Tests of Taxonomic and Biogeographic Predictivity: Resistance to Disease and Insect Pests in Wild Relatives of Cultivated Potato

David M. Spooner,* Shelley H. Jansky, and Reinhard Simon

ABSTRACT

A major justification for taxonomic and biogeographic research is its assumed ability to predict the presence of traits in a group for which the trait has been observed in only a representative subset of the group. Such predictors are regularly used by breeders interested in choosing potential sources of disease and pest resistant germplasm for cultivar improvement, by genebank managers to organize the collection, and by germplasm collectors planning to gather maximum diversity. The present study tests taxonomic and biogeographic associations with 10,738 disease and pest evaluations, derived from the literature and genebank records, of 32 pest and diseases in five classes of organisms (bacteria, fungi, insects, nematodes, and virus). The data show that ratings for only Colorado potato beetle [*Leptinotarsa decemlineata* (Say)] and one pathogen (Potato M Carlavirus) are reliably predicted both by host taxonomy and climatic variables. While it is logical to initially take both taxonomy and geographic origin into account while screening genebank materials for pest and disease resistances, such associations will hold for only for a small subset of resistance traits. Based on these results, a more effective strategy than taxonomic and biogeographic prediction is probably careful screening of core collections.

D.M. Spooner and S.H. Jansky, USDA-ARS, Dep. of Horticulture, Univ. of Wisconsin, 1575 Linden Dr., Madison, WI 53706-1590; R. Simon, International Potato Center, P.O. Box 1558, La Molina, Lima 12, Peru. Received 18 Apr. 2008. *Corresponding author (david.spooner@ars.usda.gov).

Abbreviations: EBN, endosperm balance number; GIS, geographic information systems; RF, RandomForest method to test environmental associations; SVM, Support Vector Machine to test environmental associations.

A WIDELY HELD ASSUMPTION is that taxonomically related organisms, or those found in geographic proximity, are likely to share traits (Warburton, 1967; Daly et al., 2001; Spooner et al., 2003). This concept arises from knowledge that plant populations are not randomly arranged assemblages of genotypes but are structured in space, time, and history, resulting from the combined effects of mutation, migration, selection, and drift (Loveless and Hamrick, 1984). Assumptions about the distribution of heritable traits among populations have important ramifications when planning research. For example, plant breeders sometimes use taxonomy to choose germplasm based on such assumptions as "species X is resistant to a disease Y." As well, genetic similarity among populations is thought to diminish with geographic distance, and thus, germplasm collectors are advised to collect from as many geographically distant sites as possible to maximize genetic variation among samples (Marshall and Brown, 1975; Chapman, 1989).

Taxonomic associations to traits have been well documented (Daly et al., 2001), and there are many examples of associations of genetic distance to geographic distance such as in *Chamaecytisus* (Francisco-Ortega et al., 1993), pistachio (*Pistacia vera* L.) (Hormanza et al., 1994), *Solanum* sect. *Etuberosum* (Spooner et al., 1995), sunflower

Published in Crop Sci. 49:1367–1376 (2009).
doi: 10.2135/cropsci2008.04.0211

677 S. Segoe Rd., Madison, WI 53711 USA

(*Helianthus annuus* L.) (Cronn et al., 1997), lentil (*Lens culinaris* Medik.) (Ferguson et al., 1998), and *Solanum verrucosum* Schlechtd. (del Rio and Bamberg, 2004). However, there are many other examples of no or only very weak associations of genetic distance to geographic distance, as in studies of *Fagus* (Gallois et al., 1998), maize (*Zea mays* L.) (Moeller and Schaal, 1999), *Malus* (Lamboy et al., 1996), Emmer wheat (*Triticum dicoccoides* Koern.) (Fahima et al., 1999), wild potatoes (del Rio et al., 2001; del Rio and Bamberg, 2002; McGregor et al., 2002; Ghislain et al., 2006), and *Arabidopsis* (Jørgensen and R. Mauricio, 2004). As pointed out by del Rio et al. (2001), the low size of potato populations (generally <100 plants per population) may make other stochastic events more important in the partitioning of genetic diversity. These events include environmental changes, demographic factors (i.e., chance differences among individuals in survivorship or fecundity), and genetic drift.

The relationship between disease resistance and geographic distance likewise includes examples of both associations and nonassociations. For example, genotypes sometimes cluster geographically, perhaps driven by a combination of parallel selection and gene flow, as shown in resistances of *Barley yellow dwarf virus* in barley (*Hordeum vulgare* L.) (Qualset, 1975), *Barley yellow dwarf virus* in *Avena fatua* L. (Rines et al., 1980) and in *A. sterilis* L. (Comeau, 1982), European corn borer [*Ostrinia nubilalis* (Hübner)] resistance in maize (Reid et al., 1990), net blotch (caused by *Pyrenophora teres* Drechsler) in various *Hordeum* species (Sato and Takeda, 1997), and root-knot nematode (*Meloidogynes* spp.) in peanut (*Arachis hypogaea* L.) (Holbrook et al., 2000). Comeau (1982) attributed the clustering of resistant phenotypes of *Barley yellow dwarf virus* in *A. sterilis* to similar climates.

Other studies, however, fail to find associations between disease resistance and geographic distance. For instance, Jarosz and Burdon (1991) showed the lack of geographic association to resistance to the rust fungus *Melampsora lini* (Ehrenb.) Lév. in wild flax (*Linum marginale* Planch.), which they attributed to a combination of stochastic forces of annual bottlenecks in population size and periodic extinctions, reducing the chances of the rust to adapt to the host resistance present at any one site. Westman et al. (1999) failed to find geographical or taxonomic associations of *Alternaria brassicicola* (Schwein.) Wiltshire and *Xanthomonas campestris* pv. *campestris* (Pammel) Dowson to a range of *Brassica* species. In wild potato, Jansky et al. (2006) failed to find geographic associations to white mold resistance [caused by *Sclerotinia sclerotiorum* (Lib.) de Bary] (2006) and Jansky et al. (2008) to early blight resistance [caused by *Alternaria solan*i (Ellis and G. Martin) L.R. Jones and Grout].

