Am. J. Potato Res. (2015) 92:684–696
DOI 10.1007/s12230-015-9484-2

# Colorado Potato Beetle Resistance in *Solanum oplocense* X *Solanum tuberosum* Intercross Hybrids and Metabolite Markers for Selection

**Helen H. Tai**[1] · **Kraig Worrall**[1,2] · **David De Koeyer**[1] · **Yvan Pelletier**[1] · **George C. C. Tai**[1] · **Larry Calhoun**[2]

Published online: 5 November 2015

**Abstract** *S. oplocense* Hawkes, a wild relative of the potato *S. tuberosum* L. and source of resistance against the Colorado potato beetle *Leptinotarsa decemlineata* (Say) (CPB), was intercrossed with *S. tuberosum*. Backcross clones carried varying levels of resistance. Differences in foliar metabolites between resistant and susceptible clones were analyzed using liquid chromatography-mass spectrometry (LC-MS). Supervised machine learning classification methods uncorrelated shrunken centroids (USC), k-nearest neighbor (KNN) and support vector machines (SVM) were applied to develop algorithms that can classify resistant and susceptible plants using the metabolite data. Five metabolites were found to have a low error rate of prediction of CPB resistance. The five metabolites included two glycoalkaloids previously associated with resistance and susceptibility to CPB, dehydrocommersonine and solanine, respectively. Resistance was associated with a change in composition of glycoalkaloids to higher ratios of dehydrocommersonine over solanine.

**Resumen** *S. oplocense* Hawkes, un pariente silvestre de la papa *S. tuberosum* L., y fuente de resistencia contra el escarabajo de Colorado *Leptinotarsa decemlineata* (Say) (CPB), se intercruzó con *S. tuberosum*. Los clones de la retrocruza conservaron diversos niveles de resistencia. Se analizaron las diferencias en los metabolitos foliares entre los clones resistentes y susceptibles usando espectrometría de cromatografía líquida de masas (LC-MS). Métodos supervisados de clasificación de aprendizaje de máquina no correlacionados con centroides encogidos (USC), k-cercanía de vecinos (KNN) y máquinas de respaldo de vector (SVM) se aplicaron para desarrollar algoritmos que pueden clasificar plantas resistentes y susceptibles usando los datos de los metabolitos. Se encontró que cinco metabolitos tenían un nivel bajo de error de predicción de la resistencia al CPB. Los cinco metabolitos incluyeron dos glicoalcaloides asociados previamente con resistencia y susceptibilidad al CPB, la deshidrocommersonina y la solanina, respectivamente. La resistencia se asoció con un cambio en la composición de los glicoalcaloides a altas proporciones de deshidrocommersonina sobre la solanina.

**Keywords** Colorado potato beetle resistance · Untargeted metabolite profiling · Potato · Solanum oplocense · Supervised machine learning classification

## Introduction

*Leptinotarsa decemlineata* (Say) (CPB) causes potato yield losses of 30–50 % (McLeod and Tolman 1987; Stemeroff and George 1983) and is controlled through use of neonicotinoid insecticides. CPB populations with resistance to insecticides have emerged (Alyokhin et al. 2008; Szendrei et al. 2012) increasing the need to develop alternative strategies including breeding for resistant potato germplasm. The domesticated potato, *S. tuberosum* Hawkes, has a narrow genetic base and most commercial potato varieties are susceptible hosts for CPB. Wild *Solanum* species that can be

**Electronic supplementary material** The online version of this article (doi:10.1007/s12230-015-9484-2) contains supplementary material, which is available to authorized users.

✉ Helen H. Tai
Helen.Tai@canada.ca

[1] Agriculture and Agri-Food Canada Potato Research Centre, P. O. Box 20280, 850 Lincoln Rd., Fredericton, N. B., Canada E3B 4Z7

[2] Department of Chemistry, University of New Brunswick, Fredericton, N. B., Canada

intercrossed with *S. tuberosum* are valuable sources of genetic diversity. There are a number with resistance to CPB, including some that can be introgressed with *S. tuberosum* (Flanders et al. 1992; Jansky et al. 2009; Pelletier 2007; Pelletier et al. 2011; Pelletier and Tai 2001). Metabolites produced in the foliage of wild species can function as anti-feedants and semiochemicals that affect host plant selection by mobile CPB adults and successful establishment of larvae on foliage (Pelletier et al. 2011; Pelletier and King 1987). The glandular trichome-containing wild species *S. berthaultii* Hawkes, has been shown to use chemical defense. It was crossed with *S. tuberosum* to produce germplasm with increased CPB resistance (Yencho and Tingey 1994). Glandular trichomes from *S. berthaultii* were shown to produce exudates containing sesquiterpenes (Carter et al. 1989) and sucrose fatty acid esters (King et al. 1986) that were associated with CPB resistance in the potato (Pelletier and Smilowitz 1990). *S. chacoense* Hawkes is another wild species that is high in leptine glycoalkaloids. Introgression of this species has also resulted in increased CPB resistance (Sanford et al. 1998; Tingey and Yencho 1994). Analysis of plant foliar metabolites of six wild *Solanum* species with CPB resistance demonstrated that increased tetraose over triose glycoalkaloids was associated with CPB resistance in addition to increases in phenylpropanoid metabolites (Tai et al. 2014).

Metabolite-based markers can be applied to selection and breeding in plants (Zabotina 2013). Selection for CPB resistance currently involves field and/or laboratory screening assays for CPB feeding. Metabolite marker screening would provide an alternative lower cost screening compared to CPB feeding assays. Selection for foliar leptines has been demonstrated to be an effective screen for CPB resistance. Leptine levels in the F2 of *S. tuberosum* (4×)×*S. chacoense* (4×) potato progenies were highly regressed with leaf disk consumption and field defoliation (Yencho et al. 2000). There are a number of technologies available for the discovery of metabolite markers (Fernie and Schauer 2009). Targeted metabolite profiling is optimized for analysis of selected compounds, whereas in untargeted metabolite profiling the entire range of compounds is analyzed (Vinayavekhin and Saghatelian 2010). Untargeted metabolite profiling results in highly complex profiles of peaks and requires use of computer algorithms to analyze mass spectra to identify compounds. We have successfully applied untargeted profiling using LC-MS to identify metabolites associated with CPB resistance in six wild *Solanum* species (Tai et al. 2014). One of the wild species, *S. oplocense*, has been cross-hybridized with *S. tuberosum*. We describe here application of untargeted metabolite profiling and supervised machine learning classification for identification of metabolite markers for CPB resistance using clones from backcross generation 1 (BC1) and 2 (BC2) carrying *S. oplocense* genetic material. Supervised machine learning classification involves using a set of data (training data) from individuals that have been pre-classified into groups to train an algorithm to classify other individuals. The data used for classification in this study were varying levels of metabolites analyzed by LC-MS. This study used USC, KNN and SVM machine learning classification methods. The metabolites identified for use in classification have application as markers for genetic mapping and selection and breeding of *S. oplocense* derived germplasm with CPB resistance.

## Materials and Methods

### *S. oplocense* X *S. tuberosum* Intercross

Pelletier et al. (2001) identified *S. oplocense* as a new source of resistance to Colorado potato beetle (CPB). To incorporate this resistance into cultivated potato, *S. oplocense* accession PI 473368 was crossed with three *S. tuberosum* breeding lines to produce $F_1$ hybrids. Evaluation of these hybrids under field conditions from 1998 to 2002 demonstrated that most of the hybrid clones were resistant to CPB (<10 % defoliation). In the next generation, backcross hybrids ($BC_1$) were produced by crossing 10 elite breeding lines with five different $F_1$ hybrid clones (Table 1). One set of $BC_1$ progeny were evaluated for CPB resistance and agronomic characteristics using an accelerated selection scheme. In this scheme, mini-tubers from greenhouse grown seedlings were grown simultaneously at the Potato Breeding sub-station in Benton, New Brunswick, Canada and at the Potato Research Centre in Fredericton, New Brunswick, Canada. A second set of $BC_1$ hybrids were grown in the first field season in single hill plots at Benton only. Clones showing acceptable adaptation and tuber characteristics were harvested and evaluated for CPB resistance (see below) at Fredericton in subsequent field seasons. In 2005 and 2006, superior $BC_1$ clones with CPB resistance were crossed again with adapted *S. tuberosum* germplasm to improve maturity and other agronomic traits in the $BC_2$ generation. Pedigrees and other descriptive information are given in Table 1 for the $BC_1$and $BC_2$ clones used in this study.

