Terpenes associated with resistance against the gall wasp, *Leptocybe invasa*, in *Eucalyptus grandis*

Sanushka Naidoo[1], Nanette Christie[1], Juan José Acosta[2], Makobatjatji M. Mphahlele[3], Kitt G. Payn[3], Alexander A. Myburg[1], Carsten Külheim[4]

[1]Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private bag x20, Pretoria, 0028, South Africa, [2]Camcore, Department of Forestry and Environmental Resources, University of North Carolina, Raleigh NC 27695-8008, USA, [3]Mondi Forests, Trahar Technology Centre, P.O. Box 12, Hilton, 3245, South Africa, [4]Research School of Biology, Australian National University, 46 Sullivans Creek Road, Canberra, 2601, ACT, Australia

*Correspondence: S. Naidoo. Email: Sanushka.Naidoo@up.ac.za; Tel: +27 12 420 4974

Word count: 7740

Key words: GC-MS, NIR, plant defence, attractant, repellent

Abstract

Leptocybe invasa is an insect pest causing gall formation on oviposited shoot tips and leaves of *Eucalyptus* trees leading to leaf deformation, stunting and death in severe cases. We previously observed different constitutive and induced terpenes, plant specialised metabolites that may act as attractants or repellents to insects, in a resistant and susceptible clone of *Eucalyptus* challenged with *L. invasa*. We tested the hypothesis that specific terpenes are associated with pest resistance in a *Eucalyptus grandis* half-sib population. Insect damage was scored over two infestation cycles and leaves harvested for near-infrared reflectance (NIR) and terpene measurements. We used Bayesian model averaging (BMA) for terpene selection and obtained partial least squares NIR models to predict terpene content and *L. invasa* infestation damage. In our optimal model, 29% of the phenotypic variation could be explained by seven terpenes and the monoterpene combination, limonene, α-terpineol and 1,8-cineole, could be predicted with an NIR prediction ability of 0.67. BMA supported α-pinene, γ-terpinene and iso-pinocarveol as important for predicting *L. invasa* infestation. Susceptibility was associated with increased γ-terpinene and α-pinene, which may act as a pest attractant, while reduced susceptibility was associated with iso-pinocarveol, which may act to recruit parasitoids or have direct toxic effects.

Introduction

Eucalyptus species, native to Australia and the surrounding islands, have been adopted as plantation species in various parts of the world for timber, pulp and paper production (Gomes et al., 2014; Javaregowda and Prabhu, 2010). Some *Eucalyptus* species, used in essential oil production, have been earmarked as a potential source of speciality biofuels (Iqbal et al., 2011; Mewalal et al., 2017). The high economic value of *Eucalyptus* plantations is jeopardised by various pests and diseases (Wingfield et al., 2008). One such pest is the blue gum chalcid, *Leptocybe invasa* Fisher & La Salle (Hymenoptera: Eulophidae) that causes significant losses to eucalypts (Chang et al., 2012; Javaregowda and Prabhu, 2010; Mendel et al., 2004).

The *L. invasa* gall wasp was first reported in Israel in 2000 (Mendel et al., 2004) and since then, the occurrence of *L. invasa* is concomitant with *Eucalyptus* production having been recorded in countries in the Mediterranean basin, Africa, America and Asia (Mutitu, 2003; Nyeko, 2005; Wiley and Skelley, 2008; Zhu et al., 2012). In South Africa, the insect pest was first reported in 2007 (Wylie and Speight, 2012) and has devastating impact on nursery seedlings and young plantations.

The adult female wasp, approximately 1.2 mm in length, oviposits on the shoot tips, petioles, midribs and stems of young *Eucalyptus* trees, seedlings and coppice growth (Mendel et al., 2004; Quang Thu et al., 2009; Zhu et al., 2012). Galls begin to develop two weeks after oviposition. The bump shaped galls become pink to red in colour and coalesce. After a period of five months,

the adult wasps emerge from the gall to re-infest the *Eucalyptus* plants (Mendel et al., 2004; Quang Thu et al., 2009; Zhu et al., 2012). Symptoms on the host range from evidence of oviposition with no gall development, contortion of leaves and shoot deformation, loss of apical dominance, die back (Kelly et al., 2012; Mendel et al., 2004) and in severe cases, death (Nyeko et al., 2009; Wylie and Speight, 2012).

Despite investigation into chemical and silvicultural control strategies, biological control is identified as a promising means of limiting this pest (Dittrich-Schröder et al., 2014; Kim et al., 2008; Kulkarni, 2010). This practice, coupled with the use of resistant genotypes, would be useful to limit the incidence of *L. invasa* as *Eucalyptus* species, hybrids and clones show marked variation in resistance against *L. invasa* (Dittrich-Schröder et al., 2012; Durand et al., 2011; Nyeko and Nakabonge, 2008; Quang Thu et al., 2009). For example, *E. nitens* x *E. grandis* and *E. grandis* x *E. camaldulensis* hybrids showed high rates of infestation while pure species *E. dunni*, *E. nitens*, *E. grandis*, *E. urophylla* and the hybrid *E. saligna* x *E. urophylla* were considered resistant in a greenhouse experiment (Dittrich-Schröder et al., 2012). In field, significant variation for resistance against the insect pest is observed within some species e.g. *E. grandis* (Arnulf Kanzler personal communication, Sappi Shaw Research Center, Hilton, KZN).

Terpenes are a large group of plant specialised metabolites implicated in various ecological interactions (Moore et al., 2014). They are a highly diverse group with over 20,000 known compounds (Degenhardt and Gershenzon, 2003) and can be separated through the number of

isoprene units they contain into hemi- (C5), mono- (C10), sesqui- (C15), di- (C20), tri- (C30) and tetra-terpenes (C40) (Dudareva et al., 2005; Webb et al., 2014). Examples of ecological interactions include direct plant defence through toxic effect on herbivores (Keefover-Ring et al., 2009), effects on reproduction of insect herbivores (Edwards et al., 1990, 1993; Morrow and Fox, 1980; Stone and Bacon, 1994), indirect defence against herbivores through the attraction of herbivore parasites (De Moraes et al., 1998; Giamakis et al., 2001; Turlings et al., 1995), cues that indicate the presence of other toxic constituents (Lawler et al., 1999), allelopathic agents (Alves et al., 2004), and mediators of resistance to fungal infection (Eyles et al., 2003).

Species of the genus *Eucalyptus* typically contain large amounts of foliar essential oils which are dominated by mono- and sesqui-terpenes (Coppen, 2003), and which are stored in schizogenous cavities (Carr and Carr, 1970). There is a lot of within species variation of terpenes in eucalypts, both quantitatively (Kainer et al., 2017; Wallis et al., 2011) and qualitatively (Padovan et al., 2014). Among the most common terpenes in eucalypts are the monoterpenes α-pinene and 1,8-cineole, also known as eucalyptol (Padovan et al., 2014), however individuals with over 100 different foliar terpenes have been found (Wong et al., 2017). This variety is the result of three effects: 1) Eucalypts have the largest gene family of terpene synthases known to date, which are responsible for the final biosynthetic step in terpene production (Külheim et al., 2015) , 2) these enzymes often produce multiple products (Külheim et al., 2015; Padovan et al., 2017; Schnee et al., 2002), and 3) terpenes are often further modified by enzymes such as cytochrome P450 (Pateraki et al., 2015) and glycosyl transferases (Rivas et al., 2013) leading to the vast array of terpenes found in this genus. Most constitutive terpenes are produced during ontogenesis of

leaves, exported into the extracellular cavities and are believed to be stable there for the lifetime of the leaf (Carr and Carr, 1970). Previous studies have found no indication of induction of essential oils in eucalypts by either wounding or application of methyl jasmonate (Henery et al., 2008). We have recently discovered that *L. invasa* oviposition and larvae development leads to changes in the profile of essential oils in clones of both resistant *E. grandis* (TAG5) and susceptible *E. camaldulensis* x *E. grandis* (GC540) (Oates et al., 2015). Compared to susceptible individuals, resistant individuals had approximately three times higher constitutive levels of α-pinene and less than half the amount of 1,8-cineole. Seven days post oviposition, leaves of susceptible plants contained significantly lower amounts of 1,8-cineole and α-terpinolene (Oates et al., 2015). The terpene content was concordant with changes in expression of genes involved in terpene biosynthetic pathways. Such observations suggest that specific terpene profiles or individual terpenes may be associated with resistance against the insect pest *L. invasa.*

Apart from terpenes, induced responses due to oviposition included other responses such as phytohormones in the two genotypes (Oates et al., 2015). This suggests that other chemicals may also play a role in the defence against the insect pest. Although terpene content is typically determined by Gas Chromatography-Mass Spectroscopy, near-infrared reflectance (NIR) spectra provide an indication of various chemicals including the terpenes. NIR spectra and modeling were used to predict the 1,8-cineole content in extracted Eucalyptus oil with 0.899 accuracy (Wilson et al., 2001) and foliar 1,8-cineole proportion in *Melaleuca cajuputi* tea tree oil with an accuracy of 0.92 (Schimleck and Rimbawanto, 2003). In the latter example, total foliar terpene

content was also estimated based on NIR modeling to an accuracy of 0.65 (Schimleck and Rimbawanto, 2003).

We hypothesised that resistance to *L. invasa* is attributed to chemical variation, and terpenes in particular, in *E. grandis*. The aim of this study was to determine NIR and terpene profiles associated with resistance against *L. invasa* in *E. grandis*, which could be adopted as a tool to predict resistant genotypes and to identify terpenes for further characterisation in this interaction.

Materials and Methods

Study sites

An *E. grandis* progeny trial series was established on three coastal sites in KwaZulu Natal, South Africa forming part of Mondi's tree breeding program. The three sites, namely Siya Qubeka (SQF), Mtunzini (MTZ) and Nyalazi (NYL), comprised 126 half-sib families planted in single-tree plots with 15 blocks (replications) per site in a randomised complete block design. Site environmental data are indicated in Supplementary Table 1.

Scoring of host susceptibility

The *E. grandis* trials were established in August 2012. In October 2013, when the trees were 14-months old, trees were inspected for *L. invasa* infestation. Symptoms were scored visually using

the following scale: 0- not infested, 1- infested showing evidence of oviposition but no gall development, 2- infested with galls on leaves, mid-ribs or petioles and 3- stunting and lethal gall formation. Each tree was thus categorised as either 0, 1, 2 or 3. This first round of scoring was referred to as *L. invasa* screening 1 (LS1) and included all trees within each trial. Supplementary Figure 1 shows images of representative scores. In October 2014, the trees sampled for NIR and terpene analyses were scored a second time using the same scoring regime. *L. invasa* screening 2 (LS2) was then calculated as the sum of LS1 and the second score for this sub-population.