The purpose of this and our other studies (Jansky et al., 2006, 2008) is to (i) summarize prior screening data in potato to aid potato breeders in choice of germplasm, and (ii) to use these data for broader-scale tests of the taxonomic and biogeographic associations to disease and pest resistance. We chose a disease test because the wild potato species provide a wide range of resistances to diseases, which are readily transferred to the cultivated potato. Disease resistances in wild species provide one of the major justifications for collecting and storing germplasm in genebanks, and species-specific statements about breeding values are common with wild potatoes (Ross, 1986; Hawkes, 1990; Ruiz de Galerreta et al., 1998), as well as other crops, such as *Allium* (Kik, 2002), *Brassica* (Ellis et al., 2000), Cucurbitaceae (Robinson and Decker-Walters, 1997), *Daucus* (Simon, 2000), or *Phaseolus* (Freytag and Debouck, 2002). Our prior studies (Jansky et al., 2006, 2008) designed empirical tests of taxonomic and biogeographic predictors to resistance to the fungal pathogens that cause white mold and early blight and found no consistent associations to species, taxonomic series, molecular-based clades, ploidy, breeding system, geographic distance, or climatic parameters. While these studies were superior in control of geographically well-spaced germplasm and the use of standardized disease tests, they lacked the numbers of accessions and broad range of diseases for broad-scale tests. The present study tests taxonomic and biogeographic associations of 32 pests and diseases in five classes of organisms (bacteria, fungi, insects, nematodes, and virus), with 10,738 disease and pest evaluations, derived from the literature and genebank records.

MATERIALS AND METHODS

Database Construction

Hanneman and Bamberg (1986) summarized accession-specific evaluation data for the U.S. Potato Genebank. These data came from (i) the published literature, (ii) unpublished theses, (iii) unpublished research reports arising from USDA-ARS National Plant Germplasm System Horticultural Evaluation Grants, and (iv) unpublished notes submitted to the genebank. The published references are listed in Hanneman and Bamberg (1986), but there is no linking of accession-specific evaluations to references.

We gathered evaluation data and references from the published and unpublished references listed in References, below. Screening results for a particular trait sometimes indicated intra-accession segregation or different resistance designations based on different methods or locations of screening. We entered all of these data into a Microsoft Access database (Microsoft Corp., Redmond, WA) using the actual scores of the evaluator, and linked them to three meta-data fields: (i) the reference, (ii) a thumbnail summary of the methods, and (iii) a description of the scores. This last field was entered in a standardized way with entries of the evaluation units in which scores are expressed, the number of observations, the number of species tested, and the minimum, maximum, median, and mean of resistance scores. All published and unpublished data sources from 1986 to 2001 are linked to a specific references, allowing breeders to provide crosschecks to the evaluations.

We combined the following resistance designations contained in Hanneman and Bamberg (1986), and our new data, into three classes (i) resistant (combining highly resistant, hypersensitive, hypersensitive/immune, immune, immune/medial, immune/

resistant, immune/susceptible, immune/tolerant, medial/resistant, resistant, resistant/susceptible, resistant/tolerant, symptomless carrier, tolerant/resistant, very resistant), (ii) medial (medial, medial/susceptible, medial/tolerant, moderately resistant, tolerant, tolerant/susceptible), (iii) susceptible (highly susceptible, moderately susceptible, sensitive, susceptible, susceptible/tolerant, very sensitive). Because of widely varying methods to designate accessions into resistance categories, we did not enter the type of resistance into our database from 1986 to 2001 unless the authors made these determinations, but rather entered "not noted."

The entire database contained 57,167 records. The individual disease resistance data have been entered into the Germplasm Resources Information Network (GRIN) and the database for analysis is deposited at the NRSP-6 Potato Genebank in Sturgeon Bay Wisconsin. For analysis, we shortened this database to 10,738 records by including only observations with resistance designations, from wild (not cultivated) potato species, and only those with reliable georeferenced data. We grouped together pathotypes within *Meloidogyne*, *Erwinia*, and *Verticillium*, *Streptomyces scabies* (Thaxt.) Lambert and Loria, and *S. turgidiscabies* Takeuchi, leaving a final set of resistance scores for 32 pest and diseases (Table 1) in 96 *Solanum* species.

We updated the species names to conform to Spooner and Salas (2006), and grouped the species into four classes: (i) taxonomic series of Hawkes (1990), (ii) clade relationships of Spooner et al. (1993, 2008), Rodríguez and Spooner (1997), Spooner and Castillo (1997), and Castillo and Spooner (1997), (iii) endosperm balance numbers of 1, 2, and 4, which predict crossability among *Solanum* species (Hanneman, 1994), and (iv) breeding system (inbreeding, outbreeding).

Table 1. The traits evaluated for associations to taxonomic and environmental variables in this study

Trait group	Trait	No. of observations	Resistant	Medial	Susceptible
Bacteria	*Clavibacter michiganensis sepedonicus*	46	14	13	19
	Corynebacterium sepedonicum	650	136	36	478
	Erwinia spp.	159	46	33	80
	Pseudomonas solanacearum	343	34	15	294
	Streptomyces scabies and *S. turgidiscabies*	98		18	80
	Sum	1296	230	115	951
Fungi	*Alternaria solani*	522	95	71	356
	Fusarium spp.	71	8	40	23
	Phytophthora infestans	179	60	77	42
	Synchytrium endobioticum	117	87		30
	Verticillium spp.	1321	601	148	572
	Sum	2210	851	336	1023
Insect	*Empoasca fabae*	212	105	72	35
	Leptinotarsa decemlineata	910	454	47	409
	Myzus persicae	967	158	258	551
	Potato aphid	546	151	147	248
	Potato flea beetle	218	127	40	51
	Tetranychus spp.	30	11		19
	Sum	2883	1006	564	1313
Nematode	*Diplotheca rhizophila*	77	64	10	3
	Globodera pallida	357	182		175
	Meloidogyne spp.	214	145	8	61
	Sum	648	391	18	239
Virus	Henbane Mosaic Potyvirus	25	9		16
	Potato A Potyvirus	52	19		33
	Potato Leaf Roll Luteovirus	2658	392	62	2204
	Potato M Carlavirus	99	33	16	50
	Potato S Carlavirus	87	22		65
	Potato Spindle Tuber Viroid	40	5	12	23
	Potato X Potexvirus	194	33		161
	Potato X Potexvirus cp	21	3		18
	Potato Y Potyvirus	186	74		112
	Potato Y Potyvirus n	28	10		18
	Potato Y Potyvirus ntn	260	129		131
	Potato Y Potyvirus o	23	6		17
	Tobacco Rattle Virus	28	20		8
Virus	Sum	3701	755	90	2856
Grand total		10,738	3233	1123	6382