### Field Defoliation Scoring for CPB Resistance

Defoliation of potato vines was used as an indicator of CPB activity and resistance. The percentage of defoliation for 33 $BC_1$ and $BC_2$ clones were evaluated in the field at the Potato Research Centre in years 2007 and 2008. In 2007 two duplicate plots were set up in different fields and planted around 2 weeks apart for a total of four replicate plots. In 2008 four replicate plots were planted on the same date. Three to five seed pieces from the same clone were planted in each of the plot and were randomized within each plot. Approximately 10 % of the plants in each plot consisted of the variety Russet Burbank (control)

**Table 1** *S. oplocense* X *S. tuberosum* backcross clones

| clone ID | Generation[a] | female parent | male parent | 2007 % defoliation[b] | 2008 % defoliation[c] | BLUP | # plants[d] |
|---|---|---|---|---|---|---|---|
| 13959-19 | BC1 | A84420-5 | 13213-07[e] | 47.0 | 13.3 | −9.8 | 3 |
| 13960-22 | BC1 | A087277-6 | 13213-07 | 61.5 | 24.8 | 1.9 | 3 |
| 13962-31 | BC1 | F88042 | 13213-07 | 69.5 | 34.0 | 10.1 | 4 |
| 13963-37 | BC1 | Gem Russet | 13213-07 | 4.0 | 5.5 | −22.8 | 3 |
| 14040-02 | BC1 | F58050 | 13597-06[e] | 18.5 | 8.0 | −21.9 | 3 |
| 14044-05 | BC1 | A7816-14 | 13597-06 | 21.0 | 10.5 | −19.6 | 3 |
| 14052-05 | BC1 | A84420-5 | 13597-06 | 49.5 | 13.8 | −9.1 | 4 |
| 14055-03 | BC1 | A087277-6 | 13597-01[e] | 37.5 | 9.5 | −15.6 | 3 |
| 14058-02 | BC1 | A087277-6 | 13597-11[e] | 59.0 | 13.5 | −6.1 | 6 |
| 14067-04 | BC1 | Gem Russet | 13597-01 | 50.5 | 25.0 | −1.4 | 3 |
| 14067-05 | BC1 | Gem Russet | 13597-01 | 3.5 | 6.3 | −27.6 | 3 |
| 14073-01 | BC1 | Innovator | 13597-07[e] | 24.5 | 8.8 | −19.6 | 6 |
| 14081-02 | BC1 | Norvalley | 13597-11 | 65.5 | 28.0 | 5.1 | 3 |
| 15313-06 | BC2 | A8411-8-3 | 13966-34 | 121.7 | 98.0 | 62.1 | 3 |
| 15313-09 | BC2 | A8411-8-3 | 13966-34 | 70.7 | 21.5 | 4.6 | 3 |
| 15313-13 | BC2 | A8411-8-3 | 13966-34 | 43.5 | 43.5 | 11.6 | 3 |
| 15314-05 | BC2 | A9014-2 | 13957-18 | 54.1 | 38.5 | 4.8 | 3 |
| 15314-08 | BC2 | A9014-2 | 13957-18 | 53.2 | 11.5 | −7.7 | 3 |
| 15314-16 | BC2 | A9014-2 | 13957-18 | 59.8 | 41.0 | 8.5 | 3 |
| 15315-05 | BC2 | A9014-2 | 13960-22 | 7.7 | 12.0 | −28.0 | 3 |
| 15316-05 | BC2 | A9014-2 | 13966-34 | 41.2 | 43.5 | 1.3 | 3 |
| 15316-14 | BC2 | A9014-2 | 13966-34 | 79.4 | 105.0 | 46.2 | 3 |
| 15318-07 | BC2 | F88042 | 13966-34 | 43.4 | 31.0 | −3.4 | 3 |
| 15320-09 | BC2 | Norvalley | 13960-22 | 45.0 | 49.0 | 5.5 | 7 |
| 15321-11 | BC2 | Norvalley | 13960-34 | 68.6 | 17.5 | 1.9 | 3 |
| 15321-13 | BC2 | Norvalley | 13960-34 | 60.1 | 58.5 | 16.5 | 3 |
| 15322-09 | BC2 | 13957-18 | A84118-3 | 165.5 | 162.0 | 110.7 | 3 |
| 15323-03 | BC1F1 | 13957-18 | 13959-19 | 15.6 | 3.5 | −28.3 | 6 |
| 15327-02 | BC2 | 13963-37 | A84118-3 | 55.6 | 18.0 | −10.1 | 3 |
| 15328-08 | BC2 | 13963-37 | A9014-2 | 63.5 | 16.0 | −1.1 | 3 |
| 15328-22 | BC2 | 13963-37 | A9014-2 | 27.3 | 28.5 | −11.7 | 3 |
| 15328-26 | BC2 | 13963-37 | A9014-2 | 15.9 | 5.5 | −3.7 | 3 |
| 15332-02 | BC1F1 | 13966-06 | 13957-18 | 63.7 | 9.0 | −4.2 | 3 |

[a] BC1 is backcross generation 1, BC2 is backcross generation 2, BC1F1 is the progeny of a cross between two BC1 parents

[b] and [c] The defoliation assays were carried out on field plants grown in Frederict on

[d] Plants grown in the greenhouse were used for metabolite analysis and the number of plants propagated is indicated

[e] F1 clone

and were evenly distributed within each plot to assess the uniform distribution of the Colorado potato beetle. A plastic lined trench (Boiteau et al. 1994) was installed around each plot to prevent adult CPB to colonize the field by walking and have a uniform distribution within plot. The percentage of defoliation relative to Russet Burbank was visually estimated at the end of July, when around 90 % of the larvae moved to in the ground to pupate and before adult emergence.

The data on defoliation of potato vines in the field was used to classify the clones as resistant and susceptible. The level of CPB infestation varied year to year and this affected the defoliation scores in addition to the variation between clones. Defoliation scores were subjected to statistical analysis by mixed models using GenStat 17 (VSN International) to obtain the best linear unbiased predictor (BLUP), an estimate of the CPB resistance breeding value (Henderson 1984). The mixed model incorporated effects of year of defoliation scoring and clonal variation to derive the BLUP. Let $Y_{ij}$ be the defoliation score of jth clone scored in the ith sample. The mixed model used for the observed defoliation data of the clones is

$$Y_{ij} = X\beta + Z\mu + e_{ij}$$

Where a total of $n$=194 records from a p x 1 vector of fixed effects $\boldsymbol{\beta}$ due to years ($p$=2) and a q x 1 vector of random effects (q=33 clones) $\boldsymbol{\mu}$ for the "breeding values" for clones and **e** is random error effects. **X** and **Z** are (n x p) and (n x q) "design matrices" with 1's and 0's for the years and clonal lines, respectively. The mixed model equations are solved to estimate the fixed effects **β** and random effects **μ**. The values for **μ** represent the BLUP of breeding values of the clones. Residual maximum likelihood (REML) was used to solve mixed model equations (Henderson 1984; Lynch and Walsh 1998; Tai et al. 2009).

### Plant Material for Metabolite Analysis

Multiple replicates per clone from the 13 $BC_1$ and 20 $BC_2$ clones (Table 1) were grown from seed tubers in pots in the greenhouse with potting mix and weekly fertilizer application of 20-20-20. A total of 114 plants were used for analysis. Sampling of foliage was done on fully grown plants at 12 weeks after planting. The apical leaflet from 5 leaves for each plant was pooled in 15 ml conical tubes and flash frozen in liquid nitrogen. Samples were stored at −80 °C prior to extraction.

