UNTARGETED METABOLIC FINGERPRINTING REVEALS IMPACT OF GROWTH STAGE AND LOCATION ON COMPOSITION OF SEA BUCKTHORN (*Hippophaë rhamnoides*) LEAVES

Short title: Metabolic Fingerprinting of Sea Buckthorn Leaves

Raghunath Pariyani[a], Maaria Kortesniemi[a], Jaana Liimatainen[a,1], Jari Sinkkonen[b], Baoru Yang[a*]

[a] Food Chemistry and Food Development, Department of Biochemistry, University of Turku, FI-20014 Turun yliopisto, Finland

[b] Instrument Centre, Department of Chemistry, University of Turku, FI-20014 Turun yliopisto, Finland

Author for correspondence

Professor Baoru Yang

Food Chemistry and Food Development, Department of Biochemistry, University of Turku, FI-20014 Turun yliopisto, Finland

E-mail: baoru.yang@utu.fi

Mobile: +35845273788

Word count: 7169

[1] Present address: Department of Food and Nutrition, P.O. Box 66, FI-00014, University of Helsinki, Finland

ABSTRACT

Sea buckthorn (*Hippophaë rhamnoides*) is increasingly cultivated to produce raw materials for food and nutraceuticals. There is little knowledge on composition of sea buckthorn leaves (SBLs) and the key factors influencing the composition. This research aims to unravel the metabolic profile of SBLs and the effects of cultivar, location and stage of growth, and climatic conditions on the metabolic profile of SBLs. Leaves of two sea buckthorn cultivars grown in the south and north of Finland during two consecutive growth seasons were studied using untargeted NMR metabolomics. The highest variance in the metabolic profile was linked to the growth stage, wherein, leaves from the first seven weeks of harvest were characterized with higher abundance of polyphenols, while relatively higher abundance of carbohydrates and sugars was observed in the later weeks. The growth location attributed for the second highest variation, wherein, the north–south comparison identified fatty acids and sugars as discriminatory metabolites, and the potential association of metabolome to natural abiotic stressors was revealed. An inverse correlation between carbohydrate/sugar content as well as fatty acids of higher carbon chain length with the temperature variables was evident. The supervised chemometric models with high sensitivity and specificity classified and predicted the samples based on growth stage and location, and cultivar. Non-targeted NMR-metabolomics revealed the metabolic profile of SBLs and their variation associated with various biotic and abiotic factors. Cultivar and growth stage are key factors to consider when harvesting SBLs for use in food and nutraceuticals.

Keywords: *Hippohaë rhamnoides*; Sea buckthorn leaves; Growth stage; NMR metabolomics; Weather conditions.

Practical Application

Globally Sea buckthorn cultivation has been rapidly increasing due to the known health-promoting benefits of the berries and leaves of the plant. The current research obtained new comprehensive information on the compositional profile of sea buckthorn leaves as well as the impact of major contributory factors such as cultivars, the advancement of growth stage, geographical location, and weather parameters. The findings of this research provide new knowledge and guidance for plant breeding, cultivation and commercial utilization of sea buckthorn leaves as raw materials for food, feed, and nutraceuticals.

1. INTRODUCTION

Sea buckthorn (*Hippophaë rhamnoides* L. of Elaeagnaceae) is native to Eurasia and currently cultivated in many regions of the world, especially in Europe, Asia and the North America, owing to its broad potential in various applications, especially within food and nutraceutical industry (Fatima et al., 2015; Suryakumar & Gupta, 2011). Traditionally, the sea buckthorn leaves (SBLs) have been used as ethnobotanical medicine and horse feed (Suryakumar & Gupta, 2011; Singh et al., 2005). From the beginning of this century, aqueous infusions of SBLs have been in use as herbal tea in the Asian and European regions, thus bringing the otherwise under-utilised by-product of the plant to the frontline. The SBLs tea powder has been shown to decrease the levels of murine hepatic triglycerides and cholesterol as well as elevate fecal lipid excretion (Lee et al., 2011). The potential of SBLs extracts in protection against oxidative stress has been demonstrated in various *in vitro* studies (Cho et al., 2017; Kim et al., 2017). In addition, various pharmacological roles of SBLs, such as, anti-inflammatory (Ganju et al., 2005), anti-cataract (Dubey, Deep & Singh, 2016), anti-visceral obesity (Lee et al., 2011), anti-tumor (Zhamanbaeva, Murzakhmetova, Tuleukhanov, & Danilenko 2014), and anti-microbial (Yogendra Kumar, Tirpude, Maheshwari, Bansal, & Misra, 2013; Michel, Destandau, Le Floch, Lucchesi, & Elfakir, 2012) are also reported.

SBLs are rich in polyphenols composed of monomeric *C*-glycosidic ellagitannins (ETs) and flavonoids (Tian et al., 2017; Suvanto & Salminen, 2016; Fatima et al., 2015; Yoshida, Tanaka, Chen, & Okuda, 1991). The antioxidative capacity of the SBLs tea extract correlates to the polyphenol content (Cho et al., 2014). The highest concentration of total phenolics in the SBLs of Russian origin grown in Sweden are reported to be in the leaves harvested at the end of July (Morgenstern, Ekholm, Scheewe, & Rumpunen, 2014). The total flavonols content of Romanian SBLs varied between 563–1437 mg rutin equivalents per 100 g dry weight (dw), depending on the genotype (Pop et al., 2013). The leaves of the Finnish cultivars, ′Tytti′ and ′Terhi′ are reported to have total phenolics concentration at 6047 mg and 7856 mg per 100 g fresh weight, respectively, quantified via HPLC analysis (Tian et al., 2017). Drying in moderate temperature (50–60 °C to

moisture of 1–3%) helps retaining a higher concentration of phenolic compounds and thus maintaining a high nutraceutical quality of SBLs (Guan, Cenkowski, & Hydamaka, 2005). In particular, processing technique emulating the fermentation of green tea is recommended for preparing SBLs tea, by which bifidogenicity as well as inhibitory activity on *Clostridium perfringens* (*in vitro* digestion assay) were demonstrated (Li et al., 2016). Thus, all these reports suggest that factors such as genotypes, harvesting time, and processing techniques significantly affect the phytochemical profile of SBLs and thus their bioactivity. In addition to the polyphenolics, the SBLs are characterized by high levels of ascorbic acid, tocopherols (including α-tocopherol) and carotenoids (including β-carotene) (Pop et al., 2014; Kanayama et al., 2013).

Globally sea buckthorn is known for the health promoting benefits of the berries. Lately the interest in SBLs is fast growing due to the increasing scientific reports on the bioactivities of leaves and leaf extracts. In comparison to the extensive research published on the composition of sea buckthorn berries, much less is known about the metabolic profiles of SBLs. Increasing the knowledge on the compositional profile of SBLs would promote the utilization of the leaves, which are the currently under-utilized but highly valuable materials. This will support the growth of sustainable agriculture and bioeconomy.

Metabolomics is a highly applicable tool in the phytochemical characterization of plant phenotypes and in understanding the complex networks and relations between the plant, its genetic background and the environmental factors (Kim, Choi, & Verpoorte, 2011). Nuclear magnetic resonance (NMR) spectroscopy based metabolomics has been applied to study the metabolite profiles of tea/herbal tea e.g. green tea (Lee et al., 2015; Lee et al., 2010) and Java tea (Pariyani, Ismail, Ahmad Azam, Abas, & Shaari, 2017). No previous research has been reported on the metabolomics investigation of SBLs. In our recent study, we investigated the metabolic profile of the sea buckthorn berries of two Finnish cultivars, ‘Tytti’ and ‘Terhi’, from two growth locations located at the south and north of

Finland showing strong impact of the latitude of the growth location and respective weather conditions (Kortesniemi, Sinkkonen, Yang, & Kallio, 2017).

In this research, with an objective of systematically identifying the comprehensive phytochemical profile and variation in SBLs during the growth season, we employed an untargeted metabolic fingerprinting approach combining ^{1}H NMR spectroscopy with multivariate data analysis. SBLs samples of 'Tytti' and 'Terhi' were collected in two consecutive growth years from two growing regions located in the south and north of Finland. The metabolic profile of the leaves and the pattern of variations were identified. The constructed chemometric models form a useful tool for classifying SBLs samples in the future. Furthermore, our study was aimed to unveil the association between SBLs metabolites and environmental factors, in particular, variations in the global metabolome as a result of complex interaction among the climatic variables in open fields, which has not been systematically studied to date. The research provides new comprehensive information on the SBLs phytochemicals/metabolites (variations) with respect to major contributory factors such as genotypes, geographical and weather parameters, time of harvest, as guidance for plant breeding, cultivation and commercial utilization of sea buckthorn leaves as raw materials for food, feed, and nutraceuticals.

