Newsl. 124:23–32.

Landis, J.R., and G.G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33:159–174.

Liu, Z.H., and T.L. Friesen. 2010. Identification of *Pyrenophora teres* f. *maculata*, Causal agent of spot type net blotch of barley in North Dakota. Plant Dis. 94:480–480. doi:10.1094/PDIS-94-4-0480A

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. and the 11th Wheat Breeders' Assembly, Canberra, ACT, Australia. 21–24 Sept. 2004. Cereal Chemistry Division, Royal Australian Chemical Institute (RACI), Melbourne, Victoria, Australia

MathWorks. 2009. MATLAB & Simulink student version release 2009a – Mac. Available at http://www.mathworks.com/products/matlab (verified 13 June 2011). MathWorks, Natick, MA.

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

McLean, M.S., B.J. Howlett, and G.J. Hollaway. 2009. Epidemiology and control of spot form of net blotch (*Pyrenophora teres* f. *maculata*) of barley: A review. Crop Pasture Sci. 60:303–315. doi:10.1071/CP08173

Negri, V., N. Maxted, and M. Veteläinen. 2009. European landrace conservation: An introduction. p. 1–22. *In* M. Veteläinen, V. Negri, and N. Maxted. European landraces on-farm conservation, management and use. Bioversity Technical Bulletin No. 15. Bioversity International, Rome, Italy.

Paillard, S., I. Goldringer, J. Enjalbert, M. Trottet, J. David, C. de Vallavieille-Pope, and P. Brabant. 2000. Evolution of resistance against powdery mildew in winter wheat populations conducted under dynamic management- II: Adult plant resistance. Theor. Appl. Genet. 101:457–462. doi:10.1007/s001220051503

Peeters, J.P., H.G. Wilkes, and N.W. Galway. 1990. The use of ecogeographical data in the exploitation of variance from gene banks. Theor. Appl. Genet. 80(1):110–112. doi:10.1007/BF00224023

Pessoa-Filho, M., P.H.N. Rangel, and M.E. Ferreira. 2010. Extracting samples of high diversity from thematic collections of large gene banks using a genetic-distance based approach. BMC Plant Biol. 10:127. doi:10.1186/1471-2229-10-127

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. 84:203. doi:10.1094/PDIS.2000.84.2.203B

Ramírez-Villegas, J., C. Khoury, A. Jarvis, D.G. Debouck, and L. Guarino. 2010. A gap analysis methodology for collecting crop genepools: A case study with *Phaseolus* beans. PLoS ONE 5(10):e13497. doi:10.1371/journal.pone.0013497

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

Shapiro, S.S., and M.B. Wilk. 1965. An analysis of variance test for normality (complete samples). Biometrika 52(3–4):591–611. doi:10.1093/biomet/52.3-4.591

Silvar, C., A.M. Casas, D. Kopahnke, A. Habekuß, G. Schweizer, M.P. Gracia, J.M. Lasa, F.J. Ciudad, J.L. Molina-Cano, E. Igartua, and F. Ordon. 2010. Screening the Spanish barley core collection for disease resistance. Plant Breed. 129(1):45–52. doi:10.1111/j.1439-0523.2009.01700.x

Singh, P.R., D.P. Hodson, J. Huerta-Espino, Y. Jin, P. Njau, R. Wanyera, S.A. Herrera-Foessel, and R.W. Ward. 2008. Will stem rust destroy the world's wheat crop? Adv. Agron. 98:271–309. doi:10.1016/S0065-2113(08)00205-8

Stakman, E.C. 1915. Relation between *Puccinia graminis* and plants highly resistant to its attack. J. Agric. Res. 4:193–199.

Stakman, E.C., and F.J. Piemeisel. 1917. Problems in preventing plant disease epidemics. Am. J. Bot. 44:259–267.

Steffenson, B.J. 1997. Net blotch. p. 28–31. *In* D.E. Mathre (ed.) Compendium of barley diseases The American Phytopathological Society, St Paul, MN.

Stevens, S.S. 1946. On the theory of scales of measurement. Science 103:677–680. doi:10.1126/science.103.2684.677

Stockwell, D. 2007. Niche modeling: Predictions from statistical distributions. Chapman and Hall/CRC Press, Boca Raton, FL.