### Extraction

The frozen samples were ground into a fine powder using a mortar and pestle while immersed in liquid nitrogen. 100 mg of frozen ground tissue was placed in a 1.5 mL polyethylene screw-cap tubes and kept frozen in a liquid nitrogen holding station (SPEX Sampleprep, Metuchen, NJ, USA) until all samples were ground. The ground powder was extracted with 400 μL of extraction solution (92 % methanol, 0.1 % formic acid LC-MS grade) (Sigma-Aldrich, Oakville, ON, Canada). The samples were briefly vortexed and placed on ice until all samples were prepared. Samples were then sonicated in a Branson sonicator bath for 15 min and filtered through a 0.2 μm syringe filter into an LC-MS autosampler vial. Samples were diluted ten-fold to ensure that peak intensities were in the linear range and to avoid detector saturation. The samples were allowed to equilibrate at 23 °C in the dark for 1 hour prior to analysis and were maintained conditions for the duration of the LC-MS analysis.

### LC-MS

Metabolite analysis was carried out using Acuity ultra-performance liquid chromatography- Xevo quadropole time-of-flight mass spectrometry (Waters, Milford, MA, USA) LC-MS system. Using a 5 μL loop, 0.75 μL injections were made for all samples in the study. The same volume of test mixture (rutin hydrate, caffeic acid, benzoic acid, p-coumaric acid, quercetin, L-phenylalanine, resveratrol, ferulic acid, L-tryptophan, sinapic acid, naringenin, trans-cinnamic acid, and isorhamnetin) (Sigma Aldrich, Oakville, ON, Canada) was injected. All components of the test mixture were present at approximately 35 μg/mL in 60:40 Acetonitrile/Water. All chromatographic separations were carried out on a 1 mm× 100 mm BEH C18 reverse phase column. The mobile phase was composed of LC-MS grade water with 0.1 % formic acid (phase A) and LC-MS grade acetonitrile with 0.1 % formic acid (phase B). The linear gradient consisted of six segments as follows: initial segment 95 % A, 5 % B; 13:33 min 25 % A, 75 % B; 13:53 min 5 % A, 95 % B; 18:00 min 5 % A, 95 % B; 18:01 min 95 % A, 5 % B; and 20:00 min 95 % A, 5 % B. The flow rate was 45.0 uL/min for all segments. The autosampler bed was maintained at 23 °C and the column at 35 °C. Samples were injected in a randomized fashion. Backcross clones were run separately on a fresh column. Each sample was injected in triplicate with the exception of the test mixture which was injected after every six samples to evaluate the stability of retention time and mass accuracy over the duration of the experiment. Mass spectrometry data was collected over the duration of the LC-ramp from 0 to 800 s. Masses between 100 and 1500 were detected by electrospray ionization in positive ionization mode. A lockmass solution of dilute leucine enkephalin (LE) in acetonitrile-water (50:50) was introduced via the lock-spray probe at 25 μL/min as directed by the MS manufacturer.

### Metabolite Data Processing and Analysis

Electrospray ionization was used to generate positively charged molecular ions in the mass spectrometer and the mass-to-charge ratio (*m*/z) for molecules was measured in the instrument. The molecular ion most commonly observed was the positively charged protonated adduct $[M+H]^+$. The *m*/*z* of $[M+H]^+$ was one mass one unit higher than the theoretical mass due to the addition of the proton. Mass spectrometry data was processed using Waters Markerlynx XS software. The LC-MS data were detected and noise-reduced in both the LC and MS domains such that only true analytical peaks were further processed by the software. Each peak was referred to as a feature that was identified using the *m*/*z* of the positive molecular ion and chromatographic retention time from the chromatogram (e.g., for alpha-chaconine the *m*/*z* was 852.5 and the retention time was 482 so the feature ID was 852.5/482). No retention time correction was used as retention time stability was sufficient for LC. Quantification of peak intensity was done by integrating peaks with a mean retention time in the window of 100–800 s. The retention time window was selected based on visual evaluation of chromatograms to exclude column void and washout.

The MS provided accurate *m*/*z* that could be used to search the MetLin compound mass database (https://metlin.scripps.edu/index.php). The additional mass of the proton was taken into account when searching MetLin. Compounds in the database were screened for those with a mass that was close to the *m*/*z* of the feature. The mass difference was measured in parts per million (ppm), where

$$\Delta\text{ppm} = \frac{1.0 \times 10^6(\text{measured mass}-\text{theoretical mass})}{\text{theoretical mass}}$$

Measured mass was the *m*/*z* of the feature that is corrected for the mass of the proton. Compounds in the database with Δppm of 20 or less were selected. For some features there was more than one compound in the database that matched and for others there were no matches.

Feature peak intensity data was transformed prior to analysis using supervised machine learning classification algorithms as follows: zeros were removed by adding $1\times10^{-12}$ to all feature peak intensities and data was $\log_{10}$ transformed. Varying numbers of features were used to develop classifiers for CPB resistant (BLUP<0) and susceptible (BLUP>0) using three machine learning algorithms - USC (Tibshirani et al. 2002; Yeung and Bumgarner 2003), SVM (Brown et al. 2000) and KNN (Theilhaber et al. 2002) that were part of MeV 4.7.9 (http://www.tm4.org/mev.html). Default conditions were used except for USC where #folds was increased to eight. The total number of plants used for training and validation of the classifier was 77. The classifier was then used to predict resistance and susceptibility in a separate set of 37 plants. The set of 77 training and 37 test plants included at least one plant from each clone. The predicted classification of resistant and susceptible plants was compared to actual classifications based on BLUP and percentage error of the prediction was calculated.

The Student's *t*-test was used to analyze of differences in peak intensities for features assigned to glycoalkaloids using Systat 13.1 (Systat software. San Joe, USA). The null hypothesis $H_o$: resistant=susceptible feature peak intensity was tested.

## Results

### *S. oplocense* X *S. tuberosum* Hybrids CPB Resistance Evaluation

In this study, the wild species *S. oplocense*, which has CPB resistance, was intercrossed with *S. tuberosum*. The resulting $F_1$ carried CPB resistance as determined in field defoliation assays and were backcrossed to elite breeding lines. The backcross clones were evaluated for CPB resistance using field defoliation assays. Infestation of potato plants in the field was with naturally occurring CPB that was present in the absence of chemical treatment. Defoliation of hybrid clones in field plots was visually scored (Table 1). The defoliation was measured as the percentage difference in defoliation compared with Russet Burbank. BLUP was used to generate a breeding value for defoliation using the 2 years of defoliation data. Each of the hybrid lines of potato were scored in the field for two consecutive years for % defoliation in the presence of CPB (Table 1). Variation in infestation rates occurred between the 2 years, but there was correlation between relative defoliation of clones in the 2 years (Fig. 1). BLUP breeding values for CPB resistance for each clone were calculated using the % defoliation data. A BLUP<0 corresponded to relatively low defoliation (resistant) and BLUP>0 were clones with relatively high defoliation (susceptible). Classification of resistant and susceptible for each clone for supervised machine learning was based on BLUP breeding values.

### Untargeted Metabolite Profiling

The goal of the study was to identify metabolite markers that can be used to select potato clones with resistance to CPB for the purpose of selection and breeding. Selection would be targeted for parental breeding lines and advanced selections that are propagated in the greenhouse. Additionally, CPB resistance screening in the greenhouse has an advantage that it can be done at any time of the year. For these reasons, greenhouse-grown plants were used. Previous studies had indicated that there were foliar metabolite differences between *S. oplocense* and *S. tuberosum*. In the current study backcross clones used in developing germplasm were analyzed to enhance the specificity of the metabolite marker to the breeding material in the program. Untargeted metabolite profiling using LC-MS analyzed multiple metabolites simultaneously that increased the probability of finding metabolite markers. Multiple plants were propagated for each clone. Each plant was treated as an individual in the metabolite analysis. The total number of plants was 114 and the number of plants used for each clone is indicated in Table 1. There was a total of 651 features identified in the metabolite data (Supplementary Material Table S1).