2. MATERIALS AND METHODS

2.1. Samples

The leaves of two sea buckthorn (*Hippophaë rhamnoides* L. ssp. *rhamnoides*) female genotypes ('Terhi' and 'Tytti') were collected from two locations in southern (Satava, Turku; latitude 60° 23' N, longitude 22° 09' E, altitude 1 m) and northern Finland (Tepasto, Kittilä; 68° 02' N, 24° 37' E, 210 m). An indicative map of the growth locations is shown in Fig. S1. In both locations, the sea buckthorns grew in sandy soil without administration of fertilizers or pest control. In Kittilä, supplementary watering was administered through a drip irrigation system. The leaves were

harvested weekly from two bushes per cultivar at each location during the summers of 2012 and 2013 (weeks 25–39 in the south and weeks 28–40 in the north; Table S1). The sampling was performed randomly from all sides of the foliage. The collected leaves were immediately cooled for transport and stored at –80 °C. The leaves were ground with mortar and pestle with the aid of liquid nitrogen. The lyophilised leaf powder was stored at –80 °C until extraction.

2.2. Meteorological data

The meteorological raw data was provided by the Finnish Meteorological Institute (Helsinki, Finland). The data from weather stations of Turku Artukainen (latitude 60° 27' N, longitude 22° 10' E, altitude 8 m) and Kittilä Pokka (68° 10' N, 25° 47' E, 275 m) were applied. The overall cumulative values from the start of the growth season until the final harvest (temperature sum, precipitation sum and global radiation sum), and weekly averages of relative humidity as well as sunshine hours were calculated (Table S1) and used as variables in the multivariate data analyses.

2.3. Chemicals

Methanol-d_4 (CD_3OD, 99.8 D-%) and deuterium oxide (D_2O, 99.96 D-%) were acquired from VWR International Oy (Espoo, Finland). The 3-(trimethylsilyl)propionic-2,2',3,3'-d_4 acid sodium salt (TSP, 98 D-%) was acquired from Sigma–Aldrich (St. Louis, MO).

2.4. Sample preparation

The extraction of the samples and subsequent acquisition of the NMR spectra were performed in a randomized manner, during July – September 2014. The lyophilised leaf powder (50.0 ± 0.2 mg) was extracted with 1.2 mL deuterated solvent (CD_3OD:D_2O, 8:2 v/v, with 0.02% TSP) by vortexing for 1 h and centrifuged at 9,600 × *g* for 10 min at 20 °C. The supernatant was separated, and 600 µL was taken to NMR analysis. If the extracts were not analysed on the same day, they were stored at –20 °C until analysis.

128 *2.5. NMR spectral acquisition*

129 NMR spectra were recorded using a Bruker Avance 500 NMR spectrometer (Bruker BioSpin AG,
130 Fällanden, Switzerland) operating at 500.13 MHz for proton and equipped with a broadband inverse
131 autotune probe (BBI-5mm-Zgrad-ATM). The solvent-suppressed ^{1}H NMR spectra were acquired at
132 25 °C using a double presaturation pulse programme (Bruker's pulse program *lc1pnf2*), with 256
133 scans, an acquisition time of 3.28 s, a relaxation delay of 6.70 s, and with spectral width of 10 kHz
134 consisting of 64 k data points. The parameters used in acquiring *J*-resolved (JRES) spectra were 16
135 scans, 1 k data points, 128 increments, 2 s relaxation delay, and a spectral width of 16 ppm in
136 dimensions. The heteronuclear multiple bond coherence (HMBC) spectra were obtained using 64
137 scans, 1 k data points, 1.5 s relaxation delay, 256 t1 increments at spectral width of 13 ppm and 220
138 ppm in the proton and carbon dimensions, respectively. The heteronuclear single quantum
139 coherence (HSQC) spectra were acquired using 32 scans, 1 k data points, 128 increments, 2 s
140 relaxation delay, and spectral width of 16 ppm and 165 ppm in the proton and carbon dimensions,
141 respectively.

142 *2.6. Metabolite identification*

143 The identities of NMR signals were assigned with the aid of literature, metabolite library of the
144 Chenomx NMR Suite 8.3 Professional (Chenomx Inc., Edmonton, Alberta, Canada) and querying
145 the open access web based metabolite databases such as Human Metabolome Database (HMDB,
146 http://www.hmdb.ca), and Biological Magnetic Resonance data Bank (BMRB,
147 http://www.bmrb.wisc.edu). The metabolite identities were duly confirmed by two-dimensional
148 NMR experiments such as JRES, HSQC and HMBC.

149 *2.7. Data processing and multivariate data analysis*

150 All raw ^{1}H NMR spectra were processed individually to correct the phasing, baseline, and shim
151 using Chenomx NMR Suite. All the spectra were referenced to the internal standard (TSP) at 0.00

ppm. The chemical shift region 0.0–10.0 ppm was then integrated to bins of width 0.001 ppm after total area normalization, using the Chenomx software. This dataset comprising of 10,000 bins was used to correct the misalignments of the spectra using the *icoshift* algorithm, in MATLAB platform, based on the correlational shifting of spectral intervals (Savorani, Tomasi, & Engelsen, 2010). The dataset was divided into 45 intervals and then average spectrum twice (average2) was used as the target spectrum to realign the misaligned peaks. The spectral region related to residual water (4.68–4.88) and the regions lacking signals such as δ 0.0–0.6 and δ 9.5–10.0 were removed from the aligned spectra. The newly constructed aligned binned data with 9,140 variables was then reduced to 457 chemical shift bin regions of 0.02 ppm width.

The generated dataset comprising 192 observations and 457 variables was then used in the multivariate data analysis using SIMCA-P 14.1 (Umetrics, Sartorius Stedim Biotech, Umeå, Sweden). The mean centered and Pareto scaled data was subjected to unsupervised principal component analysis (PCA) and supervised Orthogonal Partial Least Squares-Discriminant Analysis (OPLS-DA). The grouping patterns of the SBLs samples in different chemometric analysis were observed with the aid of score plots, wherein the spectra were represented as individual points along the principal components. The variables (metabolites) contributing to the characteristic grouping of the samples observed in the score plots were visualized using their corresponding loading plots.

The validation as well as the evaluation of the optimal fit of the OPLS-DA models were performed by internal validation methods of 100 permutation test, calculation of explained variation (R^2Y (cum)), predictive ability (Q^2Y (cum)), and CV-ANOVA values. In addition, external validation using prediction dataset was also carried out in order to assess the fitness and predictive ability of the generated classification models. This was achieved by randomly dividing the dataset into two; first, contains about two third of the samples of the dataset, known as training set, which is used initially to generate the model and the second known as prediction set consisting of the other one

176 third of the samples, aimed to predict the accuracy and evaluate the fitness of the model
177 independently.

178 **3. RESULTS AND DISCUSSION**

179 *3.1. Metabolic profile of sea buckthorn leaves*

180 The characteristic one-dimensional ^{1}H NMR, and two dimensional J-RES and HSQC spectra of the
181 SBLs of genotypes ′Tytti′ and ′Terhi′ extracted in 4:1 methanol water solvent system are shown in
182 Fig. 1. An untargeted metabolic profiling yielded a total of 20 chemical identities, with possible
183 nutritional and pharmacological relevance, belonging to both primary and secondary metabolite
184 classes. Table 1 lists out tentatively identified metabolites with the respective chemical shift
185 assignments, signal multiplicities and coupling constants.

186 The fatty acid signals were found to be one of the most predominant peaks in the spectra in the
187 upfield region. The presence of saturated and unsaturated fatty acids was evident by the
188 characteristic signals of terminal methyl (t, δ 0.88, 0.95), acyl groups of the hydrocarbon chain (m,
189 δ 1.2–1.3) and olefinic protons (m, δ 5.34). The presence of linoleic acid and linolenic acid was
190 confirmed from two-dimensional HSQC spectra.

191 The singlet peaks at δ 0.67 and δ 0.68 were attributed to H-18 of the phytosterols. Together with
192 these, other characteristic signals including δ 0.83 (H-27, d, J = 6.0 Hz), δ 0.84 (H-26, d, J = 6.4
193 Hz), δ 0.88 (H-29, t, J = 6.8 Hz), δ 1.01 (H-19, s), and δ 1.02 (H-21, d, J = 6.8 Hz) hinted at the
194 presence of β-sitosterol/β-stigmasterol. In addition, the connectivity of H-26 and H-27 with C-25, as
195 evident by the HMBC correlation at δ_C 29.6, further supported the identification. The presence of a
196 few more singlets in the δ 0.75–0.81 region suggests the possible presence of other phytosterols.
197 The presence of esterified sterols and trienols is already reported in SBLs (Suryakumar & Gupta,
198 2011; Guan et al., 2005). In addition to the fatty acids and phytosterols, some amino acids and
199 organic acids were also detected in the 0.5–3.0 ppm region of the spectra, as listed in Table 1.