Tissue sampling for NIR and Terpene analyses

We sampled a sub-population, comprising 61 half-sib families (180 trees from SQF, 159 trees from MTZ and 152 trees from NYL). Trees were selected within each family, where uninfested leaves from trees scored as 0, and infested leaves from trees scored as 1, 2 and 3 were harvested. From each tree a total of 3-5 mature leaves, consistently sampled from the equivalent position from a side-branch on the North side of the tree, were punched with a 1 cm cork borer and leaf disks collected into pre-weighed vials containing 5 ml of (99.7%) ethanol with tetradecane as internal standard (0.25 $mg.l^{-1}$). In addition, 5-6 leaves were collected in paper bags, their weight measured on the collection day, dried in 50ºC for three days and their dry weight recorded. Dried leaves were ground in a Foss Cyclotec 1093 mill (Foss, Höganäs, Sweden) and passed through a 1 mm sieve for NIR measurement.

Calculation of Breeding Values

Estimates of genetic parameters for *L. invasa* screenings were calculated using PROC MIX in SAS® (SAS Institute, Cary, USA). The statistical model for the analyses was as follows:

$y_{ijkl} = \mu + S_i + B(S)j(i) + F_k + FS_{ik} + E_{ijkl}$

where

y_{ijkl} is the l^{th} observation of the j^{th} block within the i^{th} site for the k^{th} family;

μ is the overall mean;

S_i is the fixed effect of the i^{th} site;

$B(S)j(i)$ is the fixed effect of the j^{th} block within the i^{th} site;

F_k is the random effect across sites of the k^{th} family = σ^2_f;

FS_{ik} is the random k^{th} family by i^{th} site interaction effect = σ^2_{fs}; and

E_{ijkl} is the error term = σ^2_e.

Family breeding values were obtained from the family best linear unbiased prediction estimates. Within family gain was calculated as: within family heritability multiplied by within family deviation. The individual breeding values (IBV), or tree gain, were calculated as family breeding values + within family gain.

Phenotypic variance was estimated as:

$$\sigma^2_p = \sigma^2_f + \sigma^2_{fs} + \sigma^2_e$$

Narrow-sense heritability was estimated as:

$$h^2 = 3\,\sigma^2_f / \sigma^2_p$$

The coefficient of relationship was assumed to be 0.33 instead of 0.25 for half-sib analysis because there is the possibility that some of open-pollinated families were not truly half-sibs, but contained some full-sibs (Squillace, 1974). Thus a coefficient of 3 instead of 4 was multiplied by the family variance in the calculation of heritability.

For each pair of trial sites, the estimates of Type B genetic correlations (r_{Bg}) were calculated as follows:

$$r_{Bg} = \sigma^2_f / (\sigma^2_f + \sigma^2_{fs})$$

Type B correlations measure the genetic correlation between the same trait expressed on two or more sites (Burdon, 1977). The parameter may range between 0 and 1 with an estimate approaching 1 giving an indication of very high correlation between family behaviour on the two sites and thus no genotype by environment interaction (GxE). Conversely a figure approaching zero suggests a high level of GxE.

Terpene measurements

Ethanol extracts of leaf tissue were separated by gas chromatography and detected by mass spectroscopy as described by Oates et al. (2015). An Agilent 6890 GC/MS using an Alltech AT-35 (35% phenyl, 65% dimethylpolyoxylane) column (Alltech, Wilmington, DE) was used with Helium as the carrier gas. The column was 60 m long with an internal diameter of 0.25 mm and with a stationary phase film thickness of 0.25 µm. The temperature regime consisted of: 100°C for 5 min, ramping to 200°C at 20°C min^{-1}, a ramp to 250°C at 5°C min^{-1}, with a hold of 250°C for 4 min. The total elution time was 25 minutes. The separate components were detected using an FID and an Agilent 5973 Mass Spectrometer dual setup through an SGE MS/FID splitter. The National Institute of Standards and Technology library (Agilent Technologies, Deerfield, IL) reference spectra enabled the identification of peaks with verification of major peaks through comparison to authentic standards. The area under each peak was determined with MSD Chemstation Data Analysis (Agilent Technologies). A relative concentration for each terpene was calculated relative to the internal standard, tetradecane. The fresh- to dry weight ratio and terpene concentrations were calculated relative to dry weight for each sample.

Near-infrared reflectance (NIR) spectroscopy

Five to six dried *E. grandis* leaves, collected in field as described above for terpene analysis, were ground to a fine powder and scanned on a desktop NIR Foss spectrophotometer (Foss Rapid Content Analyzer XM-1100), which measures absorbance of each sample between 1100

and 2498 nanometers with 2 nanometers increments. The NIR measures were repeated for each sample and averaged.

Statistical Analyses

Modeling NIR spectra to *Leptocybe* scores

We developed a programmable analysis pipeline in R (R Core Team, 2016) to process NIR spectral data and to fit chemometric models. The process of building NIR models involves the use of mathematical pre-treatments (transformations) applied to the NIR spectra. The objective of applying those transformations is to remove the scattering of diffuse reflections associated with sample particle size from the spectra to improve the subsequent regression. The most widely used transformation techniques can be divided into two categories: 1) scatter-correction methods and 2) spectral derivatives (Rinnan et al., 2009). In this study, we corrected the spectra using multiplicative scatter correction, standard normal variate and detrend from the scatter-correction methods; and a second derivative of Savitzky-Golay smoothing with two different window sizes of 5 and 7 points from the spectral derivatives methods. Additionally, we combined transformations by pairs applying scattering correction methods prior to differentiation. Pre-processing of our NIR spectral data were done using the R packages "ChemometricsWithR" (Wehrens, 2011) and "Prospectr" (Stevens and Ramirez-Lopez, 2013), we generated as outcome a total of 12 datasets of predictor variables including the raw spectra (Supplementary Table 2A).

Local outliers factors were calculated for all observations on each spectral database and used to identify outliers based on density and distance (Breunig et al., 2000). Individuals with local outliers factors values greater than 2 were excluded from the analysis, using a local outliers factors algorithm implemented in the R package "DMwR" (Torgo, 2015). The percentage of individuals classified as outliers for each set of models are given in Supplementary Table 2B. Transformed and outlier free databases were used to develop the NIR prediction models for LS1, LS2 and IBV. For this purpose, we used partial least squares implemented in the R-package "pls" (Mevik and Wehrens, 2007). Two modeling scenarios were contemplated: first, we grouped the observations by site and fitted models for each site respectively; and second, we used individual NIR spectra to fit models across all sites. For all scenarios, we evaluated the performance of our models using leave-one-out cross-validation. Desirable partial least squares NIR models are the ones that (i) maximize the coefficient of determination (R^2), (ii) maximize the percentage of the variance explained for X and Y on the training population (ExpVar_Y and ExpVar_X), (iii) minimize the standard errors of cross-validation: root mean squared error of prediction (RMSEP) and (iv) have a small number of latent variables (projection factors).

Modeling terpenes to *Leptocybe* scores

Bayesian model selection (Raftery, 1995) was performed in R, using the "bicreg" function in the Bayesian Model Averaging (BMA) package (Raftery et al., 2017), to identify which of the 48 measured terpenes (predictor variables) were the most important for predicting *L. invasa* infestation (Supplementary Table 3A). We also considered the sums of certain groups of terpenes as possible predictor variables (Supplementary Table 3B). Terpenes were combined as a result of

biological motivation or high pairwise correlations ($r > 0.6$). Biological motivation was based on either (a) shared intermediate carbocation (biosynthetically related through same intermediate precursor) or (b) the fact that terpene X is a precursor of terpene Y (biosynthetically related by 'descent'); see Keszei et al., (2008) Figure 3A.

Instead of using stepwise variable selection to choose candidate covariates, BMA accounts for uncertainty in variable selection by averaging over the best models. The Bayesian information criterion was used as a criterion for model selection and to estimate the posterior probability of a given model. Terpene variable selection was performed for the three dependent variables, LS1, LS2 and IBV, respectively, using the same two modeling scenarios mentioned above (firstly models were fit per site and secondly across all sites).

To test whether the BMA model parameters were consistent, we performed leave-one-out cross-validation: the BMA analysis was repeated *n* times, with *n* the number of individuals in the sample. For each of these different training data sets (in accordance with the leave-one-out cross-validation strategy; the data of a different individual was excluded per iteration), the model with the lowest Bayesian information criterion was used to estimate model coefficients, whereafter the *L. invasa* screening value of the excluded individual was predicted using the estimated model coefficients. Finally, a leave-one-out cross-validation R^2 value was calculated by correlating the predicted with the observed *L. invasa* screening values.

Modeling NIR spectra to terpene scores

We build terpene composition models for samples as a function of their NIR spectrum, following the same steps described above (Transformation to the spectral data, outlier identification and partial least squares modeling). Terpene models were performed only at site SQF (site at which we found the best models, see Table 1 and Supplementary Table 5) and for the subset of terpenes that we found were the most important for predicting *L. invasa* infestation. We also considered the same combinations of terpenes mentioned in the previous section (Supplementary Table 3B) for fitting partial least squares models.

Results

Leptocybe invasa infestation

The distribution of *L. invasa* screening 1 (LS1) across the three sites are indicated in Figure 1A. The sub-population was sampled for terpene and NIR measurements and the distribution of LS1, LS2 and IBV within this sub-population is shown in parts B, C and D of Figure 1. The heritability values for the *E. grandis* full population (126 families for LS1) and sub-population (61 families for LS2) at each site are indicated in Supplementary Table 4. The type B genetic correlations for LS1, for each pairwise combination of the sites, were 0.71 (MTZ:NYL), 0.84 (MTZ:SQF) and 0.74 (SQF:NYL). The type B genetic correlation for the three sites combined was 0.77, suggesting there was a relatively low level of GxE with little change in family ranking

between the sites. GxE for the sub-population could not be determined due to the smaller number of individuals per family (491 individuals across three sites and 61 families).

