Journal of Agricultural Science (2014), **152**, 906–916. © Cambridge University Press 2014
doi:10.1017/S0021859613000543

CROPS AND SOILS RESEARCH PAPER

Predicting resistance to stripe (yellow) rust (*Puccinia striiformis*) in wheat genetic resources using focused identification of germplasm strategy

A. BARI[1]*, A. AMRI[1], K. STREET[1], M. MACKAY[2], E. DE PAUW[3], R. SANDERS[3], K. NAZARI[4], B. HUMEID[5], J. KONOPKA[1] AND F. ALO[5]

[1] *International Centre for Agricultural Research in the Dry Areas, BIGM, P. O. Box 6299 Rabat-Instituts, Rabat, Morocco*
[2] *Queensland Alliance for Agricultural and Food Innovation, The University of Queensland, St Lucia Qld 4072, Australia*
[3] *International Centre for Agricultural Research in the Dry Areas, Amman, Jordan*
[4] *International Centre for Agricultural Research in the Dry Areas, Ankara, Turkey*
[5] *International Centre for Agricultural Research in the Dry Areas, Aleppo, Syrian Arab Republic*

(*Received 27 November 2012; revised 5 June 2013; accepted 24 July 2013; first published online 17 April 2014*)

SUMMARY

Stripe rust, caused by *Puccinia striiformis* f. sp. *tritici* (*Pst*), is a major wheat disease that can inflict yield losses of up to 70% on susceptible varieties under favourable environmental conditions. The timely identification of plant genetic resources likely to possess novel resistance to this disease would facilitate the rapid development of resistant wheat varieties. The focused identification of germplasm strategy (FIGS) approach was used to predict stripe rust resistance in a collection of wheat landraces conserved at ICARDA genebank. Long-term climate data for the collection sites, from which these accessions originated and stripe rust evaluation scores for one group of accessions were presented to three different non-linear models to explore the trait × collection site environment interactions. Patterns in the data detected by the models were used to predict stripe rust resistance in a second and different set of accessions. The results of the prediction were then tested against actual evaluation scores of the disease in the field. The study mimics the real scenario where requests are made to plant genetic resources curators to provide accessions that are likely to possess variation for specific traits such as disease resistance.

The models used were able to identify stripe rust-resistant accessions with a high degree of accuracy. Values as high as 0·75 for area under the curve and 0·45 for Kappa statistics, which quantify the agreement between the models' predictions and the curator's disease scores, were achieved. This demonstrates a strong environmental component in the geographic distribution of resistance genes and therefore supports the theoretical basis for FIGS. It is argued that FIGS will improve the rate of gene discovery and efficiency of mining genetic resource collections for adaptive traits by reducing the number of accessions that are normally required for evaluation to identify such variation.

INTRODUCTION

The distribution of aggressive strains of wheat stripe rust, caused by *Puccinia striiformis* f. sp. *tritici* (*Pst*), is causing high yield losses (Hovmøller *et al*. 2008; Milus *et al*. 2009) that threaten food security and the livelihoods of resource-poor farmers around the world. From 2003 onwards, farm losses caused by *Pst* have increased (Duff *et al*. 2006; Nazari *et al*. 2011; Ramdani *et al*. 2011) and it is reported to be one of the most widespread and destructive wheat diseases in areas where cool temperatures normally prevail (Dedryver *et al*. 2009). Furthermore, research has shown that recent outbreaks coincided with the spread of the two closely related aggressive *Pst* pathotypes, which are better adapted to warmer conditions (Milus *et al*. 2009; Walter *et al*. 2011). In a recent review, it was evident that *Pst* remains a significant threat in the majority of wheat-growing regions of the world with potential to inflict regular regional crop losses

* To whom all correspondence should be addressed. Email: a.bari@cgiar.org

https://doi.org/10.1017/S0021859613000543 Published online by Cambridge University Press

ranging from 0·1 to 5%, with rare events giving losses of 5–25% (Wellings 2011). However, yield losses could reach up to 70% when environmental conditions are conducive (Dedryver *et al*. 2009; Li *et al*. 2011), and in varieties where infection occurs early in the season (Beard *et al*. 2005).

The use of resistant varieties was and is still the most effective technique to curb the loss of yield and quality caused by pathogens (Qualset 1975; Evans *et al*. 2008; Duveiller *et al*. 2007). This highlights the need to identify and use novel allelic variation for currently known resistant genes as well as the discovery of new genes that can be utilized in breeding programmes. To find this novel genetic variation will necessitate a search of largely untapped plant genetic resource collections. However, to identify novel and often rare alleles from large genetic resource collections has been likened to searching for a needle in a haystack. To facilitate gene/allele mining an efficient search technique is required that can identify subsets of germplasm that will most probably contain the required trait expressions.

The core collection approach was suggested by Frankel (1984) as a way to better utilize genetic resources and has been widely adopted by the genetic resource community. The core collection, initially constructed on the basis of morphological, geographical and, more recently, molecular diversity, aims to capture as much genetic variation as possible from a source collection in a subset that is 5–10% the size of the source collection. Variations and refinements of the core collection approach include mini-cores and reference sets that further aggregate genetic variation into even smaller subsets. These representative collections aim to concentrate broad genetic variation into smaller subsets that can then be evaluated for specific traits (Upadhyaya *et al*. 2009). However, rare and adaptive alleles, most of which are thought to be functional, may not be captured by a core collection (Brown & Spillane 1999; Polignano *et al*. 2001; Gepts 2006; Dwivedi *et al*. 2008; Pessoa-Filho *et al*. 2010; Xu 2010).

The focused identification of germplasm strategy (FIGS) is a genebank search approach that seeks to maximize the likelihood of capturing the specific adaptive traits in subsets of flexible size drawn from a larger source collection (Mackay 1990, 1995, 2011; Street *et al*. 2008; Glaszmann *et al*. 2010; Kilian & Graner 2012). It exploits evolutionary and ecological drivers by filtering germplasm collections using selection criteria designed to mimic collection site environmental conditions that would impose selection pressures for the emergence of a desired trait. For example, if disease resistance is being sought then the germplasm would be filtered to capture the material from environments that favour high disease pressure which, in turn, creates a selection pressure for the emergence of host resistance. This approach has recently helped researchers to find novel alleles including new sources of resistance to powdery mildew (Bhullar *et al*. 2009, 2010), Sunn pest (El-Bouhssini *et al*. 2009, 2013) and Russian wheat aphid (El-Bouhssini *et al*. 2011, 2013).