The mid-region (3.0 to 5.5 ppm) of the spectra, primarily contributed by the characteristic peaks of carbohydrates and sugars, was heavily congested; the presence of glucose, fructose and sucrose have been confirmed by the characteristic anomeric proton doublets at δ 4.53/5.14, δ 4.13 and δ 5.40 ppm, respectively. Other metabolites identified from this region included *myo*-inositol, L-glutamic acid and gluconic acid.

The aromatic region (6.0–10.0 ppm) of the spectra showed several signals, however, they were less intense compared to other regions. Gallic acid and ellagic acid were identified from their characteristic singlets at δ 7.04 and δ 7.53, and confirmed by the HSQC and HMBC spectra. The region from δ 6.25–6.85 showed several singlets, which have been attributed to ellagitannins (ETs). Structurally, ETs are characterized with the presence of hexahydroxy diphenoyl (HHDP) unit(s) linked to sugar moiety. The H6 proton of each ring contributes distinct singlets in the δ 6.2–6.4 region, which were duly confirmed from the JRES NMR spectra. The presence of valoneoyl and galloyl groups was suggested by the presence of singlets at 7.11, 7.09, 7.06, 6.85, 6.52 and 6.27 in the JRES NMR spectra. Further, the characteristic correlations of the H6 protons with C7 were evident from the HMBC spectra (Fig. S2). In addition, the peaks in the range of δ_C 138 in HSQC are characteristic of the single bond correlation of H6 with C6. The SBLs of Finnish origin are reported to contain on an average 70–73 mg of ETs per gram dry weight (Suvanto, & Salminen, 2016). The major ETs present in quantifiable levels in SBLs include hippophaenin A, B and C, castalagin, vescalagin, pedunculagin, casuarinin and stachyurin (Suvanto et al., 2018). All these compounds are characterized with high structural similarities; for example, hippophaenin B - C, castalagin - vescalagin, and casuarinin - stachyurin are epimer pairs. Similarly, the only difference of casuarinin and stachyurin from castalagin and vescalagin is the absence of a C-C bond between the B and C rings. These close structural similarities coupled with the extensive overlap of the signals hindered the unambiguous individual identity assignment of ETs.

The doublets (J = 8.5 Hz) at δ 6.80 and δ 6.95 represented H5' of the flavonol aglycone, which was further confirmed by the HMBC correlation with δ_C 123.0 and δ_C 148.8. The doublet at δ 7.63 (J = 8.7 Hz) is correlated in HSQC with δ_C 126.2, which represents the correlation of the C6' aglycon. Similarly, the HSQC correlations at δ_C 116–118 represent characteristic shifts of C5'. Doublets with a coupling constant of 2.5 Hz at δ 6.36 and δ 6.55 further endorse the presence of flavonol glycosides. The polyphenol pattern identified by the NMR metabolic profiling of SBLs is in good agreement with previous report that ETs constitute more than 90% of the total phenolics in SBLs, and the rest constituted by flavonol glycosides (Tian et al., 2017).

The metabolic profile of SBLs differs significantly from that of berries, which is the most commonly utilized/consumed part of the sea buckthorn plant. The most differentiating metabolic feature is the polyphenolic profile. The polyphenolic profile of SBLs is reported to be constituted mostly by ETs, whereas, the flavonol glycosides composed of the isorhamnetin and quercetin glycosides represent the major phenolic compounds in berries (Tian et al., 2017). According to Fatima et al., 2015, SBLs have several-fold higher levels of phenolics with gallic acid as the predominant phenolic acid. SBLs are characterized by the presence of ellagic, sinapic, and cinnamic acids and rutin, and the absence of myricetin, whereas, the berries are rich in *p*-coumaric acid, myricetin and quercetin but lack rutin (Fatima et al., 2015). Our previous study on the NMR metabolic profile of the berries collected from the cultivars ′Tytti′ and ′Terhi′ identified unique metabolites such as L-quebrachitol, and ethyl as well as methyl β-D-glucopyranoside (Kortesniemi et al., 2017), however, they could not be identified from the spectra of leaves of the corresponding cultivars. Other constituents of nutritional and sensory significance, including, malic acid, asparagine and ascorbic acid, and several other primary metabolites were also identified from berries (Liu et al., 2017).

3.2. Unsupervised chemometric investigation using Principal Component Analysis

Principal component analysis (PCA) is the most commonly used unsupervised dimensionality reduction tool in metabolomics to investigate the main variance, detect grouping trends and outliers in a dataset. The variations of the dataset are visualized along the principal components (PC), wherein, the first PC constitutes the highest explained variation. The focal strength of PCA is that the intragroup variations and/or larger sources of variability in the dataset are highlighted (Worley & Powers, 2013).

The PCA model generated from the mean-centered and Pareto scaled data showed excellent goodness of fit ($R^2X_{(cum)}$ = 0.93) and high predictive ability ($Q^2_{(cum)}$ = 0.86). The first three PCs together constituted for a total of 69.7% of the variance involved in the dataset (Table 2). As observed from the score scatter plot (Fig. 2 A and Fig. S3), the most significant variation in the analyzed SBLs samples was brought about by the differences in the stages of growth. The samples representing the early stages of growth (samples collected from the beginning of harvest until first week of August) were clustered on the negative axis of the PC1, against those samples representing the later stages seen on the positive axis of PC1. The PC1 constituted for 30.3% of the total variance of the dataset. The discriminant grouping according to the growth stage was observed uniformly on both the studied cultivars, irrespective of the geographical location. This is further evident from the clustering of initial seven weeks of samples from northern Finland together, although the onset of growth season in the north was delayed (week 22), when compared with the south (week 18).

The corresponding loadings scatter plot (Fig. 2 B) showed that the main variables responsible for the separation on the first principal component (p[1]) were bins 1.25 and 1.29 ppm (fatty acids), and aromatic region constituted by polyphenols (bin labels 6.0–7.2 ppm) on the initial growth stages. The late growth stages were characterized with the chemical shift bins contributed by the sugars and carbohydrate regions (bin label 3.0–4.0 ppm), which are seen on the positive axis of PC1. However, the α-glucose (bin 5.15) identified from the characteristic anomeric protons was found to be higher on the initial stages of growth.

A trend in the distribution of samples according to the growth region could be seen along PC2; however, a clear discrimination of samples of the north from those of the south was brought together with PC3 (Fig. 2 C). The PC2 and PC3 explain 28% and 12% of the total variation, respectively. The chemical shift bins representing the acyl groups of fatty acids (bins 1.25, 1.29 and 2.77) seem to have the predominant contribution towards the samples in the north, against those from the sugars and carbohydrates (bin label 3.0–4.0 ppm) in the south (Fig. 2 D). A growing body of literature show that abiotic stressors such as variations in temperature, light, and salinity as well as drought cause accumulation of lipids, including fatty acids, in the plant tissues (Singer, Zou, & Weselake, 2016; De Bigault Du Granrut & Cacas., 2016). Higher abundance of fatty acids indicated by the higher intensities of acyl group signals in the SBLs from north Finland might be elicited by the prevalent unique environmental stressor such as low temperature.

The PCA identified and highlighted that among the various variable conditions possibly contributing to the differed metabolite compositions, growth stage exerts the highest effect, followed by the growth region. A separation based on the cultivars was not evident from the unsupervised analysis, suggesting that the variations in metabolic profiles of the leaves of the selected cultivars ′Tytti′ and ′Terhi′ are more subtle than the growth stages and geographical origin.

3.3. Supervised chemometric investigation using Orthogonal Projections to Latent Structures-Discriminant Analysis (OPLS-DA)

The dataset was subjected to supervised multivariate data analysis, which is particularly used in building models capable of classifying (future) samples using the available spectral data. This is performed by fitting the samples in the discriminant analysis version of OPLS models (i.e., OPLS-DA), which is a highly used tool in the field of metabolomics in recent years. The most significant variables contributing to the discrimination of the two groups under analysis were determined by a three-tier approach. At the first step, variables carrying high weighting in the differentiation of the two groups located at the two tails of the S-plot were identified. As a second tier criterion, among

298 the variables identified from the S-plot, only those with a Variable Important in Projection (VIP) ≥
299 1 were considered as discriminant markers. In addition to these, thirdly, the jack-knife bars of the
300 thus selected variables were sorted from the loading column plots. Thus, the systematic approach of
301 combining information from these three approaches ensured the identification of the discriminant
302 variables of statistical significance by filtering out those variables with shared features between the
303 groups.

