약학박사학위논문

Multi-platform metabolomics approach to evaluate the authenticity of herbal medicine

2017년 2월

서울대학교 대학원

약학과 약품분석학전공

트롱휴이

Multi-platform metabolomics approach to evaluate the authenticity of herbal medicine

지도교수 박정일

이 논문을 약학박사 학위논문으로 제출함

2017 년 2 월

서울대학교 대학원

약학과 약품분석학전공

트롱휴이

트롱휴이의 박사학위논문을 인준함

2017 년 2 월

위원장 권성원 (인)

부위원장 양현옥 (인)

위원 박정일 (인)

위원 김유선 (인)

위원 홍순선 (인)

Abstract

Herbal medicine has been traditionally and historically used worldwide. However, along with the increase usage of the herbal medicine, its market also has been suffered from the practice of adulteration, either genuinely by lack of experience or intentionally for business gains. Therefore, it is essential to develop an approach that could effectively ascertain the correct identification of the ingredients used in the herbal remedies. Here in this study, we illustrated the application of a multi-platform metabolomics approach in assisting the establishment of a better quality control method for herbal medicine. We first illustrated the capacity of UPLC-QTOFMS based metabolomics to discriminate well 4 common *Panax* species in *Panax* genus, namely *Panax ginseng*, *Panax vietnamensis*, *Panax notoginseng*, *Panax quinquefolius*. Next we further examine whether ^{1}H-NMR based metabolomics approach could evaluate the same *Panax ginseng* samples but from dissimilar locations. The results showed that not only it could well differentiate those samples, but also it could point out the possible mixing proportion in case if those samples were mixed together. Additionally, utilizing those data derived from metabolomics approach, we also discussed the possibility in connecting metabolite variation to that of phylogenetic, as for UPLC-QTOFMS data and in building a model effectively assessing the mixing proportion of intentional admixtures, as for ^{1}H-NMR data. Consequently, we believe that the ease and transferability of our approach as well as its applicability to other products could contribute to establishing a safer market and greater consumer confidence by preventing herbal medicine adulteration.

Keywords: metabolomics; ginseng; authentication; phylogenetic; QTOF-MS; ^{1}H-NMR

Student number: 2013-22583

Contents

List of figures

List of tables

Part I UPLC-QTOFMS based metabolomics followed by stepwise partial least squar e-discriminant analysis (PLS-DA) explore the possible relation between the variations in secondary metabolites and the phylogenetic divergences of the genus *Panax*

1. Introduction

Phylogenetics is the study of evolutionary relations among groups of different species. This field of study involves the classification of organisms according to evolutionary sources or environmental adaptations. The discipline traces geographical or physical occurrences of the species to generate allopatric speciation and investigates the biological scattering of many living organisms [1, 2]. Many botanists, specifically evolutionary biologists, have employed DNA sequencing to measure infra-generic dispersion to construct trees, hence forming the framework to examine and discuss similarities and dissimilarities within species [3]. The branching point of a phylogenetic tree shows both the beginning of a new lineage and the divergence of characteristics such as morphologies, behaviors and chemical properties [4, 5].

Metabolites, terminal outcomes from adaptations in response to biological surroundings, have drawn attention from many evolutionists in that metabolites directly transfer signals from the genomes, transcriptomes and proteomes of an organ in sequence to the tissues or other organs, thus furthering close relations among biological phenotypes and differences in species-specific characteristics. In fact, metabolites can be divided into two categories: primary and secondary metabolites. Primary metabolites, such as carbohydrates, amino acids, fatty acids and organic acids, are directly involved in normal growth, development and reproduction, essentially existing in all living things [6]. On the other hand, secondary metabolites are not directly involved in those processes but are considered to be produced by adaptations to the surrounding environment and have multiple ecological functions such as defense (against herbivores, microbes, viruses or competing plants) and signaling (to attract pollinating or seed-dispersing animals) [7]. Indeed, secondary metabolites provide the ultimate distinguishing traits between similar species. Currently, metabolomics has become one of the most common approaches for the intensive profiling and comparison of secondary metabolites among species and has gradually permeated into botany for the identification of changes to physiological, genotypic, or other factors that impact metabolism [8].

Despite the gradually increasing number of papers related to phylogenetics or metabolomics over the past several years, however, few papers have reported metabolomic studies from a phylogenetic perspective. Until now, researchers have focused on measuring and observing the quantity of metabolites to find phylogenetic markers [9, 10]. Normally, in the common metabolomic approach, the profiles of metabolites are acquired, and then the metabolic profiles between phenotypes are compared with the help of statistical tools such as principal component analysis (PCA) and partial least square-discriminant analysis (PLS-DA) [11]. However, all the approaches utilized to classify samples by species only reveal the expressional differences among the species, not the differences in metabolites between each clade of the phylogenetic tree. For example, Xie et al. and Chan et al. showed metabolomics-based discrimination among three *Panax* species using both LC-MS and NMR, which only revealed the chemical composition and relative abundance of the metabolites in each species. Their results could not explain the metabolomic variations in the *Panax* phylogenetic tree [12-14].

In this study, we attempted a new metabolomic approach using stepwise PLS-DA to determine the connection between phylogenetic relations and differences in secondary metabolites of the phylogenetic tree of the *Panax* genus. The phylogenetic trees were constructed based on the gene sequences of four species, *P. ginseng* (PG), *P. vietnamensis* (PV), *P. notoginseng* (PN) and *P. quinquefolius* (PQ). The differences in secondary metabolites among the four were analyzed by ultra-performance liquid chromatography-quadrupole time of flight mass spectrometry (UPLC-QTOFMS), and the resulting data were then subjected to PCA to demonstrate the common metabolomic approach. The PCA results showed that all the secondary metabolites were clustered according to species, which simply mirrored the results expected from conventional metabolomics-based studies. Then, we applied stepwise PLS-DA according to the branching point of the phylogenetic tree of these species to obtain the differences in the metabolites between clades of the tree. The results revealed that some particular secondary metabolites of these plants, commonly known as ginsenosides [15-17], can be mapped onto the plants' phylogenetic trees to better explain the divisions in branching points.

2. Materials and methods

2.1. Chemicals

HPLC grade acetonitrile, methanol and water were purchased from J.T. Baker (Phillsburg, NJ, USA). Formic acid, 2-propanol (LC/MS grade) and lithium hydroxide were purchased from Sigma-Aldrich (St. Louis, MO, USA) and used to make a lithium formate solution serving as a mass calibrant (50% 2-propanol with 1% lithium hydroxide and 0.1% formic acid). Using silica gel column chromatography and semi-preparative liquid chromatography, nine standard compounds for assigning peaks (ginsenoside Rb1, Rb2, Rc, Rd, Re, Rg1, majonoside R1, R2, and vinaginsenoside R2) were isolated from fresh PG as for Rb1, Rb2, Rc, Rd, Re, Rg1 and from fresh PV as for majonoside R1, R2 and vinaginsenoside R2. After isolation, the purity of all ginsenoside standards was determined to be over 95% by HPLC-UV-ELSD [18].

2.2. Plant material

Dried rhizomes of PN, PQ and PG were purchased in a Korean herbal market, Korea. Unlike other ginsengs, PV was directly collected at the ginseng farms in Quangnam Province, Socialist Republic of Vietnam due to its rarity on the market. After collecting, PN, PQ and PG were identified by Prof. Won Keun Oh from the Department of Pharmacognosy, Seoul National University, Korea while PV was authenticated by Prof. Minh Duc Nguyen from the Department of Pharmacognosy, University of Medicine and Pharmacy, Ho Chi Minh city, Vietnam. Voucher specimens were deposited at the Deparment of Biomedical and Pharmaceutical Analysis, Seoul National University, Korea.

2.3. Sample preparation

Two rhizomes for each sample, with a total of twenty rhizomes for ten samples of each species, were splintered into small fragments to be freeze-dried, completely eliminating moisture. After 24 hours of freeze-drying, they were pulverized using a grinder (DA700, Daesung Artlon, Seoul, Korea) and sieved to a particle size of 90–125 µm. Twenty mg of powdered rhizome was extracted with 2 mL of pure methanol and centrifuged at 13,000 g for 5 min, and the supernatant was filtered with a 0.5 µm PTFE filter (Otawa, Tokyo, Japan). A total of 5 QC samples were prepared by mixing together 100 µL aliquots of all ten samples to confirm the reproducibility of data acquisition. All standard solutions were prepared in pure methanol.

2.4. Data acquisition parameters

A UPLC (Waters, Milford, MA, USA) with micrOTOF-QII (Bruker Daltonik GmbH, Bremen, Germany) was utilized. A 5 µL aliquot of each sample extract was injected into an ACQUITY BEH C_{18} column (2.1 × 100 mm, 1.7 µm, Waters, Milford, MA, USA). A column temperature of 40 °C was employed for peak separation, using a mobile phase consisting of A: water with 0.1% formic acid; and B: acetonitrile with 0.1% formic acid. The flow rate was 0.3 mL/min; the gradient started at 82% A, changed to 75% A for 5 min, changed to 68% A for 20 min and held for 4 min, then changed to 62% A for 11 min, to 55% A for 5 min, and eventually to 0% A for 20 min, with a pre-run rinse of 82% A for 10 min. The source parameters were as follows: capillary voltage 3.5 kV, nebulizer pressure 1.2 bar, dry gas flow rate 8 L/min and dry gas temperature 200 °C. The ion transfer and collision stages were set as follows: funnel 1 RF 400 Vpp, funnel 2 RF 400 Vpp, hexapole RF 400 Vpp, quadrupole ion energy 15 eV, collision energy 10 eV, collision RF 400 Vpp, transfer time 100 µs and pre-pulse storage 5 µs. High purity nitrogen was used as a nebulizer gas, dry gas, and collision gas.

2.5. Multivariate statistical analysis

Quantitative data were extracted from the software DataAnalysis 4.0 (Bruker Daltonik GmbH, Bremen, Germany). Then, the table of all the derived data was converted to a suitable format and subsequently processed by MetaboAnalyst 2.0 (http://www.metaboanalyst.ca) and SIMCA-P$^+$ 12 (Umetrics, Umeå, Sweden) for statistical analysis, including PCA and PLS-DA. The PCA plot verified the reliability of the analysis with the clustered and centered QC data, and also displayed the suitability of the grouping based on species [19]. Then, a novel approach to find phylogenetic markers was applied: PLS-DA was performed with 3 steps classification according to phylogenetic divergence (first: PN vs. PV, PQ, and PG; second: PV vs. PQ and PG; third: PQ vs. PG). The markers were determined by VIP values of each analysis.

3. Results and discussion

3.1. Ginsenoside identification in four species

Due to the limited reproducibility of LC-MS and the extensive presence of secondary metabolites in plants, a general LC-MS database for plant secondary metabolites has not yet been established. In this context, we have constructed an in-house database of the ginsenosides referenced in studies profiling *Panax* species, including retention times obtained under similar conditions, nominal masses, *m/z* values of fragment ions and compounds' existence in each species. Then, using the library and confirmed standards, we assigned the peaks in the same manner as in our previous study of *Schisandra chinensis* [20] (Figure 1) . Consequently, 17, 13, 14, and 12 ginsenosides of PV (Table 1), PN (Table 2), PQ (Table 3) and PG (Table 4), respectively, were identified. 9 ginsenosides among them were identified using standards, while the rest were identified using information from the literature [21-26]. In negative mode, ESI produced two dominant precursor ions, [M–H]$^-$ and [M+HCOO]$^-$, which were readily identified (Figure 2).