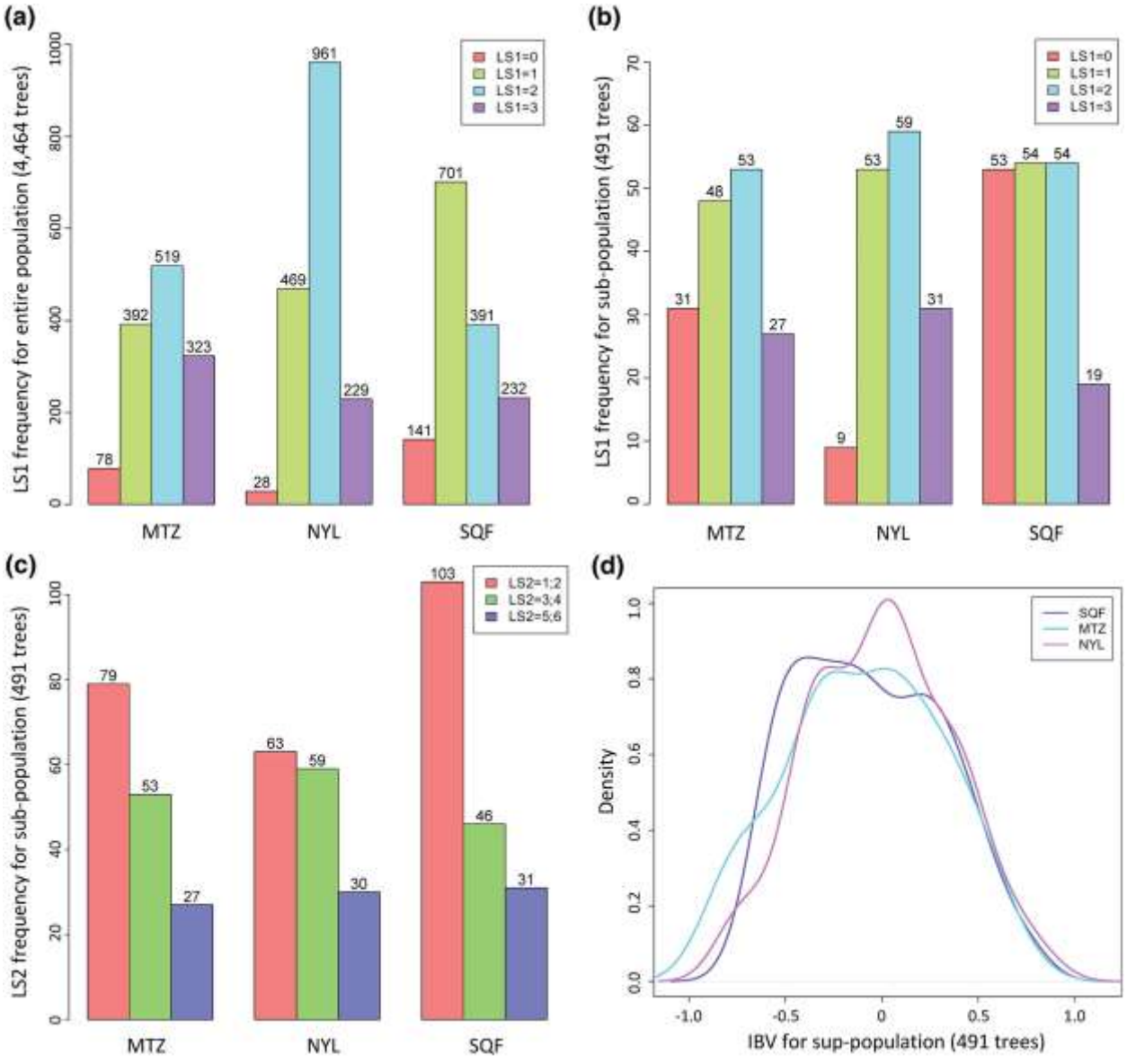


Figure 1. Distribution of *Leptocybe invasa* screenings (LS1 and LS2) and individual breeding values (IBV) per site. (A) Distribution of the LS1 scores for the entire population (4,464 trees) represented across three sites (red: LS1=0, green: LS1=1, blue: LS1=2, purple: LS1=3). (B) Distribution of the LS1 scores of the sub-population (491 trees) that were sampled for terpene and near-infrared reflectance measurements represented across three sites (red:

LS1=0, green: LS1=1, blue: LS1=2, purple: LS1=3). (C) Distribution of the LS2 scores of the sub-population represented across three sites (red: LS2=1-2, green: LS2=3-4, blue: LS2=5-6). (D) Distribution of individual breeding values (IBV) based on the LS2 scores for the sub-population per site (purple: Siya Qubeka (SQF) site, blue: Mtunzini (MTZ) site, pink: Nyalazi (NYL) site).

Table 1. The best partial least squares models, based on near-infrared reflectance (NIR) data, for *Leptocybe invasa* screenings (LS1, LS2) and individual breeding values (IBV) at the Siya Qubeka (SQF) site and across the three sites.

Site	Variable	Dataset[a]	Factors	RMSEP[b]	ExpVar_Y[c]	ExpVar_X[d]	R^2_Validation	r[e]
SQF	LS1	NIR	7	0.726	52.77	99.44	0.462	0.680
SQF	**LS2**	**SNV**	**8**	**1.017**	**68.83**	**99.06**	**0.622**	**0.789**
SQF	IBV	SNV	6	0.272	53.57	95.43	0.475	0.689
All sites	LS1	MSC	15	0.845	33.44	99.89	0.248	0.498
All sites	LS2	MSC	15	1.377	27.38	99.89	0.166	0.407
All sites	IBV	MSC	9	0.345	25.78	99.37	0.196	0.442

[a] Predictor variable dataset name after corresponding pre-processing technique was applied to the NIR spectra (acronyms are explained in Supplementary Table 2A).

[b] Root mean square error of the prediction.

[c] Percentage of the Y variable that is accounted in the model.

[d] Percentage of the NIR spectral data that is accounted in the model.

[e] Prediction ability: correlation between observed and NIR predicted values.

NIR to predict *L. invasa* infestation

NIR spectra were used to predict the level of infestation by *L. invasa*. The changes to the spectra after applying mathematical transformations are depicted in Supplementary Figure 2. Model differences were observed across sites. The best partial least squares regression models were obtained for the SQF site (Table 1), with R^2 ranging between 0.462 and 0.622. Table 1 also gives the best models across all sites, with R^2 values ranging between 0.166 and 0.248. Selected models for MTZ and NYL sites had low R^2 values ranging between 0.026 and 0.115 (Supplementary Table 5) and a low percentage of the variation in the Y variables were explained (9.30 - 16.15%).

To select the best model for each variable for each site, we compared the model statistics under each transformed and outlier-free database, and selected the one that (1) maximized the coefficient of determination (R^2_validation), the prediction ability (r) and the percentage of the variance explained for X and Y on the training population (ExpVar_Y and ExpVar_X), (2) minimize the standard errors cross-validation: root mean squared error of prediction (RMSEP) and (3) have a small number of projection factors. Table 2 shows all model statistics that were obtained for each dataset when modeling LS2 for the SQF site. For this case, a standard normal variate transformed dataset gives the best model. Model diagnostic plots were also created for each dataset and were used to select the number of latent variables (factors) of each model. Note that for LS2 under the standard normal variate dataset (model diagnostic plots in Figure 2), 8 factors give the highest R^2 and the lowest root mean squared error of prediction. NIR models under the second scenario (across all sites) did not perform well.

Table 2. Partial least squares models, based on near-infrared reflectance (NIR) data, for *Leptocybe invasa* screening 2 (LS2) at the Siya Qubeka (SQF) site.

Dataset[a]	Factors	RMSEP[b]	ExpVar_Y[c]	ExpVar_X[d]	R^2_Validation	r[e]
SNV	**8**	**1.017**	**68.83**	**99.06**	**0.622**	**0.789**
MSC	8	1.022	68.44	99.09	0.619	0.787
DT	5	1.075	62.68	93.66	0.577	0.76
SG5	5	1.079	67.33	84.29	0.574	0.757
SG7	5	1.073	65.94	86.37	0.578	0.76
SNV_SG5	5	1.084	68.02	84.51	0.569	0.755
SNV_SG7	5	1.077	66.55	86.66	0.575	0.758
MSC_SG5	5	1.085	68.01	84.52	0.569	0.754
MSC_SG7	5	1.078	66.54	86.66	0.574	0.758
DT_SG5	5	1.084	68.02	84.51	0.569	0.755
DT_SG7	5	1.077	66.55	86.66	0.575	0.758
NIR	5	1.120	60.33	98.48	0.540	0.735

[a] Predictor variable dataset name after corresponding pre-processing technique was applied to the NIR spectra (acronyms are explained in Supplementary Table 2A).

[b] Root mean square error of the prediction.

[c] Percentage of the Y variable that is accounted in the model.

[d] Percentage of the NIR spectral data that is accounted in the model.

[e] Prediction ability: correlation between observed and NIR predicted values.

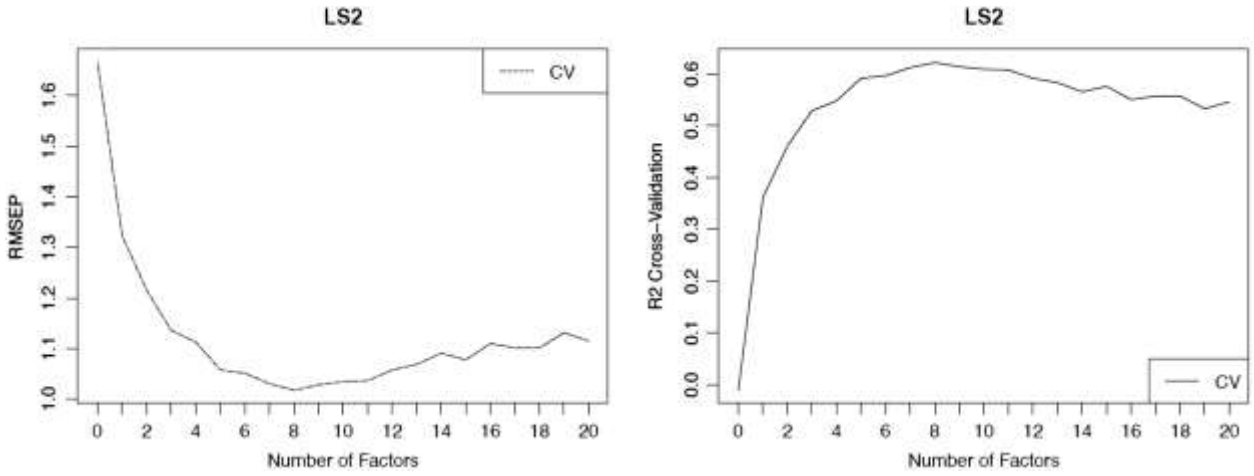


Figure 2. Model diagnostics of *Leptocybe invasa* screening 2 (LS2) with a standard normal variate (SNV) transformed database at the SQF site.

Terpenes to predict *L. invasa* infestation

A descriptive analysis on the terpene measurements (48 terpenes) was performed to explore the data. Supplementary Figures 3A, 3B and 4, respectively, show the hierarchical clustering dendrogram of the terpene measurements across all sites, a graphical display of the all-versus-all terpene correlation matrix and a principal component analysis biplot representing the relationship between the terpenes and the individual trees grouped per site. From these analyses, it is evident that there are groups of terpenes that are highly correlated, so we needed to find a subset of terpenes to fit our models that minimize the likelihood of having multicollinearity problems. Supplementary Figure 5A-C shows boxplots of terpene concentration, separated by site.