### Supervised Machine Learning Classification Using USC

Supervised machine learning classification classifies individuals into groups based on quantitative data. The method involves the use of a set of quantitative data from multiple individuals with known classifications to train an algorithm. The algorithm is cross-validated using methods that sub-sample the training data set multiple times to see if the same classification result is obtained. The trained, cross-validated algorithm can then be used to classify other individuals. The data used in this study was the peak intensity of metabolites (features) in potato foliage and the two classes were resistant or susceptible to CPB (based on BLUP). Feature intensity data from the foliage of 114 plants derived from the 33 clones in Table 1 was collected. Each of the 114 plants was classified as resistant or susceptible to CPB based on BLUP data for defoliation. Data for 77 plants were used for training and cross-validation of algorithms. The training and cross-validation plants consisted of 45 plants

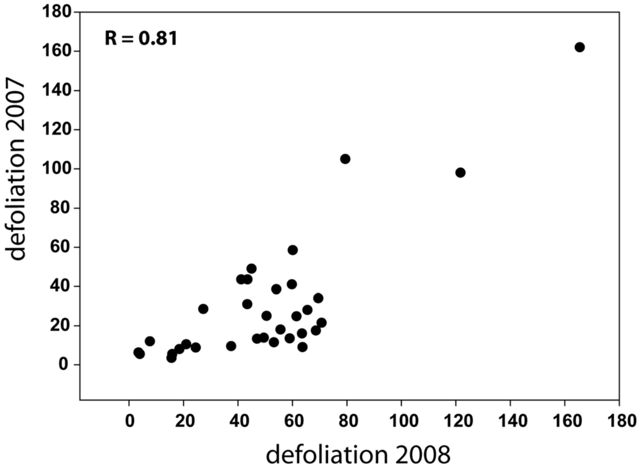

**Fig. 1** Correlation analysis between relative defoliation scores between years 2007 and 2008

**Table 2** Cross-validation of 77 training plants

| | # Features | # Errors |
|---|---|---|
| USC | 651 | 2.5 |
| USC | 35 | 2.5 |
| USC | 5 | 2.8 |
| KNN | 651 | 13 |
| KNN | 35 | 16 |
| KNN | 5 | 17 |
| SVM | 651 | 19 |
| SVM | 35 | 16 |
| SVM | 5 | 21 |

classified as resistant and 32 plants classified as susceptible. The numbers of classification errors found in the cross-validation of training samples are reported in Table 2. The trained and cross-validated algorithm was then used to classify a separate set of 37 plants (22 resistant and 15 susceptible) to test the predication accuracy of the classifiers (Table 3). The class predictions for the 37 plants were compared with actual defoliation scoring collected for the plants.

USC has an integrated classification and feature selection algorithm. The classification algorithm used m-fold cross-validation, where data from the training plants were randomly divided into m subsets of roughly equal size. Each of the m subsets was left out in turn in the cross-validation, and the other (m-1) subsets were used to train the classification algorithm. The result of the training phase was a list of various classifiers with different numbers of features and the corresponding of the average number of classification errors in cross-validation. The user then selects a desired classifier from the list using the feature selection option. The classifier with all 651 features was selected and was used to classify the separate set of 37 test plants. There were seven misclassifications of susceptible and two for resistant (Table 3). The error rate on the test plants was higher than for the cross-validation of training plants (Tables 2 and 3).

Metabolites produced at high levels were preferable for use as markers in the final end-user assay. Additional selection was applied to the data to reduce the features to those with peak intensity where $\log_{10}>4$ in at least one sample. Thirty-five features fulfilled the criteria. Table 4 lists the *m*/*z* of the 35 features and the results of the search for compounds in the MetLin database with theoretical masses that were a Δppm of 20 or less compared to the feature *m*/*z*. The molecular formula for the compounds found in the MetLin database was also presented. Many of the features present at high intensity were glycoalkaloids including chaconine, solanine, dehydrocommersonine and demissine. The average feature peak intensities for plants categorized as resistant or susceptible are listed in Table 4.

USC classification was done using the 35 high intensity features and the classifier with all 35 features was selected. The results showed that the error rate for classification of the 37 separate test samples with the 35 high intensity feature classifier was similar to the classifier using all 651 features (Tables 2 and 3). The feature selection option for USC in MeV was used to select another classifier with a smaller number of features among the 35 high intensity ones. The second selected classifier was based on five features with an average of 2.8 mistakes in the cross-validation was selected (Table 2). The five features were 188.1/285, 475.3/314, 574.4/411, 868.5/442, and 1046.6/464 (Table 4). The error rate for classification of the test samples was the same at 27 %, however, there were more plants misclassified as resistant rather than susceptible (Table 3). These results show that classifiers with smaller numbers of features can have similar error rates as those with large number of features. A smaller number of features would be advantageous in development of targeted quantitative screening assays. The USC selection of five features included 188.1/285. The feature 188.1/285 was relatively unchanged between resistant and susceptible plants on average compared with other metabolites (Table 4). Feature 475.3/314 was also among the five USC selected features and was also relatively unchanged. It was matched with five compounds with Δppm of 6 (Table 4). There were no *m*/*z* matches in MetLin with 574.4/411 (Table 4) which was on average higher in resistant over susceptible plants.

## Glycoalkaloids

Two features from the USC five-feature classifier, 868.5/442, and 1046.6/464, were a match with the glykoalkaloids, alpha-solanine and dehydrocommersonine, respectively (Table 4). The average peak intensity of 868.5/442 was 3.650 for resistant plants and 4.005 for susceptible plants,

**Table 3** Metabolite-based classification of 37 test plants

| | # Features[b] | Classification[c] | | Errors[d] | | % Error |
|---|---|---|---|---|---|---|
| | | resistant | susceptible | resistant | susceptible | |
| Actual[a] | | 22 | 15 | | | |
| USC | 651 | 16 | 21 | 2 | 8 | 27.0 |
| USC | 35 | 20 | 17 | 4 | 6 | 27.0 |
| USC | 5 | 28 | 9 | 9 | 2 | 27.0 |
| KNN | 651 | 25 | 12 | 4 | 1 | 13.5 |
| KNN | 35 | 23 | 14 | 1 | 0 | 2.7 |
| KNN | 5 | 21 | 16 | 1 | 1 | 5.4 |
| SVM | 651 | 12 | 25 | 1 | 11 | 32.4 |
| SVM | 35 | 1 | 36 | 0 | 21 | 56.8 |
| SVM | 5 | 20 | 17 | 4 | 6 | 27.0 |

[a] 37 plants from Table 1 were selected as test plants. The test plants included at least one plant from each clone. The BLUP score was used to assign the actual classes shown in the first row. BLUP<0 was resistant and BLUP>0 was resistant

[b] 651 features is the total number and 35 is the number with $\log_{10}$>4. Five is the number of features selected by the USC classifier. This was the smallest number of features that could be used for classification

[c] Classification of the test plants using USC, KNN and SVM

[d] The number clones classified incorrectly as resistant or susceptible

and for 1046.6/464 it was 5.114 and 4.400 for resistant and susceptible plants, respectively. The glycoalkaloid peaks had the highest intensity in the total ion chromatogram (Fig. 2). The base peak (peak with highest intensity) for both CPB resistant and susceptible plants was 852.5/482, which was assigned to alpha-chaconine. Feature 852.5/482 had an average peak intensity of 6.046 and 5.962 for resistant and susceptible plants, respectively, showing similar levels between classes (Table 4). Alpha-solanine eluted from the UPLC column over a longer period of time compared to alpha-chaconine as demonstrated by its assignment to four features 868.5/494, 868.5/477, 868.5/457, and 868.5/442 with MarkerLynx (Table 4). In comparison, alpha-chaconine was assigned to a single feature, 852.5/482. The dehydrocommersonine feature 1046.6/464 was also assigned to a single feature with a lower retention time compared to alpha-chaconine and alpha-solanine. The alpha-solanine features, 868.5/494 and 868.5/477, had retention times that were under a broad peak that included 852.5/482 in the total ion chromatogram (Fig. 2a and b), indicating co-elution of alpha-solanine and alpha-chaconine. The features 868.5/442 and 868.5/457 had a retention times that did not overlap with the alpha-chaconine 852.2/482 peak. However, it was noted that 868.5/457 co-eluted with 1046.6/464 (Fig. 2b), indicating that 868.5/442 was the only featured assigned to alpha-solanine that did not co-elute with another glycoalkaloid. 868.5/442 also had the highest difference in average peak intensity between resistant and susceptible plants. Interestingly, 868.5/442 was the only alpha-solanine feature that was included in the USC five-feature classifier.