In the present study, the FIGS approach was applied to stripe rust on wheat to locate sources of novel resistance in plant genetic resources collections stored at ICARDA's geneBank. The study mimics the scenario where a request is made to a plant genetic resources curator to provide accessions that are likely to have novel resistance to stripe rust. For common diseases, such as stripe rust, a proportion of the collection is likely to have been evaluated. These data can be used to determine a relationship between the geographic incidence of resistance and the environment under which it emerged. This relationship can then be used to select further germplasm subsets for evaluation.

The hypothesis being tested in the present study is that a relationship exists between the geographic distribution of stripe rust resistance and collection site climate descriptors, and thus making it possible to predict which collection sites are most likely to yield resistant genotypes.

MATERIALS AND METHODS

The modelling process

The study was conducted in two steps: first a learning phase trains a predictive model using collection site environmental data and stripe rust scores for a subset of accessions that had been previously evaluated. In effect the process quantifies the relationship, if any exists, between the disease score and a suite of environmental data. In the second phase, the model is supplied with the collection site climatic data of a further set of accessions which has also been evaluated for stripe rust (Yr) resistance. This second set of accessions is different and independent from the dataset used in the first step to build the model. The model then uses the relationships observed in the training phase to classify the second set of accessions as potentially resistant or susceptible. The results of this

https://doi.org/10.1017/S0021859613000543 Published online by Cambridge University Press

Table 1. *Confusion matrix (2 × 2 contingency table)*

		Observed/scored as	
		Resistant	Susceptible
Predicted	Resistant	A	B
	Susceptible	C	D

prediction are then validated by the actual disease evaluation scores.

The models used are based on the paradigm that the value of the stripe rust resistance state, denoted by (Y) as a response variable, for accessions depends on the climatic independent variables (X) of the collecting sites, where $X=(x_1,\ldots,x_n)$ (Bari *et al*. 2012). In the present study, the relationship was defined within a non-linear framework where the model-based learning techniques were used to overcome the problem of restrictive parametric paradigms on the one hand and the prerequisite distribution assumptions on the other (Drake *et al*. 2006). The study details both the development and validation of predictive models for stripe rust resistance using agroclimatic parameters as the predictors.

The modelling framework also included a tuning process for optimal accuracy. The quality of the model was measured using cross-validation algorithms where the data were split into two subsets: the algorithms developed on the training test as well as the validation test were then used to predict for trait states in the second set of material, whose evaluation data were not presented to the models during the prediction phase.

The accuracy of the models (or their predictive power) was measured based on their ability to accurately predict the number of times the fitted model correctly classified the response for each of the two descriptor states (resistant or susceptible). The parameters used to quantify the classification correctness were the area under the curve (AUC) of the receiver operating characteristics (ROC) plots and the Kappa parameters, which were derived from a 2 × 2 contingency table (Table 1) (Swets *et al*. 2000; Fawcett 2006). An AUC value of 0·5 would represent randomness, whereas a value of 1 would represent peak model performance (Fawcett 2006). A value of 0·7 was considered to indicate that the model is performing favourably in terms of its ability to accurately predict a trait state.

The Kappa statistic is used to assess improvement over chance and measures the specific agreement, such as the agreement between the curator's evaluation and the model's prediction. A value below 0·4 is an indication of poor agreement and a value of 0·4 and above is an indication of good agreement (Landis & Koch 1977). Thus, a high value is an indication that the model performance was adequate for prediction purposes and the overall agreement between observed and predicted values (Scott *et al*. 2002; Steyerberg *et al*. 2010). The models, detailed in Table 2, are described below.

Random forest

Random forest (RF) is a clustering algorithm developed to act like an ensemble classifier – methods that generate many classifiers and aggregate their results (Breiman 2001; Liaw & Wiener 2002). In a normal standard tree classifier, the best among all variables is used at each node; however, in RF the best split is randomly selected at each node among subset predictors. The random selection of inputs or combination of inputs at each node to grow each tree improves the overall accuracy of RF model prediction (Breiman 2001).

Two parameters for use of RF: the number of variables in the random subset at each split node (*mtry*) and the number of trees in the forest (*ntree*) (Liaw & Wiener 2002). The optimization of these two parameters is driven by monitoring the magnitude of the mean square prediction error (rate of classification error) observed in the out-of-bag (OOB) set; that is, the ability of each iteration to correctly classify an unknown germplasm accession as resistant or susceptible. Random forests have been used in gene selection and classification of microarray data and genome-wide association studies for complex human diseases (Lunetta *et al*. 2004; Díaz-Uriarte & Alvarez de Andrés 2006).

Neural networks

Neural networks (NN) are commonly used to approximate the complex functional relationships that might exist between input variables, for example climate data (X) and response variables (Y) such as reaction to disease infection; in the present case the reaction of wheat genetic resources accessions to stripe rust pathogen. Neural networks act like general linear models (GLMs) although they do not necessarily require an assumption of linearity between the trait variable (dependent variable) and climate variable

https://doi.org/10.1017/S0021859613000543 Published online by Cambridge University Press

Table 2. *Model-based learning techniques used in the study*

Model	Tuning parameters	Library (R language)	References
Random forests (RF)	Number of trees (*n.tree*) Number of predictors chosen at each node (*mtry*)	RF	(Breiman 2001; Prasad *et al*. 2006; Cutler *et al*. 2007)
Neural networks (NN)	Number of hidden layers (Number of perceptrons): size; Decay value (ε)	nnet	(Venables & Ripley 2002; Kuhn 2008)
Support vector machines (SVM)	gamma(γ), cost(C)	svm (e1071) ksvm (kernalab)	(Karatzoglou *et al*. 2006; Kuhn 2008; Dimitriadou *et al*. 2010)

(independent variables); it is the data that define the functional form of the relationship or the dependency between the trait (Y) and the climate data (X) (Warner & Misra 1996; Günther & Fritsch 2010). Analogous to clusters of neurons in a nervous system, NNs learn from the training data by iteratively adapting their parameters (weights of input data) until an approximation of the functional relationship emerges (Günther & Fritsch 2010).