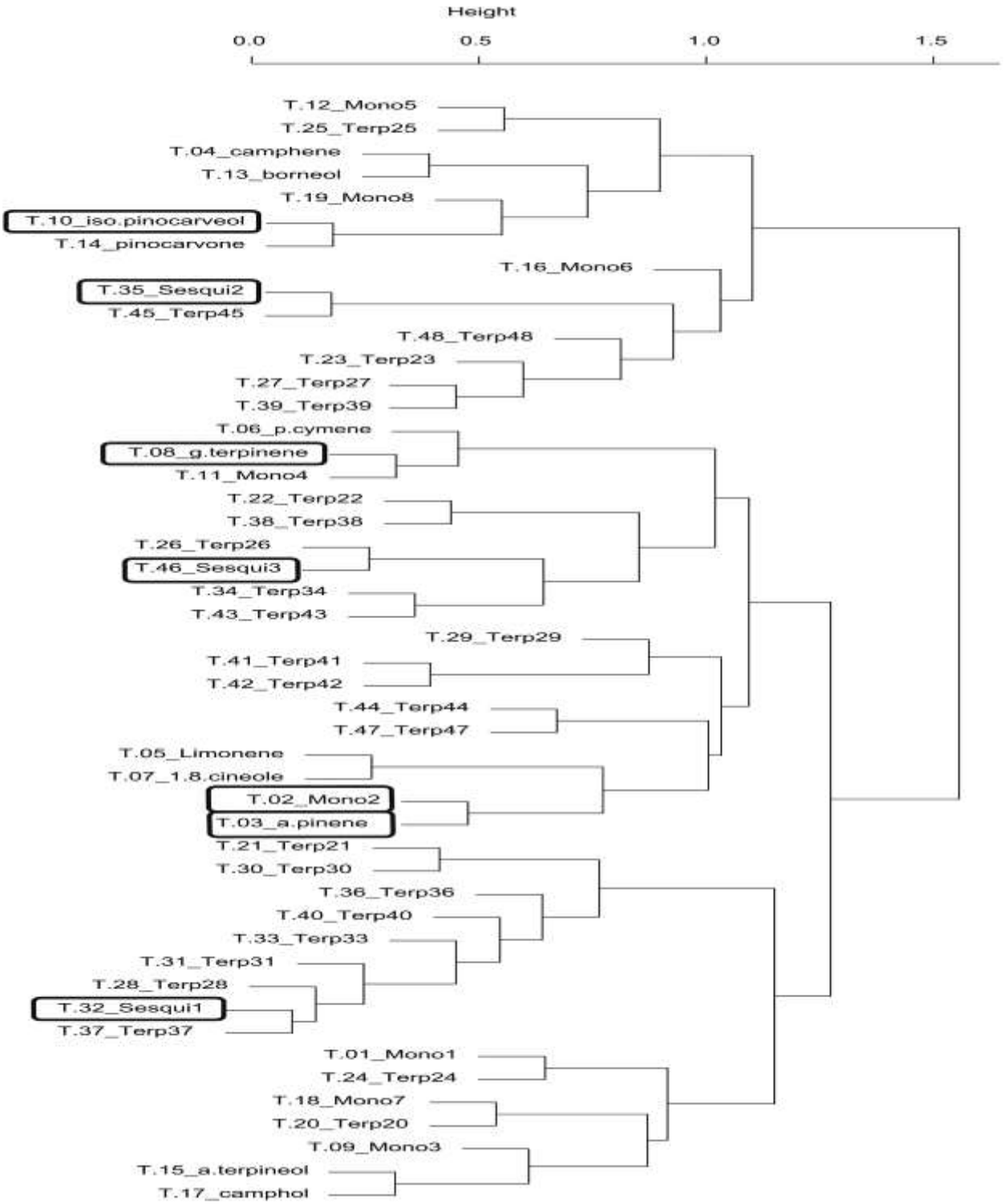


Figure 3. Hierarchical clustering dendrogram of the 48 measured terpenes at the Siya Qubeka (SQF) site. The seven terpenes selected for near-infrared reflectance modeling are boxed.

Table 3. Bayesian model selection to identify the most important terpenes for predicting *Leptocybe invasa* infestation based on *L. invasa* screenings (LS1, LS2) and individual breeding values (IBV) at the Siya Qubeka (SQF) site and across all three sites. The model with the highest R^2 value out of the top five Bayesian information criterion-ranked models are reported.

	SQF			**All sites**		
Terpene	**LS1**	**LS2**	**IBV**	**LS1**	**LS2**	**IBV**
T.2 (monoterpene 2)*	T.2	T.2	T.2	.	T.2	.
T.3 (α-pinene)*	T.3	T.3	T.3	.	T.3	.
T.8 (γ-terpinene)*	T.8	T.8	T.8	.	T.8	.
T.10 (iso-pinocarveol)*	T.10	T.10	T.10	T.10	T.10	T.10
T.12 (monoterpene 5)	.	.	T.12	.	.	.
T.19 (monoterpene 8)	.	.	.	T.19	.	.
T.22 (terpene 22)	.	.	.	.	T.22	.
T.23 (terpene 23)	.	.	.	.	T.23	T.23
T.26 (terpene 26)	.	.	.	T.26	.	.
T.32 (sesquiterpene 1)*	T.32	T.32	.	.	.	.
T.35 (sesquiterpene 2)*	T.35	T.35	T.35	.	.	.
T.45 (terpene 45)	.	.	.	T.45	T.45	T.45
T.46 (sesquiterpene 3)*	T.46	T.46	T.46	.	.	.
Model[a]	1	**1**	3	4	1	5
nVar[b]	7	**7**	7	4	7	3
R^2	0.31	**0.35**	0.30	0.06	0.14	0.05
LOO CV[c] R^2	0.25	**0.29**	0.25	0.04	0.11	0.03
BIC[d]	-29	**-40**	-28	-7	-32	-5
Post Prob[e]	0.02	**0.05**	0.01	0.04	0.05	0.03

* Terpenes selected for near-infrared reflectance (NIR) modeling.

[a] The model number out of the top five Bayesian information criterion (BIC)-ranked models.

[b] The number of variables selected for that model.

[c] Leave-one-out (LOO) cross-validation (CV) R^2 value.

[d] The Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models. The model with the lowest BIC is preferred.

[e] The posterior probabilities of the models selected.

The best BMA models were obtained for the SQF site (see Supplementary Table 6A-D for all BMA results). A summary of the most important terpenes for predicting LS1, LS2 and IBV at the SQF site and across all sites, together with the relevant model statistics, are presented in Table 3. Seven terpenes from models at the SQF site were selected for further analysis, i.e. NIR modeling to predict terpene content. Figure 3 shows the hierarchical clustering dendrogram of the terpene measurements at the SQF site, with the seven selected terpenes highlighted and scattered across different clusters: T.2 (monoterpene 2), T.3 (α-pinene), T.8 (γ-terpinene), T.10 (iso-pinocarveol), T.32 (sesquiterpene 1), T.35 (sesquiterpene 2) and T.46 (sesquiterpene 3). These terpenes were the result of the top BMA model for both LS1 (R^2 = 0.306) and LS2 (R^2 = 0.346); and six of these terpenes were included in the model with the highest R^2 value (R^2 = 0.302) for IBV (Table 3).

Performing leave-one-out cross-validation on the BMA models at the SQF site (and considering only the top Bayesian information criterion-ranked model per BMA run), four of the seven selected terpenes were present in more than 90% of the models (monoterpene 2, α-pinene, γ-terpinene and sesquiterpene 3) and the remaining three terpenes were added if the presence in

more than 50% of the models were considered (iso-pinocarveol, sesquiterpene 1, sesquiterpene 2). An additional three terpenes were added if the presence in more than 35% of the models were considered (monoterpene 5, terpene 34 and terpene 37). The average number of factors (terpenes included in a model) across the cross-validation models was 7 (min=4, max=7) and the average R^2 was 0.34 (min=0.27, max=0.36). BMA models that were obtained across the top five LS1, LS2 and IBV models from data of the MTZ site (average R^2 = 0.129), the NYL site (average R^2 = 0.08) and across all sites (average R^2 = 0.07) did not perform well and were thus not considered in further analyses.

To further improve the BMA models for LS2 in the SQF site, different combinations of predictor variables were included together with individual terpenes as input to separate BMA analyses. Note that when the sum of a group of monoterpenes was included, the separate monoterpenes (that made up the sum) were not included as predictor variables for that analysis. However, it was not possible to obtain a higher R^2 value than when the 7 individual terpenes mentioned above were included in the model (calibration R^2 = 0.35 and validation R^2 = 0.29).

Table 4. The best partial least squares models, based on near-infrared reflectance (NIR) data, to predict terpene content at the Siya Qubeka (SQF) site.

Variable	Dataset[a]	Factors	RMSEP[b]	ExpVar_Y[c]	ExpVar_X[d]	R^2 Validation	r[e]
T.2 (monoterpene 2)	SG7	7	0.0906	35.17	91.02	0.077	0.277
T.3 (α-pinene)	MSC_SG7	7	2.075	46.69	90.20	0.235	0.485
T.8 (γ-terpinene)	SG7	14	0.590	74.00	97.31	0.188	0.433
T.10 (iso-pinocarveol)	SG7	8	0.151	57.52	91.44	0.333	0.577
T.32 (sesquiterpene 1)	DT	3	0.126	7.758	80.81	0.013	0.116
T.35 (sesquiterpene 2)	MSC	5	0.135	8.90	94.40	0.003	0.051
T.46 (sesquiterpene 3)	MSC	4	0.083	14.32	87.07	0.080	0.282

[a] The model number out of the top five Bayesian information criterion (BIC)-ranked models.

[b] The number of variables selected for that model.

[c] Leave-one-out (LOO) cross-validation (CV) R^2 value.

[d] The Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models. The model with the lowest BIC is preferred.

[e] The posterior probabilities of the models selected.

Table 5. The best partial least squares models, based on near-infrared reflectance (NIR) data, to predict monoterpene combinations for the Siya Qubeka (SQF) site.

Variable	Dataset[a]	Factors	RMSEP[b]	ExpVar_ Y[c]	ExpVar_ X[d]	R^2 Validation	r[e]
sum(T10,T14)[f]	SG7	9	0.207	59.55	94.39	0.349	0.591
sum(T5,T7)[f]	**SG7**	**13**	**1.181**	**82.90**	**96.96**	**0.466**	**0.683**
sum(T7,T15)[g]	SG7	12	1.154	78.94	96.75	0.461	0.679
sum(T6,T8) [g]	DT_SG7	12	1.249	72.41	96.67	0.335	0.579
sum(T5,T7,T15) [g]	SG7	12	1.249	78.57	96.75	0.454	0.674
sum(T3-T5,T7, T10,T13-T15) [g]	SG7	8	2.958	50.97	93.15	0.231	0.481

[a] Predictor variable dataset name after corresponding pre-processing technique was applied to the NIR spectra (acronyms are explained in Supplementary Table 2A).

[b] Root mean square error of the prediction.

[c] Percentage of the Y variable that is accounted in the model.

[d] Percentage of the NIR spectral data that is accounted in the model.

[e] Prediction ability: correlation between observed and NIR predicted values.

[f] Reason for combining terpenes: high pair-wise correlation ($r > 0.6$).