Figure 1. Chromatograms of 4 species and QC.

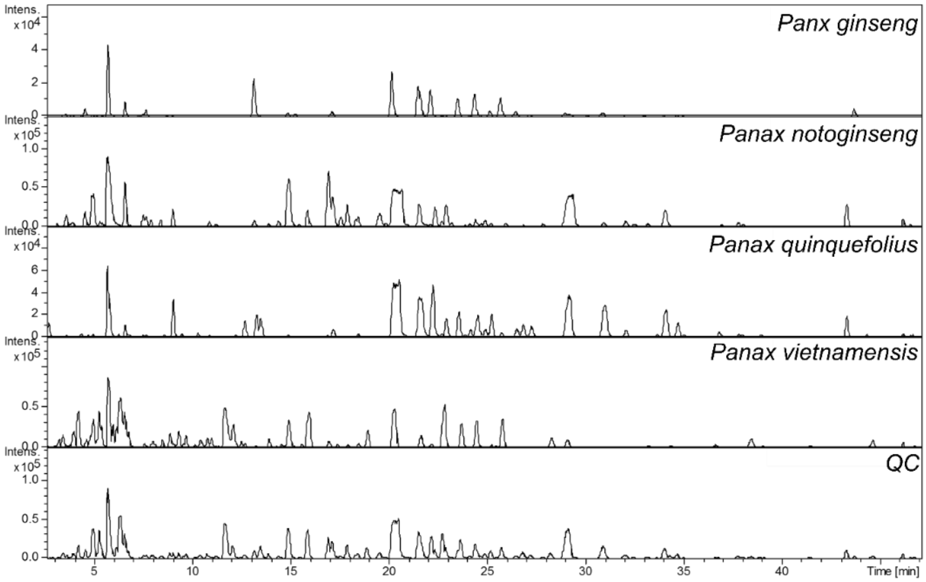


Table 1. Assigned ginsenosides of *Panax vietnamensis*

Retention time (min)	*m/z*			Formula	Error (mDa)	Identification
	Adduct ion	Exact	Measured			
2.3	$[M+HCOO]^-$	717.4425	717.4424	$C_{36}H_{64}O_{11}$	-0.1	Vinaginsenoside R12
3.4	$[M-H]^-$	801.4630	801.4641	$C_{41}H_{70}O_{15}$	1.1	Vinaginsenoside R14
4.9	$[M-H]^-$	931.5260	931.5244	$C_{47}H_{80}O_{18}$	-1.6	Notoginsenoside R1
5.2	$[M-H]^-$	815.4787	815.4768	$C_{42}H_{72}O_{15}$	-1.9	Majonoside R1
5.7	$[M+HCOO]^-$	845.4898	845.4882	$C_{42}H_{72}O_{14}$	-1.6	Ginsenoside Rg1
5.7	$[M+HCOO]^-$	991.5477	991.5447	$C_{48}H_{82}O_{18}$	-3	Ginsenoside Re
6.3	$[M-H]^-$	785.4681	785.4636	$C_{41}H_{70}O_{14}$	-4.5	Majonoside R2
9.6	$[M+HCOO]^-$	1033.5583	1033.5559	$C_{50}H_{84}O_{19}$	-2.4	Pseudoginsenoside Rs1
11.6	$[M-H]^-$	827.4787	827.4753	$C_{43}H_{72}O_{15}$	-3.4	Vinaginsenoside R2
11.8	$[M-H]^-$	841.4943	841.4912	$C_{44}H_{74}O_{15}$	-3.1	Vinaginsenoside R1
18.4	$[M+HCOO]^-$	638.4370	638.4369	$C_{36}H_{62}O_9$	-0.1	Ginsenoside Rh1
20.2	$[M-H]^-$	1107.5945	1107.5956	$C_{54}H_{92}O_{23}$	1.1	Ginsenoside Rb1
24.4	$[M-H]^-$	1077.5840	1077.5803	$C_{53}H_{90}O_{22}$	-3.7	Ginsenoside Rc
25.1	$[M-H]^-$	1077.5840	1077.5849	$C_{53}H_{90}O_{22}$	0.9	Ginsenoside Rb2
27.3	$[M-H]^-$	1149.6051	1149.6050	$C_{56}H_{94}O_{24}$	-0.1	Quinquenoside R1
29.2	$[M-H]^-$	945.5417	945.5388	$C_{48}H_{82}O_{18}$	-2.9	Ginsenoside Rd
34.1	$[M+HCOO]^-$	991.5477	991.5472	$C_{48}H_{82}O_{18}$	-0.5	Gypenoside XVII

Table 2. Assigned ginsenosides of *Panax notoginseng*

Retention time (min)	*m/z*			Formula	Error (mDa)	Identification
	Adduct ion	Exact	Measured			
4.5	$[M-H]^-$	961.5366	961.5373	$C_{48}H_{82}O_{19}$	0.7	20-O-Glucoginsenoside Rf
4.9	$[M-H]^-$	931.5260	931.5291	$C_{47}H_{80}O_{18}$	3.1	Notoginsenoside R1
5.7	$[M+HCOO]^-$	845.4898	845.4912	$C_{42}H_{72}O_{14}$	1.4	Ginsenoside Rg1
5.7	$[M+HCOO]^-$	991.5477	991.5496	$C_{48}H_{82}O_{18}$	1.9	Ginsenoside Re

14.9	$[M-H]^-$	769.4732	769.4758	$C_{41}H_{70}O_{13}$	2.6	Notoginsenoside R2
17.1	$[M-H]^-$	783.4889	783.4890	$C_{42}H_{72}O_{13}$	0.1	Ginsenoside Rg2
20.2	$[M-H]^-$	1107.5945	1107.5946	$C_{54}H_{92}O_{23}$	0.1	Ginsenoside Rb1
23.1	$[M+HCOO]^-$	683.4370	683.4383	$C_{36}H_{62}O_9$	1.3	Ginsenoside F1
24.4	$[M-H]^-$	1077.5840	1077.5844	$C_{53}H_{90}O_{22}$	0.4	Ginsenoside Rb2
25.1	$[M-H]^-$	1077.5840	1077.5849	$C_{53}H_{90}O_{22}$	0.9	Ginsenoside Rb3
29.0	$[M-H]^-$	945.5417	945.5414	$C_{48}H_{82}O_{18}$	-0.3	Ginsenoside Rd
34.0	$[M+HCOO]^-$	991.5477	991.5484	$C_{48}H_{82}O_{18}$	0.7	Gypenoside XVII
43.2	$[M+HCOO]^-$	829.4949	829.4995	$C_{42}H_{72}O_{13}$	4.6	Ginsenoside F2

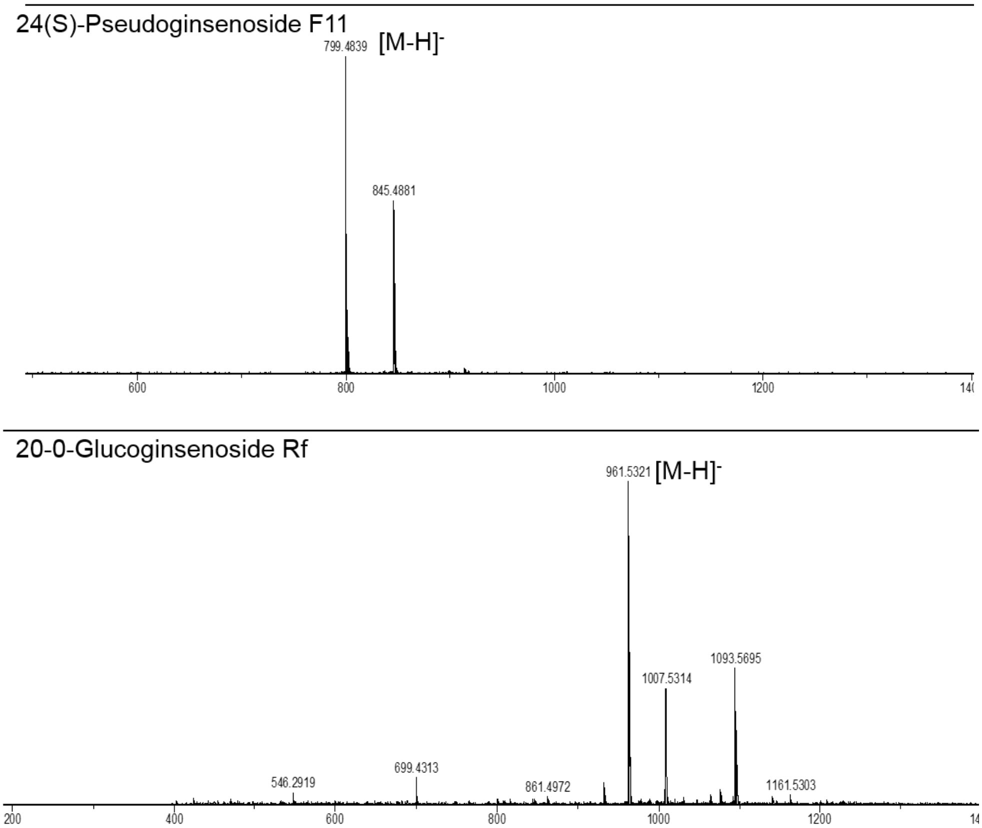
Table 3. Assigned ginsenosides of *Panax quinquefolius*

Retention time (min)	***m/z***			**Formula**	**Error (mDa)**	**Identification**
	Adduct ion	**Exact**	**Measured**			
4.9	$[M-H]^-$	931.5260	931.5261	$C_{47}H_{80}O_{18}$	0.1	Notoginsenoside R1
5.7	$[M+HCOO]^-$	845.4898	845.4858	$C_{42}H_{72}O_{14}$	-4	Ginsenoside Rg1
5.7	$[M+HCOO]^-$	991.5477	991.5435	$C_{48}H_{82}O_{18}$	-4.2	Ginsenoside Re
13.4	$[M-H]^-$	799.4838	799.4839	$C_{42}H_{72}O_{14}$	0.1	24(S)-Pseudoginsenoside F11
14.9	$[M-H]^-$	769.4732	769.4771	$C_{41}H_{70}O_{13}$	3.9	Notoginsenoside R2
17.1	$[M-H]^-$	783.4889	783.4847	$C_{42}H_{72}O_{13}$	-4.2	Ginsenoside Rg2
20.2	$[M-H]^-$	1107.5945	1107.5976	$C_{54}H_{92}O_{23}$	3.1	Ginsenoside Rb1
22.2	$[M-H]^-$	1077.5840	1077.5841	$C_{53}H_{90}O_{22}$	0.1	Ginsenoside Rc
24.3	$[M-H]^-$	1077.5840	1077.5844	$C_{53}H_{90}O_{22}$	0.4	Ginsenoside Rb2
25.1	$[M-H]^-$	1077.5840	1077.5849	$C_{53}H_{90}O_{22}$	0.9	Ginsenoside Rb3
27.2	$[M-H]^-$	1149.6051	1149.6049	$C_{56}H_{94}O_{24}$	-0.2	Quinquenoside R1
29.0	$[M-H]^-$	945.5417	945.5436	$C_{48}H_{82}O_{18}$	1.9	Ginsenoside Rd
34.0	$[M+HCOO]^-$	991.5477	991.5471	$C_{48}H_{82}O_{18}$	-0.6	Gypenoside XVII
43.2	$[M+HCOO]^-$	829.4949	829.4910	$C_{42}H_{72}O_{13}$	-3.9	Ginsenoside F2