In the present study, the climate data were fed as inputs to the NN model (*nnet* model R library (Venables & Ripley 2002)). The *nnet* is a feed-forward NN with a single hidden layer where the information is processed from the input nodes, through the hidden layer to the output nodes of the network. Prior to the use of the *nnet* model the tuning process was performed to define the number of neurons (size) and the decay parameter (ε) that measures the trade-offs between the weights (w_k) and the prediction error. The weights keep changing through the back-propagation iterative process until the reduction in error is optimized. The input weights are estimated based on the training set and by the sum-of-squared errors.

Support vector machines

Support vector machines (SVM) is also a learning-based technique that maps input data to a high-dimensional space, and then optimally separates it into respective classes by isolating those inputs, which fall close to the data boundaries (Cortes & Vapnik 1995; Principe *et al*. 2000). The mapping of input data to high-dimensional space is carried out through the processes called kernel functions (equivalent to basis functions) that allow the separation of data.

In the present study, the radial basis function (RBF) was used as the kernel function for the SVM model. It uses Gaussian transfer functions, the centres and widths of which are determined by unsupervised learning rules. The SVM–RBF first assigns a 'score' or a label to the combination of input variables, in the present case, the site climatic variables, during the unsupervised procedure and then the value of the label is matched with the actual Y-value, in the present case the trait value (presence or absence of resistance). The results of this comparison are fed back to the system and adjustments are made to the labels until the error between the predicted and the actual are minimized. A Gaussian RBF kernel was tuned using a 10-fold cross-validation on the training data, with the best value of C (cost). The tuning process also involved performing a grid search of values to adjust the two parameters (γ) of the kernel function (RBF) and the cost (C) values.

Support vector machines are rapid and tend to distinguish optimally between groups and predict entities while minimizing the loss of information (Guo *et al*. 2005; Karatzoglou *et al*. 2006).

Germplasm and disease data

For each model the training dataset consisted of collection site data and stripe rust (Yr) scores following a modified Cobb scale (Peterson *et al*. 1948) for 725 wheat landrace accessions, conserved in the ICARDA geneBank, which had been evaluated for stripe rust resistance on different occasions previous to 2010. The predictive stage was applied to collection site information associated with 2910 accessions also conserved in the ICARDA geneBank. These accessions were planted in the field during the 2010/11 cropping season at Tel Hadya (36°00′N 36°56′E, 284 m a.s.l.), ICARDA and artificially inoculated using the *Yr27* virulent pathotype that was identified during the 2009/10 epidemic in Syria.

Climatic datasets

The climate data were extracted for each site from geographical information system (GIS) climatic surface maps developed at ICARDA (Table 3). The process of generating these surface maps from the meteorological

https://doi.org/10.1017/S0021859613000543 Published online by Cambridge University Press

Table 3. *The climate variables extracted from geographical information system (GIS) surfaces*

Variable name	Variable description	Unit	Number of variables
pet	Monthly potential evapotranspiration	mm	12
ari	Monthly moisture index (ari)	–	12
prec	Monthly precipitation cm	mm	12
tmin	Monthly minimum temperature	°C	12
tmax	Monthly maximum temperature	°C	12

climatic station data were carried out by co-splining (Hutchinson 1995; Hutchinson & Corbett 1995), which is a local interpolation method as implemented in the ANUSPLIN software (Hutchinson 2000). The process converts the climate data from meteorological stations into climate surfaces and the interpolation process included the use of terrain variables as auxiliary variables, whereby these variables were first converted into digital elevation models (DEM), which led to an increase in the precision of the interpolated values (De Pauw *et al*. 2000). The climate data for sites where the accessions were collected were then extracted from these surfaces to generate data sets using the geographical location as a link between the accessions and climate surfaces (Endresen *et al*. 2011, 2012; Bari *et al*. 2012).

RESULTS

The models used were all able to identify stripe rust-resistant accessions with a high degree of accuracy. The ROC plots for all models were well above the diagonal line (Fig. 1*a* to *c*) with a vertical trend towards the left-hand side, indicating high AUC values and that the models classified the resistant accessions more correctly with less false positive error. A straight diagonal line in the graph is expected when a model is no better than random. All the models yielded plots that demonstrate a better than random performance, achieving values as high as 0·75 for AUC.

The Kappa values obtained were also mostly adequate to conclude that there is a trait–environment relationship and overall agreement. For the first set of data (in which the trait scores were made known to the model) both the SVM and NN models attained average scores of 0·4 and 0·41 with upper and lower scores ranging from 0·35 to 0·48. The mean score for the RF model was 0·37 with the scores ranging from 0·3 to 0·44. For the second set of data, in which the models constructed on the first data set were used to predict the trait state, the Kappa scores for all models achieved a mean value of 0·4 and above, with ranges of 0·36–0·46.

The histograms in Fig. 1(*d* to *f*) further illustrate the extent of separation between the two trait states. The values for AUC and Kappa and Fig. 1 all illustrate that the models were able to correctly classify sites that yield either resistant or susceptible genotypes with a success rate ranging from 0·70 to 0·78 in the validation set (Table 4) and 0·76 in the test set (Table 5). In other words, an accession predicted to be resistant based on climate data is likely to be scored as resistant by the curator or evaluator.