[g] Reason for combining terpenes: biological motivation; based on (a) shared intermediate carbocation (biosynthetically related through same intermediate precursor) or (b) terpene X is precursor of terpene Y (biosynthetically related by 'descent'). This is based on Keszei et al., (2008) Figure 3A.

NIR to predict terpenes (SQF site)

Prediction models for terpene concentration were run for the seven terpenes selected under BMA (see list of terpenes above). For those terpenes, the best models were obtained for T.10 (iso-pinocarveol), T.3 (α-pinene) and T.8 (γ-terpinene) with R^2 values ranging between 0.188 and 0.333 and with prediction abilities between 0.433 and 0.577 (Table 4). Table 5 shows summary statistics of the best models obtained when terpenes were combined based on either biological

motivation or high pairwise correlations ($r > 0.6$). Selection of best models was made according to the following conditions: small number of latent variables (factors) that minimize the root mean square of the prediction (RMSEP), maximize the proportion of variation explained for both the dependent and independent variables, and maximize the cross-validation R^2. Prediction abilities of those models ranged between 0.481 for sum(T3-T5,T7,T10,T13-T15) and 0.683 for sum(T05,T07), the latter being the terpene combination for which we obtained the best model with a 47 % of the trait variation.

Discussion

We sought to associate terpene profiles with *Leptocybe* damage in a sub-population of an *E. grandis* breeding trial. There was phenotypic variation of *Leptocybe* damage in the first year of *L. invasa* infestation (Figure 1A) with the NYL site showing more score 2 phenotypes (with galls) and the SQF site showing more score 0 phenotypes (absence of galls). Within the sub-population, the LS1 scores in NYL showed a higher frequency of 3 and lower frequency of 0 than the other two sites (Figure 1B). In the second round of phenotyping after re-infestation by *L. invasa*, all individuals showed the presence of galls (score 1, Figure 1C) with SQF appearing to contain more of the healthy phenotype (higher frequency of score 1 and 2) compared to the other sites.

We detected 48 terpenes in the *E. grandis* individuals, however only a subset could be identified and were previously identified in *Eucalyptus* species (Kainer et al., 2017; Padovan et al., 2012;

Wallis et al., 2011). Eucalypts often contain distinct foliar chemical variation within a species, termed 'chemotypes', where the foliar chemical profile of one sub-population is dominated by one or few chemicals, while another sub-population is dominated by different chemicals (Keszei et al., 2008; Padovan et al., 2014). Ecologically, this type of variation is important as some pest or herbivores preferentially eat only one chemotype (Moore et al., 2014; Padovan et al., 2010). While there has been no previous record of chemotypic variation in *E. grandis*, two closely related species, *E. pelita* and *E. urophylla* have two described monoterpene chemotypes each, which are dominated by 1,8-cineole and α-pinene and 1,8-cineole and *p*-cymene, respectively (Padovan et al., 2014). We tested whether this progeny trial contained distinct chemotypes, but could not identify any. The terpene profile of every individual was dominated by α-pinene. Previous attempts to identify chemotypic variation in *E. grandis* did not test many individuals (Brophy and Southwell, 2002), which prompted us to test for such variation in a larger population, however, due to some degree of introgression due to the ongoing inbreeding programme of *E. grandis* in South Africa, the potential full chemical variation of this species has still not been elucidated.

Using the phenotypes captured for the 491 *E. grandis* individuals over two infestation seasons, we successfully developed models to predict *L. invasa* infestation scores based on NIR spectra (Table 1, 2) and terpene content (Table 3). We also related NIR to terpenes (Table 4, 5). The selected models, indicated in bold text in Table 2, 3 and 5, explain 62, 29 and 47% of the trait variation, respectively. The performance of models were evaluated using leave-one-out cross-validation. In general, models with root mean squared error of prediction values smaller than 0.3 indicate very good predictive models (Veerasamy et al., 2011).

There was a marked discrepancy in site for models that explained the NIR and terpene association with *L. invasa* scores. In both cases, i.e. using NIR and terpene data, the only site which passed our criteria of acceptable models was SQF. This is in agreement with our calculations of heritability for LS2 per site. For MTZ and NYL we found negligible heritability values, so that the variation of the *L. invasa* score is mainly due to environmental conditions or random experimental variation. In contrast, SQF had a heritability value of 0.16, being the only test in which genetic variation was found between individuals (Supplementary Table 4) contributing to better models. GxE could not be determined for the 491 individuals, but a low GxE was estimated for the full *E. grandis* population. Supplementary Table 1 indicates that there were some slight differences in the environmental data for the three sites. SQF and MTZ had higher moisture content than NYL. The MTZ site contains a hardy grass which is thought to compete with *E. grandis* growth during establishment. The percentage stocking at 4-year was 88% for SQF, 81% for NYL and 67% for MTZ. The average diameter at breast height was highest at site SQF (Supplementary Table 1). Collectively this suggests that SQF has the best growth conditions out of the sites sampled. Interestingly, SQF had a higher proportion of score 0 and 1 for LS1 and score 1 and 2 for LS2 than the other two sites - indicating a more resistant phenotype (Figure 1A and 1C respectively).

When modeling terpene effects on *L. invasa* score, we found that the terpenes that most contributed to the models act in opposing directions (see Supplementary Table 3C and coefficients in Supplementary Table 6A). One group, including α-pinene, γ-terpinene and sesquiterpene 1, showed increasing damage to trees with increasing concentration of terpenes. α-Pinene has a relatively high vapour pressure (3 mm Hg at 20°C) compared to other

monoterpenes and is therefore more volatile than most monoterpenes commonly found in eucalypts. Ladybeetles are attracted to α-pinene from persimmon (*Diospyros kaki*) (Zhang et al., 2009) while trap catch of the invasive pine bark beetle *Hylurgus ligniperda* was increased to over 200-fold when α-pinene was used as an attractant (Kerr et al., 2017), therefore we could expect α-pinene to act as a volatile cue for *L. invasa* oviposition. γ-Terpinene levels were constitutively higher in the susceptible clone GC540 compared to the resistant *E. grandis* clone TAG5 and was induced to much higher levels upon insect oviposition (Oates et al., 2015). Interestingly, the levels of γ-terpinene decreased in the resistant genotype after infestation (Oates et al., 2015). It is feasible that γ-terpinene may play a role in promoting susceptibility to the insect pest, however, this remains to be demonstrated.

The other group of terpenes (including monoterpene 2, iso-pinocarveol, sesquiterpene 2 and sesquiterpene 3) acting in the opposite direction may play a direct role in defence against *L. invasa*, where higher concentrations of the compound lead to reduced damage by *L. invasa*. Evidence from other systems indicate that this could be achieved through different ways, such as direct toxic effect on larvae either leading to death or reduced growth of larvae (McLean et al., 1993), or through indirect defences by attracting parasites through tritrophic ways (reviewed in Gershenzon and Dudareva, (2007)). Several parasitoids of *L. invasa* have been identified with some being adopted for biological control (e.g. *Seletrichoides neseri*, *Ophelimus maskelli*, *Seletrichoides kyceri*; reviewed in Zheng et al., (2014)) however the volatile cues that attract these parasitic wasps have not been investigated. In a study by Visser et al., (2015), artificial inoculation of the *E. grandis* clone TAG5 with the fungal pathogen *Chrysoporthe austroafricana*

led to the induction of iso-pinocarveol systemically, in leaf tissue. This *E. grandis* clone was also found to be resistant to *L. invasa* (Oates et al., 2015).

In summary, we produced models for terpene to *L. invasa* infestation, NIR to terpenes and NIR to *L. invasa* interactions in *E. grandis* that explained 29% (Table 3, bold text), 47% (Table 5, bold text) and 62% (Table 2, bold text) of the trait variation, respectively. These methods developed in this study can be utilised as a guideline to model other plant-insect interaction systems as NIR may be a more cost effective approach to modeling resistance. One approach to improve the model would involve setting up a similar experiment in a controlled environment where the best and worst performing individuals were cloned and exposed to *L. invasa* so that robust phenotypes may be observed. In this manner, stronger associations could be derived for terpenes and resistance to *L. invasa* revealing important cues that could act as attractants and repellents against the insect pest.

Acknowledgments

The authors acknowledge funding from the National Research Foundation (NRF) South Africa Bioinformatics and Functional Genomics Programme (Grant ID 89669) and the Department of Science and Technology Eucalyptus genomics platform grant. We thank Ms Jessie Au and Dr Amanda Padovan for assistance with the leaf sample preparation and NIR. The authors declare no competing interests.

References

Alves, M. D. C. S., Filho, S. M., Innecco, R., and Torres, S. B. (2004). Alelopatia de extratos voláteis na germinação de sementes e no comprimento da raiz de alface. *Pesqui. Agropecu. Bras.* 39, 1083–1086.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). LOF: identifying density-based local outliers. *ACM sigmod Rec.* 29.

Brophy, J. J., and Southwell, I. A. (2002). "Eucalyptus chemistry," in *Eucalyptus – the genus Eucalyptus.*, ed. J. J. W. Coppen (Taylor and Francis: London), 102–160.

Burdon, R. D. (1977). Genetic correlation as a concept for studying genotype-environment interaction in forest tree breeding. *Silvae Genet.* 26, 168–175.

Carr, D. J., and Carr, S. G. M. (1970). Oil glands and ducts in *Eucalyptus* L'Herit. II: development and structure of oil glands in the embryo. *Aust. J. Bot.* 18, 191–212.

Chang, R., Arnold, R., and Zhou, X. (2012). Association between enzyme activity levels in Eucalyptus clones and their susceptibility to the gall wasp, *Leptocybe Invasa,* in South China. *J. Trop. For. Sci.* 24, 256–264.

Coppen, J. J. W. (2003). *Eucalyptus: The genus Eucalyptus.* CRC Press LLC.

De Moraes, C. M., Lewis, W. J., Paré, P. W., Alborn, H. T., and Tumlinson, J. H. (1998). Herbivore-infested plants selectively attract parasitoids. *Nature* 393, 570–573.

Degenhardt, J., and Gershenzon, J. (2003). "Terpenoids," in *Encyclopedia of Applied Plant Sciences*, eds. T. Brian, D. Murphy, and B. Murray (Amsterdam, Elsevier), 500–504.

Dittrich-Schröder, G., Harney, M., Neser, S., Joffe, T., Bush, S., Hurley, B. P., et al. (2014). Biology and host preference of *Selitrichodes neseri*: A potential biological control agent of the Eucalyptus gall wasp, *Leptocybe invasa*. *Biol. Control* 78, 33–41.

Dittrich-Schröder, G., Wingfield, M. J., Hurley, B. P., and Slippers, B. (2012). Diversity in *Eucalyptus* susceptibility to the gall forming wasp *L. invasa*. *Agric. For. Entomol.* 14, 419–427.

Dudareva, N., Andersson, S., Orlova, I., Gatto, N., Reichelt, M., Rhodes, D., et al. (2005). The nonmevalonate pathway supports both monoterpene and sesquiterpene formation in snapdragon flowers. *Proc. Natl. Acad. Sci.* 102, 933–938.