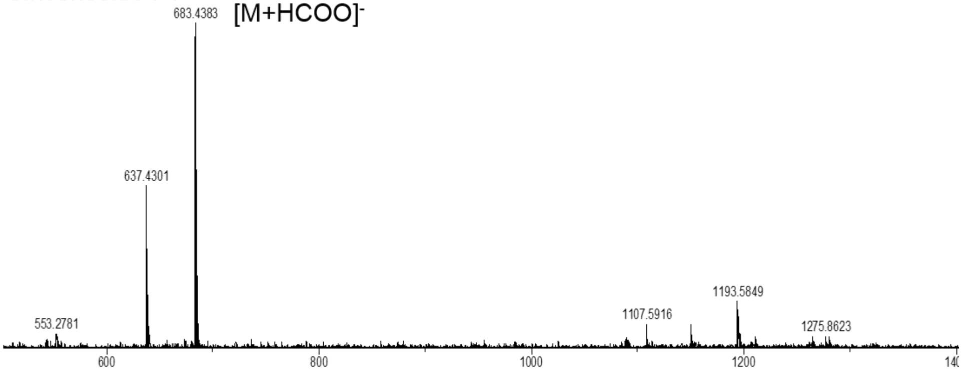
Table 4. Assigned ginsenosides of *Panax ginseng*

Retention time (min)	***m/z***			**Formula**	**Error (mDa)**	**Identification**
	Adduct ion	**Exact**	**Measured**			
4.5	$[M-H]^-$	961.5366	961.5321	$C_{48}H_{82}O_{19}$	-4.5	20-O-Glucoginsenoside Rf
5.7	$[M+HCOO]^-$	845.4898	845.4945	$C_{42}H_{72}O_{14}$	4.7	Ginsenoside Rg1
5.7	$[M-H]^-$	945.5417	945.5433	$C_{48}H_{82}O_{18}$	1.6	Ginsenoside Re
13.1	$[M-H]^-$	799.4838	799.4841	$C_{42}H_{72}O_{14}$	0.3	Ginsenoside Rf
14.8	$[M-H]^-$	769.4732	769.4701	$C_{41}H_{70}O_{13}$	-3.1	Notoginsenoside R2
17.1	$[M-H]^-$	783.4889	783.4877	$C_{42}H_{72}O_{13}$	-1.2	Ginsenoside Rg2
20.1	$[M-H]^-$	1107.5945	1107.5917	$C_{54}H_{92}O_{23}$	-2.8	Ginsenoside Rb1
22.1	$[M-H]^-$	1077.5840	1077.5868	$C_{53}H_{90}O_{22}$	2.8	Ginsenoside Rc
24.3	$[M-H]^-$	1077.5840	1077.5839	$C_{53}H_{90}O_{22}$	-0.1	Ginsenoside Rb2
25.1	$[M-H]^-$	1077.5840	1077.5849	$C_{53}H_{90}O_{22}$	0.9	Ginsenoside Rb3
27.1	$[M-H]^-$	1149.6051	1149.6028	$C_{56}H_{94}O_{24}$	-2.3	Quinquenoside R1
28.9	$[M-H]^-$	945.5417	945.5454	$C_{48}H_{82}O_{18}$	3.7	Ginsenoside Rd

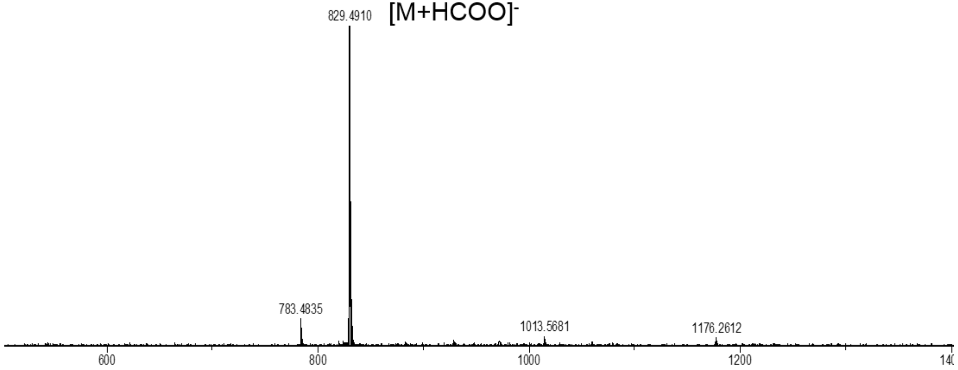
Figure 2. Mass spectra of 26 ginsenosides. Precursor ions for assignment were checked with ionization patterns.