Analytical Methods

For the analysis of pest and disease associations with taxonomic classifications, we cross- tabulated the data and tested for associations of frequency of each rating (susceptible, medial, resistant) with each taxonomic class using a chi-square cut-off value of 5% to reject the null hypotheses of no association of frequency counts in the categories. The chi-square test was not adjusted with conservative multiple test adjustment methods for three reasons: (i) from previous studies as in Jansky et al. (2006, 2008) the expectation to find associations was low; (ii) the variability of the data could not be controlled as strictly as in experimental work since it is based on literature review; and (iii) the objective of this study is to complement previous work and generate hypotheses for further follow-up. The strength of any significant associations was tested with the lambda value, which ranged from 0 to 1 (Michael and Lewis-Beck, 1995). A lambda value of 0.5 means the prediction error was reduced by 50% for that independent or predictive variable. The lambda value can be compared across tabulations and is comparable to a correlation coefficient. We report the results for all significant combinations of the response or dependent variable (RATING) for a certain disease with one of the taxonomic variables (species, series, clade, ploidy, endosperm balance number [EBN], and breeding system) in Table 2. We

Table 2. Overview of strength of associations as measured by the lambda statistic† between pest and disease ratings with taxonomic variables.

Trait group	Trait	Species	Series	Clade	Ploidy	EBN‡	Breeding system
Bacteria	*Clavibacter michiganensis sepedonicus*	0.56	0.41	0.19	0.11	NS	NS
	Corynebacterium sepedonicum	0.07	0.00	0.00	0.00	0.00	NS
	Erwinia spp.	0.46	0.13	NS	NS	NS	NS
	Pseudomonas solanacearum	0.16	0.10	0.02	0.00	0.00	0.00
	Streptomyces scabies and *S. turgidiscabies*	0.22	0.00	NS	NS	NS	NS
Fungi	*Alternaria solani*	0.20	0.10	0.08	0.00	0.00	0.00
	Fusarium spp.	NA	NS	NS	NS	NS	NS
	Phytophthora infestans	*0.60*	0.31	0.06	0.06	0.06	0.20
	Synchytrium endobioticum	NA	NS	0.03	NS	NS	NS
	Verticillium spp.	0.28	0.18	0.12	0.09	0.07	0.13
Insects	*Empoasca fabae*	0.32	0.10	NS	0.00	0.05	0.00
	Leptinotarsa decemlineata	*0.60*	0.27	0.10	0.13	0.05	0.14
	Myzus persicae	0.34	0.13	0.05	0.00	0.03	0.00
	Potato aphid	0.32	0.14	0.04	0.02	0.01	0.02
	Potato flea beetle	0.33	0.08	0.02	NS	NS	NS
	Tetranychus spp.	NA	NS	NS	0.00	NS	NS
Nematodes	*Diplotheca rhizophila*	NA	NS	NS	NS	NS	NS
	Globodera pallida	0.57	0.53	0.11	0.42	0.37	0.46
	Meloidogyne spp.	NA	NS	0.06	NS	NS	NS
Virus	Henbane Mosaic Potyvirus	NA	NS	NS	0.00	NS	0.00
	Potato A Potyvirus	NA	0.53	0.26	0.26	NS	0.06
	Potato Leaf Roll Luteovirus	0.01	0.00	0.00	0.00	0.00	0.00
	Potato M Carlavirus	*0.67*	0.22	NS	NS	NS	NS
	Potato S Carlavirus	*0.68*	0.36	0.10	0.00	NS	NS
	Potato Spindle Tuber Viroid	NA	NS	NS	NS	NS	NS
	Potato X Potexvirus	0.42	0.18	0.15	0.00	NS	NS
	Potato X Potexvirus cp	NA	NS	NS	NS	NS	NS
	Potato Y Potyvirus	NA	0.20	0.09	0.09	0.08	0.03
	Potato Y Potyvirus n	NA	NS	NS	NS	NS	NS
	Potato Y Potyvirus ntn	0.58	0.55	0.55	0.47	0.46	0.03
	Potato Y Potyvirus o	NA	*0.67*	NS	NS	0.33	NS
	Tobacco Rattle Virus	NA	NS	NS	0.25	0.25	0.00

†For each association the reported lambda values were calculated independently. Strong associations (>0.6) are highlighted in italic. Nonsignificant associations (5% level) are reported as NS.

‡EBN, endosperm balance number.

also combined data of all variables into five groups of organisms (bacteria, fungi, insects, nematodes, and viruses) and tested associations of each group with taxonomic variables using the same scheme. We considered values above 0.8 as very strong, between 0.6 and 0.8 as strong, between 0.4 and 0.6 as moderate, between 0.2 and 0.4 as weak, and between 0 and 0.2 as very weak.

Greene et al. (1999a, 1999b) demonstrated the utility of geographic information systems (GIS) analysis, combining physical distance and environmental data, in exploring sampling protocols. We tested associations of disease and pest resistances to climatic variables following our same methodologies in Jansky et al. (2006, 2008). We used the geo-references of the accessions to extract altitude and climatic variables that are monthly averages for minimum and maximum temperatures (TMIN1–12, TMAX1–12) and rainfall (RAIN1–12). The different seasonality of the data north and south of the equator was corrected by replacing values for January with values from July, for February with values from August, and so forth for the northern hemisphere.

We used two statistical approaches to test environmental associations: (i) the Support Vector Machine (SVM; Guo et al., 2005) approach and (ii) the RandomForest method (RF; Thuiller et al., 2005). The SVM approach is a flexible method that finds an optimal separation between data in hyperspace. It can work with continuous and nominal response variables and can incorporate relationships (colinearities) between the predictor variables. Thus, it is not necessary to account for spatial autocorrelation. The RF approach is based on classification and regression trees. Instead of one decision tree, it uses a majority of decision trees to classify each object, and to make the final classification. Whereas SVM uses separability, RF uses proximity between objects. To assess the potential of predictivity for both SVM and RF, we divided the dataset in two thirds for

Table 3. Summary of associations of pest and disease ratings with geographic and climatic variables as measured using predictive statistical methods (Support Vector Machines and RandomForest). Only traits for which resistance could be predicted with 80% accuracy in both methods are reported. The table also reports the top five important independent variables used for prediction.