Durand, N., Rodrigues, J. C., Mateus, E., Boavida, C., and Branco, M. (2011). Susceptibility variation in *Eucalyptus* spp in relation to *Leptocybe invasa* and *Ophelimus maskelli*, two invasive gall wasps occurring in Portugal. *Silva Lusit.*, 19–31.

Edwards, P. B., Wanjura, W. J., and Brown, W. V. (1993). Selective herbivory by Christmas beetles in response to intraspecific variation in *Eucalyptus* terpenoids. *Oecologia* 95, 551–557.

Edwards, P. B., Wanjura, W. J., Brown, W. V, and Dearn, J. M. (1990). Mosaic resistance in plants. *Nature* 347, 434.

Eyles, A., Davies, N. W., Yuan, Z. Q., and Mohammed, C. (2003). Host responses to natural infection by Cytonaema sp. in the aerial bark of *Eucalyptus globulus*. *For. Pathol.* 33, 317–331.

Gershenzon, J., and Dudareva, N. (2007). The function of terpene natural products in the natural

world. *Nat. Chem. Biol.* 3, 408–414.

Giamakis, A., Kretsi, O., Chinou, I., and Spyropoulos, C. G. (2001). *Eucalyptus camaldulensis*: volatiles from immature flowers and high production of 1,8-cineole and β-pinene by in vitro cultures. *Phytochemistry* 58, 351–355.

Gomes, V. J., Longue, D., Colodette, J. L., and Ribeiro, R. A. (2014). The effect of eucalypt pulp xylan content on its bleachability, refinability and drainability. *Cellulose* 21, 607–614.

Henery, M. L., Wallis, I. R., Stone, C., and Foley, W. J. (2008). Methyl jasmonate does not induce changes in *Eucalyptus grandis* leaves that alter the effect of constitutive defences on larvae of a specialist herbivore. *Oecologia* 156, 847–859.

Iqbal, Z., Akhtar, M., Qureshi, T. M., Akhter, J., and Ahmad, R. (2011). Variation in composition and yield of foliage oil of *Eucalyptus polybractea*. *J. Chem. Soc. Pakistan* 33, 183–187.

Javaregowda, J., and Prabhu, S. T. (2010). Susceptibility of eucalyptus species and clones to gall wasp, *Leptocybe invasa* Fisher and La Salle (Eulophidae: Hymenoptera) in Karnataka. *Karnataka J. Agric. Sci.* 23, 220–221.

Kainer, D., Bush, D., Foley, W. J., and Külheim, C. (2017). Assessment of a non-destructive method to predict oil yield in *Eucalyptus polybractea* (blue mallee). *Ind. Crops Prod.* 102, 32–44.

Keefover-Ring, K., Thompson, J. D., and Linhart, Y. B. (2009). Beyond six scents: Defining a seventh *Thymus vulgaris* chemotype new to southern France by ethanol extraction. *Flavour Fragr. J.* 24, 117–122.

Kelly, J., La Salle, J., Harney, M., Dittrich-Schroder, G., Hurley, B. P., and Undefined, O. (2012). *Selitrichodes neseri* n. sp, a new parasitoid of the eucalyptus gall wasp *Leptocybe invasa* Fisher & La Salle (Hymenoptera: Eulophidae: Tetrastichinae). *Zootaxa* 3333, 50–57.

Kerr, J. L., Kelly, D., Bader, M. K. F., and Brockerhoff, E. G. (2017). Olfactory cues, visual cues, and semiochemical diversity interact during host location by invasive forest beetles. *J. Chem. Ecol.* 43, 17–25.

Keszei, A., Brubaker, C. L., and Foley, W. J. (2008). A molecular perspective on terpene variation in *Australian Myrtaceae*. *Aust. J. Bot.* 56, 197–213.

Kim, I. K., Mendel, Z., Protasov, A., Blumberg, D., and La Salle, J. (2008). Taxonomy, biology, and efficacy of two Australian parasitoids of the eucalyptus gall wasp, *Leptocybe invasa* Fisher & La Salle (Hymenoptera: Eulophidae: Tetrastichinae). *Zootaxa* 1910, 1–20.

Külheim, C., Padovan, A., Hefer, C., Krause, S. T., Köllner, T. G., Myburg, A. A., et al. (2015). The *Eucalyptus* terpene synthase gene family. *BMC Genomics* 16, 450.

Kulkarni, H. (2010). Screening eucalyptus clones against *Leptocybe invasa* Fisher and La Salle (Hymenoptera: Eulophidae). *Karnataka J. Agric. Sci.* 23, 87–90.

Lawler, I. R., Stapley, J., Foley, W. J., and Eschler, B. M. (1999). Ecological example of conditioned flavor aversion in plant-herbivore interactions: Effect of terpenes of *Eucalyptus* leaves on feeding by common ringtail and brushtail possums. *J. Chem. Ecol.* 25, 401–415.

McLean, S., Foley, W. J., Davies, N. W., Brandon, S., Duo, L., and Blackman, A. J. (1993). Metabolic fate of dietary terpenes from *Eucalyptus radiata* in common ringtail possum (*Pseudocheirus peregrinus*). *J. Chem. Ecol.* 19, 1625–1643.

Mendel, Z., Protasov, A., Fisher, N., and Salle, J. La (2004). Taxonomy and biology of *Leptocybe invasa* gen. & sp. n. (Hymenoptera: Eulophidae), an invasive gall inducer on *Eucalyptus*. *Aust. J. Entomol.* 43, 101–113.

Mevik, B.-H., and Wehrens, R. (2007). The pls package: principal component and partial least squares regression in R. *J. Stat. Softw.* 18, 1–23.

Mewalal, R., Rai, D. K., Kainer, D., Chen, F., Külheim, C., Peter, G. F., et al. (2017). Plant-Derived Terpenes: A Feedstock for Specialty Biofuels. *Trends Biotechnol.* 35, 227–240.

Moore, B., Andrew, R., Külheim, C., and Foley, W. (2014). Explaining intraspecific diversity in plant secondary metabolites in an ecological context. *New Phytol.* 201, 733–750.

Morrow, P. A., and Fox, L. R. (1980). Effects of variation in *Eucalyptus* essential oil yield on insect growth and grazing damage. *Oecologia* 45, 209–219.

Mutitu, K. E. (2003). A pest threat to *Eucalyptus* species in Kenya. *KEFRI Tech. Rep.*, 12.

Nyeko, P. (2005). The cause, incidence and severity of a new gall damage on *Eucalyptus* species at Oruchinga refugee settlement in Mbarara district, Uganda. *Uganda J. Agric. Sci.* 11, 47–50.

Nyeko, P., Mutitu, E. K., and Day, R. K. (2009). Eucalyptus infestation by *Leptocybe invasa* in Uganda. *Afr. J. Ecol.* 47, 299–307.

Nyeko, P., and Nakabonge, G. (2008). Occurence of pests and diseases in tree nurseries and plantations in Uganda. *Sawlog Prod. Grant Scheme, Kampala, Uganda.*

Oates, C. N., Külheim, C., Myburg, A. A., Slippers, B., and Naidoo, S. (2015). The transcriptome and terpene profile of *Eucalyptus grandis* reveals mechanisms of defense

against the insect pest, *Leptocybe invasa*. *Plant Cell Physiol.* 56, 1418–1428.

Padovan, A., Keszei, A., Köllner, T. G., Degenhardt, J., and Foley, W. J. (2010). The molecular basis of host plant selection in *Melaleuca quinquenervia* by a successful biological control agent. *Phytochemistry*.

Padovan, A., Keszei, A., Külheim, C., and Foley, W. J. (2014). The evolution of foliar terpene diversity in Myrtaceae. *Phytochem. Rev.* 13, 695–716.

Padovan, A., Keszei, A., Wallis, I. R., and Foley, W. J. (2012). Mosaic Eucalypt trees suggest genetic control at a point that influences several metabolic pathways. *J. Chem. Ecol.* 38, 914–923.

Padovan, A., Webb, H., Mazanec, R., Grayling, P., Bartle, J., Foley, W. J., et al. (2017). Association genetics of essential oil traits in *Eucalyptus loxophleba*: explaining variation in oil yield. *Mol. Breed.* 37, 73.

Pateraki, I., Heskes, A., and Hamberger, B. (2015). "Cytochromes P450 for terpene functionalization and metabolic engineering," in *Biotechnology of Isoprenoids*, ed. J. Schrader, J. Bohlmann (Cham, Springer International Publishing), 107–139.

Quang Thu, P., Dell, B., and Isobel Burgess, T. (2009). Susceptibility of 18 eucalypt species to the gall wasp *Leptocybe invasa* in the nursery and young plantations in Vietnam. *ScienceAsia* 35, 113–117.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raftery, A. A., Hoeting, J., Volinsky, C., Painter, I., and Yeung, K. Y. (2017). Bayesian model

averaging. Available at: https://cran.r-project.org/web/packages/BMA/BMA.pdf.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociol. Methodol.* 25, 111–163.

Rinnan, Å., Berg, F. van den, and Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Trends Anal. Chem.* 28, 1201–1222.

Rivas, F., Parra, A., Martinez, A., and Garcia-Granados, A. (2013). Enzymatic glycosylation of terpenoids. *Phytochem. Rev.* 12, 327–339.

Schimleck, L. R., and Rimbawanto, A. (2003). Near infrared spectroscopy for cost effective screening of foliar oil characteristics in a *Melaleuca cajuputi* breeding population. *J. Agric. Food Chem.* 51, 2433–2437.

Schnee, C., Kollner, T. G., Gershenzon, J., and Degenhardt, J. (2002). The maize gene *terpene synthase 1* encodes a sesquiterpene synthase catalyzing the formation of *(E)*-β-farnesene, *(E)*-nerolidol, and *(E,E)*-farnesol after herbivore damage. *Plant Physiol.* 130, 2049–2060.

Squillace, A. E. (1974). Average genetic correlations among offspring from open-pollinated forest trees. *Silvae Genet.* 23, 149–156.

Stevens, A., and Ramirez-Lopez, L. (2013). An introduction to the prospectr package. R package Vignette R package version 0.1.3. Available at: https://cran.r-project.org/web/packages/prospectr/vignettes/prospectr-intro.pdf.

Stone, C., and Bacon, P. E. (1994). Relationships among moisture stress, insect herbivory, foliar cineole content and the growth of river red gum *Eucalyptus camaldulensis*. *J. Appl. Ecol.* 31, 604–612.

Torgo, L. (2015). Functions and data for “Data Mining with R”. Available at: https://cran.r-project.org/web/packages/DMwR/DMwR.pdf.

Turlings, T. C., Loughrin, J. H., McCall, P. J., Rose, U. S., Lewis, W. J., and Tumlinson, J. H. (1995). How caterpillar-damaged plants protect themselves by attracting parasitic wasps. *Proc. Natl. Acad. Sci.* 92, 4169–4174.