Although all models yielded plots that demonstrated a better than random performance, there is a difference in the performance statistics between the three models used. The values generated by the NN model were slightly higher than those of the SVM model, with the RF model yielding the lowest. Despite this, however, as previously mentioned all three models supported the conclusion that the long-term climatic conditions at a collection site influence the emergence of *Pst* resistance due to selection pressure exerted by the pathogen and this relationship or dependency is detectable by models such as those used in the present study. This dependency was used as *a priori* information in turn for the prediction of stripe resistance in independent data with high accuracy and agreement values. These are the first published results where such a comparison was made between the predictions and the actual evaluation or scores for a disease such as stripe rust. The results support the assertion that the approach could be an effective tool to further enhance the discovery of new genes of resistance to biotic stresses.

DISCUSSION

High accuracy and agreement values across the models indicate the presence of strong relationships between the response to *Pst Yr27*-virulent strains and the collection site environments of the accessions tested. That is, the present study confirms that environmental data for accession collection sites can be used to predict for stripe rust (*Yr*) resistance. The strength of the predictions is even higher than in a previous study where FIGS was used to detect the presence of such

https://doi.org/10.1017/S0021859613000543 Published online by Cambridge University Press

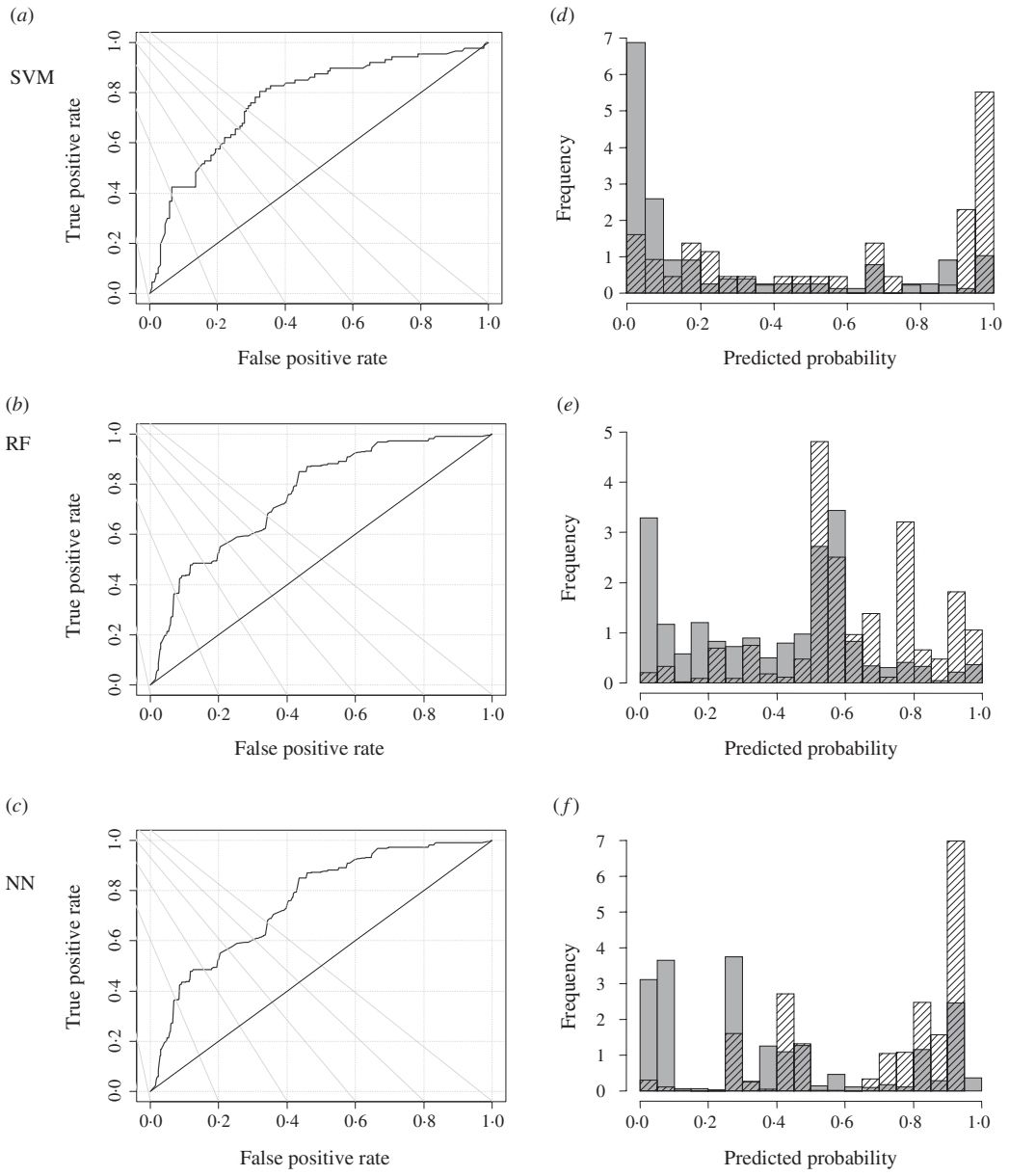


Fig. 1. (*a*–*c*) Receiver-operating characteristic (ROC) plots for the support vector machines (SMV), random forests (RF) and neural networks (NN) models applied to the test set of accessions (evaluation data not made known to the model during the prediction phase). The diagonal line represents the result if the predictions by the models resembled a random classification. The ROC curve to the left of the diagonal plot is the true positive rate *v*. false positive rate. (*d*–*f*) density histograms for prediction of resistance and susceptibility for the SMV, RF and NN models applied to the test set of accessions (evaluation data not made known to the model during the prediction phase); black cross-hatched bars indicates the probability density distribution for resistance and grey solid bars indicates susceptibility.

relationships for response to stem rust infection (Bari *et al*. 2012). Of further significance, unlike previous studies, the accessions used to test the models were not used in their development (so were unknown to the models); however the accuracy of the predictions was higher. This suggests that the *Pst* may have a more specific and narrow set of climate conditions where it thrives than the causal agent of stem rust race Ug99 (*Puccinia graminis* f. sp. *tritici; Pgt*) studied previously (Bari *et al*. 2012). Yellow rust incidence occurs mostly in areas with cool temperatures during the vegetative cycle, whereas stem rust is more prevalent in areas with high temperatures during the period from flowering to maturity (Roelfs *et al*. 1992).