Method†	Trait	*Clavibacter michiganensis sepedonicus*	*Leptinotarsa decemlineata*	*Meloidogyne* spp.	*Phytophthora infestans*	Potato flea beetle	Potato M Carlavirus
	n	46	910	214	179	218	99
RF	accuracy	0.93	0.87	0.83	0.87	0.83	0.95
	kappa	0.90	0.77	0.58	0.80	0.71	0.92
	R%	1.00	0.90	0.97	0.85	0.87	0.94
	Var 1	*TMIN9*§	LAT	*TMIN6*	PREC6	PREC1	*TMAX11*
	Var 2	ALT	PREC12	PREC4	PREC9	PREC10	LAT
	Var 3	LAT	PREC10	*TMIN12*	LAT	*TMAX7*	*TMIN10*
	Var 4	*TMAX7*	PREC1	*TMAX11*	PREC12	PREC7	PREC11
	Var 5	*TMAX6*	*TMIN1*	*TMIN2*	PREC7	LAT	*TMAX9*
SVM							
	svm.cost	4	1	1	2	1	4
	svm.gamma	0.25	1	1	1	0.125	1
	accuracy	0.89	0.82	0.82	0.85	0.69	0.98
	kappa	0.84	0.67	0.55	0.77	0.37	0.94
	R%	0.86	0.80	0.97	0.90	0.94	0.91
	Var 1	*TMIN6*	PREC10	*TMIN12*	PREC7	*TMAX6*	ALT
	Var 2	PREC7	*TMIN9*	PREC12	*TMAX7*	*TMIN2*	PREC2
	Var 3	PREC3	*TMIN7*	*TMIN3*	ALT	*TMIN11*	PREC11
	Var 4	*TMIN8*	*TMIN11*	*TMAX7*	*TMIN2*	*TMAX2*	*TMIN11*
	Var 5	*TMAX5*	*TMAX2*	PREC9	*TMAX10*	PREC9	*TMAX1*

†RF, RandomForest method; SVM, Support Vector Machine.

§Italic type indicates temperature-related variables (e.g., TMIN9 = average minimum monthly temperature in September). Underlining indicates rainfall (precipitation) related variables (e.g., PREC12 = average monthly precipitation in December). TMIN1 to TMIN12 are average monthly minimum temperatures; TMAX1 to TMAX12 are average monthly maximum temperatures; and PREC1 to PREC12 are average monthly precipitation data. LAT, latitude; ALT, altitude; Accuracy, average percent of correctly predicted ratings from test set using a 10-fold cross-validation; Kappa, the average accuracy value corrected for chance; R%, the correctly predicted resistant ratings based on using the whole dataset; Var1 to Var5, the top five most important variables in order as determined by a leave-one-out approach; Svm.cost, a standard parameter for the SVM after optimization; Svm.gamma, a standard parameter for the SVM after optimization.

training and one third for testing with 10-fold cross-validation. We further considered only those pest and disease predictions that were at least predictable at the 80% accuracy level (following the same classification scheme as described above for the lambda value) on the whole dataset for each pest or disease and were among the top 10 results by both methods.

Cohen's kappa (Cohen, 1960) values were calculated to assess the agreement of predicted ratings versus observed ratings for both SVM and RF in the cross-validation step. The kappa value measures the agreement between two raters and corrects for agreement by chance. Kappa is a dimensionless value varying between 0 and 1. The interpretation of the kappa value is similar to the lambda value, which is 0.8 to 1.0, and corresponds to almost perfect agreement whereas values below 0 indicate no agreement, 0.0 to 0.2 slight agreement, etc.

All calculations were conducted using the statistical software R (R Development Core Team, 2008) including several of its add-on packages. Maps were constructed using the packages maps (Brownrigg et al., 2007) and maptools (Lewin-Koh et al., 2008). The RandomForest method was used as implemented in the package RandomForest (Liaw et al., 2008); the method SVM was used as implemented in the package e1071 (Dimitriadou et al., 2008). The kappa value was calculated as implemented in package e1071.

RESULTS

Taxonomic Results

Analyses of association of the three disease resistance ratings (resistant, medial, and susceptible) within the five trait group classes (bacteria, fungi, insects, nematodes, and viruses) and the five taxonomic variables (species, series, clade, EBN, and breeding system) did not yield any strong associations (the strongest lambda value was 0.4 for the association of nematode pest ratings with species), and we did not explore taxonomic associations to trait groups further. Analyses based on the three disease resistance ratings with the 32 individual traits and the five taxonomic classifications are presented in Table 2. Strong associations were found for one fungal, one insect, and three viral diseases. Four out of the strong (see above) associations {*Phytophthora infestans* (Mont.) de Bary, Colorado potato beetle [*Leptinotarsa decemlineata* (Say)], Potato M and S Carlavirus} were found with potato species and one (Potato Y Potyvirus O) with taxonomic series. For *P. infestans* the three species with most resistant ratings were *S. verrucosum* Schltdl., *S. microdontum* Bitter, and *S. pinnatisectum* Dunal. For *L. decemlineata* the three species with most resistant

ratings were *S. berthaultii* Hawkes, *S. chacoense* Bitter, and *S. tarijense* Hawkes. For Potato M Carlavirus the three species with most resistant ratings were *S. palustre* Poepp., *S. leptophyes* Bitter, and *S. berthaultii*. For Potato S Carlavirus the one species with most resistant ratings was *S. palustre*. For Potato Y Potyvirus the one series with most resistant ratings was the outgroup of potato, section *Etuberosum*.

Geographic and Climatic Results

The summary results of RF and SVM analyses of environmental associations to disease resistance are shown in Table 3 including only those traits for which more than 80% accuracy of predicting resistant accessions was achieved by both methods. The results showed that the RF method was slightly more accurate, on average, in predicting resistant accessions, as estimated by the kappa value. Thus, predictive maps for distribution of resistant, susceptible, and intermediate ratings (the latter only where applicable, since some pest or diseases had no intermediate ratings) were produced using the method that overall produced the better results and using the whole dataset to calibrate the model for each pest or disease of the final set. We analyzed associations between the rating response variable at individual trait level as well as at trait group level.