304 The OPLS-DA analyses were performed to classify and predict SBLs samples based on the growth
305 stage and growth location, which were identified to be responsible for 70% of the total variance
306 involved in the whole dataset from the PCA model. In addition to that, the ability to discriminate
307 the cultivars (′Tytti′ and ′Terhi′) was also investigated.

308 The validation of the supervised multivariate models, such as OPLS-DA, is particularly important
309 to eliminate the potential risk of over fitting. The plots generated from the permutation tests (100 on
310 both variables) are shown in Fig. S4. All three models proved to have good validity based on their
311 R2Y- and Q2Y-intercept values being lower than 0.3–0.4 and 0.05, respectively (Eriksson et al.,
312 2013). More detailed validation features including R2Y, Q2Y and CV-ANOVA are presented in
313 Table 2. In addition, an external validation method was employed to estimate how well the
314 generated model will perform when applied to new samples, as described in section 3.3.1.

315 The Fig. 3- A, C and E showed a clear discrimination of samples of early vs late stages of growth,
316 north vs south growth locations, and cultivars ′Tytti′ vs ′Terhi′, respectively, by component 1. The
317 model diagnostics including the fit and robustness of the specific OPLS-DA models are presented in
318 Table 2. It is evident from Fig. 3 B that the early growth stage was characterized with the presence
319 of bins representing the ETs (bins 7.15, 7.13, 7.11, 7.09, 6.67, and 6.65), flavonol glycosides (bins
320 6.93 and 6.73), and fatty acids (bins 2.15, 1.97, 1.25 and 1.23), hinting at their higher abundance in
321 the initial stages of growth compared to the rest of the weeks in the growth period. It was reported
322 that the levels of flavonols and flavonol glycosides in SBLs of Russian origin grown in Sweden

decrease as the season advances towards the end of July, while the levels of procyanidins and hydrolysable tannins increase (Morgenstern et al., 2014). It could be suggested from the current findings that the general trend of polyphenolic profile, including the flavonols and ellagitannins, in Finnish SBLs reached their peak by the first week of August in the south, while in the north it remained on the higher side until the end of August during the two years investigated. The bins characteristic of the sugars and carbohydrates on the positive axis of the S-plot suggested their higher abundance in the later growth stages. With regard to the metabolites responsible for the discrimination of SBLs samples between north and south growth location, it is clear from Fig. 3 D that the samples from north were characterized with the presence of bins representing the sugars hinting at their higher abundance. The α- (bin 5.15) and β- (bin 4.54) glucose are particularly identified to be the discriminating metabolites in the north. On the other hand, samples from south had higher presence of resonance signals at the aromatic regions, in particular those of ETs (bins 7.07, 6.83, 6.41). Succinate (bin 2.33) was identified to be another significant discriminating marker in samples from south. Comparing the two cultivars, the higher abundance of fatty acids (bins 1.25 and 1.29) in cultivar ′Terhi′ was clearly evident from the Fig. 3 F. Other classes of metabolites found to be present in higher quantities in ′Terhi′ included carbohydrates and sugars. On the other hand, the samples from ′Tytti′ had higher presence of ETs (bins 7.13, 7.11, 7.05, and 6.27) and flavonol glycosides (bins 6.95, 6.93, 6.83, and 6.79).

3.3.1. Prediction and external validation of the OPLS-DA models

The *Y*-predicted score scatter plots of the external validation sets of the OPLS-DA models specific to the discrimination of SBLs based on growth stages, growth locations and cultivars are shown in Fig. 4- A, C and E, respectively. Table 2 shows that the models have demonstrated high correct classification rates ranging from 95–100% based on the criteria of assigning to the class with nearest *Y*-prediction score values, and 85– 94% with a more stringent class assignment criteria set at a *Y*-prediction score value of 0.65.

In addition, all three prediction models (on growth stage, growth location, and cultivars) were able to classify the samples with high accuracy indicated by the high sensitivity and specificity summarized via Receiver Operating Characteristic (ROC) plots in Fig. 4- B, D and F. The ROC plot represents the trade-off between the sensitivity (true positives) and specificity (false positives) on the Y and X axis, respectively. The area under the ROC curve (AUC) is an estimate of the accuracy of the binary classification, where a value equal to 1.0 represents the complete separation of the two classes (Alonso, Marsal, & Julià, 2015).

3.4. Correlation of the identified discriminatory metabolites with the weather parameters

A PLS-DA model was used to understand the association of metabolic composition of SBLs with the characteristic weather parameters in the north and south growth locations in Finland. The variables of the dataset comprised the chemical shift bins of the identified discriminatory metabolites, i.e., fatty acids, sugars and carbohydrates, polyphenolics, and the chosen weather parameters (Table S1). It is evident from the weather data that north Finland differs from south in having shorter growth period with a cooler growth environment marked with low- temperature, sunshine hours and radiation. The temperature sum of the growth season in the north is about 40% lower compared to that of the south, in both growth years. However, there was no large difference in the relative humidity between the north and the south.

The constructed PLS-DA model from the UV scaled data was optimally good as shown by the validation via 100 permutation test (Y-intercepts of R^2 and Q^2 0.099 and −0.406, respectively), as well as $R^2Y_{(cum)} = 0.643$ and $Q^2_{(cum)} = 0.469$, respectively. Fig. S5 presents the biplot obtained from PLS-DA, which is a combination of the score and loading plots represented in a single plane. The most striking observation is that all chemical shift bins representing sugars have clustered along with the northern SBLs samples, situated directly opposite to the weather variables related to higher temperature. This suggested an inverse correlation between the abundance of sugars/carbohydrates in SBLs and the temperature. In addition, sugars showed an inverse correlation with radiation and

sunshine hours and a direct correlation with relative humidity. The soluble sugar concentrations of the *Eucalyptus tereticornis* leaves decreased at daily ambient air temperature above 20 °C, whereas, it remained relatively constant in the temperature window of 10–20 °C (Aspinwall et al., 2016). In another study on the cold acclimation of the tea plants, the total soluble sugar content was found significantly rising during the winter, reaching the peak on the day with the lowest recorded temperature. A general up-regulation of the genes involved in the starch, sucrose and raffinose metabolism, as well as sugar transporters was also observed (Yue et al., 2015). The protective role of sugars against dehydration and freezing caused by various environmental stressors are shown to be mediated by their osmoprotectant and interactive ability with the phospholipid bilayer (Wingler, Stangberg, Saxena, & Mistry, 2012; Shao, He, Bao, & Mao, 2009).

The observed pattern of fatty acids suggested that their abundance in SBLs is primarily governed by the developmental cue, as all the characteristic signals were clustered near to the samples representing early growth stages, from both south and north locations. The decline in the levels of fatty acids during the later growth stages could be related to the onset of leaf senescence. Previous research on the leaves of Arabidopsis (*Arabidopsis thaliana*), *Brachypodium distachyon*, and switchgrass (*Panicum virgatum*) reported that the leaf senescence brought about a decline in the levels of fatty acids, culminated at about 80% reduction at the end of the growth season (Yang & Ohlrogge, 2009). Chemical shift bins of the acyl moiety (1.25 and 1.23) representing the fatty acid chain length were clustered differently from other fatty acid signals on the quadrant opposite to the temperature, radiation and sunshine. This could be suggestive of the accumulation of fatty acids with increased chain length in response to abiotic stressors. Fatty acid carbon chain length is found to be influenced by factors such as temperature as evident by the increased production of erucic acid, a very long chain fatty acid (C22:1), in rapeseed at low temperature (Singer et al., 2016). The underlying mechanism for the influence of temperature on carbon chain length has not been fully elucidated yet (De Bigault Du Granrut & Cacas., 2016).

The correlation analysis of metabolites with the weather variables performed in this study was aimed at deriving a primary insight on how the metabolite pattern in SBLs is affected by real time environmental stressors. However, the study is limited by the inclusion of data from only two growth years, which is inferior in its ability to propose a concrete correlation. Hence, monitoring several years' data or studies in controlled growth environments are recommended to establish more specific correlations.