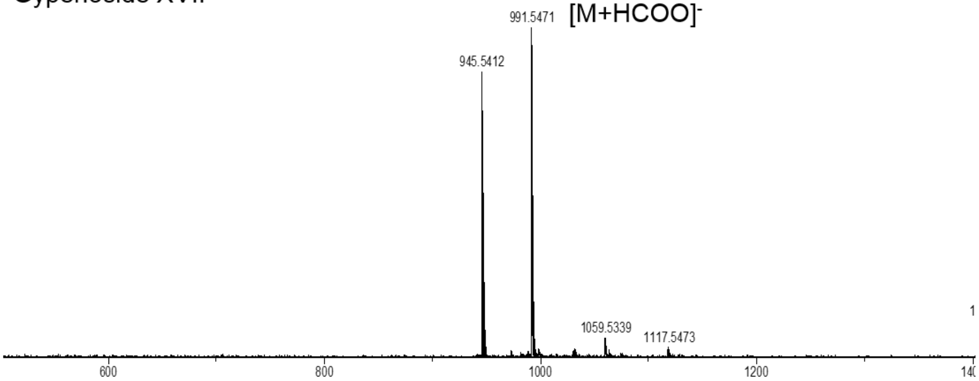
Ginsenoside F1



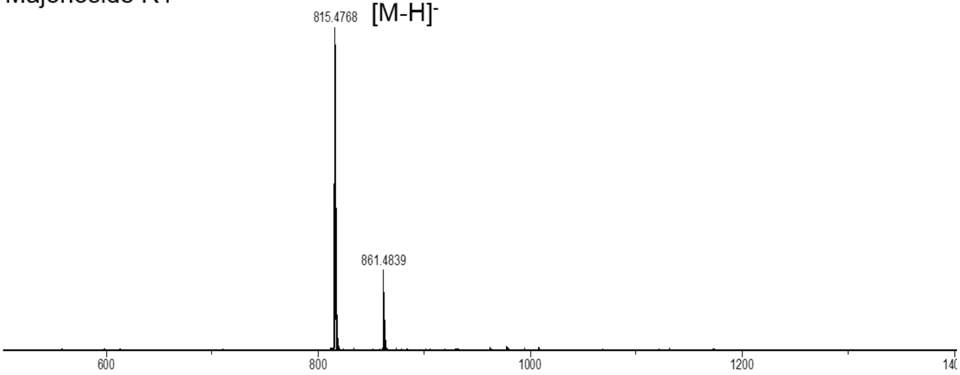
Ginsenoside F2



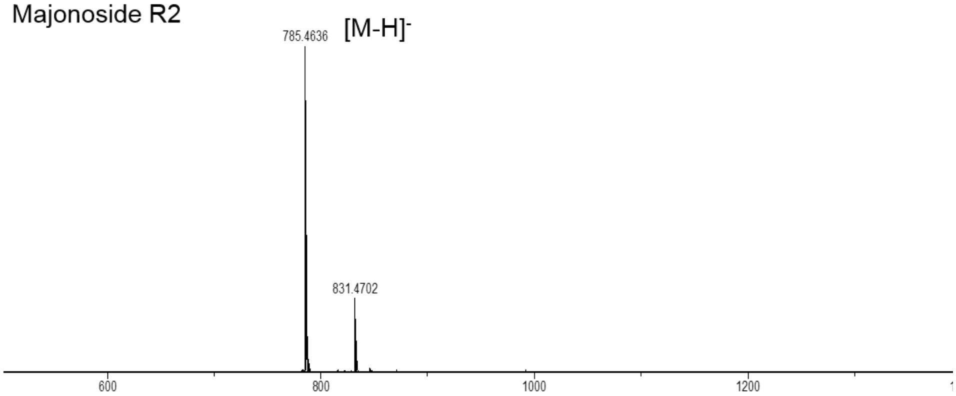
Gypenoside XVII



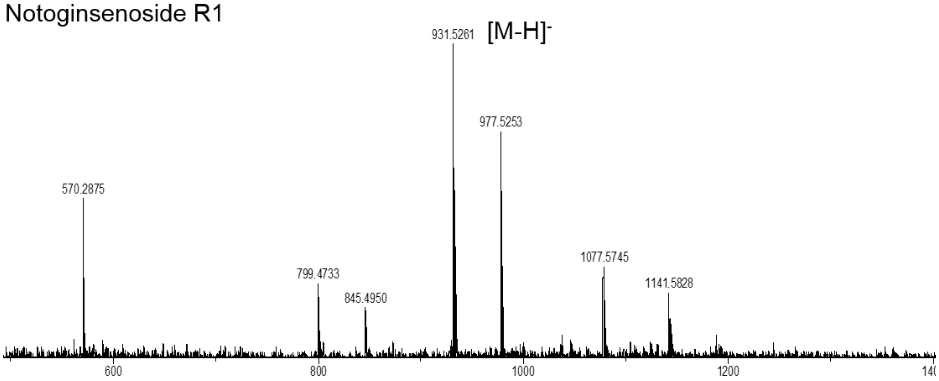
Majonoside R1

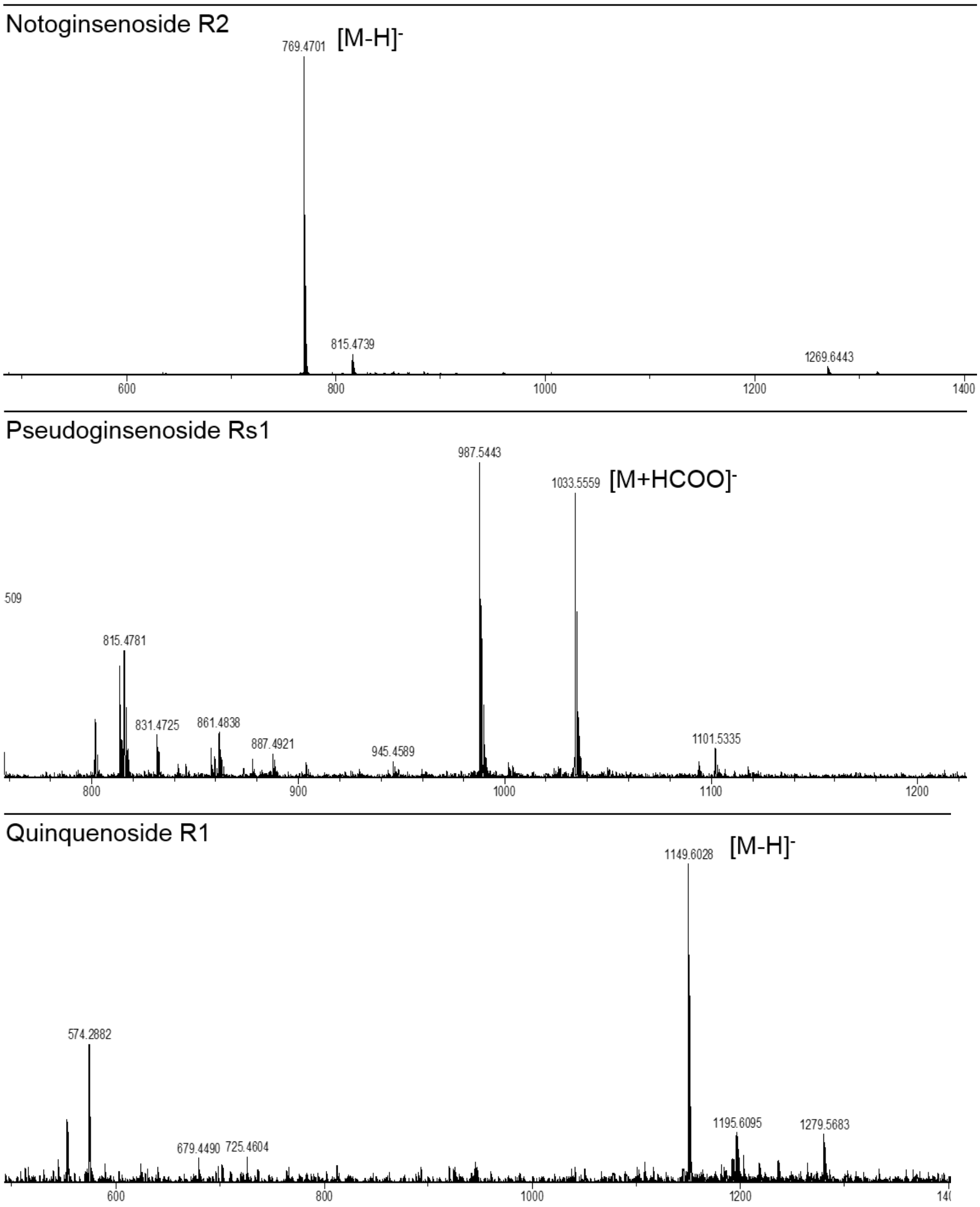


Majonoside R2

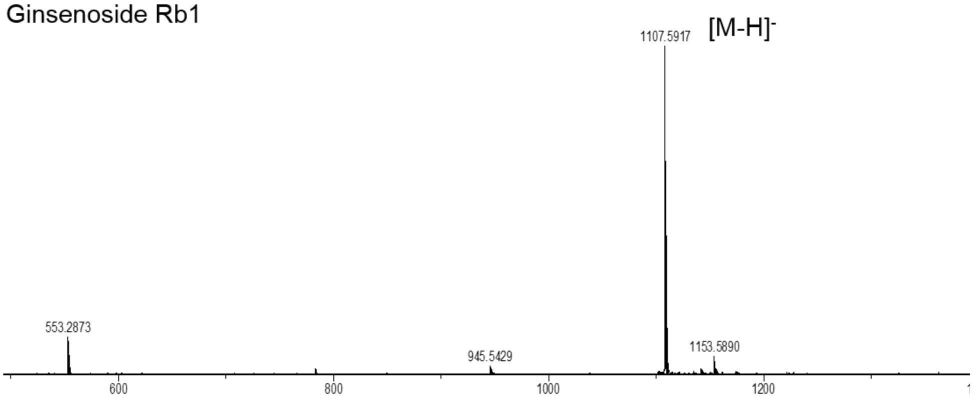


Notoginsenoside R1

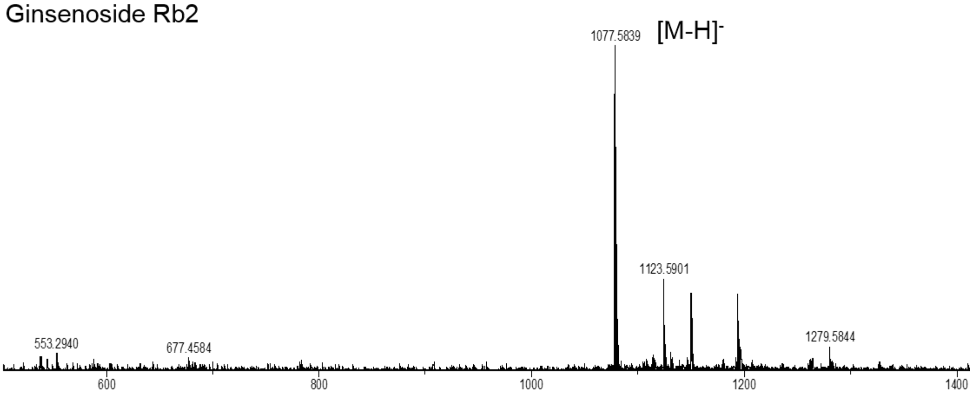



Notoginsenoside R2
769.4701
[M-H]-
815.4739
1269.6443
600
800
1000
1200
1400
Pseudoginsenoside Rs1
987.5443
1033.5559
[M+HCOO]-
509
815.4781
831.4725
861.4838
887.4921
945.4589
1101.5335
800
900
1000
1100
1200
Quinquenoside R1
1149.6028
[M-H]-
574.2882
679.4490
725.4604
1195.6095
1279.5683
600
800
1000
1200
14(