Results for both RF and SVM methods are reported in Table 3. Out of 32 pests and pathogens analyzed for associations with geographic or climatic variables only six could consistently be predicted using the two complementary data mining methods. These pests and pathogens are *Clavibacter michiganensis* ssp. *sepedonicus* (Spieckermann and Kothoff) Davis et al., Colorado potato beetle, *Meloidogyne* spp., *Phytophthora infestans*, potato flea beetle (*Epitrix subcrinata* LeConte), and Potato M Carlavirus. As an example, Fig. 1 shows the predicted maps of probable sites for resistant, intermediate, and susceptible accessions with regard to *C. michiganense sepedonicus*. As Fig. 1 shows, often accessions from all three resistance classes superficially appear to be spatially very close, a situation similar for the other pests and diseases. However, the spatial and environmental (temperature and precipitation) data (Table 3) are distinctive enough that they can be used to predict pest and disease ratings, and these can then be used to suggest areas where similar ratings can be found elsewhere. As can be seen from Table 3 on average the RF method gives slightly better predictions (higher %R values). However, the set of most important predictive geographic and climatic variables shows no simple overall pattern. Geographic and rainfall variables are more often used by the RF method while the SVM uses more temperature-related variables for the set of six most reliably predictable pests and diseases. Despite this, 6 out of the top 10 predictable disease ratings are common in both methods. The geographic variable most frequently used by the RF method was latitude for the set of six most reliably predictable pests and diseases while the SVM had no clear preference.

DISCUSSION

Predictive taxonomic and climatic associations were found for only 4 to 6 of 32 pests and diseases. Taxonomic variables predicted the distribution of resistance to *Phytophthora infestans* (causal organism of late blight), Colorado potato beetle, Potato M and S Carlavirus. Climatic variables predicted the distribution of resistance to *Clavibacter michiganensis sepedonicus* (ring rot), Colorado potato beetle, root-knot nematode, *Phytophthora infestans*, potato flea beetle, and Potato M Carlavirus. There was no consistent pattern of any single significant climatic variable associated with resistances to these diseases or pests. The most frequent predictive taxonomic variable was the species, followed by series. Thus, species are the single best predictor among the 38 environmental and five taxonomic variables analyzed. Consequently, it may be appropriate to make species-specific statements about disease or pest resistance traits, as long as it is understood that significant variability exists among accessions within a species and plants within an accession. The geographic variable latitude was the second best predictor of the distribution of disease and pest resistance traits. This might be broadly related to the geographic distribution of pathogens and pests.

The climatic variables show no simple prediction pattern. This is probably because climatic niches must accommodate both host and pathogen preferences, and the combination of plant, pathogen, and environment determines whether disease develops. The data used for GIS analyses are based on means across several years. They do not take into account the year-to-year variation that often influences the development of disease epidemics. A famous example of this phenomenon is the Irish potato famine, which was caused by late blight. Before 1845, late blight was not a major disease in Ireland even though it had probably been in Europe for several years. However, during the cool, wet summer of 1845, the pathogen (*P. infestans*) thrived, leading to the first major crop failure (Burt, 1995). In 1846, the planting of infected tubers, followed by a second cold, wet summer, had disastrous consequences essentially leading to a complete crop failure. Scenarios such as this, in which short periods of extreme weather lead to severe pathogen–pest selection pressures, are likely to occur with *Solanum* species in the wild. However, the mean annual temperature and precipitation data in GIS databases would not detect these events. Similarly, microclimate variation in temperature and rainfall patterns can exert a strong influence on interactions between plants and other biotic agents, but these differences are not reflected in climatic variables at our disposal for analysis. Many of the wild *Solanum* species are found in the Andes, where populations separated by short distances are likely to experience very different climates and consequently different selection pressures.

Agricultural fields cause another type of local selection pressure not detected in our analyses. By studying habitat notes from taxonomists who collected the accessions used in our early blight predictivity study (Jansky et al., 2008), we found a preponderance of resistant accessions collected in or adjacent to crop production fields. Disease pressure is expected to be stronger near agricultural sites, and resistant populations may evolve more quickly there. Unfortunately, detailed collection site information is lacking for most accessions, so it is of limited value for predicting sources of disease resistance genes. Limited data on the localized distribution of pathogens and pests also impedes our ability to identify accessions with resistance traits. Furthermore, pathotypes are likely to vary across geographic regions and may not be the same as those used during disease resistance evaluations in the United States.

General screening data from genebank records, like the data on which this study is based, have been shown to be useful in association mapping studies (Gebhardt et al., 2004), suggesting that such evaluation data are meaningful. Our conclusions stated here need to be evaluated with further tests using accessions spread throughout the geographic range of potato, with comparable evaluation data using the same screening methodology, and with better intraspecific and intra-accession replicates, as in the studies of Jansky et al. (2006, 2008). Many wild potato species are often obligate outcrossers and represent highly heterogeneous gene pools, which contributes to heterogeneity in evaluation results. In addition, evaluations have been performed with different methodologies and designations of trait resistances vary among investigators, complicating comparisons across experiments. For example, the U.S. Potato Genebank data indicate that the *S. chacoense* accession 320285 was evaluated for resistance to the Colorado potato beetle in six studies. That