Selection for glycoalkaloids in breeding can be problematic as high glycoalkaloid levels are toxic to humans. Therefore additional analysis of glycoalkaloids was done as the five feature classifier included two features that matched with alpha-solanine and dehydrocommersonine in the MetLin database. A *t*-test was done to test for differences in peak intensities between resistant and susceptible plants for features that were assigned to glycoalkaloids in Table 4. The results show that most of the glycoalkaloids are significantly increased in resistant plants (Table 5). An exception was the alpha-solanine feature 868.5/442 which showed a significant decrease in resistant plants. The other glycoalkaloid in the five feature classifier, dehydrocommersonine 1046.6/464, was increased in resistant plants. These results indicate resistant plants have a high ratio of 1046.6/464 to 868.5/442. A *t*-test of the ratio of 1046.4/464:868.5/442 demonstrated that the ratio was significantly different in resistant and susceptible plants (Table 5). These results suggest that a selection strategy for CPB resistance can use a high ratio of dehydrocommersonine to alpha-solanine. In addition to 868.5/442, there were three other features that were assigned to alpha-solanine. The ratio of the peak intensity for 1046.6/464 to the total peak intensity for all four alpha-solanine features was compared between resistant and susceptible plants using the *t*-test and significant differences were also found (Table 5). These results indicate that selection for resistance can target a change in the composition of glycoalkaloids to increase dehydrocommersonine over alpha-solanine rather than increases in any one glycoalkaloid. This selection strategy would off-set selection of clones with high levels of glycoalkaloids.

**Table 4** Features with high peak intensity ($\log_{10}$ peak intensity>4 in at least one sample)

| | | | | | | Average BLUP[b] | |
|---|---|---|---|---|---|---|---|
| | | | | | | −13.825 | 19.250 |
| feature ID | [M+H]+*m/z* | putative compound | theoretical mass | Δppm | formula | resistant[c] | susceptible[d] |
| 188.1/285[a] | 188.0704 | Deethylatrazine | 187.0625 | 3 | $C_6H_{10}ClN_5$ | 4.434 | 4.488 |
| | | 3-amino-2-naphthoic acid | 187.0633 | 1 | $C_{11}H_9NO_2$ | | |
| | | Indoleacrylic acid | 187.0633 | 1 | $C_{11}H_9NO_2$ | | |
| 398.3/477 | 398.3457 | Verazine | 397.3344 | 9 | $C_{27}H_{43}NO$ | 5.388 | 5.257 |
| | | Solanidine | 397.3345 | 9 | $C_{27}H_{43}NO$ | | |
| 398.3/464 | 398.3464 | Verazine | 397.3344 | 11 | $C_{27}H_{43}NO$ | 5.240 | 4.665 |
| | | Solanidine | 397.3345 | 11 | $C_{27}H_{43}NO$ | | |
| 399.4/474 | 399.3507 | tetracosanedioic acid | 398.3396 | 9 | $C_{24}H_{46}O_4$ | 4.873 | 4.737 |
| | | lauroyl peroxide | 398.3396 | 9 | $C_{24}H_{46}O_4$ | | |
| | | axillarenic acid | 398.3396 | 9 | $C_{24}H_{46}O_4$ | | |
| 445.7/477 | 445.7483 | | | | | 4.789 | 4.748 |
| 474.3/314 | 474.2608 | | | | | 4.399 | 4.114 |
| 475.3/314[a] | 475.2658 | gitoxigenin diacetate | 474.2618 | 6 | $C_{27}H_{38}O_7$ | 3.865 | 3.574 |
| | | diterpenoid EF-D | 474.2618 | 6 | $C_{27}H_{38}O_7$ | | |
| | | lucidenic acid L | 474.2618 | 6 | $C_{27}H_{38}O_7$ | | |
| | | lucidenic acid I | 474.2618 | 6 | $C_{27}H_{38}O_7$ | | |
| | | lucidenic acid B | 474.2618 | 6 | $C_{27}H_{38}O_7$ | | |
| | | 3alpha,7alpha,12alpha-trihydroxy-5alpha-cholan-24-yl sulfate | 474.2651 | 13 | $C_{24}H_{42}O_7S$ | | |
| 519.8/473 | 519.7647 | | | | | 3.318 | 2.860 |
| 534.8/465 | 534.7727 | | | | | 5.265 | 4.820 |
| 535.3/465 | 535.2742 | pyropheophorbide a | 534.2631 | 7 | $C_{33}H_{34}N_4O_3$ | 5.034 | 4.601 |
| | | GV 150013X | 534.2631 | 7 | $C_{33}H_{34}N_4O_3$ | | |
| | | pyrophaeophorbide a | 534.2631 | 7 | $C_{33}H_{34}N_4O_3$ | | |
| | | 7,8-dihydrovomifoliol 9-[rhamnosyl-(1->6)-glucoside] | 534.2676 | 1 | $C_{25}H_{42}O_{12}$ | | |
| | | 3-hydroxy-beta-ionol 3-[glucosyl-(1->6)-glucoside] | 534.2676 | 1 | $C_{25}H_{42}O_{12}$ | | |
| 560.4/479 | 560.4002 | gamma-chaconine | 559.3873 | 9 | $C_{33}H_{53}NO_6$ | 5.329 | 5.242 |
| 560.4/465 | 560.4003 | gamma-chaconine | 559.3873 | 10 | $C_{33}H_{53}NO_6$ | 5.329 | 5.242 |
| 561.4/465 | 561.4056 | | | | | 5.313 | 4.897 |
| 574.4/411[a] | 574.3774 | | | | | 3.269 | 2.457 |
| 706.5/478 | 706.4612 | | | | | 4.849 | 4.729 |
| 722.5/477 | 722.4535 | gamma2-solamarine | 721.4401 | 8 | $C_{39}H_{63}NO_{11}$ | 4.375 | 4.221 |
| | | beta-solanine | 721.4401 | 8 | $C_{39}H_{63}NO_{11}$ | | |
| 852.5/482 | 852.5217 | alpha-chaconine | 851.5031 | 13 | $C_{45}H_{73}NO_{14}$ | 6.046 | 5.962 |
| 853.0/481 | 853.0197 | | | | | 4.651 | 4.527 |
| 853.5/482 | 853.5263 | | | | | 5.751 | 5.659 |
| 854.5/482 | 854.5310 | | | | | 5.239 | 5.138 |
| 866.5/420 | 866.4962 | | | | | 4.110 | 3.494 |
| 867.5/420 | 867.4996 | | | | | 3.806 | 3.196 |
| 868.5/494 | 868.5111 | solamargine | 867.4980 | 6 | $C_{45}H_{73}NO_{15}$ | 4.941 | 4.843 |
| | | alpha-solanine | 867.4980 | 6 | $C_{45}H_{73}NO_{15}$ | | |
| | | beta-solamarine | 867.4980 | 6 | $C_{45}H_{73}NO_{15}$ | | |
| 868.5/457 | 868.5127 | solamargine | 867.4980 | 8 | $C_{45}H_{73}NO_{15}$ | 3.713 | 3.619 |
| | | alpha-solanine | 867.4980 | 8 | $C_{45}H_{73}NO_{15}$ | | |