4. CONCLUSION

An untargeted analysis using NMR spectroscopy coupled with various chemometric methods identified a wide range of metabolites in SBLs and revealed the variation in metabolic composition of SBLs in respect to the time of harvest, growth location, and cultivar. Leaves harvested early in the summer contain higher proportions of lipids and potential bioactive components, whereas late harvest yielded leaves with relatively higher stores of carbohydrates and sugars. The leaves of cultivar ′Terhi′ had relatively more lipids compared to ′Tytti′, although these two cultivars shared a closely similar metabolic profile. The correlation analysis of the SBLs metabolites with the environmental factors revealed that abiotic stress conditions, primarily low temperature, promote the accumulation of fatty acids with higher carbon chain length as well as carbohydrates and sugars in the SBLs.

As a polyphenolics-rich side-stream of the berry crop, the SBLs may hold the potential as raw material for food and nutraceutical industry. Our study demonstrated that the phenolic content is the highest in sea buckthorn leaves collected in the early part of the growth period. This should be taken into consideration when harvesting the sea buckthorn leaves for applications in food, feed and nutraceuticals. To our best knowledge, this is the first metabolomics study of SBLs using non-targeted NMR-metabolomics method. The systematic information on the metabolic characterization of SBLs contributed by this study could guide not only in strategizing the collection and effective utilization of SBLs but also in confirming their authenticity/quality control.

Acknowledgments

Funding from Tekes – the Finnish Innovation Funding Agency (decision 40055/13 for the project "Finnish-Indian Ingredients for Improving Food Safety and Health – FINDfood") is gratefully acknowledged. The authors wish to thank Hannu Lappalainen and Seppo Lappalainen for maintaining the field experiments, Thomas Josserand for technical assistance, and everyone having taken part in the leaf sampling.

Author Contributions

Conception and design of study: JL, MK and BY. Acquisition of data: JL, RP and MK. Analysis and interpretation of data: RP, MK, BY. Drafting the manuscript: RP. Revising the manuscript critically for important intellectual content: MK, BY, JL, JS.

Conflict of interest

The authors declare no conflicts of interest.

435 **REFERENCES**

436 Alonso, A., Marsal, S., & Julià, A. (2015). Analytical methods in untargeted metabolomics: state of
437 the art in 2015. *Frontiers in bioengineering and biotechnology*, *3*, 23.

438 Aspinwall, M.J., Drake, J.E., Campany, C., Vårhammar, A., Ghannoum, O., Tissue, D.T., Reich,
439 P.B., & Tjoelker, M.G. (2016). Convergent acclimation of leaf photosynthesis and respiration to
440 prevailing ambient temperatures under current and warmer climates in *Eucalyptus tereticornis*. *New*
441 *Phytologist*, *212*(2), 354-367.

442 Cho, C.H., Jang, H., Lee, M., Kang, H., Heo, H.J., & Kim, D.O. (2017). Sea buckthorn (*Hippophae*
443 *rhamnoides* L.) leaf extracts protect neuronal PC-12 cells from oxidative stress. *Journal of*
444 *Microbiology and Biotechnology*, *27*(7), 1257-1265.

445 Cho, H., Cho, E., Jung, H., Yi, H. C., Lee, B., & Hwang, K. T. (2014). Antioxidant activities of sea
446 buckthorn leaf tea extracts compared with green tea extracts. *Food Science and Biotechnology*,
447 *23*(4), 1295-1303.

448 De Bigault Du Granrut, A., & Cacas, J.L. (2016). How Very-Long-Chain Fatty Acids Could Signal
449 Stressful Conditions in Plants?, *Frontiers in plant science*, *7*, 1490.

450 Dubey, S., Deep, P., & Singh, A.K. (2016). Phytochemical characterization and evaluation of anti-
451 cataract potential of sea buckthorn leaf extract. *Veterinary ophthalmology*, *19*(2), 144-148.

452 Eriksson, L., Byrne, T., Johansson, E., Trygg, J., & Vikström, C. (2013). Multi- and megavariate
453 data analysis – basic principles and applications (3rd ed.,). Malmö, Sweden: MKS Umetrics AB.

454 Fatima, T., Kesari, V., Watt, I., Wishart, D., Todd, J.F., Schroeder, W.R., Paliyath, G., & Krishna,
455 P. (2015). Metabolite profiling and expression analysis of flavonoid, vitamin C and tocopherol
456 biosynthesis genes in the antioxidant-rich sea buckthorn (*Hippophae rhamnoides*
457 L.). *Phytochemistry*, *118*, 181-191.

458 Ganju, L., Padwad, Y., Singh, R., Karan, D., Chanda, S., Chopra, M.K., Bhatnagar, P., Kashyap, R.,
459 & Sawhney, R.C. (2005). Anti-inflammatory activity of Seabuckthorn (*Hippophae rhamnoides*)
460 leaves. *International Immunopharmacology*, *5*(12), 1675-1684.

461 Guan, T.T., Cenkowski, S., & Hydamaka, A. (2005). Effect of drying on the nutraceutical quality of
462 sea buckthorn (*Hippophae rhamnoides L. ssp. sinensis*) leaves. *Journal of food science*, *70*(9),
463 E514-E518.

464 Kanayama, Y., Sato, K., Ikeda, H., Tamura, T., Nishiyama, M., & Kanahama, K. (2013). Seasonal
465 changes in abiotic stress tolerance and concentrations of tocopherol, sugar, and ascorbic acid in sea
466 buckthorn leaves and stems. *Scientia Horticulturae*, *164*, 232-237.

467 Kim, S.J., Hwang, E., Yi, S.S., Song, K.D., Lee, H.K., Heo, T.H., Park, S.K., Jung, Y.J., & Jun,
468 H.S. (2017). Sea buckthorn leaf extract inhibits glioma cell growth by reducing reactive oxygen
469 species and promoting apoptosis. *Applied Biochemistry and Biotechnology*, *182*(4), 1663-1674.

470 Kim, H. K., Choi, Y. H., & Verpoorte, R. (2011). NMR-based plant metabolomics: Where do we
471 stand, where do we go?. *Trends in Biotechnology*, *29*(6), 267-275.

472 Kortesniemi, M., Sinkkonen, J., Yang, B., & Kallio, H. (2017). NMR metabolomics demonstrates
473 phenotypic plasticity of sea buckthorn (*Hippophaë rhamnoides*) berries with respect to growth
474 conditions in Finland and Canada. *Food Chemistry*, *219*, 139-147.

475 Lee, H.I., Kim, M.S., Lee, K.M., Park, S.K., Seo, K.I., Kim, H.J., Kim, M.J., Choi, M.S., & Lee,
476 M.K. (2011). Anti-visceral obesity and antioxidant effects of powdered sea buckthorn (*Hippophae*
477 *rhamnoides* L.) leaf tea in diet-induced obese mice. *Food and chemical toxicology*, *49*(9), 2370-
478 2376.

479 Lee, J. E., Lee, B. J., Chung, J. O., Hwang, J. A., Lee, S. J., Lee, C. H., & Hong, Y. S. (2010).
480 Geographical and climatic dependencies of green tea (*Camellia sinensis*) metabolites: a ^{1}H NMR-
481 based metabolomics study. *Journal of Agricultural and Food Chemistry*, *58*(19), 10582-10589.

Lee, J.E., Lee, B.J., Chung, J.O., Kim, H.N., Kim, E.H., Jung, S., Lee, H., Lee, S.J., & Hong, Y.S. (2015). Metabolomic unveiling of a diverse range of green tea (*Camellia sinensis*) metabolites dependent on geography. *Food chemistry*, *174*, 452-459.

Li, G., Zhang, J., Liu, E., Wang, F. H., Qi, S., Xiang, X., & Du, W. (2016). Effects of different sea buckthorn leaf tea processing technologies on nutrient level and fecal microflora in vitro. *Journal of Food & Nutrition Research*, *55*(3), 205-213.

Liu, Y., Fan, G., Zhang, J., Zhang, Y., Li, J., Xiong, C., Zhang, Q., Li, X., & Lai, X. (2017). Metabolic discrimination of sea buckthorn from different *Hippophaë* species by ^{1}H NMR based metabolomics. *Scientific Reports*, *7*(1), 1585.

Michel, T., Destandau, E., Le Floch, G., Lucchesi, M. E., & Elfakir, C. (2012). Antimicrobial, antioxidant and phytochemical investigations of sea buckthorn (*Hippophaë rhamnoides* L.) leaf, stem, root and seed. *Food Chemistry*, *131*(3), 754-760.

Morgenstern, A., Ekholm, A., Scheewe, P., & Rumpunen, K. (2014). Changes in content of major phenolic compounds during leaf development of sea buckthorn (*Hippophaë rhamnoides* L.). *Agricultural and Food Science*, *23*, 207-219.