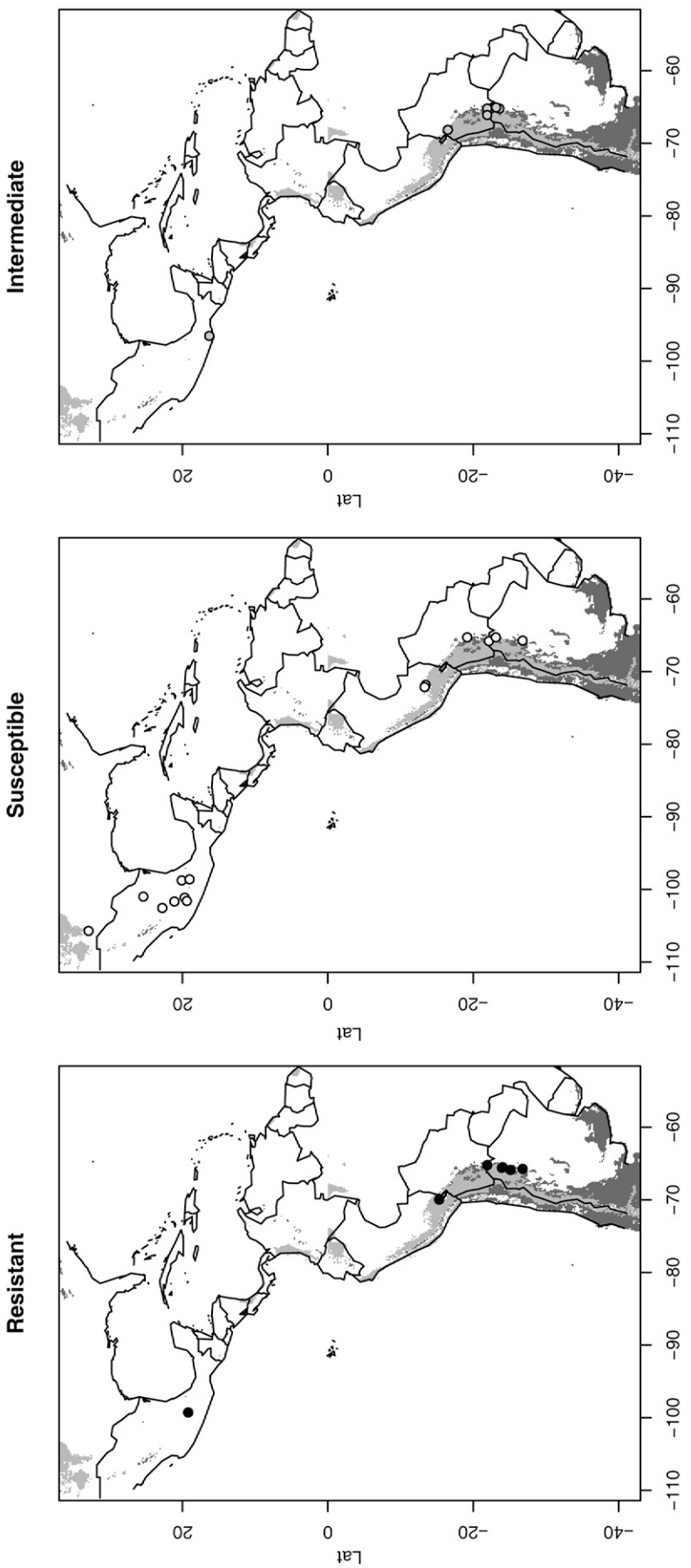

Figure 1. The geographic distribution of resistant, susceptible and intermediate wild potato accessions with regard to *Clavibacter michiganensis sepedonicus*. White areas are predicted by the RandomForest model to contain susceptible accessions, light-gray for intermediate accessions, and dark-gray for resistant accessions. Dots correspond to the accessions used to calibrate the RandomForest method. White dots are susceptible; light gray, intermediate; and dark gray, dots resistant.

single accession was scored as very resistant (one study), resistant (two studies), susceptible (two studies), and very susceptible (one study). Resistance evaluation methods were reported for four of these studies. They were performed in the same time period (three in 1984 and one in 1985) and scores were based on the extent of defoliation in field trials. However, there are many reasons for the variable scores: (i) plants were scored one, two, or three times during the summer, so resistance data vary depending on what stage of the insect (larval or adult) was feeding before evaluation (Pelletier et al., 2007; Sinden et al., 1991); (ii) evaluations were performed in four different states (New Jersey, Minnesota, Virginia, and New York), which may contain genetically different populations of the Colorado potato beetle as a result of differential selection for resistance to pesticides (Alyokhin et al., 2007); and (iii) there is genetic variability among plants within *S. chacoense* for the major resistance mechanism, glycoalkaloid production (Sinden et al., 1986; Torka, 1950). Despite these unavoidable pitfalls, breeders continue to rely on broad germplasm screens when choosing accessions for enhancement of breeding clones. It is worthwhile noting that resistance ratings for only one pest (Colorado potato beetle) and one pathogen (Potato M Carlavirus) were reliably predicted both by host taxonomy and climatic variables. Taking the present data as representative, our study implies that it is worthwhile to take both taxonomy and geographic origin into account while screening genebank materials for pest and disease resistances. However, such associations will hold only for a small subset of resistance traits.

If the results of Jansky et al. (2006, 2008) and the present data are representative of results across diverse crops and diseases, it justifies more focused evaluations of germplasm collections using the core collection approach. This strategy assembles a workable subset of the entire collection with a minimum of redundancy with in-depth accession-specific evaluations (Spooner et al., 2005).

Acknowledgments

We thank the USDA-ARS National Plant Germplasm System Horticultural Evaluation Grants program for funding, Sarah Stephenson for technical assistance, Felipe de Mendiburu for revision of statistical methods, and two anonymous reviewers for helpful comments. This research was supported by the USDA, CIP, and by NSF DEB 0316614 to David Spooner.

References

Alyokhin, A., G. Dively, M. Patterson, C. Castaldo, D. Rogers, M. Mahoney, and J. Wollam. 2007. Resistance and cross-resistance to imidacloprid and thiamethoxam in the Colorado potato beetle *Leptinotarsa decemlineata*. Pest Manage. Sci. 63:32–41.

Brownrigg, R., T.P. Minka, R.A. Becker, and A.R. Wilks. 2007. Maps library for R. Available at cran.r-project.org/web/packages/maps/index.html (accessed 21 Mar. 2008; verified 6 Apr. 2009). R Foundation for Statistical Computing, Vienna, Austria.

Burt, P.J.A. 1995. The potato and the pathogen: The Irish potato famine of 1845. Weather 50:342–346.

Castillo, R., and D.M. Spooner. 1997. Phylogenetic relationships of wild potatoes, *Solanum* series *Conicibaccata* (sect. *Petota*). Syst. Bot. 22:45–83.

Chapman, C.G.D. 1989. Collection strategies for the wild relatives of crops. p. 263–279. *In* A.D.H. Brown et al. (ed.) The use of plant genetic resources. Cambridge Univ. Press, Cambridge, UK.

Comeau, A. 1982. Geographic distribution of resistance to barley yellow dwarf virus in *Avena sterilis* wild oat lines in Europe, Africa, Turkey and Iran. Can. J. Plant Pathol. 4:147–151.

Cronn, R., M. Brothers, K. Klier, P.K. Bretting, and J.F. Wendel. 1997. Allozyme variation in domesticated annual sunflower and its wild relatives. Theor. Appl. Genet. 95:532–545.