**Table 4** (continued)

| | | | | | | Average BLUP[b] | |
|---|---|---|---|---|---|---|---|
| | | | | | | −13.825 | 19.250 |
| | | beta-solamarine | 867.4980 | 8 | $C_{45}H_{73}NO_{15}$ | | |
| 868.5/442[a] | 868.5128 | solamargine | 867.4980 | 8 | $C_{45}H_{73}NO_{15}$ | 3.650 | 4.005 |
| | | alpha-solanine | 867.4980 | 8 | $C_{45}H_{73}NO_{15}$ | | |
| | | beta-solamarine | 867.4980 | 8 | $C_{45}H_{73}NO_{15}$ | | |
| 868.5/477 | 868.5172 | solamargine | 867.4980 | 13 | $C_{45}H_{73}NO_{15}$ | 5.614 | 5.487 |
| | | alpha-solanine | 867.4980 | 13 | $C_{45}H_{73}NO_{15}$ | | |
| | | beta-solamarine | 867.4980 | 13 | $C_{45}H_{73}NO_{15}$ | | |
| 869.5/494 | 869.5162 | koryoginsenoside R1 | 868.5184 | 10 | $C_{46}H_{76}O_{15}$ | 4.569 | 4.467 |
| 869.5/477 | 869.5213 | koryoginsenoside R1 | 868.5184 | 10 | $C_{46}H_{76}O_{15}$ | 5.316 | 5.189 |
| 870.5/476 | 870.5258 | | | | | 4.770 | 4.640 |
| 1016.6/471 | 1016.5588 | delta5-demissine | 1015.5352 | 16 | $C_{50}H_{81}NO_{20}$ | 3.072 | 2.539 |
| 1017.6/470 | 1017.5630 | | | | | 2.662 | 2.098 |
| 1018.6/469 | 1018.5694 | demissine | 1017.5508 | 11 | $C_{50}H_{83}NO_{20}$ | 2.195 | 1.672 |
| 1046.6/464[a] | 1046.5711 | dehydrocommersonine | 1045.5458 | 17 | $C_{51}H_{83}NO_{21}$ | 5.114 | 4.400 |
| 1047.6/465 | 1047.5730 | | | | | 4.912 | 4.368 |
| 1048.6/464 | 1048.5792 | | | | | 4.465 | 3.956 |

[a] Selected by the USC algorithm for five-feature classifier

[b] There were 66 plants in the study that were classified as resistant (BLUP<0) and 48 classified as susceptible (BLUP >0). The average BLUP was calculated for resistant and susceptible plants

[c] The average $\log_{10}$ peak intensity for the feature over the 66 resistant plants

[d] The average $\log_{10}$ peak intensity for the feature over the 48 susceptible plants

## Supervised Machine Learning Classification Using KNN and SVM

KNN was also used to classify plants using all 651 features. Leave-one-out cross-validation was used for training the KNN algorithm meaning that in each round of cross-validation one sample is left out to do the classificiation. The numbers of errors in cross-validation of the training samples were higher than for the USC cross-validation for the 651 features (Table 2). However, the results from the separate set of 37 test plant using

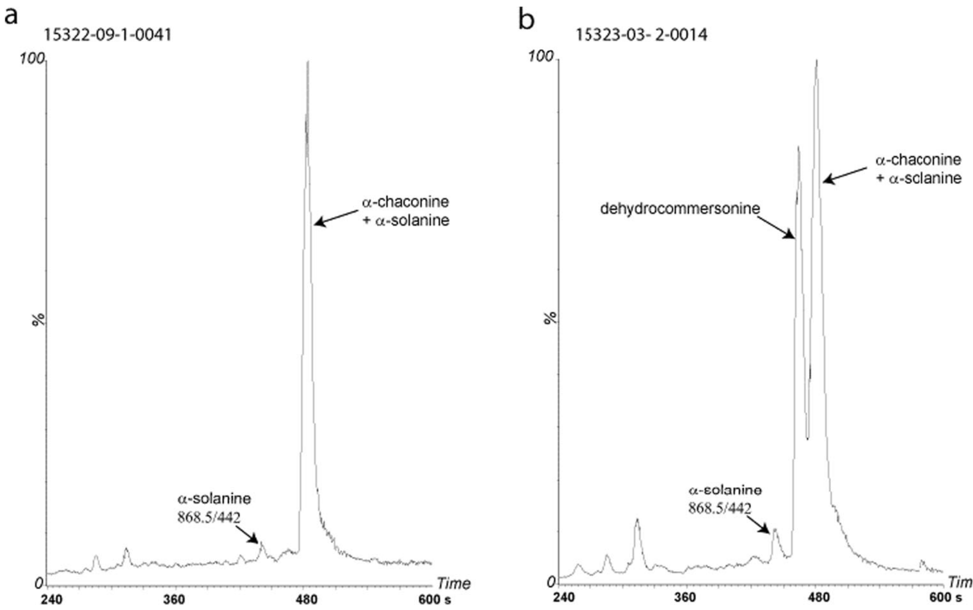

**Fig. 2** Total ion chromatograms for two plants a) 15322-09-1-0041 (base peak intensity 577566) with susceptbility to CPB and b) 15232-03-2-0014 (base peak intensity 1176298) with resistance to CPB. The base peak feature is 852.2/482 (alpha-chaconine) for both a) and b). 100 % peak intensity for each was set at the intensity of the base peak intensity

**Table 5** *T*-test of differences in glycoalkaloid between CPB resistant and susceptible plants

| Feature ID | glycoalkaloid | CPB resistance | Avg peak intensity[a] | t-statistic[b] | p-value[c] |
|---|---|---|---|---|---|
| 560.4/479 | gamma-chaconine | resistant<br>susceptible | 5.329<br>5.242 | 1.835 | 0.070 |
| 560.4/465 | gamma-chaconine | resistant<br>susceptible | 5.339<br>5.140 | 3.976 | 0.000** |
| 722.5/477 | beta-solanine | resistant<br>susceptible | 4.381<br>4.227 | 2.510 | 0.014* |
| 852.5/482 | alpha-chaconine | resistant<br>susceptible | 6.046<br>5.962 | 2.092 | 0.039* |
| 868.5/494 | alpha-solanine | resistant<br>susceptible | 4.941<br>4.843 | 1.986 | 0.050* |
| 868.5/457 | alpha-solanine | resistant<br>susceptible | 3.713<br>3.619 | 0.877 | 0.382 |
| 868.5/442 | alpha-solanine | resistant<br>susceptible | 3.650<br>4.005 | −2.624 | 0.010** |
| 868.5/477 | alpha-solanine | resistant<br>susceptible | 5.614<br>5.487 | 2.182 | 0.031* |
| 1016.6/471 | delta5-demissine | resistant<br>susceptible | 3.072<br>2.539 | 3.149 | 0.002** |
| 1018.6/469 | demissine | resistant<br>susceptible | 2.195<br>1.672 | 2.940 | 0.004** |
| 1046.6/464 | dehydrocommersonine | resistant<br>susceptible | 5.114<br>4.400 | 3.927 | 0.000** |
| 1046.6/464:868.5/442[d] | resistant<br>susceptible | 65.118<br>22.301 | 4.392 | 4.392 | 0.000** |
| 1046.6/464:total alpha-solanine[e] | resistant<br>susceptible | 0.307<br>0.161 | 4.318 | 4.318 | 0.000** |

[a] The average $\log_{10}$ peak intensity for the feature over the 66 resistant and 48 susceptible plants. In the lower part of the table the average of the ratio of the peak intensities is listed

[b] The null hypothesis tested $H_o$: resistant=susceptible feature peak intensities. The degrees of freedom was 112

[c] Significance at $p \leq 0.05$ was indicated by * and $p \leq 0.01$ by **

[d] The ratio of the two features was calculated for resistant and susceptible plants and tested for significant differences

[e] The feature peak intensities for all the features assigned to alpha-solanine were added. The ratio of the peak intensity for 1046.6/464 over the total of the alpha-solanine peak intensities was calculated for resistant and susceptible plants and tested for significant differences

KNN for classification showed low error rates compared with the other two algorithms (Table 3). The 35 high peak intensity features were also used for classification using KNN. When the 37 test plants were classified using the 35-feature KNN classifier, only a single error in classification was found resulting in an error rate of 2.7 % (Table 3). The five-feature classifier selected by the USC algorithm was also used for classification using KNN. There was also a low rate of classification error for the 37 test plants of 5.4 % (Table 3). The error rate in the cross-validation of training plants was higher than the classification error round in the separate set of 37 test plants for KNN. Moreover, the error rate with KNN classification was the lowest among the three algorithms tested.