Pariyani, R., Ismail, I. S., Ahmad Azam, A., Abas, F., & Shaari, K. (2017). Identification of the compositional changes in *Orthosiphon stamineus* leaves triggered by different drying techniques using ^{1}H NMR metabolomics. *Journal of the Science of Food and Agriculture*, *97*(12), 4169-4179.

Pop, R. M., Weesepoel, Y., Socaciu, C., Pintea, A., Vincken, J. P., & Gruppen, H. (2014). Carotenoid composition of berries and leaves from six romanian sea buckthorn (*Hippophae rhamnoides* L.) varieties. *Food Chemistry*, *147*, 1-9.

Pop, R.M., Socaciu, C., Pintea, A., Buzoianu, A.D., Sanders, M.G., Gruppen, H., & Vincken, J.P. (2013). UHPLC/PDA–ESI/MS analysis of the main berry and leaf flavonol glycosides from different Carpathian *Hippophaë rhamnoides* L. varieties. *Phytochemical Analysis*, *24*(5), 484-492.

506 Savorani, F., Tomasi, G., & Engelsen, S.B. (2010). Icoshift: A versatile tool for the rapid alignment
507 of 1D NMR spectra. *Journal of Magnetic Resonance*, *202*(2), 190-202.

508 Shao, Y., He, Y., Bao, Y., & Mao, J. (2009). Near-infrared spectroscopy for classification of
509 oranges and prediction of the sugar content. *International Journal of Food Properties*, *12*(3), 644-
510 658.

511 Singer, S.D., Zou, J., & Weselake, R.J. (2016). Abiotic factors influence plant storage lipid
512 accumulation and composition. *Plant Science*, *243*, 1-9.

513 Singh, V., Yang, B., Kallio, H., Bala, M., Sawhney, R.C., Gupta, R.K., Mörsel, J.T., Lu, R., &
514 Tolkachev, O.N. (2005). Seabuckthorn (*Hippophae* L.): A Multipurpose Wonder Plant, Vol. II:
515 Biochemistry and Pharmacology. India: Daya Publishing House.

516 Suryakumar, G., & Gupta, A. (2011). Medicinal and therapeutic potential of sea buckthorn
517 (*Hippophae rhamnoides* L.). *Journal of Ethnopharmacology*, *138*(2), 268-278.

518 Suvanto, J., & Salminen, J. P. (2016). Quantification of the main ellagitannins of *Hippophae*
519 *rhamnoides* leaves. *Planta Medica*, *81*(S 01), S317.

520 Suvanto, J., Tahtinen, P., Valkamaa, S., Engström, M.T., Karonen, M., & Salminen, J.P. (2018).
521 Variability in Foliar Ellagitannins of *Hippophaë rhamnoides* L. and Identification of a New
522 Ellagitannin, Hippophaenin C. *Journal of Agricultural and Food Chemistry*, *66*(3), 613-620.

523 Tian, Y., Liimatainen, J., Alanne, A.L., Lindstedt, A., Liu, P., Sinkkonen, J., Kallio, H., & Yang, B.
524 (2017). Phenolic compounds extracted by acidic aqueous ethanol from berries and leaves of
525 different berry plants. *Food chemistry*, *220*, 266-281.

526 Wingler, A., Stangberg, E. J., Saxena, T., & Mistry, R. (2012). Interactions between temperature
527 and sugars in the regulation of leaf senescence in the perennial herb *Arabis alpina* L. F. *Journal of*
528 *Integrative Plant Biology*, *54*(8), 595-605.

529 Worley, B., & Powers, R. (2013). Multivariate analysis in metabolomics. *Current*
530 *Metabolomics*, *1*(1), 92-107.

531 Yang, Z., & Ohlrogge, J.B. (2009). Turnover of fatty acids during natural senescence of
532 Arabidopsis, Brachypodium, and switchgrass and in Arabidopsis β-oxidation mutants. *Plant*
533 *Physiology*, *150*(4), 1981-1989.

534 Yogendra Kumar, M. S., Tirpude, R. J., Maheshwari, D. T., Bansal, A., & Misra, K. (2013).
535 Antioxidant and antimicrobial properties of phenolic rich fraction of sea buckthorn (*Hippophae*
536 *rhamnoides* L.) leaves in vitro. *Food Chemistry*, *141*(4), 3443-3450.

537 Yoshida, T., Tanaka, K., Chen, X., & Okuda, T. (1991). Tannins from *Hippophae rhamnoides*.
538 *Phytochemistry*, *30*(2), 663-666.

539 Yue, C., Cao, H.L., Wang, L., Zhou, Y.H., Huang, Y.T., Hao, X.Y., Wang, Y.C., Wang, B., Yang,
540 Y.J., & Wang, X.C. (2015). Effects of cold acclimation on sugar metabolism and sugar-related gene
541 expression in tea plant during the winter season. *Plant Molecular Biology*, *88* 6), 591-608.

542 Zhamanbaeva, G. T., Murzakhmetova, M. K., Tuleukhanov, S. T., & Danilenko, M. P. (2014).
543 Antitumor activity of ethanol extract from *Hippophae rhamnoides* L. leaves towards human acute
544 myeloid leukemia cells in vitro. *Bulletin of Experimental Biology and Medicine*, *158*(2), 252-255.

Table 1. Metabolites identified from the ^{1}H and 2D NMR spectra of SBLs extracts with the signal assignments, chemical shift (δ_H, ppm), multiplicity (s, singlet; d, doublet; t, triplet; dd, doublet of doublets; m, multiplet) and scalar coupling constant (*J*, Hz) values

No.	Metabolite	Assignment	δ_H	Multiplicity	*J* (Hz)
1	Phytosterols β-sitosterol/β-stigmasterol		0.67–0.81	s	-
		H18	0.68	s	-
		H27	0.83	d	6.0
		H26	0.84	d	6.4
		H19	1.01	s	-
		H21	1.02	d	6.8
2	Fatty acids				
		CH_3 (acyl group), linoleic (ω6)	0.88	t	7.0
		CH_3 (acyl group), α-linolenic (ω3)	0.95	t	7.5
		-(CH_2)n- (acyl group)	1.28	m	-
		-OCO-CH_2-CH_2- (acyl group)	1.59	m	-
		-CH_2CH = CH (acyl group), unsaturated fatty acids	2.05	m	-
		-OCO-CH_2- (acyl group)	2.33	m	-
			2.35	t	7.1
		= $CHCH_2CH$ = (acyl group), linolenic and linoleic	2.78	m	-

		-CH=CH-, unsaturated fatty acids	5.34	m	-
3	Threonine	CH_3	1.32	d	6.3
		H4	3.62	d	4.9
4	Alanine	β-CH_3	1.47	d	7.5
		α-CH	3.64	m	-
5	Acetate	CH_3	1.90	s	-
6	Succinate	α, β-CH_2	2.33	s	-
7	γ amino butyric acid	β-CH_2	1.89		
		α-CH_2	2.26	t	7.55
		γ-CH_2	2.98	t	7.05
8	Choline	CH_3	3.16	s	-
9	Myo-inositol	CH	3.21	t	9.6
			3.59	t	9.6
			4.02	t	2.3
10	Gluconic acid	CH_2	3.64		
		CH	3.77	m	-
		CH	4.12	d	3.7
11	L-glutamic acid		2.05	m	-
			2.34	m	-
			3.74	dd	7.2, 4.8
12	Fructose	CH	4.13	d	8.3
13	β-glucose	C_1H	4.53	d	8.0

14	α-glucose	C_1H	5.14	d	3.9
15	Sucrose	Glc-C_1H	5.40	d	3.85
16	Ellagitannins	Galloyl H6	7.09	s	-
			7.06	s	-
			6.85	s	-
		Valoneoyl H	7.11	s	-
			6.52	s	-
			6.27	s	-
17	Flavonol glycosides	Aglycon H5′	6.95	d	8.5
		H6′	7.63	d	8.7
		H6	6.36	d	2.5
			6.54	d	2.5
18	Gallic acid	H2, H6	7.04	s	-
19	Ellagic acid		7.53	s	-
20	Trigonelline	H1	9.14	s	-
		H3	8.85	m	-
			8.05	m	-
			4.41	s	-