The current study presents further evidence that the FIGS approach can accurately identify accessions

https://doi.org/10.1017/S0021859613000543 Published online by Cambridge University Press

Table 4. *Accuracy parameters for support vector machines (SMV), random forests (RF) and neural networks (NN) self-learning models used to predict stripe rust resistance in the training set of accessions (725). The models were presented with long term climatic descriptors for accession collection sites and full evaluation data for reaction to the stripe rust pathogen. The models arrived at optimal parameter definition by an interactive bootstrapping process that sets tunes of the models by checking how accurately predictions conform to the actual evaluation scores*

Model		AUC	Sensitivity	Specificity	Proportion correct	Kappa
SVM	**mean**	**0·72**	**0·65**	**0·78**	**0·74**	**0·40**
	lower	0·69	0·61	0·74	0·72	0·35
	upper	0·74	0·69	0·82	0·77	0·45
RF	**mean**	**0·70**	**0·64**	**0·76**	**0·73**	**0·37**
	lower	0·67	0·61	0·71	0·69	0·30
	upper	0·73	0·67	0·81	0·76	0·44
NN	**mean**	**0·73**	**0·69**	**0·77**	**0·74**	**0·41**
	lower	0·70	0·58	0·69	0·70	0·35
	upper	0·76	0·79	0·85	0·78	0·48

Table 5. *Accuracy parameters for support vector machines (SMV), random forests (RF) and neural networks (NN) self-learning models used to predict stripe rust resistance in 2910 accessions of wheat (evaluation data not made known to the model during the prediction phase). The models were previously tuned and validated on a stripe rust training data set (Table 4) were presented with long term climatic descriptors for accession collection sites of a wheat subset whose stripe rust evaluation scores were not made known to the model in the prediction phase of the algorithm. The evaluation scores were then used to evaluate the performance of the models predictive ability using Kappa an area under the curve (AUC) parameters*

Model		AUC	Sensitivity	Specificity	Proportion correct	Kappa
SVM	**Mean**	**0·72**	**0·67**	**0·78**	**0·75**	**0·41**
	Lower	0·71	0·64	0·74	0·73	0·36
	Upper	0·74	0·70	0·81	0·76	0·45
RF	**Mean**	**0·71**	**0·63**	**0·80**	**0·75**	**0·40**
	Lower	0·70	0·58	0·77	0·73	0·36
	Upper	0·73	0·67	0·84	0·77	0·45
NN	**Mean**	**0·74**	**0·74**	**0·74**	**0·73**	**0·41**
	Lower	0·72	0·65	0·67	0·69	0·37
	Upper	0·76	0·83	0·81	0·76	0·46

with desirable trait expression and contribute to the proof of concept of the approach. It also confirms the findings of other FIGS studies that used agroclimatic data to predict the presence of resistance to disease or pest organisms (Bhullar *et al*. 2010; El-Bouhssini *et al*. 2011; Endresen *et al*. 2011, 2012; Bari *et al*. 2012). In fact, the results concur with intuitive expectation since the accessions that have been found to possess resistance to stripe rust in the past are often from Ethiopia, Afghanistan, parts of Turkey and Iran in areas where environmental conditions are conducive for the epidemics of the disease and for recurrent stripe rust outbreaks. For example, the stripe rust resistance gene *Yr45* was found in a common spring wheat genotype (PI 181434) from Afghanistan (Li *et al*. 2011).

Geographical patterns of disease resistance have been shown by other studies in wheat (Bonman *et al*. 2007) and for powdery mildew where winter wheat populations with the highest level of resistance were encountered in sites where powdery mildew (*Pm*) pressure was high (Paillard *et al*. 2000). In a more applied study, also demonstrating a geographic pattern to trait distribution, FIGS was used to search for accessions resistant to powdery mildew. It turned out that 0·4 of the collection sites chosen yielded

https://doi.org/10.1017/S0021859613000543 Published online by Cambridge University Press

genotypes resistant to *Pm*, and molecular studies on the *Pm3* locus revealed seven new functional resistance alleles (Bhullar *et al*. 2010).

Use of *a priori* information, based on the FIGS approach, can not only detect the presence of relationships between a given trait and the climatic/environmental parameters under which it evolved, but could also help to assign trait values for accessions not evaluated. Here, FIGS was used to assign scores (predicting resistance or susceptibility) with the aim of constructing subsets containing the desired trait expression at a higher probability than if the material was selected at random, thus reducing the cost of evaluating accessions lacking the desired expression (Koo & Wright 2000). This issue of evaluation has been a major constraint hindering the effective use of plant genetic resources (FAO 2010). Paradoxically though, genetic resource collections have been a source of scores of specific genes with impressive impacts on wheat breeding such as the 'dwarfing' genes, *Rht1* and *Rht2*, found in a wheat landrace from Japan (Hoisington *et al*. 1999). The lack of evaluation data in particular has been highlighted as the single most prevalent and long-standing complaint in the management of plant genetic resources collections vis-à-vis plant improvement (Koo & Wright 2000). As shown in the present study there is, however, the potential to use limited evaluation to improve the effectiveness and efficiency with which plant genetic resources can be used.

The models used in the present work, as applied to genetic resource diversity research, represent a novel approach and deserve particular comment. Firstly, the RF, NN, SVM are perhaps more robust and easier to use since they have fewer assumptions and thus require less data preparation when compared with linear approaches (Drake *et al*. 2006; Tirelli *et al*. 2009). For example, these data do not have to be checked for, and if necessary transformed, to ensure normal distribution. Secondly, the use of these types of recursive partitioning algorithms is gaining momentum in areas where the data are too highly dimensional for standard regression methods such as principal components analysis (PCA), in which the decomposition of variables into reduced components leads to the loss of their individual effects, thus rendering the important variables unidentifiable in the interpretation (Strobl *et al*. 2009).