Classification using SVM with 651 features with leave-one-out cross-validation was done. Error rates were the highest for SVM (Tables 2 and 3). The 37 test plants were classified using SVM and there was a 32.4 % error rate with 11 misclassification of susceptible and 1 of resistant (Table 3). Classification using the 35 high peak intensity features was also done. Cross-validation error of the training plants was similar to KNN, but the classification of the 37 test plants had the highest error rate of 56.8 % with 21 plants misclassified as susceptible (Table 3). The five-feature classifier selected by the USC algorithm was also used for SVM classification. There were high error rates for cross-validation of the 77 training plants with 21 mistakes in assignments (Table 2). The 37 test plants were classified using the five-feature SVM classifier and it had an error of 27 % (Table 3).

## Discussion

The goal this work was two-fold: first, to develop new germplasm resources for CPB resistance and second, to develop a more cost effective way to phenotype CPB resistance. Wild *Solanum* species are a resource for many different resistance traits for potato including CPB resistance. In this study, the wild species *S. oplocense* was intercrossed with *S. tuberosum*. $F_1$ generated were evaluated in field defoliation assays and were found to carry CPB resistance and further backcross clones also carried resistance. A challenge in potato breeding is phenotyping the CPB resistance trait. Quantifying insect feeding through scoring for defoliation in the field as a result of natural infestation is typically used. Alternatively, laboratory feeding assays can be done. Both methods are time consuming and laborious. In this study we investigate the feasibility of using metabolite markers to provide a prediction for CPB resistance. Several lines of evidence indicate that CPB resistance is dependent on foliar metabolite composition (Pelletier and King 1987; Rangarajan et al. 2000; Tai et al. 2014; Tingey 1984; Yencho and Tingey 1994). Discovery of metabolites conferring CPB resistance would enable selection of resistance using metabolite markers. Approaches to finding metabolite markers have included application of untargeted metabolomics (Fernie and Schauer 2009; Zabotina 2013). The advantage of the untargeted metabolomics approach is the large number of metabolites that can be screened at the same time. Strategies to find diagnostic metabolite markers include application of supervised machine learning methods. These techniques discover and identify patterns and relationships between hundreds of metabolites in a dataset from individuals that are classified into distinct groups (Kourou et al. 2015). The outcome is a prediction of the group an unknown individual belongs to using metabolite markers. These methods involve using a set a training data where each individual has a set of untargeted metabolite profiling data and a classification. Supervised machine learning classification algorithms were used successfully with untargeted metabolomics to develop classifiers of organic and conventional production for wheat (Kessler et al. 2015). In this study we applied supervised machine learning methods to classify *S. oplocense*-containing germplasm as resistant or susceptible to CPB based on foliar metabolites. The method developed used field defoliation by CPB as the criteria for classification as resistant or susceptible. However, foliage from greenhouse-grown plants as opposed to plants grown in the field were used for analysis, since environmental variability in glycoalkaloid production in field-grown plants was previously reported (Valcarcel et al. 2014). Additionally, there were practical advantages to screening plants propagated indoors for the breeding program. Plants grown indoors could be screened at any time during the year and germplasm propagated using in vitro tissue culture could be transferred to pots for indoor growth directly, whereas field growth required tuber production.

Three supervised machine learning classification methods were compared – USC, KNN and SVM. All three were able to classify plants as CPB resistant or susceptible, but with varying error. Overall, the KNN algorithm performed better than USC and SVM in that order. Another finding was that a smaller subset of metabolites can be as effective as or better than the entire LC-MS metabolite profile in classification. The five-feature classifier selected by the USC algorithm had error rates that were similar to or lower than a classifier based on the entire metabolite profile. The design of targeted assays for screening large numbers of clones is feasible with a small number of metabolites so it is desirable to identify metabolite profiles with few metabolites.

The five-feature classifier contained two glycoalkaloids previously associated with CPB resistance and susceptibility, dehydrocommersonine and alpha-solanine, respectively (Tai et al. 2014). This result was supportive of the biological relevance of the five-feature classifier. The other features include 188.1/285 which could be matched to deethylatrazine, 3-amino-2-napthoic acid or indoleacrylic acid. However, deethylatrazine is an environmental degradation product of the synthetic herbicide atrazine (Shipitalo and Owens 2003) and 3-amino-2-napthoic acid is also a synthetic compound (Allen and Bell 1942) indicating that they were not likely produced as a natural metabolite in foliage. Indoleacrylic acid, on the other hand, is a known plant growth regulator that functions similarly to auxin (Marklová 1999) and 188.1/285 was assigned to this compound. Another of the five features was 475.3/314 which could be matched with gitoxigenin diacetate, a synthetic acetate of a naturally occurring cardenolide (Hashimoto et al. 1986); diterpenoid EF-D, a plant isoterpenoid (Baxter et al. 1999) or three lucidenic acids, derived from the mushroom, *Ganoderma lucidum*, (Weng et al. 2007). The fungal-derived lucidenic acids were less likely to be metabolites of *Solanum* foliage and gitoxigenin diacetate was a synthetic product, indicating that diterpenoid EF-D, also known as 12-deoxy-phorbol-13-alpha-methylbutyrate-20-acetate, was the most likely identity of 475.3/314. Diterpenoids have been found by others to be effective anti-feedants against CPB (Bozov et al. 2014), which provides support for the assignment of diterpenoid EF-D to 475.3/314. There was little change in the average levels of 188.1/285 and 475.3/314 between resistant and susceptible plants. However, removal of these features from the classifier increased the error rates of prediction (data not shown). There were no *m*/*z* matches with 574.4/411, which was higher is resistant compared to susceptible plants.

The toxicity of glycoalkaloids to humans are a concern, therefore, selection to avoid high levels of glycoalkaloids is desirable. The five feature classifier included two features that were a match with glycoalkaloids dehydrocommersonine (1046.6/464) and alpha-solanine (868.5/442). However, the selection strategy for resistance will be for a change in the composition of glycoalkaloids to a higher ratio of 1046.6/464:868.5/442. This selection strategy will be compatible with selection of low overall levels of glycoalkaloids.

The study has demonstrated that the *S. oplocense* germplasm generated has CPB resistance. Additionally, five metabolites were found that can serve as markers for selection of

CPB resistance, which has advanced development of lower cost screening tools for CPB resistance in *S. oplocense*-carrying potato germplasm.

**Acknowledgments** The authors would like to thank Charlotte Davidson, Catherine Clark and Katherine Douglass who provided technical assistance. The work was funded by Agriculture and Agri-Food Canada Agricultural Bioproducts Innovation Program and the Developing Innovative Agri-products program.

## References

Allen, C.F.H., and A. Bell. 1942. 3-amino-2-naphthoic acid. *Organic Syntheses* 22: 19.

Alyokhin, A., M. Baker, D. Mota-Sanchez, G. Dively, and E. Grafius. 2008. Colorado potato beetle resistance to insecticides. *American Journal of Potato Research* 85(6): 395–413.

Baxter, H., J.B. Harborne, and G.P. Moss. 1999. *Phytochemical dictionary: a handbook of bioactive compounds from plants*. Philidelphia: Taylor & Francis.

Boiteau, G., Y. Pelletier, G. C. Misener, and G. Bernard. 1994. Development and evaluation of a plastic trench barrier for protection of potato from walking adult Colorado potato beetles (Coleoptera: Chrysomelidae). *Journal of Economic Entomology* 87: 1325–1331.

Bozov, P.I., T.A. Vasileva, and I.N. Iliev. 2014. Structure and antifeedant activity relationship of neo-clerodane diterpenes against Colorado potato beetle larvae. *Chemistry of Natural Compounds* 50(4): 762–764.

Brown, M.P.S., W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares, and D. Haussler. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences* 97(1): 262–267.

Carter, C.D., T.J. Gianfagna, and J.N. Sacalis. 1989. Sesquiterpenes in glandular trichomes of a wild tomato species and toxicity to the Colorado potato beetle. *Journal of Agricultural and Food Chemistry* 37(5): 1425–1428.

Fernie, A.R., and N. Schauer. 2009. Metabolomics-assisted breeding: a viable option for crop improvement? *Trends in Genetics* 25(1): 39–48.