Veerasamy, R., Rajak, H., Jain, A., Sivadasan, S., Varghese, C. P., and Agrawal, R. K. (2011). Validation of QSAR Models - Strategies and Importance. *Int. J. Drug Des. Disocovery* 2, 511–519.

Visser, E. A., Wegrzyn, J. L., Steenkmap, E. T., Myburg, A. A., and Naidoo, S. (2015). Combined *de novo* and genome guided assembly and annotation of the *Pinus patula* juvenile shoot transcriptome. *BMC Genomics* 16, 1057.

Wallis, I. R., Keszei, A., Henery, M. L., Moran, G. F., Forrester, R., Maintz, J., et al. (2011). A chemical perspective on the evolution of variation in *Eucalyptus globulus*. *Perspect. Plant Ecol. Evol. Syst.* 13, 305–318.

Webb, H., Foley, W. J., and Külheim, C. (2014). The genetic basis of foliar terpene yield: Implications for breeding and profitability of Australian essential oil crops. *Plant Biotechnol.* 31, 363–376.

Wehrens, R. (2011). Chemometrics with R - multivariate data analysis in the natural sciences and life sciences. Available at: https://cran.r-project.org/web/packages/ChemometricsWithR/ChemometricsWithR.pdf.

Wiley, J., and Skelley, P. (2008). A Eucalyptus pest, *Leptocybe invasa* Fisher and LaSalle

(Hymenoptera: Eulophidae), genus and species new to Florida and North America. *Florida Dep. Agric. Consum. Serv.*, 38870–38871.

Wilson, N. D., Watt, R. A., and Moffat, A. C. (2001). A near-infrared method for the assay of cineole in eucalyptus oil as an alternative to the official BP method. *J. Pharm. Pharmacol.* 53, 95–102.

Wingfield, M., Slippers, B., Hurley, B., Coutinho, T., Wingfield, B., and Roux, J. (2008). Eucalypt pests and diseases: growing threats to plantation productivity. *South. For. a J. For. Sci.* 70, 139–144.

Wong, Y. F., Perlmutter, P., and Marriott, P. J. (2017). Untargeted metabolic profiling of *Eucalyptus* spp. leaf oils using comprehensive two-dimensional gas chromatography with high resolution mass spectrometry: Expanding the metabolic coverage. *Metabolomics* 13, 46.

Wylie, F., and Speight, R. (2012). *Insect pests in tropical forestry.* CABI Publishing, Wallingford, UK.

Zhang, Y., Xie, Y., Xue, J., Peng, G., and Wang, X. (2009). Effect of volatile emissions, especially α-pinene, from persimmon trees infested by Japanese wax scales or treated with methyl jasmonate on recruitment of ladybeetle predators. *Environ. Entomol.* 38, 1439–1445.

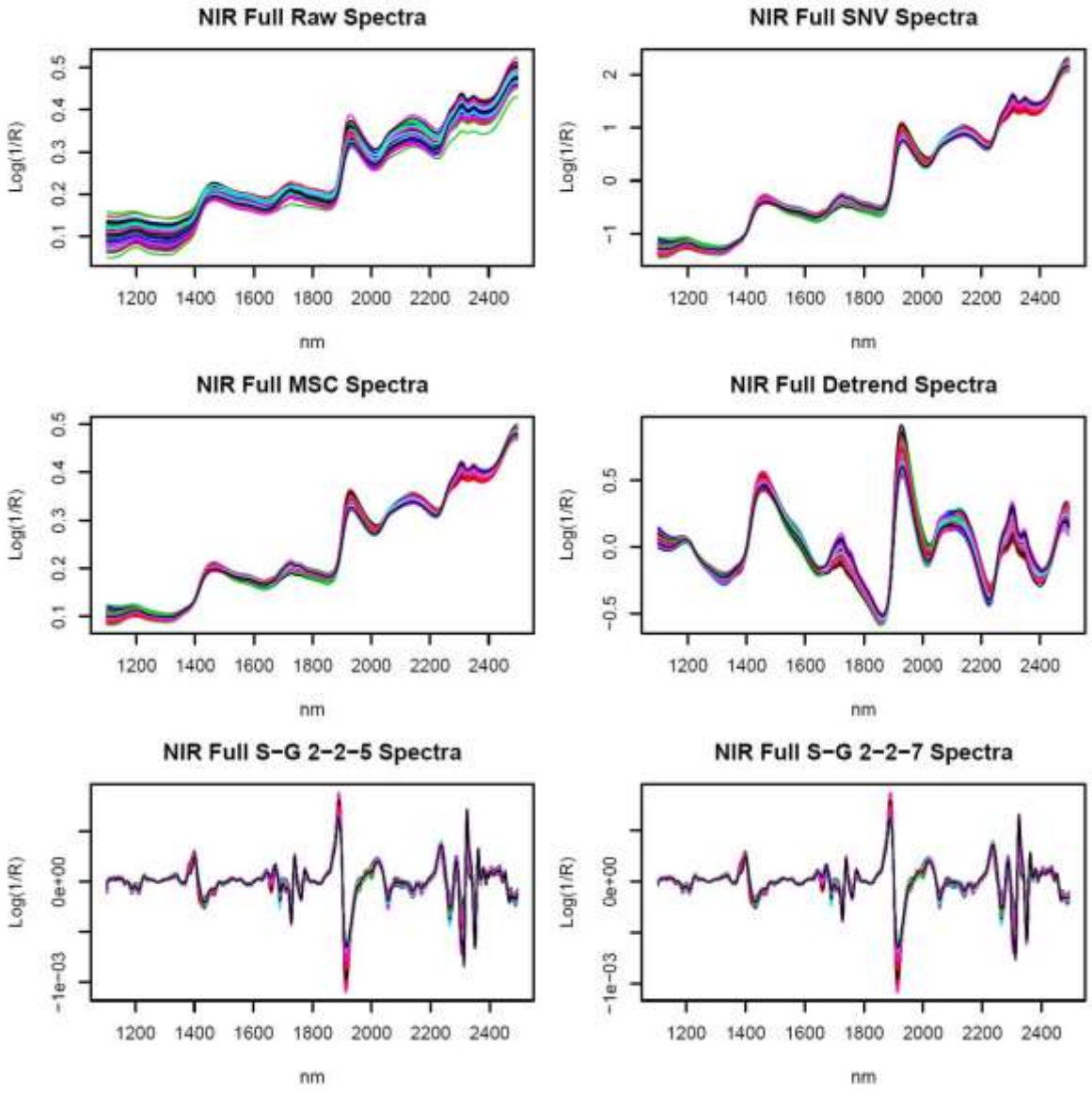
Zheng, X. L., Li, J., Yang, Z. D., Xian, Z. H., Wei, J. G., Lei, C. L., et al. (2014). A review of invasive biology, prevalence and management of *Leptocybe invasa* Fisher & La Salle (Hymenoptera: Eulophidae: Tetrastichinae). *African Entomol.* 22, 68–79.

Zhu, F. li, Ren, S., Qiu, B., Huang, Z., and Peng, Z. (2012). The abundance and population

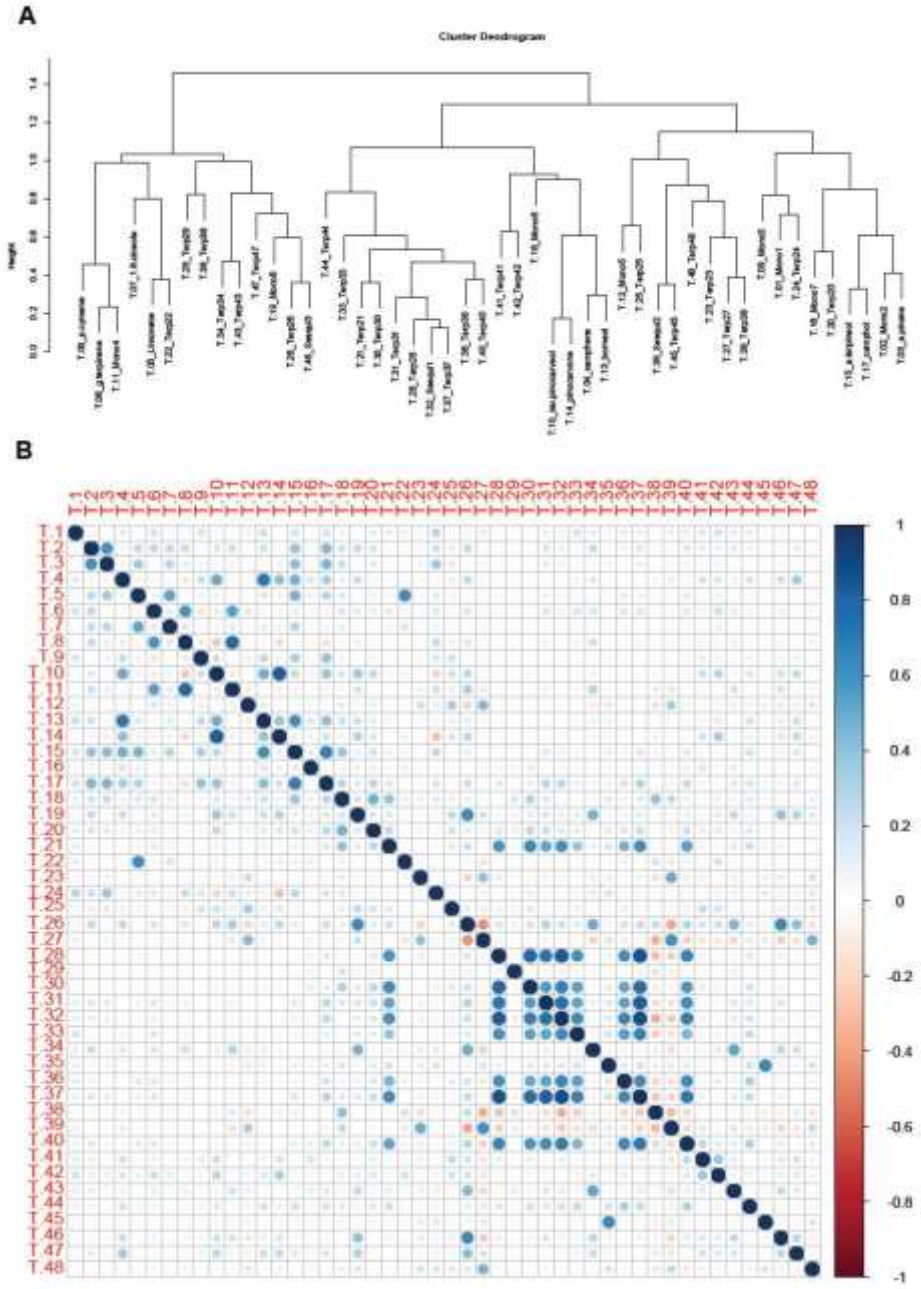
dynamics of *Leptocybe invasa* (Hymenoptera: Eulophidae) galls on *Eucalyptus* spp. in China. *J. Integr. Agric.* 11, 2116–2123.