Stukenbrock, E.H., and B.A. McDonald. 2008. The origins of plant pathogens in agro-ecosystems. Annu. Rev. Phytopathol. 46:75–100. doi:10.1146/annurev.phyto.010708.154114

Upadhyaya, H.D., K.N. Reddy, M. Irshad Ahmed, and C.L.L. Gowda. 2009. Identification of geographical gaps in the pearl millet germplasm conserved at ICRISAT genebank from West and Central Africa. Plant Genet. Resources: Characterization and Utilization 8(1):45–51. doi:10.1017/S147926210999013X

USDA-ARS. 2010a. Germplasm Resources Information Network (GRIN), National Plant Germplasm System (NPGS). Available at http://www.ars-grin.gov/npgs (last updated 25 Mar. 2010, verified 14 June 2011). USDA-ARS, National Germplasm Resources Laboratory, Beltsville, MD.

USDA-ARS. 2010b. Trait: Net blotch (NETBLOTCH). Available at http://www.ars-grin.gov/cgi-bin/npgs/html/desc.pl?1041 (verified 13 June 2011). USDA-ARS, National Germplasm Resources Laboratory, Beltsville, MD.

USDA-ARS. 2010c. Trait: Stem rust adult (STEMRUSTAD). Available at http://www.ars-grin.gov/cgi-bin/npgs/html/desc.pl?65049 (verified 13 June 2011). USDA-ARS, National Germplasm Resources Laboratory, Beltsville, MD.

van Hintum, Th.J.L., A.H.D. Brown, C. Spillane, and T. Hodgkin. 2000. Core collections of plant genetic resources. IPGRI Technical Bulletin No. 3. International Plant Genetic Resources Institute, Rome, Italy.

Vavilov, N.I. 1920. Zakon gomologicheskikh ryadov v nasledstvennoj izmenchivosti. [The law of homologous series in variation]. (In Russian.) p. 3–20. Proc. III All-Russian Plant Breed. Conf., Saratov, USSR. 4–13 June 1920. Gubpoligrafotdel (Governmental Printing Division), Saratov, USSR.

Vavilov, N.I. 1922. The law of homologous series in variation. J. Genet. 12(1):47–89. doi:10.1007/BF02983073

Vavilov, N.I. 1957. Mirovye resurcy sortov chlebnych zlakov, zernovych bobovych, l'na i ich ispol'zovanie v selekcii. Opyt agroklimatičeskogo obozrenija važnejŝich polevych kultur. [World resources of cereals, grain leguminous crops and flax and their utilization in plant breeding. Agroecological survey of the principal field crops]. (In Russian.) Izdatel'stvo Akademii Nauk SSR, Moskva, Leningrad.

Vavilov, N.I. 1992a. Origin and geography of cultivated plants. Cambridge Univ. Press, Cambridge, UK.

Vavilov, N.I. 1992b. Problems concerning new crops. p. 256–285 *In* V.F. Dorofeyev (ed.) N.I. Vavilov: Origin and geography of cultivated plants. University Press, Cambridge, UK.

Vavilov, N.I. 1992c. The phyto-geographical basis for plant breeding. Studies of the original material used for plant breeding. p. 316–366. *In* V.F. Dorofeyev (ed.) Origin and geography of cultivated plants. Cambridge Univ. Press, Cambridge, UK.

Wold, H. 1966. Estimation of principal components and related models by iterative least squares. p. 391–420. *In* P.R. Krishnaiaah (ed.) Multivariate analysis. Academic Press, New York, NY.

Wold, S. 1976. Pattern recognition by means of disjoint principal

component models. Patt. Recog. 8:127–139. doi:10.1016/0031-3203(76)90014-5

Wold, S., A. Ruhe, H. Wold, and W.J. Dunn. 1984. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comp. 5:735–743. doi:10.1137/0905052

Wold, S., and M. Sjostrom. 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.

WorldClim, 2010. WorldClim – Global climate data. Available at http://www.worldclim.org/current (verified 13 June 2011). Museum of Vertebrate Zoology, University of Califormia, Berkeley, CA.

Wright, S., and B. Gaut. 2005. Molecular population genetics and the search for adaptive evolution in plants. Mol. Biol. Evol. 22:506–519. doi:10.1093/molbev/msi035

Xu, Y. 2010. Plant genetic resources: Management, evaluation and enhancement. p. 151–194. *In* Molecular plant breeding. CAB International, Wallingford, UK.