Flanders, K., J. Hawkes, E. Radcliffe, and F. Lauer. 1992. Insect resistance in potatoes: sources, evolutionary relationships, morphological and chemical defenses, and ecogeographical associations. *Euphytica* 61(2): 83–111.

Hashimoto, T., H. Rathore, D. Satoh, J.F. Griffin, A.H.L. From, K. Ahmed, D.S. Fullerton, and G. Hong. 1986. Cardiac glycosides. 6. Gitoxigenin C16 acetates, formates, methoxycarbonates, and digitoxosides. Synthesis and Na+, K+−ATPase inhibitory activities. *Journal of Medicinal Chemistry* 29(6): 997–1003.

Henderson, C.R. 1984. *Applications of linear models in animal breeding*. Guelph: University of Guelph.

Jansky, S.H., R. Simon, and D.M. Spooner. 2009. A test of taxonomic predictivity: resistance to the Colorado potato beetle in wild relatives of cultivated potato. *Journal of Economic Entomology* 102(1): 422–431.

Kessler, N., A. Bonte, S.P. Albaum, P. Mäder, M. Messmer, A. Goesmann, K. Niehaus, G. Langenkämper, and T.W. Nattkemper. 2015. Learning to classify organic and conventional wheat – A machine learning driven approach using the MeltDB 2.0 metabolomics analysis platform. *Frontiers in Bioengineering and Biotechnology* 3: 1–10.

King, R. R., Y. Pelletier, R. P. Singh, and L. A. Calhoun. 1986. 3,4-Di-O-isobutyryl-6-O-caprylsucrose: the major component of a novel sucrose ester comples from the type B glandular trichomers of *Solanum berthaultii* Hawkes (PI473340). *Journal of the Chemical Society*, *Chemical Communication* 14: 1078–1079.

Kourou, K., T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, and D.I. Fotiadis. 2015. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* 13: 8–17.

Lynch, M., and B. Walsh. 1998. *Genetics and analysis of quantitative traits*, 990. Sunderland: Sinauer Associates Inc.

Marklová, E. 1999. Where does indolylacrylic acid come from. *Amino Acids* 17(4): 401–413.

McLeod, C. and J.H. Tolman. (eds.) 1987. Evaluation of losses in potatoes. In: *Potato Pest Management in Canada-Lutte contre les parasites de la pomme de terre au Canada.* Boiteau, G., Singh, R. and Parry R. Agriculture and Agri-Food Canada. Fredericton, NB, Canada, pp. 363–373

Pelletier, Y. 2007. Level and genetic variability of resistance to the Colorado potato beetle (leptinotarsa decemlineata (say)) in wild *Solanum* species. *American Journal of Potato Research* 84(2): 143–148.

Pelletier, Y. and R.R. King. 1987. Semiochemicals and Potato pests: review and perspective for crop protection. In *Potato Pest Management in Canada*, eds. Boiteau, G., R. P. Singh, R. H. Parry. Proceedings of a Symposium on Improving Potato Pest Protection, Fredericton, NB, Canada.

Pelletier, Y., and Z. Smilowitz. 1990. Effect of trichome B exudate of *Solanum berthaultii* Hawkes on consumption by the Colorado potato beetle, *Leptinotarsa decemlineata* (Say). *Journal of Chemical Ecology* 16(5): 1547–1555.

Pelletier, Y., and G.C.C. Tai. 2001. Genotypic variability and mode of action of Colorado potato beetle (coleoptera: chrysomelidae) resistance in seven *Solanum* species. *Journal of Economic Entomology* 94(2): 572–578.

Pelletier, Y., C. Clark, and G.C. Tai. 2001. Resistance of three wild tuber-bearing potatoes to the Colorado potato beetle. *Entomologia Experimentalis et Applicata* 100(1): 31–41.

Pelletier, Y., F.G. Horgan, and J. Pompon. 2011. Potato resistance to insects. *The Americas Journal of Plant Science and Biotechnology* 5(Special issue1): 37–52.

Rangarajan, A., A.R. Miller, and R.E. Veilleux. 2000. Leptine glycoalkaloids reduce feeding by Colorado potato beetle in diploid *Solanum* sp. Hybrids. *Journal of the American Society for Horticultural Science* 125(6): 689–693.

Sanford, L., S. Kowalski, C. Ronning, and K. Deahl. 1998. Leptines and other glycoalkaloids in tetraploid *Solanum tuberosum* x *Solanum chaocoense F1 & F2* Hybrid and Backcross Families. *American Journal of Potato Research* 75(4): 167–172.

Shipitalo, M.J., and L.B. Owens. 2003. Atrazine, deethylatrazine, and deisopropylatrazine in surface runoff from conservation tilled watersheds. *Environmental Science & Technology* 37(5): 944–950.

Stemeroff, M., and J.A. George. 1983. *The benefits and costs of controlling destructive insects on onions, apples and potatoes in Canada, 1960–80*. Ottawa: Entomological Society of Canada.

Szendrei, Z., E. Grafius, A. Byrne, and A. Ziegler. 2012. Resistance to neonicotinoid insecticides in field populations of the Colorado potato beetle (Coleoptera: Chrysomelidae). *Pest Management Science* 68(6): 941–946.

Tai, G.C.C., A.M. Murphy, and X. Xiong. 2009. Investigation of long-term field experiments on response of breeding lines to common scab in a potato breeding program. *Euphytica* 167(1): 69–76.

Tai, H.H., K. Worrall, Y. Pelletier, D. De Koeyer, and L.A. Calhoun. 2014. Comparative metabolite profiling of *Solanum tuberosum* against six wild *Solanum* species with Colorado potato beetle resistance. *Journal of Agricultural and Food Chemistry* 62(36): 9043–9055.

Theilhaber, J., T. Connolly, S. Roman-Roman, S. Bushnell, A. Jackson, K. Call, T. Garcia, and R. Baron. 2002. Finding genes in the C2C12 osteogenic pathway by k-nearest-neighbor classification of expression data. *Genome Research* 12(1): 165–176.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences* 99(10): 6567–6572.

Tingey, W. 1984. Glycoalkaloids as pest resistance factors. *American Journal of Potato Research* 61(3): 157–167.

Tingey, W.M., and G.C. Yencho. 1994. Insect resistance in potato: a decade of progress. In *Advances in potato pest biology and management*, ed. G.W. Zehnder, R.K. Jansson, M.L. Powelson, and K.V. Raman, 405–425. St. Paul: APS Press.

Valcarcel, J., K. Reilly, M. Gaffney, and N. O'Brien. 2014. Effect of genotype and environment on the glycoalkaloid content of rare, heritage, and commercial potato varieties. *Journal of Food Science* 79(5): T1039–T1048.

Vinayavekhin, N. and A. Saghatelian. 2010. Untargeted metabolomics. Curr Protoc Mol Biol Chapter 30:Unit 30 1 1–24.

Weng, C.J., C.F. Chau, K.D. Chen, D.H. Chen, and G.C. Yen. 2007. The anti-invasive effect of lucidenic acids isolated from a new *Ganoderma lucidum* strain. *Molecular Nutrition & Food Research* 51(12): 1472–1477.

Yencho, G.C., and W.M. Tingey. 1994. Glandular trichomes of *Solanum berthaultii* alter host preference of the Colorado potato Beetle, *Leptinotarsa decemlineata*. *Entomologia Experimentalis et Applicata* 70(3): 217–225.

Yencho, G.C., S. Kowalski, G. Kennedy, and L. Sanford. 2000. Segregation of leptine glycoalkaloids and resistance to Colorado potato beetle (*Leptinotarsa decemlineata* (Say)) in F2 *Solanum tuberosum* (4×)×*S. chacoense* (4×) potato progenies. *American Journal of Potato Research* 77(3): 167–178.

Yeung, K., and R. Bumgarner. 2003. Multiclass classification of microarray data with repeated measurements: application to cancer. *Genome Biology* 4(12): R83.

Zabotina, O.A. 2013. Metabolite-based biomarkers for plant genetics and breeding. In *Diagnostics in plant breeding*, ed. T. Lübberstedt and R.K. Varshney, 281–309. Dordrecht: Springer.