Supporting Information

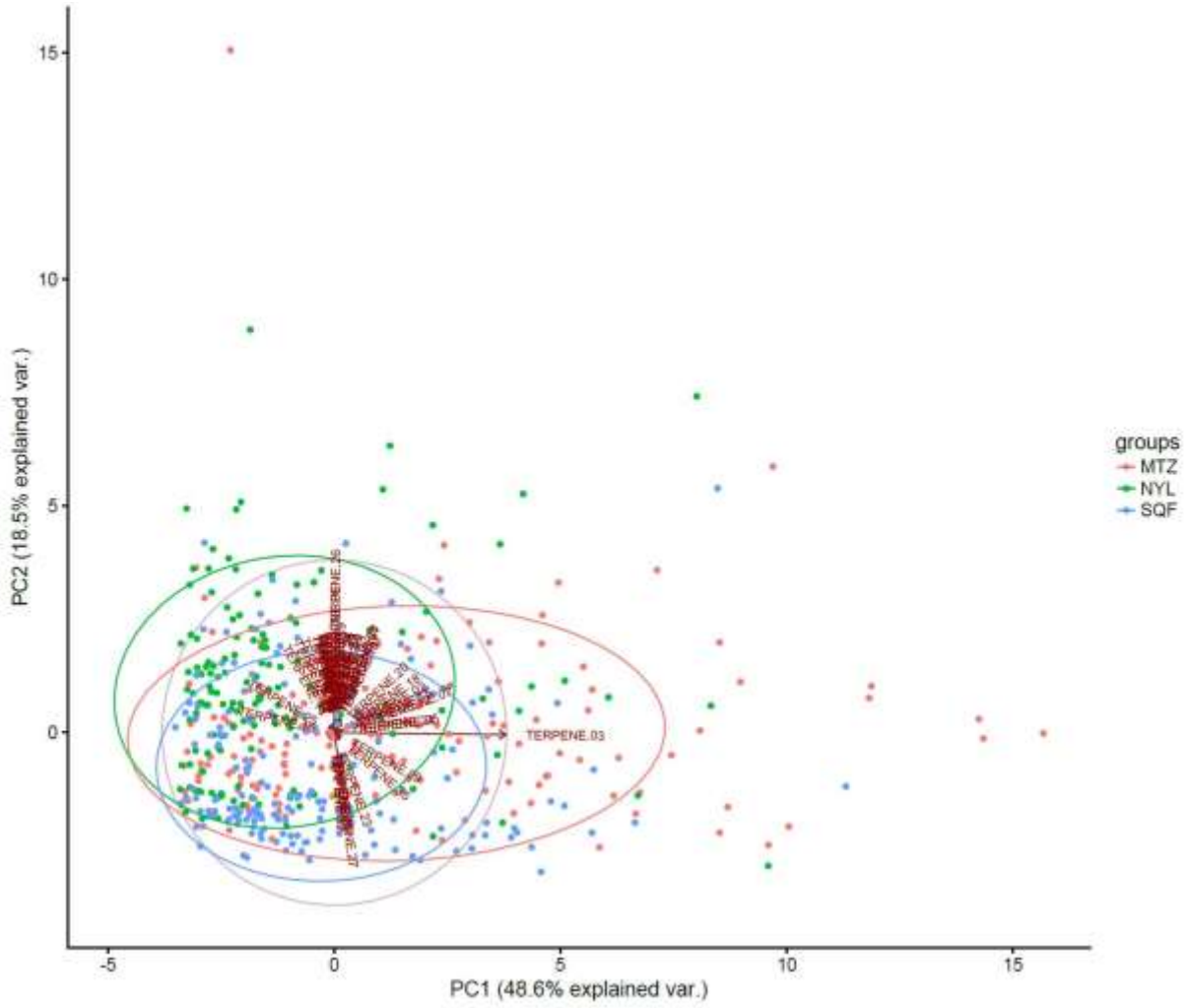
Supplementary Figure 1. Scores for screening *Leptocybe invasa* infestation on *Eucalyptus grandis*. (A) Score 0 – not infested. (B) Score 1 – infested without galls. (C) Score 2 – infested with galls. (D) Score 3 – stunting and lethal gall formation.



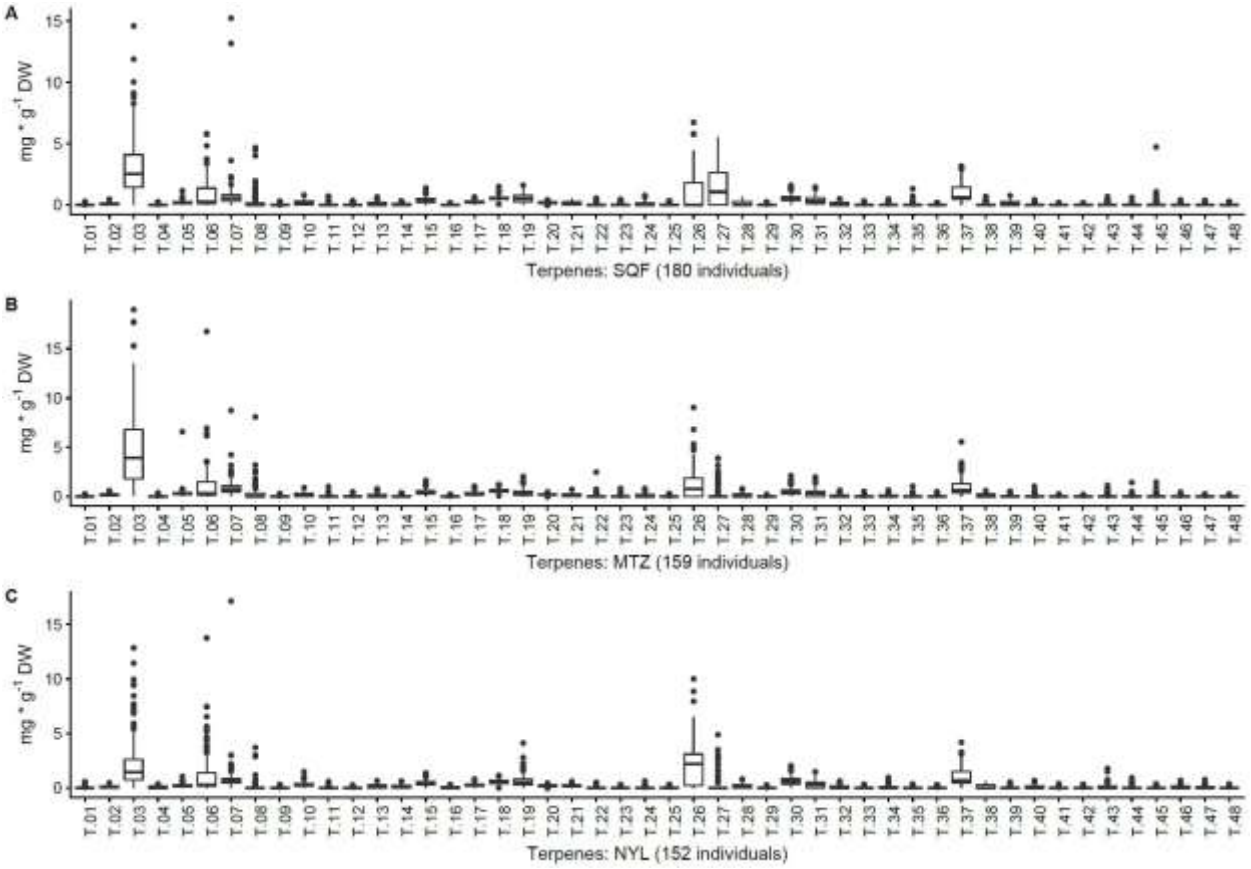
Supplementary Figure 2. The near-infrared reflectance (NIR) spectra, after applying six different mathematical transformations: (1) Raw spectral data, (2) Standard normal variate (SNV), (3) Multiplicative scatter correction (MSC), (4) Detrend, (5) Savitzky-Golay, 5-point smoothing, second-order polynomial, second derivative (S-G 2-2-5) and (6) Savitzky-Golay, 7-point smoothing, second-order polynomial, second derivative (S-G 2-2-7).



Supplementary Figure 3. (A) Hierarchical clustering dendrogram of the 48 measured terpenes across all sites. (B) Graphical display of the all-versus-all terpene correlation matrix, calculated across all sites. The plot was generated using the "corrplot" function in R, using the data across all sites. Positive correlations are displayed in blue and negative correlations in red. Color intensity and the size of circles are proportional to correlation coefficients.



Supplementary Figure 4. Principal components (PC) analysis biplot representing the relationship between the terpenes and the individual trees grouped per site: Siya Qubeka (SQF) in blue; Mtunzini (MTZ) in red; Nyalazi (NYL) in green.



Supplementary Figure 5. Boxplots of terpene measurements, per terpene. (A) Boxplots of terpene measurements for the Siya Qubeka (SQF) site. (B) Boxplots of terpene measurements for the Mtunzini (MTZ) site. (C) Boxplots of terpene measurements for the Nyalazi (NYL) site.

Supplementary Table 1. Environmental and phenotype data (full population) for the three *Eucalyptus grandis* sites surveyed for *Leptocybe invasa* infestation. [Excel file]

Supplementary Table 2. Predictor variable datasets and outlier detection prior to partial least squares modeling. (A) Predictor variable datasets used for outlier detection and partial least squares modeling. (B) The proportion of samples classified as outliers (and thus trimmed) for each set of models, prior to partial least squares modeling. [Excel file]

Supplementary Table 3. The 48 measured terpenes. (A) The name, major ions and retention time of the 48 measured terpenes. (B) The motivation for combining groups of terpenes. (C) The correlation of terpenes with the *Leptocybe invasa* screenings (LS1, LS2) and individual breeding values (IBV) for the Siya Qubeka (SQF) site. [Excel fie]

Supplementary Table 4. *Leptocybe invasa* heritability estimates for *L. invasa* screening 1 (LS1) and *L. invasa* screening 2 (LS2) for the *Eucalyptus grandis* population across sites. [Excel file]

Supplementary Table 5. Summary of the best partial least squares models, based on near-infrared reflectance (NIR) data, for *Leptocybe invasa* screenings (LS1, LS2) and individual breeding values (IBV) at the Mtunzini (MTZ) and Nyalazi (NYL) sites. [Excel file]

Supplementary Table 6. Bayesian model selection results to identify the most important terpenes for predicting *Leptocybe invasa* infestation. (A) Bayesian model selection results at the Siya Qubeka (SQF) site. (B) Bayesian model selection results at the Mtunzini (MTZ) site. (C) Bayesian model selection results at the Nyalazi (NYL) site. (D) Bayesian model selection results across all three sites. [Excel file]