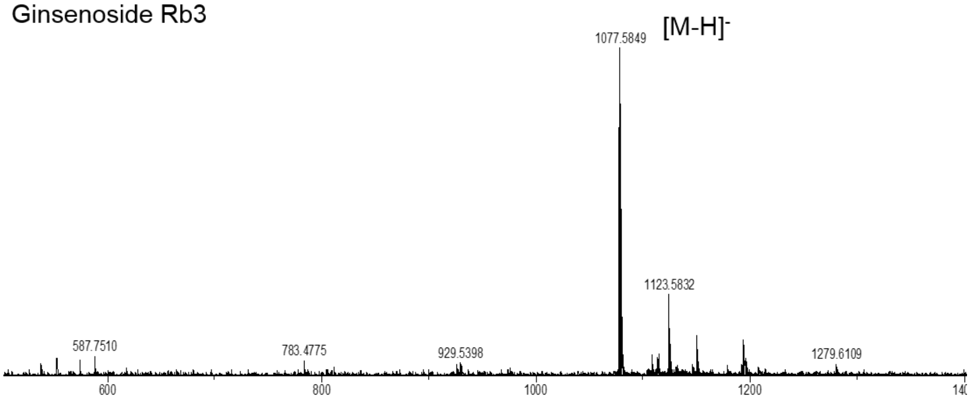
Ginsenoside Rb1



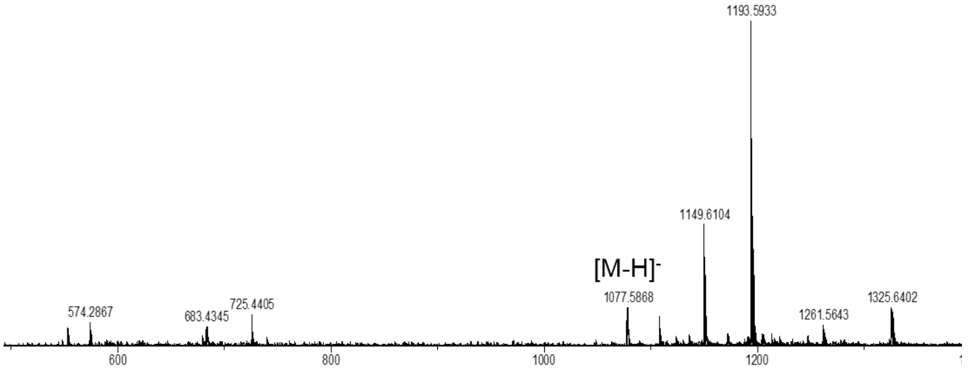
Ginsenoside Rb2



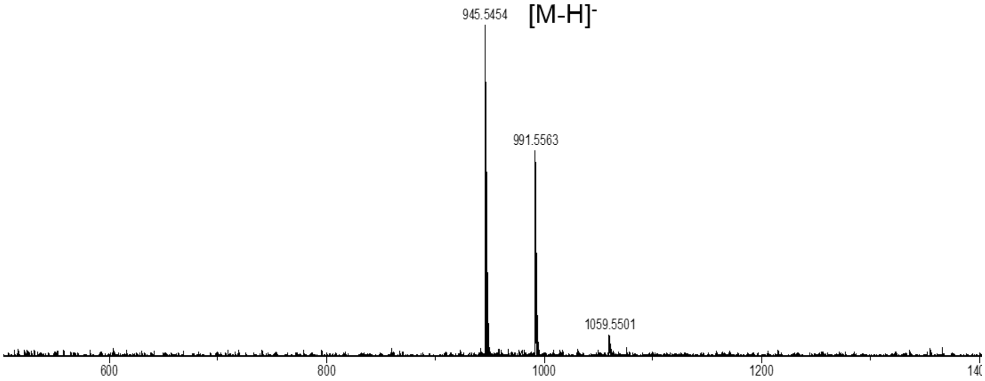
Ginsenoside Rb3



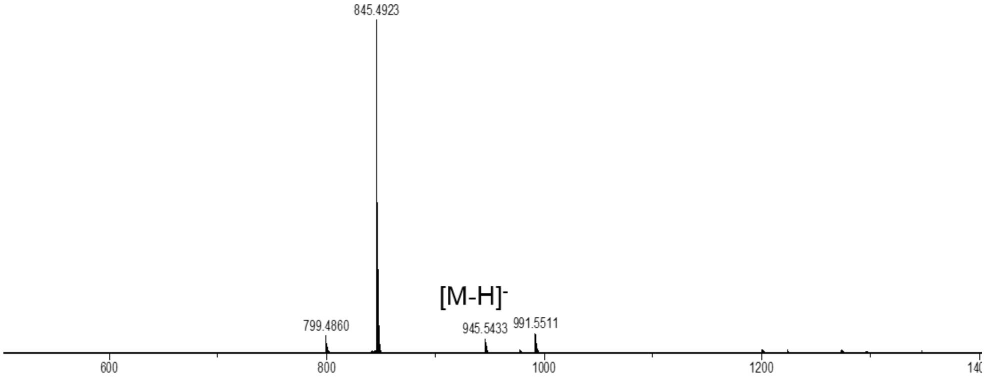
Ginsenoside Rc



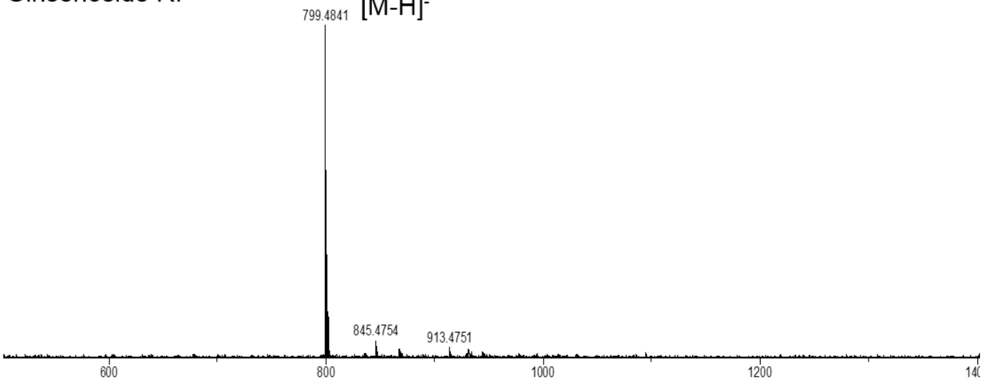
Ginsenoside Rd



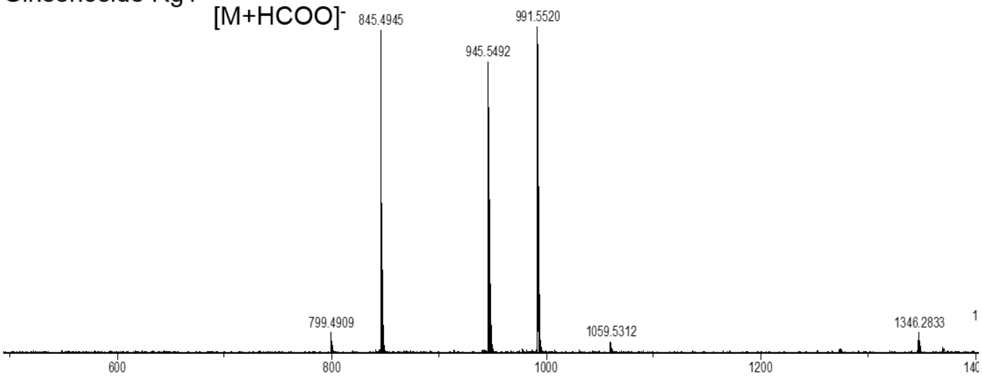
Ginsenoside Re



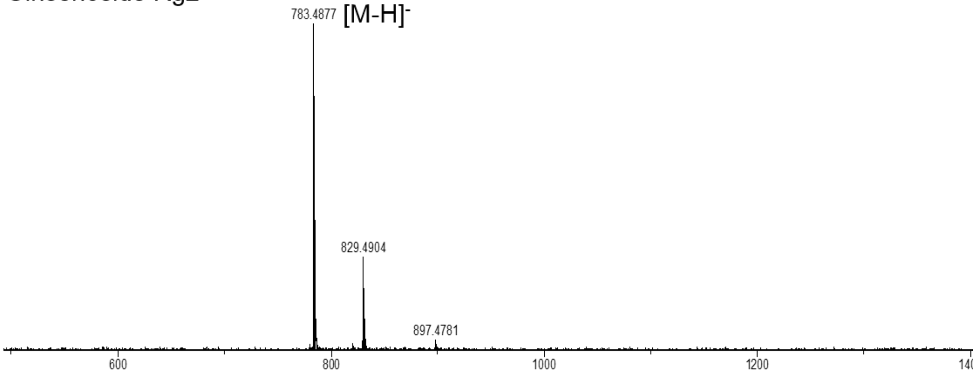
Ginsenoside Rf



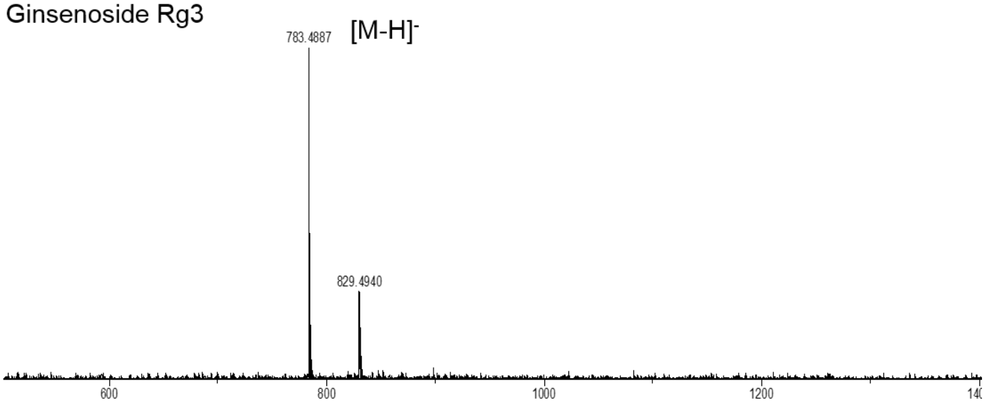
Ginsenoside Rg1



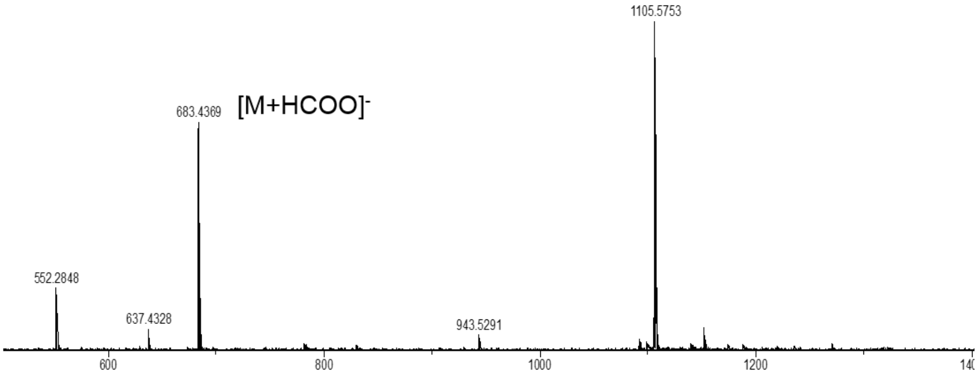
Ginsenoside Rg2



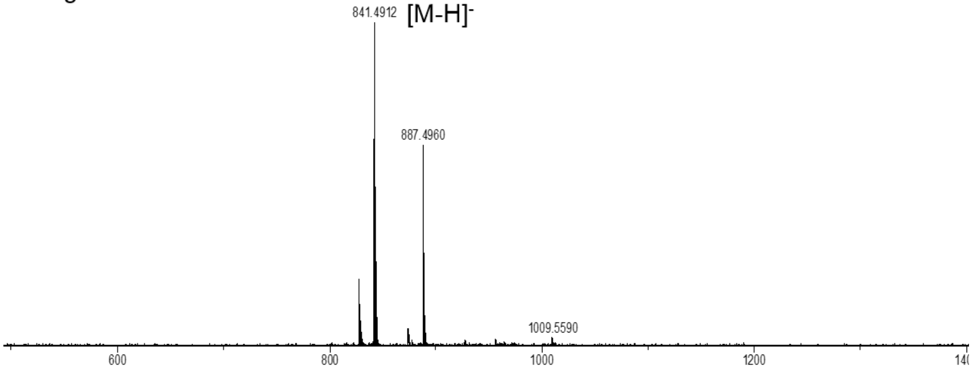
Ginsenoside Rg3



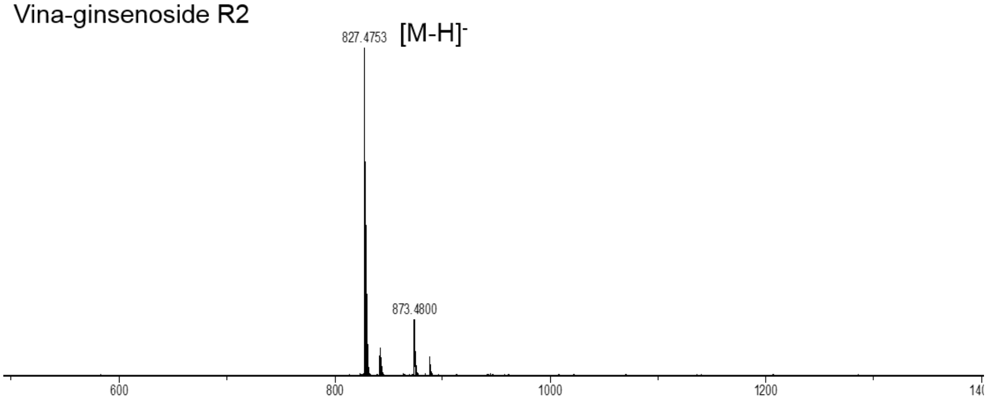
Ginsenoside Rh1



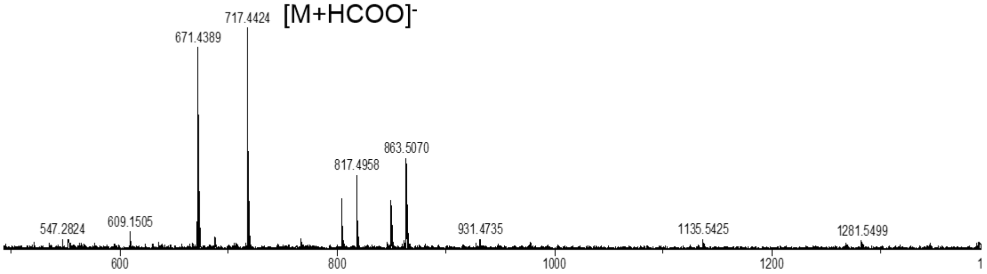
Vina-ginsenoside R1



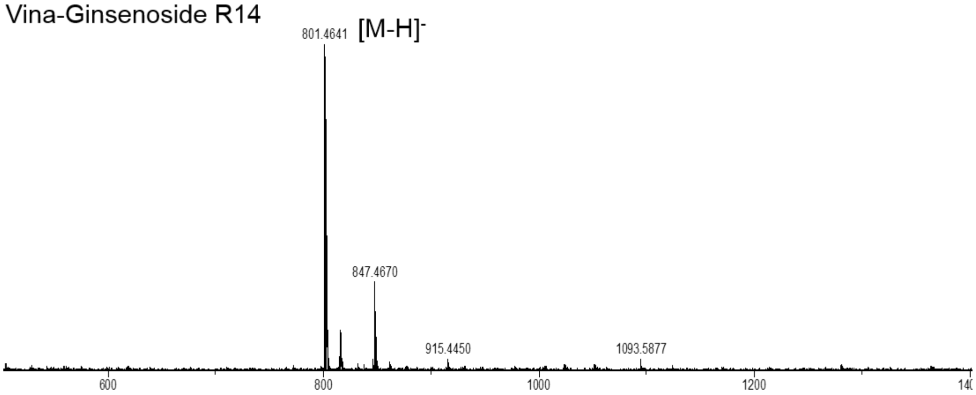
Vina-ginsenoside R2



Vina-Ginsenoside R12



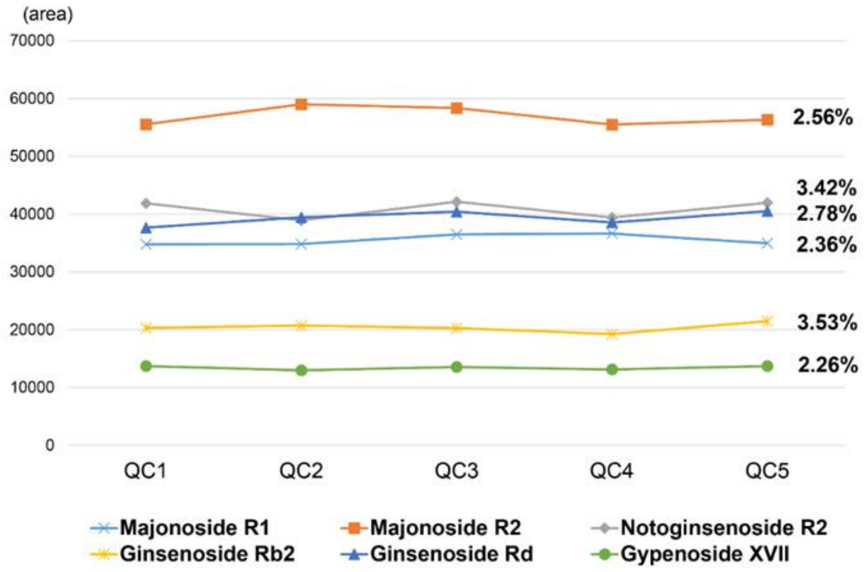
Vina-Ginsenoside R14



3.2. Method validation

Our analysis process had reproducible retention times, and *m/z* values were calibrated with lithium formate before every sample to minimize deviations in mass accuracy during data acquisition. To calculate the reliability of our results, mass precision was evaluated for six ginsenosides in the QC (quality control) data in conjunction with the retention time and *m/z* alterations. The variations in the *m/z* values and retention times of the six ginsenosides showed that significant errors during analysis were unlikely. Then, the areas of the peaks with verified mass and retention time precision were utilized to confirm the sensitivity of the mass detector (Figure 3). All the peaks showed a coefficient of variation lower than 15%, indicating that unintended factors that might have interfered with the collection of reliable data were negligible in our experiments.

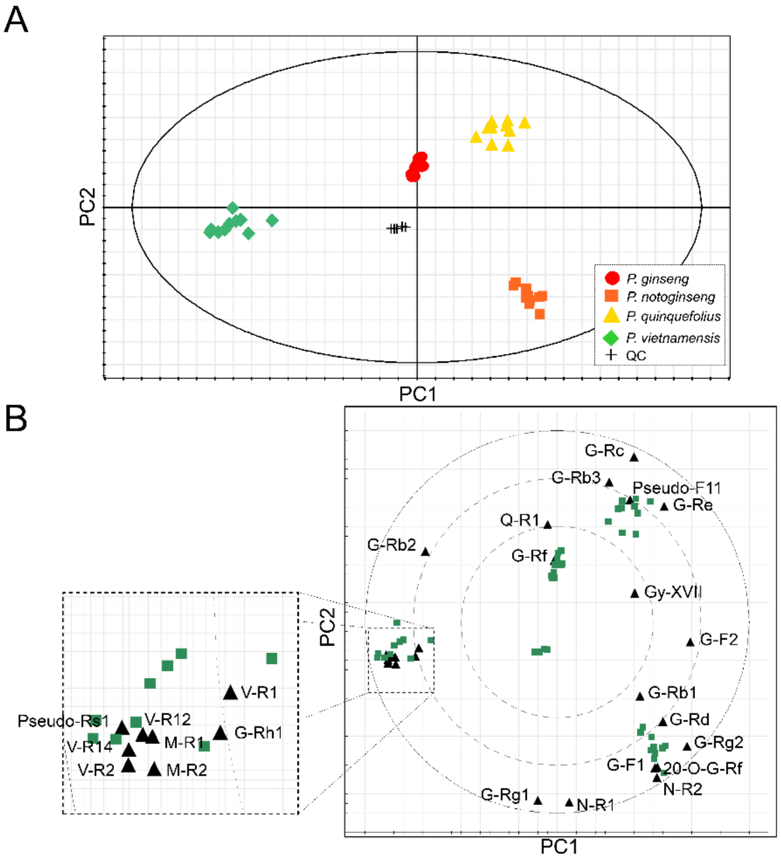
Figure 3. Validation of reproducibility within the experiment. Six peaks dispersed on the retention time for representative of the other ginsenosides. CV values under 15% exhibit the acquired data were credible.



3.3. Discriminating the four species with common metabolomic approaches

Principal component analysis, an unsupervised analysis, was performed on the four species. The score plots showed good clustering, and PC1 and PC2 were determined as good criteria for discriminating the four species (Fig. 4A). Additionally, the centering of QC samples in the plot showed the reproducibility of the analytical methods and statistical analysis [19]. The loading plots related metabolite abundance with each species (Fig. 4B), and bar graphs of all the ginsenosides are included in Fig. S3. Majonosides, vinaginsenosides, and some ginsenosides such as G-Rh1 and pseudo-Rs1 were biased toward PV. Notoginsenosides were present in PN at high concentrations, and PN contained the largest number of highly expressed metabolites (13 metabolites) among the four species. However, the results showed the limitations of the conventional approach, the mere determination of which metabolites are highly expressed in a specific species. Furthermore, it is widely known among ginseng experts that ginsenosides occur with varying abundances. Thus, quantitative differences in ginsenosides and the interpretation of such differences by common metabolomic approaches provide no novel information from a phylogenetic perspective. Therefore, we introduced the stepwise PLS-DA approach as an attempt to better understand the phylogenetic relationship among *Panax* species.

Figure 4. Score plot (A) and bi-plot (B) of PCA. (B): green dots: metabolites detected in the experiment; black dots: metabolite markers



3.4. Construction of phylogenetic trees