545

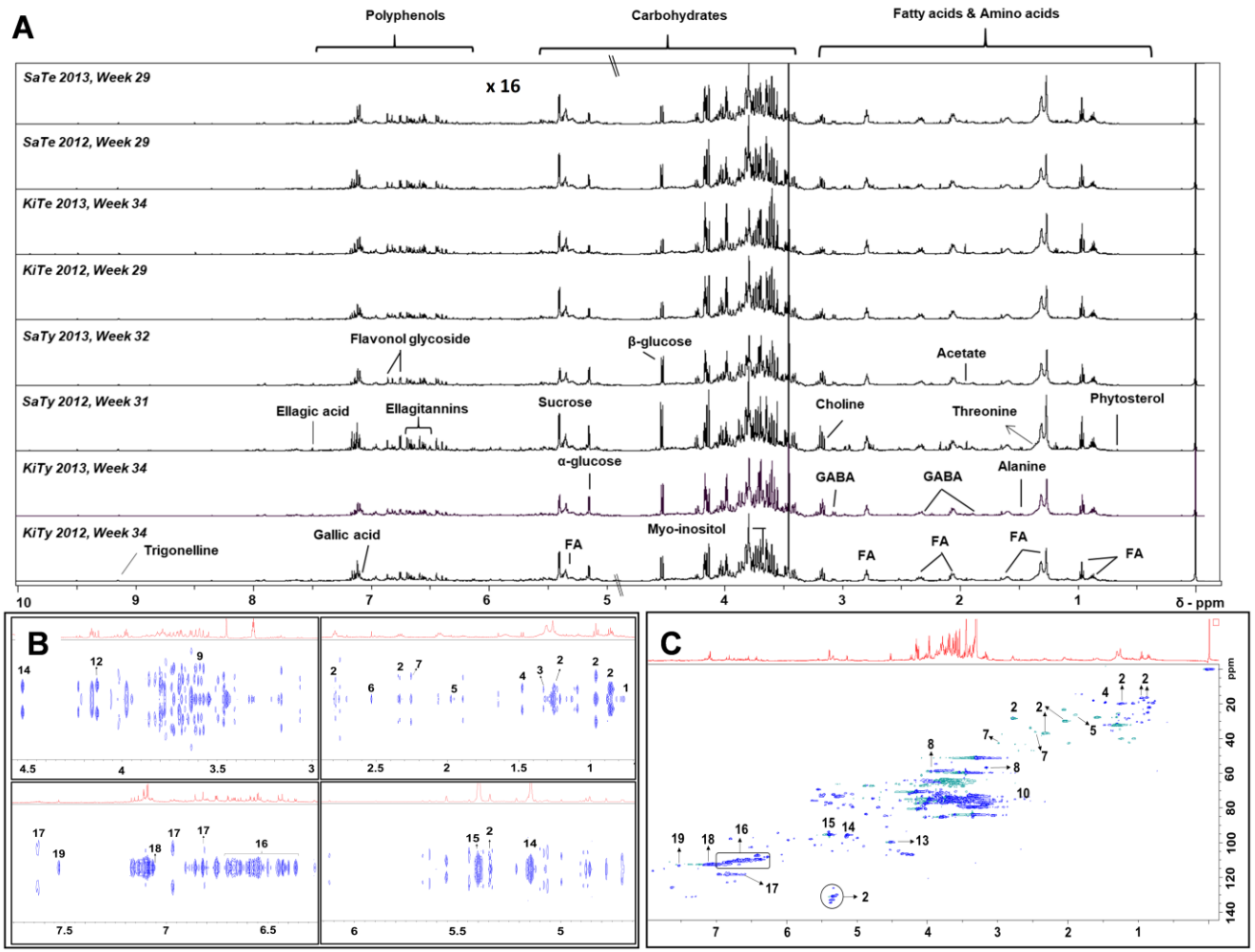
Table 2. Summary of various chemometric models used in the analysis of ^{1}H NMR spectra of SBLs samples

Parameter of study	Chemometric model	Number of components	R^2X (cum)	R^2Y (cum)	Q^2 (cum)	CV-ANOVA	Number of samples		External validation classification rate (%)	
							Training	Prediction[b]	Nearest class[c]	Y_{pred} > 0.65
General trends and outliers	PCA	18	0.926	-	0.859	-	-	-	-	-
	PCA	3	0.697	-	0.640	-	-	-	-	-
Growth stage	OPLS-DA	1+3+0[a]	0.632	0.830	0.763	6.92e-035	132	60	95	85
Growth location	OPLS-DA	1+2+0[a]	0.459	0.826	0.771	2.26e-037	131	61	100	93.44
Cultivars	OPLS-DA	1+3+0[a]	0.674	0.834	0.785	4.09e-037	131	61	96.72	88.52

[a] Predictive component + orthogonal in *X* component + orthogonal in *Y* component

[b] Separately chosen samples in random from the main dataset

[c] Class membership assigned based on the proximity of Y_{pred} values



549

550 **Figure 1.** NMR spectra of SBLs samples, A: Representative one dimensional ^{1}H NMR spectrum of SBLs collected from south (Sa) and north (Ki)
551 growing regions of Finland, belonging to the cultivars ´Terhi' (Te) and 'Tytti' (Ty), during the summers of 2012 and 2013, B: Expanded regions of
552 JRES spectrum, C: HSQC spectrum. Refer to Table 1 to interpret the metabolite numbers in B and C. Abbreviations: FA- Fatty acid, GABA- γ amino
553 butyric acid.