The FIGS approach addresses the lack of *ex ante* evaluation as well as reducing the cost and time it takes to identify the constraint then identify the required genetic variation and its introgression into appropriate cultivars. This is an issue highlighted by Koo & Wright (2000), who found that the speed with which resistant varieties were developed depended on how quickly novel resistance to diseases could be identified. In a similar study, Gollin *et al*. (2000) used models to examine the efficiency of breeding cultivars resistant to diseases and insects; they concluded that the search for a desirable trait is of equal importance to the process of transferring it into improved backgrounds. Gollin *et al*. (2000) found that the use of a focused search approach, which they called a specialized knowledge case, had a positive impact on crop improvement because it increased the probability of finding the desirable plant material while shortening the time taken to develop an improved variety.

Although the present study and previous ones have shown the potential for FIGS to enhance the discovery of novel sources of variation, there is certainly room for improvement. Although an increase in accuracy of these, or alternative models, to define relationships and dependencies between trait expression and the environmental descriptors used thus far may not be expected, an increase in accuracy can be expected as further environmental descriptors are included, or the environmental data are presented in different formats. For example, the modelling done in the present paper used monthly weather data expressed as 12 separate monthly attributes, for each of maximum and minimum temperatures, precipitation, evapotranspiration and aridity/moisture variables (Table 3), corresponding to the calendar year; January to December. Further studies will experiment with daily data that are arranged and aligned so it is expressed from the onset to the end of growing period for each collection site, thus describing the actual conditions during the growing season of the target crop more accurately. In a similar vein, environmental variables for specific crop development phases could be derived so that conditions across the collections sites at critical stages of development could be more accurately compared or used as selection criteria.

Further, by using a landscape genetics approach, FIGS not only improves the scope for utilizing accessions that lack evaluation data, but also accessions that may possess the desired variation but which lack sections of the passport data in particular the accession geographical location. That is, if the general location of the collection site is known, relationships that might exist between landscape features and genetic variation between and within plant

https://doi.org/10.1017/S0021859613000543 Published online by Cambridge University Press

populations could be explored and used as predictors (Heywood 1991; Podolsky & Holtsford 1995; Sharbel *et al*. 2000; Manel *et al*. 2003; Hakes & Cronin 2011).

CONCLUSION

This research contributes to the proof of concept for the use of the FIGS approach to mine genetic resource collections for adaptive traits and demonstrates an effective modelling process that could be used for discovering novel or new sources of resistance to other pests and diseases.

The successful use of non-linear algorithms demonstrated here suggests wider applications in studies of ecogeographic genetic variation where the relationships between variables are likely to be complex and multi-dimensional.

While the models used were effective at predicting stripe rust resistance, improvements in predictive power of such models are likely as further ecogeographic and co-evolutionary descriptors are included or current variables are redefined to reflect conditions within the crop-growing seasons.

REFERENCES

Bari, A., Street, K., Mackay, M., Endresen, D. T. F., De Pauw, E. & Amri, A. (2012). Focused identification of germplasm strategy (FIGS) detects wheat stem rust resistance linked to environmental variables. *Genetic Resources and Crop Evolution* **59**, 1465–1481.

Beard, C., Loughman, R., Thomas, G. & Jayasena, K. (2005). *Managing Stripe Rust and Leaf Rust in Wheat*. South Perth, Australia: Department of Agriculture and Food. Farmnote 43/2005, replaces 102/01. Available from: http://www.agric.wa.gov.au/PC_92983.html (verified 11 June 2013).

Bhullar, N. K., Street, K., Mackay, M., Yahiaoui, N. & Keller, B. (2009). Unlocking wheat genetic resources for the molecular identification of previously undescribed functional alleles at the *Pm3* resistance locus. *Proceedings of the National Academy of Sciences USA* **106**, 9519–9524.

Bhullar, N. K., Zhang, Z., Wicker, T. & Keller, B. (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 Biology* **10**, 88.

Bonman, J. M., Bockelman, H. E., Jin, Y., Hijmans, R. J. & Gironella, A. (2007). Geographic distribution of stem rust resistance in wheat landraces. *Crop Science* **47**, 1955–1963.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.

Brown, A. H. D. & Spillane, C. (1999). Implementing core collections – principles, procedures, progress, problems and promise. In *Core Collections for Today and Tomorrow* (Eds R. C. Johnson & T. Hodgkin), pp. 1–10. Madison, Wisconsin, USA: Crop Science Society of America.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning* **20**, 273–297.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology* **88**, 2783–2792.

Dedryver, F., Paillard, S., Mallard, S., Robert, O., Trottet, M., Negre, S., Verplancke, G. & Jahier, J. (2009). Characterization of genetic components involved in durable resistance to stripe rust in the bread wheat 'Renan'. *Phytopathology* **99**, 968–973.

De Pauw, E., Göbel, W. & Adam, H. (2000). Agrometeorological aspects of agriculture and forestry in the arid zones. *Agricultural and Forest Meteorology* **103**, 43–58.

Díaz-Uriarte, R. & Alvarez De Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3. DOI: 10.1186/1471-2105-7-3.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. & Weingessel, A. (2010). *R Library (e1071): Misc Functions of the Department of Statistics*. Vienna: The R Foundation for Statistical Computing.

Drake, J. M., Randin, C. & Guisan, A. (2006). Modelling ecological niches with support vector machines. *Journal of Applied Ecology* **43**, 424–432.

Duff, C., Hamblin, P. & Poole, N. (2006). *Stripe Rust Management in Wheat*. Kingston, ACT, Australia: Grains Research & Development Corporation (GRDC).

Duveiller, E., Singh, R. P. & Nicol, J. M. (2007). The challenges of maintaining wheat productivity: pests, diseases, and potential epidemics. *Euphytica* **157**, 417–430.

Dwivedi, S. L., Upadhyaya, H. D., Stalker, H. T., Blair, M. W., Bertioli, D. J., Nielen, S. & Ortiz, R. (2008). Enhancing crop gene pools with beneficial traits using wild relatives. In *Plant Breeding Reviews vol. 30* (Ed. J. Janick), pp. 179–230. Hoboken, NJ, USA: John Wiley & Sons Inc.