The sequences of 18S rDNA and trnK of each species were downloaded from GenBank as files in FASTA format [27]. The detailed accession numbers of the FASTA files and species can be found in Table 5. Because the combination of trnK and 18S rRNA sequences is considered to provide better accuracy than the data from one sequence [27, 28], we combined the downloaded 18S rDNA and trnK data using the ShortRead package in R software. However, the combined data varied considerably in length, making it difficult to compare different sets of data. Hence, we cut the terminal sequence of the combined data to a consistent length, 4345 base pairs. Finally, all the data were exported as FASTA format files using the seqinR package in R software and used to reconstruct the phylogenetic trees [29].

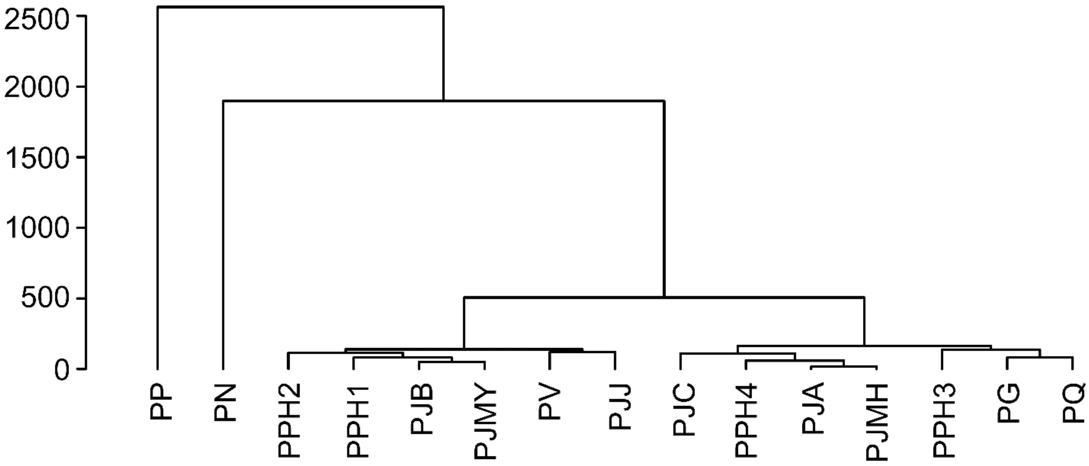
The phylogenetic trees were reconstructed using a simplistic and widely used hierarchical clustering algorithm. In fact, it is one of the methods using for inferring the phylogenetic tree even though it is not commonly used to construct phylogenetic trees in terms of ginseng research. The core principle of the algorithm is based on highlighting the distance of surrounding or remote objects to one object; the use of the method parallels the basic idea of phylogenetics, which reflects the evolutionary relationships among species [30-32]. Consequently, two phylogenetic trees were reconstructed (Figure 5). One tree showed the relations of 15 species, while the other focused on the relations of the four species used in this study. Both phylogenetic trees showed similarities in the phylogenetic relations of the four species, thus confirming the consistency of our method of constructing phylogenetic trees.

Table 5. Sequences downloaded from GenBank and their accession numbers

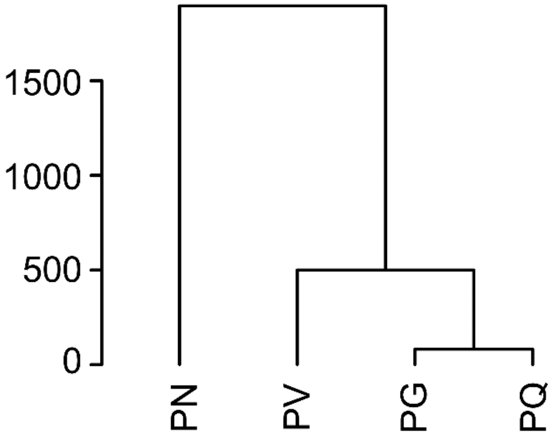
Species	**Abbreviation**	**GenBank Accession**	
		trnK	18S rDNA
Panax ginseng C. A. Meyer	PG	AB087999	D83275
Panax quinquefolius L.	PQ	AB088001	D85172
Panax japonicus C. A. Meyer (Japan)	PJJ	AB088000	D84100
Panax japonicus C.A. Meyer (China)	PJC	AB088006	AB088018
Panax japonicus C.A. Meyer var. *major* C.Y. Wu et Feng	PJMH	AB088008	AB088020
Panax japonicus C.A. Meyer var. *angustifolius* (Burk.) Cheng et Chu	PJA	AB088007	AB088019
Panax japonicus C.A. Meyer var. *bipinnatifidus* (Seem.) C.Y. Wu et Feng	PJB	AB088010	AB088021
Panax notoginseng (Burk.) F.H. Chen.	PN	AB088002	D85171
Panax vietnamensis Ha et Grushv.	PV	AB088003	AB033635
Panax pseudoginseng Wall.	PP	AB088016	AB088026
Panax pseudoginseng Wall. subsp. *himalaicus* Hara Mayodia, Arunachal Pradesh State, India	PPH1	AB088011	AB044902
Panax pseudoginseng Wall. subsp. *himalaicus* Hara KMPG (originated from Chame, Nepal)	PPH2	AB088012	AB088022
Panax pseudoginseng Wall. subsp. *himalaicus* Hara KMPG (originated from Langtang, Nepal)	PPH3	AB088013	AB088023
Panax pseudoginseng Wall. subsp. *himalaicus* Hara KMPG(originated from Gokyo, Nepal)	PPH4	AB088014	AB088024

Figure 5. Phylogenetic tree of 15 *Panax* plants (A) and 4 *Panax* species (B)

A



B



3.5. Stepwise PLS-DA for revealing markers of phylogenetic divergences

As described above, the conventional metabolomic approach with PCA or PLS-DA derives the expressional differences among species, not between clades of the phylogenetic tree. Thus, we suggested that, after stepwise classification according to divergences on the phylogenetic tree, PLS-DA, a supervised method, be applied at each step to obtain phylogenetic markers (Figure 6). Samples were classified according to the criteria of each branching and projected into latent spaces. From the result, the variable importance of projection (VIP) values determined the branching markers, and finally the loading plots revealed an increase or decrease in metabolite expression.