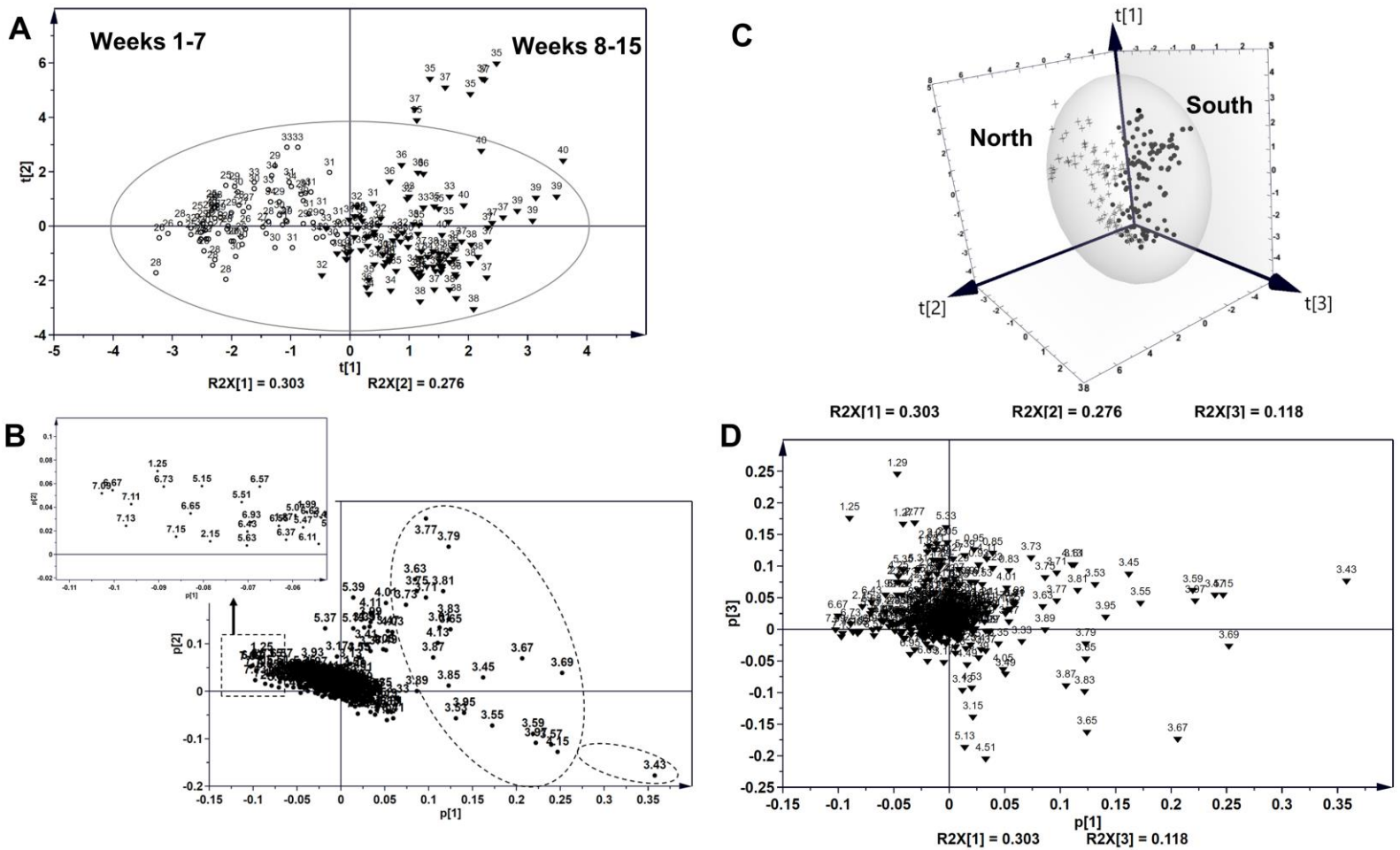
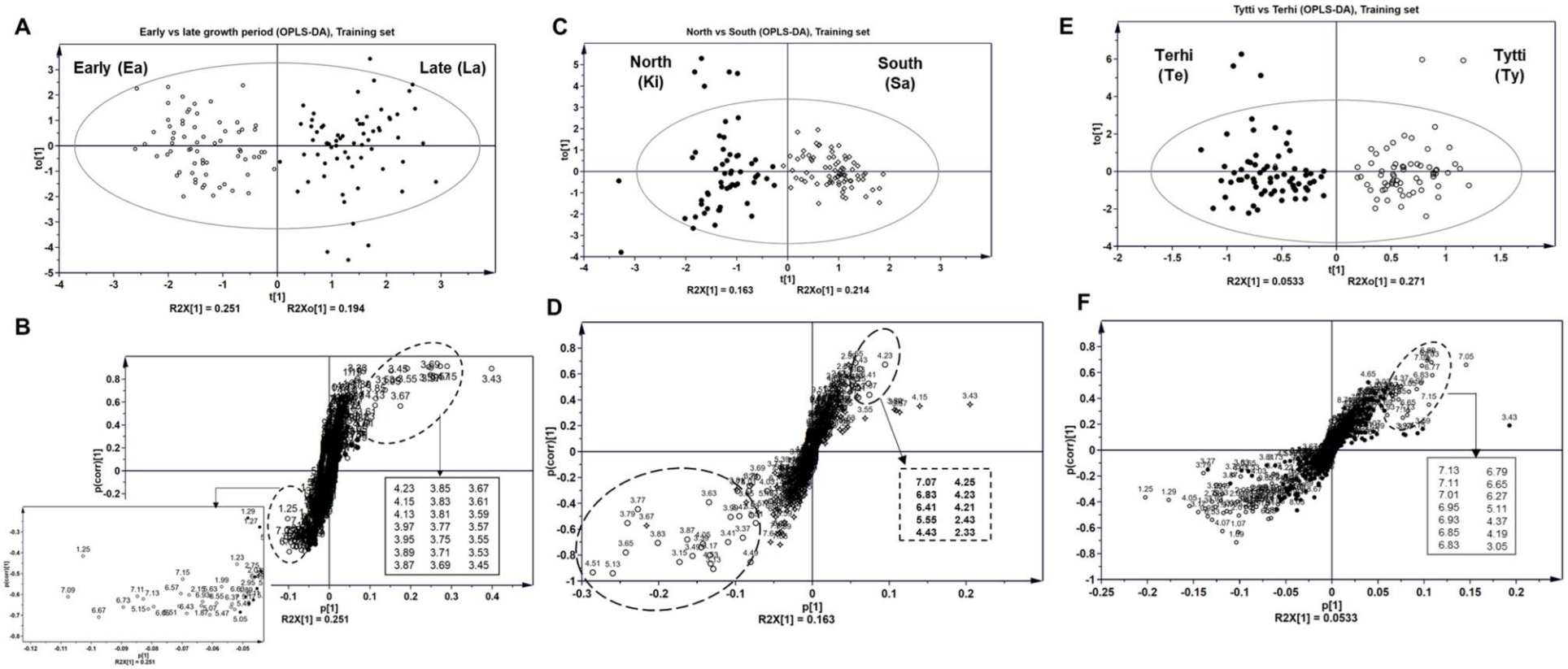


Figure 2. PCA of the binned data generated from the NMR spectra of all the collected samples of SBLs, (192 observations x 457 variables), A: Score scatter plot, B: Loading scatter plot PC1 vs PC2, C: 3D score plot PC1 vs PC2 vs PC3, D: Loading scatter plot of PC1 vs PC3. Symbols of hollow circles and inverted triangles in (A) represented samples from early and later growth stages, respectively. Symbols of stars and dark circles in (C) represented samples from north and south growth locations in Finland, respectively.



559

560 **Figure 3.** OPLS-DA of the SBLs samples. A, C, and E represent the score scatter plots of training set samples, t(1) vs to(1), of the analysis on growth
561 stage, growth location and cultivars, respectively. B, D, and F represent the corresponding S-plots of A, C, and E. The hollow circles in B, D, and F
562 represent variables identified to be significant by combining information from S-plot, VIP plot (VIP ≥ 1), and loading column plot.

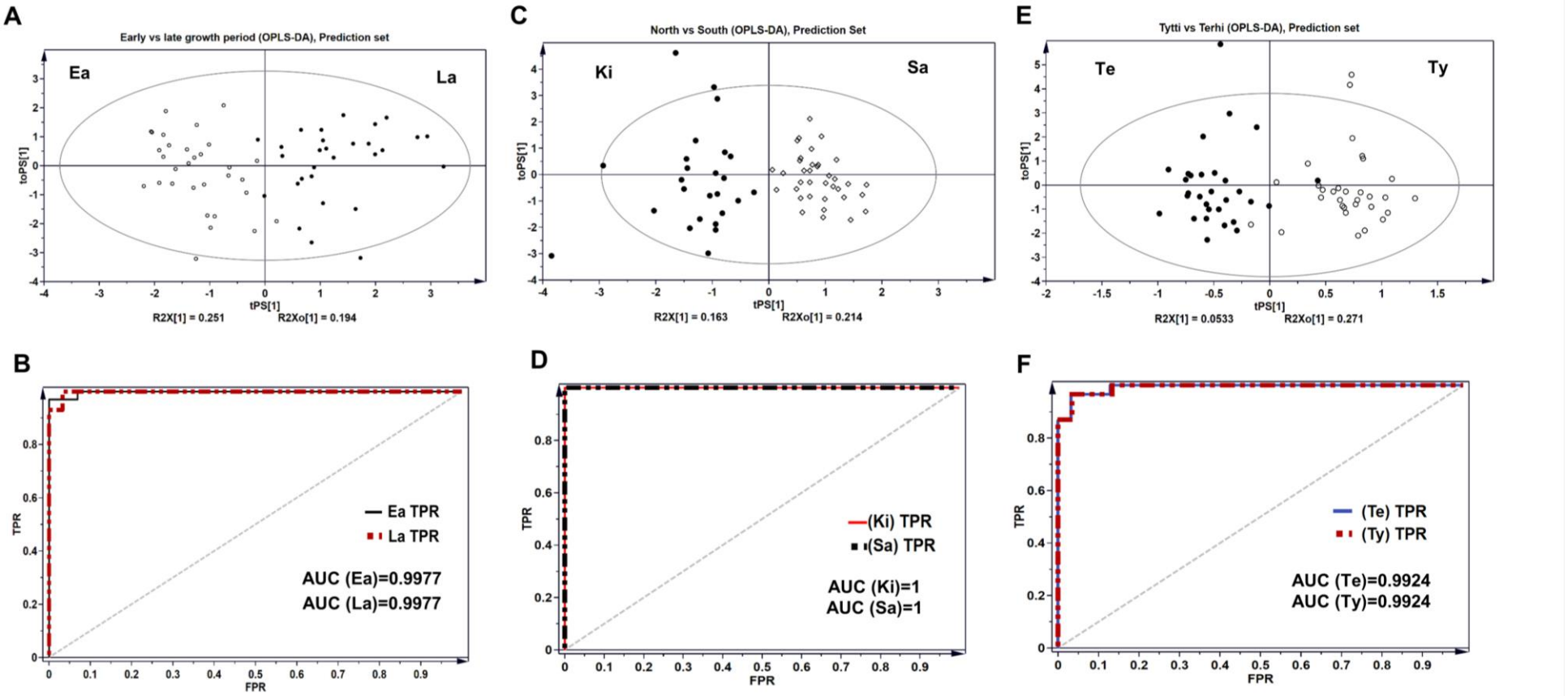


Figure 4. OPLS-DA analysis of the SBLs samples. A, C, and E represent the score scatter plots of prediction set samples, tPS(1) vs toPS(1), of the analysis on growth stage- early (Ea) vs late (La), growth location- north (Ki) vs south (Sa), and cultivars- 'Terhi' (Te) vs 'Tytti (Ty), respectively. B, D, and F represent the corresponding Receiver Operating Characteristic (ROC) plots of A, C, and E.