El-Bouhssini, M., Street, K., Joubi, A., Ibrahim, Z. & Rihawi, F. (2009). Sources of wheat resistance to Sunn pest, *Eurygaster integriceps* Puton, in Syria. *Genetic Resources and Crop Evolution* **56**, 1065–1069.

El-Bouhssini, M., Street, K., Amri, A., Mackay, M., Ogbonnaya, F. C., Omran, A., Abdalla, O., Baum, M., Dabbous, A. & Rihawi, F. (2011). Sources of resistance in bread wheat to Russian wheat aphid (*Diuraphis noxia*) in Syria identified using the Focused Identification of Germplasm Strategy (FIGS). *Plant Breeding* **130**, 96–97.

El-Bouhssini, M., Ogbonnaya, F. C., Chen, M., Lhaloui, S., Rihawi, F. & Dabbous, A. (2013). Sources of resistance in primary synthetic hexaploid wheat (*Triticum aestivum* L.) to insect pests: Hessian fly, Russian wheat aphid and Sunn pest in the Fertile Crescent. *Genetic Resources and Crop Evolution* **60**, 621–627.

Endresen, D. T. F., Street, K., Mackay, M., Bari, A. & De Pauw, E. (2011). Predictive association between biotic stress traits and ecogeographic data for wheat and barley landraces. *Crop Science* **51**, 2036–2055.

https://doi.org/10.1017/S0021859613000543 Published online by Cambridge University Press

Endresen, D. T. F., Street, K., Mackay, M., Bari, A., Amri, A., De Pauw, E., Nazari, K. & Yahyaoui, A. (2012). Sources of resistance to stem rust (ug99) in bread wheat and durum wheat identified using focused identification of germplasm strategy (FIGS). *Crop Science* **52**, 764–773.

Evans, N., Baierl, A., Semenov, M. A., Gladders, P. & Fitt, B. D. L. (2008). Range and severity of a plant disease increased by global warming. *Journal of the Royal Society: Interface* **5**, 525–531.

FAO (2010). *The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture*. Rome, Italy: FAO.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861–874.

Frankel, O. H. (1984). Genetic perspective of germplasm conservation. In *Genetic Manipulations: Impact on Man and Society* (Eds W. Arber, K. Limensee W. J. Peacock & P. Stralinger), pp. 161–170. Cambridge, UK: Cambridge University Press.

Gepts, P. (2006). Plant genetic resources conservation and utilization: the accomplishments and future of a societal insurance policy. *Crop Science* **46**, 2278–2292.

Glaszmann, J. C., Kilian, B., Upadhyaya, H. D. & Varshney, R. K. (2010). Accessing genetic diversity for crop improvement. *Current Opinion in Plant Biology* **13**, 167–173.

Gollin, D., Smale, M. & Skovmand, B. (2000). Searching an *ex situ* collection of wheat genetic resources. *American Journal of Agricultural Economics* **82**, 812–827.

Günther, F. & Fritsch, S. (2010). Neuralnet: training of neural networks. *The R Journal* **2**, 30–38.

Guo, Q., Kelly, M. & Graham, C. H. (2005). Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling* **182**, 75–90.

Hakes, A. S. & Cronin, J. T. (2011). Environmental heterogeneity and spatiotemporal variability in plant defense traits. *Oikos* **120**, 452–462.

Heywood, J. S. (1991). Spatial analysis of genetic variation in plant populations. *Annual Review of Ecology and Systematics* **22**, 335–355.

Hoisington, D., Khairallah, M., Reeves, T., Ribaut, J. M., Skovmand, B., Taba, S. & Warburton, M. (1999). Plant genetic resources: what can they contribute toward increased crop productivity? *Proceedings of the National Academy of Sciences USA* **96**, 5937–5943.

Hovmøller, M. S., Yahyaoui, A. H., Milus, E. A. & Justesen, A. F. (2008). Rapid global spread of two aggressive strains of a wheat rust fungus. *Molecular Ecology* **17**, 3818–3826.

Hutchinson, M. F. (1995). Interpolating mean rainfall using thin plate smoothing splines. *International Journal of Geographical Information Systems* **9**, 385–403.

Hutchinson, M. F. (2000). *ANUSPLIN version 4.1. User Guide*. Canberra, Australia: Center for Resource and Environmental Studies, Australian National University.

Hutchinson, M. F. & Corbett, J. D. (1995). Spatial interpolation of climatic data using thin plate smoothing splines. In *Co-ordination and Harmonisation of Databases and Software for Agroclimatic Applications* (Ed. FAO), pp. 211–224. FAO Agrometeorology Series no. 13. Rome: FAO.

Karatzoglou, A., Meyer, D. & Hornik, K. (2006). Support vector machines in R. *Journal of Statistical Software* **15**, 1–28. Available from: http://www.jstatsoft.org/v15/i09 (verified 12 June 2013).

Kilian, B. & Graner, A. (2012). NGS technologies for analyzing germplasm diversity in genebanks. *Briefings in Functional Genomics* **11**, 38–50.

Koo, B. & Wright, B. D. (2000). The optimal timing of evaluation of genebank accessions and the effects of biotechnology. *American Journal of Agricultural Economics* **82**, 797–811.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software* **28**, 1–26. Available from: http://www.jstatsoft.org/v28 (verified 12 June 2013).

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.

Li, Q., Chen, X. M., Wang, M. N. & Jing, J. X. (2011). Yr45, a new wheat gene stripe resistance on the long arm of chromosome 3D. *Theoretical and Applied Genetics* **122**, 189–197.

Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. *R News* **2**, 18–22.

Lunetta, K. L., Hayward, L. B., Segal, J. & Van Eerdewegh, P. (2004). Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics* **5**, 32. doi: 10.1186/1471-2156-5-32.

Mackay, M. C. (1990). Strategic planning for effective evaluation of plant germplasm. In *Wheat Genetic Resources: Meeting Diverse Needs* (Eds J. P. Srivastava & A. B. Damania), pp. 21–25. Chichester, UK: John Wiley & Sons.