We applied metabolomics-based phylogenetics to these species by classifying specimens according to their divergence on the tree using PLS-DA. Each step showed good discrimination based on the divergence, but the determination of the branching markers needed a scaling step [33]. Among several scaling methods, we selected auto-scaling, given that this method did not omit any branching marker candidates (Table 6). After scaling, components with VIP values higher than 1.00 for all the three branching points were determined to be the markers.

By loading the averaged value of each metabolite in PLS-DA, we showed not only increases on the left or right side but also alterations (importance) for each metabolite, which can explain the degree of differences according to the evolutionary divergences of the phylogenetic tree. Furthermore, the levels of alterations were translated to color saturations, which formed a heat map related to the phylogenetic tree (Figure 7). The first branching specified eleven ginsenosides that are highly expressed in PN. This result revealed that this approach was capable of identifying metabolites with species-specific abundance, but the conventional approach also enabled us to find abundant metabolites in each species. Other ginsenosides with clear abundances in only one species showed results from the two approaches that matched the phylogenetic tree. However, metabolites that decreased only in one species contradicted the previously established phylogenetic markers. More than two branching points exhibited the same metabolites as markers, G-Rg2, 20-O-G-Rf, N-R1, G-Rd, G-Rc, G-Re and G-Rg1, which had more than two alterations during a down flow on the tree. Hence, the conventional metabolomic approach cannot be applied to these metabolites.

The size of alterations can also be evaluated in this approach. Furthermore, this approach is more likely to assimilate many experimental results about metabolites related to the phylogenetic tree, thus making a database for a whole phylogenetic mapping of metabolic alterations.

Figure 6. The flow of stepwise PLS-DA for phylogenetic mapping of metabolites

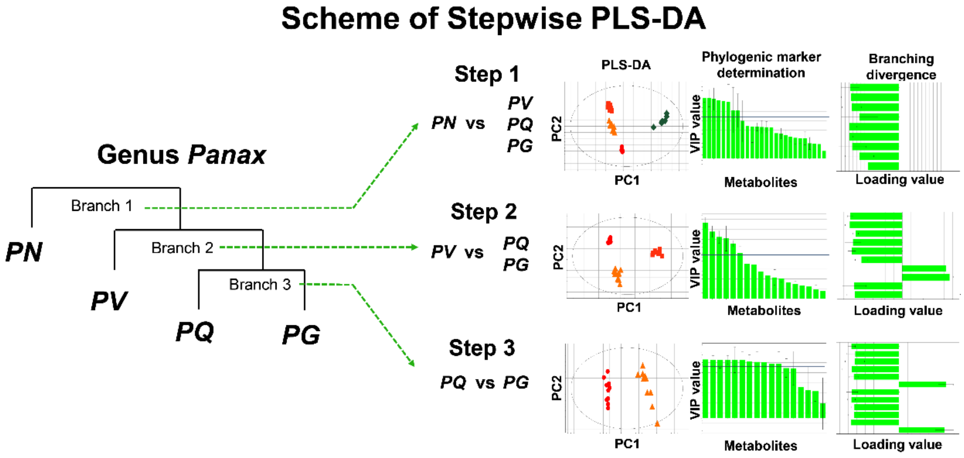
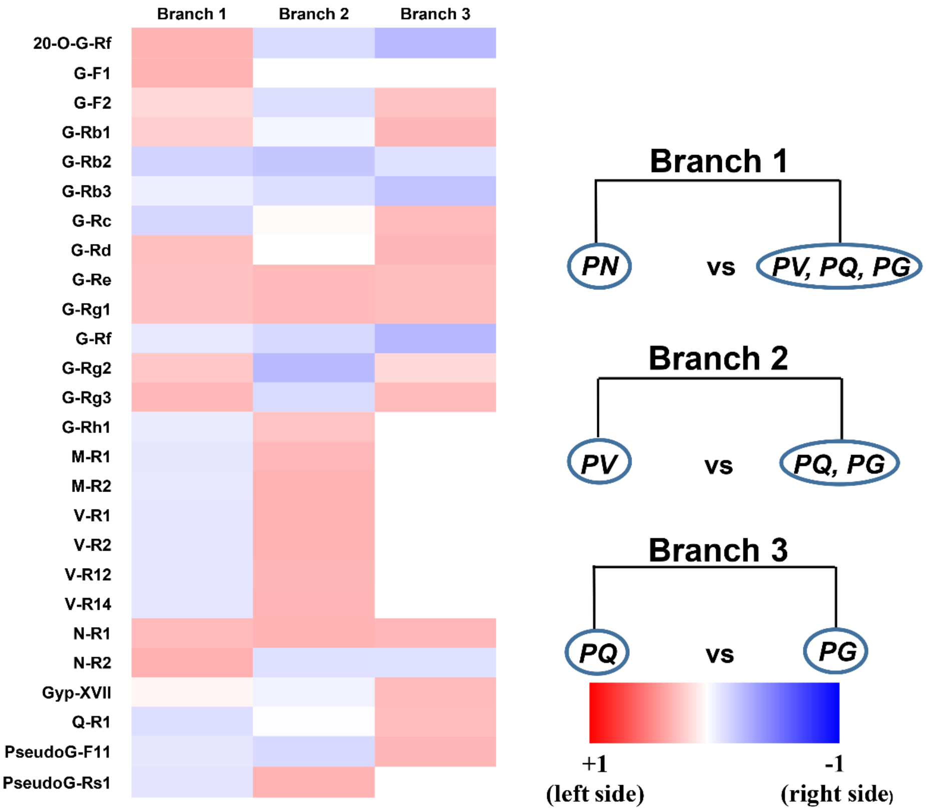


Figure 7. Heat map of metabolite alterations according to divergences on the phylogenetic tree. The level of alterations is shown as color saturation.



3.6. Distribution of secondary metabolites from the phylogenetic point of view

After mapping all the variations in secondary metabolites onto the phylogenetic tree, we can more closely examine the differences among the four species, such as alterations to gene sequences within many species or species differentiation caused by the impact of their surroundings. These *Panax* species were originally cultivated in different regions and thus in different climates. Therefore, we assume that the environmental factor has to be taken into account. According to the Köppen–Geiger climate classification, the Northern United States and Canada, in which PQ is cultivated, and South

Korea, in which PG grows, are regions of similar climate conditions characterized by a cold and snowy winter and warm-to-hot summer [34]. Therefore, PQ and PG belong to one clade. PV is cultivated in the middle region of Vietnam, dominated by a distinctive tropical climate, which might explain why the species has its unique ocotillol ginsenosides. Finally, PN is mostly cultivated in Yunnan Province in China, a region that has the characteristics of both of the two aforementioned climate conditions. Thus, PN belongs to the clade consisting of the two clades of the above species. It is suggested that, during the differentiation process from a common ancestor, the secondary metabolite profile of each species has somehow actively changed to adapt to its specific climate. As a consequence, the metabolite profile of one species might have differentiated from those of others in the form of some key metabolites despite the fact that the related species share some similar metabolites inherited from a common ancestor. This phenomenon may be accounted for by "switch on" and "switch off" gene expression, which means that, in certain cases, the gene responsible for enzymes producing a given structure or structure skeleton might have been switched off and then switched on at some later point of the evolution process [35]. Therefore, these chemotaxonomic discussions, which come from research concerning both phylogenetics and metabolomics, could expedite research in phylogenetics and evolution, with the adoption of the fast and easy protocol of stepwise PLS-DA.

4. Conclusion

In brief, our results proved that compared to the conventional metabolomics approach with PCA or PLS-DA only deriving the expressional differences among species, this proposed method might be handy to study the phylogenetic relationships and differences in secondary metabolites of plants by applying stepwise PLS-DA along the phylogenetic tree, thus highlighting the expressional differences between clades and within clade and possibly enabling further discussion of environmental evolution and life strategies embedded in it.

Part II A ^{1}H-NMR-based metabolomics approach to evaluate the geographical authenticity of *Panax ginseng* and its application in building a model effectively assessing the mixing proportion of intentional admixtures.

1. Introduction

Ginseng, the root of *Panax ginseng* is one of the most commonly used herbal medicine in the world, especially in Korea. Ginseng contains many primary metabolites such as amino acids, carbohydrates and minor elements [36], while its secondary metabolites, dammarane saponins (generally known as ginsenosides) [37], have been reported to exhibit numerous pharmacological effects such as anti-aging [38], memory enhancement [39], vasodilation [40], cognitive performance enhancement [41], antioxidant activity and cancer prevention [42]. Due to its popularity and the enormous size of the market, ginseng has long been the target of falsification, especially falsification of the cultivation region.

Ginseng currently available on the market comes mainly from Korea and China and has similar shape regardless of different cultivated locations. However, in the practice of pricing and grading herbal food products, the origin does play a major role, resulting in a higher price for the ginseng products originating from a marketable site and vice versa. For instance, if blended samples are intentionally produced by mixing ginseng roots of a lower market value with those of a higher market value, it will be unfeasible for diagnostic morphological examination to detect the proportion of mixing. This issue also holds true for other prepared forms such as powder and extract. Consequently, the ginseng market particularly and the herbal food market generally are extremely likely to suffer from contamination from this kind of adulteration practice, therefore decreasing the actual value of the product. In fact, Wallace *et al.* did report that 50% of the ginseng products labeled as "Korean ginseng (*Panax ginseng*)" were replaced by "American ginseng (*Panax quinquefolius*)" [43], and many others also expressed concerns over similar issues for other herbal food or plant species [44-48]. According to WHO Traditional Medicine Strategy of 2002-2005 and 2014-2023, this falsification practice is a threat to consumer safety and may also erode consumer confidence, thus calling for the establishment of a better authentication approach.

Undoubtedly, the traditional morphological inspection by an expert, which is subjective and lacks reproducibility, needs to be replaced by a better technique. As a result, it is essential to develop a scientific method which can effectively authenticate the origin of cultivation and, if possible, point out the proportions of blending. The genetic diversities or differences in DNA genomics are considered as a more reliable solid approach for botanical identification than morphological observation. However, this approach will be effective only if the genetic makeup of the plants is distinctive. When the genetic makeup of plants cultivated at different sites is uniform, the finding of the gene markers still remains a great challenge. Metabolomics, a research field that has recently and rapidly been developed, might provide the answer.