Daly, D.C., K.M. Cameron, and D.M. Stevenson. 2001. Plant systematics in the age of genomics. Plant Physiol. 127:1328–1333.

del Rio, A.H., and J.B. Bamberg. 2002. Lack of association between genetic and geographical origin characteristics for the wild potato *Solanum sucrense*. Am. J. Potato Res. 79:335–338.

del Rio, A.H., and J.B. Bamberg. 2004. Geographical parameters and proximity to related species predict genetic variation in the inbred potato species *Solanum verrucosum* Schltdl. Crop Sci. 44:1170–1177.

del Rio, A.H., J.B. Bamberg, Z. Huamán, A. Salas, and S.E. Vega. 2001. Association of ecogeographical variables and RAPD marker variation. Wild potato populations of the USA. Crop Sci. 41:870–878.

Dimitriadou, E., K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. 2008. Misc Functions of the Department of Statistics (e1071), TU Wien. Available at cran.r-project.org/web/packages/e1071/index.html (verified 6 Apr. 2009). R Foundation for Statistical Computing, Vienna, Austria.

Ellis, P.R., B.B. Kift, D.A.C. Pink, P.L. Jukes, J. Lynn, and G.M. Tatchell. 2000. Variation in resistance to cabbage aphid between and within wild and cultivated *Brassica* species. Genet. Resour. Crop Evol. 47:395–401.

Fahima, T., G.L. Sun, A. Beharav, T. Krugman, A. Beiles, and E. Nevo. 1999. RAPD polymorphism of wild emmer wheat populations, *Triticum dicoccoides*, in Israel. Theor. Appl. Genet. 98:434–447.

Ferguson, M.E., B.V. Ford-Lloyd, L.D. Robertson, N. Maxted, and H.J. Newbury. 1998. Mapping the geographical distribution of genetic variation in the genus *Lens* for the enhanced conservation of plant genetic diversity. Mol. Ecol. 7:1743–1755.

Francisco-Ortega, J.H., J. Newbury, and B.V. Ford-Lloyd. 1993. Numerical analysis of RAPD data highlight the origin of cultivated tagasaste (*Chamaecytisus proliferus* ssp. *palmensis*) in the Canary Islands. Theor. Appl. Genet. 87:264–270.

Freytag, G.F., and D.G. Debouck. 2002. Taxonomy, distribution, and ecology of the genus *Phaseolus* (Leguminosae-Papilionoideae) in North America, Mexico and Central America. Sida 23:1–300.

Gallois, A., J.C. Audran, and M. Burus. 1998. Assessment of genetic relationships and population discrimination among *Fagus sylvatica* L. by RAPD. Theor. Appl. Genet. 97:211–219.

Gebhardt, C., A. Ballvora, B. Walkemeier, P. Oberhagemann, and K. Schüler. 2004. Assessing genetic potential in germplasm collections of crop plants by marker-trait association: A case study for potatoes with quantitative variation of resistance to late blight and maturity type. Mol. Breed. 13:93–102.

Ghislain, M., D. Andrade, F. Rodríguez, R.J. Hijmans, and D.M. Spooner. 2006. Genetic analysis of the cultivated potato

Solanum tuberosum L. Phureja Group using RAPDs and nuclear SSRs. Theor. Appl. Genet. 113:1515–1527.

Greene, S.L., T.C. Hart, and A. Afonin. 1999a. Using geographic information to acquire wild crop germplasm for *ex situ* collections: I. Map development and field use. Crop Sci. 39:836–842.

Greene, S.L., T.C. Hart, and A. Afonin. 1999b. Using geographic information to acquire wild crop germplasm for *ex situ* collections: II. Post-collection analysis. Crop Sci. 39:843–849.

Guo, Q., M. Kelly, and C.H. Graham. 2005. Support vector machines for predicting distribution of sudden oak death in California. Ecol. Modell. 182:75–90.

Hanneman, R.E., Jr. 1994. Assignment of endosperm balance numbers to the tuber-bearing Solanums and their close non-tuber-bearing relatives. Euphytica 74:19–25.

Hanneman, R.E., Jr., and J.B. Bamberg. 1986. Inventory of tuber-bearing *Solanum* species. Wisc. Agric. Exp. Stn. Bull. 533:1–216.

Hawkes, J.G. 1990. The potato: Evolution, biodiversity, and genetic resources. Belhaven Press, Oxford.

Holbrook, C.C., M.G. Stephenson, and A.W. Johnson. 2000. Level and geographic distribution of resistance to *Meloidogyne arenaria* in the U.S. peanut germplasm collection. Crop Sci. 40:1168–1171.

Hormanza, J.I., L. Dollo, and V.S. Polito. 1994. Determination of relatedness and geographical movements of *Pistacia vera* (*Pistachio*: Anacardiaceae*)* germplasm by RAPD analysis. Econ. Bot. 48:349–358.

Jansky, S.H., R. Simon, and D.M. Spooner. 2006. A test of taxonomic predictivity: Resistance to white mold in wild relatives of cultivated potato. Crop Sci. 46:2561–2570.

Jansky, S.H., R. Simon, and D.M. Spooner. 2008. A test of taxonomic predictivity: Resistance to early blight in wild relatives of cultivated potato. Phytopathology 98:680–687.

Jarosz, A.M., and A.A. Burdon. 1991. Host-pathogen interactions in natural populations of *Linum marginale* and *Melampsora lini*. II. Local and regional variation in patterns of resistance and racial structure. Evolution 45:1618–1627.

Jørgensen, S., and R. Mauricio. 2004. Neutral genetic variation among wild North American populations of the weedy plant *Arabidopsis thaliana* is not geographically structured. Mol. Ecol. 13:3403–3413.

Kik, G. 2002. Exploitation of wild relatives for breeding of cultivated *Allium* species. *In* H.H. Rabinowitch and L. Currah (ed.) Advances in *Allium* science. CABI Publ., New York.

Lamboy, W.F., J. Yu, P.L. Forsline, and N.F. Weeden. 1996. Partitioning of allozyme diversity in wild populations of *Malus sieversii* L. and implications for germplasm collection. J. Am. Soc. Hortic. Sci. 121:982–987.

Lewin-Koh, N.J., E.J. Pebesma, E. Archer, S. Dray, D. Forrest, P. Giraudoux, D. Golicher, V. Gómez Rubio, P. Hausmann, T. Jagger, S.P. Luque, D. MacQueen, A. Niccolai, T. Short, and R. Bivand. 2008. Maptools library for R. Available at cran.r-project.org/web/packages/maptools/index.html (verified 6 Apr. 2009). R Foundation for Statistical Computing, Vienna, Austria.