Mackay, M. C. (1995). One core collection or many? In *Core Collections of Plant Genetic Resources* (Eds T. Hodgkin, A. H. D. Brown, Th. J. L. Van Hintum & E. A. V. Morales), pp. 199–210. Chichester, UK: John Wiley & Sons.

Mackay, M. C. (2011). *Surfing the genepool: the effective and efficient use of plant genetic resources*. Ph.D. Thesis, The Swedish University of Agricultural Sciences, Alnarp, Sweden.

Manel, S., Schwartz, M. K., Luikart, G. & Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology and Evolution* **18**, 189–197.

Milus, E. A., Kristensen, K. & Hovmøller, M. S. (2009). Evidence for increased aggressiveness in a recent widespread strain of *Puccinia striiformis* f. sp. *tritici* causing stripe rust of wheat. *Phytopathology* **99**, 89–94.

Nazari, K., Hodson, D. & Hovmøller, M. (2011). Yellow rust in CWANA in 2010 & 2011: the situation and measures for control. In *Proceedings of the Borlaug Global Rust Initiative 2011 Technical Workshop* (Ed. R. McIntosh), p. 24. Saint Paul, Minnesota, USA: The Borlaug Global Rust Initiative.

Paillard, S., Goldringer, I., Enjalbert, J., Trottet, M., David, J., De Vallavieille-Pope, C. & Brabant, P. (2000). Evolution of resistance against powdery mildew in winter wheat

https://doi.org/10.1017/S0021859613000543 Published online by Cambridge University Press

populations conducted under dynamic management. II. Adult plant resistance. *Theoretical and Applied Genetics* **101**, 457–462.

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

Peterson, R. F., Campbell, A. B. & Hannah, A. E. (1948). A diagramatic scale for estimating rust intensity of leaves and stem of cereals. *Canadian Journal of Research Section C* **26**, 496–500.

Podolsky, R. H. & Holtsford, T. P. (1995). Population structure of morphological traits in *Clarkia dudleyana*. I. Comparison of FST between allozymes and morphological traits. *Genetics* **140**, 733–744.

Polignano, G. B., Uggenti, P. & Scippa, G. (2001). Diversity analysis and core collection formation in Bari faba bean germplasm. *PGR Newsletter: FAO/Bioversity* **125**, 33–38.

Prasad, A. M., Iverson, L. R. & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* **9**, 181–199.

Principe, J. C., Euliano, N. R. & Lefebvre, W. C. (2000). *Neural and Adaptive Systems*: *Fundamentals through Simulations*. New York: John Wiley & Sons.

Qualset, C. O. (1975). Sampling germplasm in a center of diversity: an example of disease resistance in Ethiopian barley. In *Crop Genetic Resources for Today and Tomorrow* (Eds O. H. Frankel & J. G. Hawkes), pp. 81–96. Cambridge, UK: Cambridge University Press.

Ramdani, A., Nazari, C., Hodson, D., Lhaloui, S., Abbad-Andaloussi, F. & Nsarellah, N. (2011). Status of wheat diseases in Morocco during the 2009–10 growing season: yellow rust is becoming a more dangerous disease. In *Proceedings of the Borlaug Global Rust Initiative 2011 Technical Workshop* (Ed. R. McIntosh), p. 155. Saint Paul, Minnesota, USA: The Borlaug Global Rust Initiative.

Roelfs, A. P., Singh, R. P. & Saari, E. E. (1992). *Rust Diseases of Wheat: Concepts and Methods of Disease Management*. Mexico, DF: CIMMYT.

Scott, J. M., Heglund, P. J., Morrison, M. L., Haufler, J. B., Raphael, M. G., Wall, W. A. & Samson, F. B. (2002). *Predicting Species Occurrences: Issues of Accuracy and Scale*. Washington, DC, USA: Island Press.

Sharbel, T. F., Haubold, B. & Mitchell-Olds, T. (2000). Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Molecular Ecology* **9**, 2109–2118.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J. & Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128–138.

Street, K., Mackay, M., Zuev, E., Kaur, N., El Bouhssini, M., Konopka, J. & Mitrofanova, O. (2008). Diving into the genepool: a rational system to access specific traits from large germplasm collections. In *Proceedings of the 11th International Wheat Genetics Symposium* (Eds R. Appels, R. Eastwood, E. Lagudah, P. Langridge, M. Mackay, L. Mcintyre & P. Sharp), pp. 28–31. Brisbane, Australia: Sydney University Press.

Strobl, C., Malley, J. & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* **14**, 323–348.

Swets, J. A., Dawes, R. M. & Monahan, J. (2000). Better decisions through science. *Scientific American* **283**, 82–87.

Tirelli, T., Pozzi, L. & Pessani, D. (2009). Use of different approaches to model presence/absence of *Salmo marmoratus* in Piedmont (Northwestern Italy). *Ecological Informatics* **4**, 234–242.

Upadhyaya, H. D., Pundir, R. P. S., Dwivedi, S. L. & Gowda, C. L. L. (2009). *Mini Core Collections for Efficient Utilization of Plant Genetic Resources in Crop Improvement Programs*. Information Bulletin 78. Patancheru, Andhra Pradesh, India: ICRISAT.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*, 4th edn. New York: Springer.

Walter, S., Fejer-Justesen, A., De Vallavieille-Pope, C. & Hovmøller, M. S. (2011). Global distribution of aggressive wheat yellow rust strains. In *Proceedings of the Borlaug Global Rust Initiative 2011 Technical Workshop* (Ed. R. McIntosh), p. 155. Saint Paul, Minnesota, USA: The Borlaug Global Rust Initiative.

Warner, B. & Misra, M. (1996). Understanding neural networks as statistical tools. *The American Statistician* **50**, 284–293.

Wellings, C. R. (2011). Global status of stripe rust: a review of historical and current threats. *Euphytica* **179**, 129–141.

Xu, Y. (2010). Plant genetic resources: Management, evaluation and enhancement. In *Molecular Plant Breeding* (Ed. Y. Xu), pp. 151–194. Wallingford, UK: CAB International.

https://doi.org/10.1017/S0021859613000543 Published online by Cambridge University Press