Liaw, A., M. Wiener, L. Breiman, and A. Cutler. 2008. RandomForest: Breiman and Cutler's random forests for classification and regression. Available at cran.r-project.org/web/packages/randomForest/index.html (verified 6 Apr. 2009). R Foundation for Statistical Computing, Vienna, Austria.

Loveless, M.D., and J.L. Hamrick. 1984. Ecological determinants of genetic structure in plant populations. Annu. Rev. Ecol. Syst. 15:65–95.

Marshall, D.R., and A.D.H. Brown. 1975. Optimum sampling strategies in genetic conservation. p. 53–80. *In* O.H. Frankel and J.G. Hawkes (ed.) Crop genetic resources for today and tomorrow. Cambridge Univ. Press, Cambridge, UK.

McGregor, C.E., R. van Treuren, R. Hoekstra, and T.J.L. van Hintum. 2002. Analysis of the wild potato germplasm of the series *Acaulia* with AFLPs: Implications for ex situ conservation. Theor. Appl. Genet. 104:146–156.

Michael, S., and M. Lewis-Beck. 1995. Data analysis: An introduction. Quantitative Applications in Social Sciences Series, A Sage University Paper. Thousand Oaks, London.

Moeller, D.A., and B.A. Schaal. 1999. Genetic relationships among Native American maize accessions of the Great Plains assessed by RAPDs. Theor. Appl. Genet. 99:1061–1067.

Pelletier, Y., C. Clark, and D. De Koeyer. 2007. Level and genetic variability of resistance to the Colorado potato beetle (*Leptinotarsa decemlineata* (Say)) in wild *Solanum* species. Am. J. Potato Res. 84:143–148.

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. International Biological Programme 2:81–96. Cambridge Univ. Press, UK.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reid, L., J.T. Arnason, C. Nozzolillo, and R. Hamilton. 1990. Resistance of maize germplasm to European corn borer, *Ostrinia nubilalis*, as related to geographical origin. Can. J. Bot. 68:311–316.

Rines, H.W., D.D. Stuthman, L.W. Briggle, V.L. Youngs, H. Jedlinski, D.H. Smith, J.A. Webster, and P.G. Rothman. 1980. Collection and evaluation of *Avena fatua* for use in oat improvement. Crop Sci. 20:63–68.

Robinson, R.E., and D.S. Decker-Walters. 1997. Cucurbits. CABI Publ., New York.

Rodríguez, A., and D.M. Spooner. 1997. Chloroplast DNA analysis of *Solanum bulbocastanum* and *S. cardiophyllum*, and evidence for the distinctiveness of *S. cardiophyllum* subsp. *ehrenbergii* (sect. *Petota*). Syst. Bot. 22:31–43.

Ross, H. 1986. Potato breeding: Problems and perspectives. Adv. Plant Breed. Suppl. 13.

Ruiz de Galerreta, J.I., A. Carrasco, A. Salazar, I. Barrena, E. Iturritxa, R. Marquinez, F.J. Legorburu, and E. Ritter. 1998. Wild *Solanum* species as resistance sources against different pathogens of potato. Potato Res. 41:57–68.

Sato, K., and K. Takeda. 1997. Net blotch resistance in wild species of *Hordeum*. Euphytica 95:179–185.

Simon, P.W. 2000. Domestication, historical development, and modern breeding of carrot. Plant Breed. Rev. 19:157–190.

Sinden, S.L., L.L. Sanford, W.W. Cantelo, and K.L. Deahl. 1986. Leptine glycoalkaloids and resistance to the Colorado potato beetle (Coleoptera: Chrysomelidae) in *Solanum chacoense*. Environ. Entomol. 15:1057–1062.

Sinden, S.L., L.L. Sanford, W.W. Cantelo, and K.L. Deahl. 1991. Alleleochemically mediated host resistance to the Colorado potato beetle, *Leptinotarsa decemlineata* (Say) (Coleoptera: Chrysomelidae). Mem. Ent. Soc. Can. 157:19–28.

Spooner, D.M., G.J. Anderson, and R.K. Jansen. 1993. Chloroplast DNA evidence for interrelationships of tomatoes, potatoes, and pepinos (Solanaceae). Am. J. Bot. 80:676–688.

Spooner, D.M., and R.T. Castillo. 1997. Reexamination of series relationships of South American wild potatoes (Solanaceae: *Solanum* sect. *Petota*): Evidence from chloroplast DNA restriction site variation. Am. J. Bot. 84:671–685.

Spooner, D.M., W.L.A. Hetterscheid, R.G. van den Berg, and W. Brandenburg. 2003. Plant nomenclature and taxonomy: An horticultural and agronomic perspective. Hortic. Rev. (Am Soc Hortic Sci) 28:1–60.

Spooner, D.M., F. Rodríguez, Z. Polgár, H.E. Ballard, Jr., and S.H. Jansky. 2008. Genomic origins of potato polyploids: GBSSI gene sequencing data. Crop Science 48(S1):S27–S36.

Spooner, D.M., and A. Salas. 2006. Structure, biosystematics, and genetic resources. p. 1–39. *In* J. Gopal and S.M. Paul Khurana (ed.) Handbook of potato production, improvement, and post-harvest management. The Haworth Press, Binghamton, NY.

Spooner, D.M., J. Tivang, J. Nienhuis, J.T. Miller, D.S. Douches, and A. Contreras-M. 1995. Comparison of four molecular markers in measuring relationships among the wild potato relatives *Solanum* section *Etuberosum* (subgenus *Potatoe*). Theor. Appl. Genet. 92:532–540.

Thuiller, W., J. Vayreda, J. Pino, S. Sabate, S. Lavorel, and C. Gracia. 2005. Large-scale environmental correlates of forest tree distributions in Catalonia (NE Spain). Glob. Ecol. Biogeogr. 12:313–325.

Torka, M. 1950. Breeding potatoes with resistance to the Colorado potato beetle. Am. Potato J. 27:263–271.

Warburton, F.E. 1967. The purposes of classifications. Syst. Zool. 16:241–245.

Westman, A.L., S. Kresovick, and M.H. Dickson. 1999. Regional variation in *Brassica nigra* and other weeds crucifers for disease reaction of *Alternaria brassicicola* and *Xanthomonas campestris* pv. *campestris*. Euphytica 106:253–259.