Metabolomics inspects the composition of an organism or biological system; thus, it can easily describe all metabolites, which can be regarded as the terminal response of one organism to its surroundings [49]. Using metabolomics, the metabolic profiles of ginseng cultivated in dissimilar regions will be different because each region has its own distinctive environmental factors. Therefore, investigating the differences in the metabolic profiles of ginseng from different regions might be meaningful in assisting the prevention of origin counterfeiting. In fact, many successful applications of metabolomics in tracking the physiological responses of plants to surrounding, in performing metabolic fingerprinting as well as in the characterization of various plants or foods, have been reported [50-54]. Among the analytical platforms often used in metabolomics techniques such as a chromatographic and spectroscopic platform, a proton nuclear magnetic resonance (^{1}H-NMR)-based metabolomics approach possesses advantages of highly reproducible, non-destructive, widely applicable methodology [55, 56]. Moreover, the use of ^{1}H-NMR-based metabolomics also simplifies the process of sample preparation and reduces the analysis time, thus making it suitable for high-throughput analysis.

In this paper, the uniformity in genetic makeup of 60 ginseng roots collected in Korea and China was first examined by a DNA-based technique, which revealed that the 60 ginseng roots were all genetically similar, thus demonstrating the very narrow genetic diversity within the ginseng samples from different geographical populations. As a result, ^{1}H-nuclear magnetic resonance (^{1}H-NMR)-based metabolomics approach, combined with a statistical method was used to effectively differentiate ginseng roots collected in Korea and China. In addition, as an attempt to recreate the adulteration practice in reality, we prepared numerous blended samples representing different ratios of Korea samples to China samples. Subsequently, the mixed samples were effectively estimated the proportion of mixing by a constrained least-squares statistical method constructed by our own laboratory, hence indicating the practical application of our method.

2. Experimental

2.1. Solvents and Chemicals

Analytical grade deuterium oxide (D_2O) was purchased from Euriso-top (France), 3-(trimethylsilyl)-propionic-2,2,3,3-d_4 acid sodium salt (TMSP), sodium deuteroxide (NaOD) and deuterium chloride (DCl) were obtained from Sigma (St. Louis, MO, USA). Monopotassium phosphate (KH_2PO_4) and dipotassium phosphate (K_2HPO_4) were obtained from American Bio (Natick, MA, USA).

2.2. Sample collection and preparation

As the origin of cultivation is critical in this experiment, whole dried roots of ginseng were directly purchased from herbal market in Seoul, Korea and Ji Lin, China, regardless of their ages to ascertain the samples currently available on the market. The morphology of the samples was authenticated by Professor Oh Won Keun at the Department of Pharmacognosy, Seoul National University, Korea. Samples were kept in the plastic bags and stored at -20 °C until use. Voucher specimens were deposited at the Department of Biomedical and Pharmaceutical Analysis, Seoul National University, Korea.

The collected samples were cut into small pieces and subsequently freeze-dried to completely eliminate moisture. Then, samples were pulverized and sieved. The powdered samples that passed between 125- 300 μm sieves were used for ^{1}H-NMR analysis.

Mixed samples were prepared as follows: 100 mg from each of 30 samples from Korea were mixed carefully together to make one representative Korea sample. The same procedure was applied to make a representative China sample. Then, those two representative samples were used to make 7 different mixing ratios: 0, 10, 25, 50, 75, 90 and 100% of the Korea sample. Samples of each ratio were prepared in triplicate.

2.3. DNA extraction and Polymerase Chain Reaction (PCR) analysis

Total DNA was extracted from the samples according to the reported protocol [57]. The intergenic spaces of the chloroplast genome were amplified by 23 primer pairs (Table 6) [58]. PCR was performed in a 30 μL reaction mixture consisting of 40 ng template, 20 pmol of primer pair (Cosmo Genetech, Seoul, Korea), 2.5 mM of dNTP and 0.4 unit of *Tag* polymerase (TaKaRa Shuzo). The PCR conditions are as follows: first, a denaturing step at 94 ℃ for 5 min, followed by 20 cycles of a DNA denaturing step at 95 ℃ for 30 s, a primer annealing step at 55 ℃ for 30 s and a *Tag* polymerase activation and DNA extension step at 72 ℃ for 30 s. Finally, the complete extension of DNA was done at 72 ℃ for 5 min. The PCR products were separated and plated on agarose gels (1.5-3%).

Table 6. Sequences and locations of the primers used for DNA analysis.

Primer ID	Location		Primer sequence
pgcpir013	*psaC - ndhE*	F	AGGCTCGGACACATTGAGTA
		R	TGAAGCAGCTATTGGACTGG
pgcpir014	*rrn5 - trnR* (ACG)	F	CTGCGGAAAAATAGCTCGAC
		R	GCCACGTGCTCTAATCCTCT
pgcpir015	*trnV* (GAC) - *rrn16*	F	ACCTTGACGRGGRGGAAGTC
		R	TGAGCCAGGATCGAACTCTC
pgcpir016	*psbC - trnS* (UGA)	F	TTCCATGACCCCTCTTAATTG
		R	TTCGAATCCCTCTCTCTCCT
pgcpir017	*trnG* (UCC) - *trnR* (UCU)	F	AAGCTAACGATGCGGGTTC
		R	CAAAGGTTTAGAAGACCTCTGTCC
pgcpir018	*rps12 - clpP*	F	GGGATTTCGTGACATTTCTGA
		R	TGTTGATCTTGTAGCGGTTGA
pgcpir019	*rpl14 - rpl16*	F	CCGCTGTTATCCGCTACATT
		R	TCGTCTAAAATGCCTATACGAACTC
pgcpir020	*trnA* (UGC) - *rrn23*	F	TTCGAGTCCGCTTATCTCCA
		R	ATCCACCGTAAGCCTTTCCT
pgcpir034	*ndhB-rps7*	F	TCATTCTGTACATGCCAGTTCAT
		R	GCCATACGCAAAAAGGAAGA
pgcpir035	*rps2-rpoC2*	F	CTCTTCCAAATTGATGTTCCAA
		R	TCCATGATACACCAGAACAATCA
pgcpir036	*psbA-trnK* (UUU)	F	CGCTTTCGCGTCTCTCTAAA
		R	ATCCGACTAGTTCCGGGTTC
pgcpir037	*trnG* (GCC)-*trnfM* (CAU)	F	TCTTTGCCAAGGAGAAGACG
		R	GGTTCATGCATGTTTGTTGC
pgcpir038	*rpl23-trnI* (CAU)	F	CTGCATATTTGATTCCATCCA
		R	ATTGGCGAATTCGTAGGTTC
pgcpir039	*petG-trnW* (CCA)	F	TACAGGCGTGGTGATCAGTT
		R	GGTAGAACGTGGGTCTCCAA
pgcpir042	*ndhG-ndhI*	F	CCCGATCCCAGAAAGACTAA
		R	CCGATGGCAGTAATTGACG
pgcpir043	*clpP-psbB*	F	TCCAGGACTTCGAAAGGGTA
		R	ACACGATACCAAGGCAAACC
pgcpir047	*psbK-psbI*	F	TGTTTGGCAAGCTGCTGTAA
		R	AAACGAAAAGTTTGAGAGTAAGCA
pgcpir050	*trnN* (GUU)-*ycf1*	F	CTCTACCACTGAGCTACTGAGGA
		R	TTCATGCATAAGGATACTAGATTACC
pgcpir055	*rrn16 -trnI* (GAU)	F	GGTAGCCGTACTGGAAGGTG
		R	AGGCACAACGACGCAATTAT
pgcpir056	*rps18-rpl20*	F	GAGTCGACCGCTAGAACTGC
		R	GGAAAGAATCCACCGGAATAA
pgcpir058	*cemA-petA*	F	TCGTGTTTCTCCGTCACTTG
		R	TCGAGTAATCTGTTCCTTTATCCA
pgcpir060	*rrn4.5-rrn5*	F	AGGCATCCTAACAGACCGATA
		R	CCTCTACGCCTAGGACACCA
pgcpir062	*psbB-psbT*	F	GCATTCCAAAAACTGGGAGA
		R	GGAATGTATAAACCAATGCTTCC

2.4. Metabolite extraction

Forty milligrams of powdered material were vortexed in 1 mL pH 7 buffer comprising monopotassium phosphate and dipotassium phosphate in D_2O (containing 0.0025% TMSP as the internal chemical shift standard), which was then extracted by ultrasonication for 40 min at 30 ℃. After extraction, the sample was centrifuged at 16,000 x *g* for 30 min. Subsequently, the supernatant was filtered using a cellulose membrane (0.45 µm). The filtrate was transferred into a 5 mm NMR tube for analysis.

2.5. NMR spectroscopy

To ensure that the samples were all analyzed under identical instrument conditions and parameters, all ^{1}H-NMR spectra of samples, including blended samples, were measured on a 600 MHz JEOL NMR ECA 600 spectrometer, equipped with a TH5 probe (JEOL, Tokyo, Japan) following the acquisition parameters: 5.7 µs (45 ℃) pulse width, 12018.6 Hz spectral width, number of scans equal to 32 and 5 s relaxation delay. During the relaxation delay, the water suppression process was enforced to eliminate the unwanted signals from residual water. Fourier transformation, phase and baseline correction were applied to the data. Calibration of the data was carried out by shifting the TMSP signal to 0.0 ppm using MestReNova (version 6.0), and the intensities of the peaks were normalized to TMSP.

Peaks in ^{1}H-NMR were first tentatively assigned by comparing the chemical shifts and the coupling constants of peaks to those of standards referred frompprevious papers [55] or in the freely available database [59, 60], using MestReNova (version 6.0). Afterwards, those tentatively identified peaks were further checked by comparing with standard prepared in our lab. After assignment, the confirmation of peak identities was achieved by two-dimensional NMR such as ^{1}H-correlation spectroscopy (^{1}H-^{1}H COSY) and heteronuclear multiple bond coherence (HMBC). Finally, all the intensities of all the spectra were derived and saved as ASCII format data for data processing and analysis.

2.6. NMR data processing and analysis

The residual water signal region (δ 4.6-5 ppm) was eliminated from the raw spectra. Then, the spectra were divided into bin steps at every δ 0.01 and aligned, producing a total of 980 bins. All the integrated values were normalized to the intensity of the TMSP signal. The normalized integrated data were pre-treated with the pareto scaling method prior to multivariate statistical analysis. This scaling reduces the relative importance of large value metabolites, thus effectively increasing the importance of low value metabolites while keeping the structure of the data partially intact. Therefore, the scaling permits both low and high value metabolites to contribute to constructing the pattern [33].

The multivariate statistical analysis was done using SIMCA-P^{+} software (version 12.0, Umetrics, Umea, Sweden) while univariate statistical analysis was performed on MetaboAnalyst 2.0 [61].

3. Results and discussion

3.1. Identification of the species using DNA-based technique

Molecular markers have long been utilized as a means of performing the estimation of the genetic diversity and genetic similarity. Chloroplasts are intracellular organelles of plants that are mainly responsible for photosynthesis. The genome of the chloroplast has long been used to construct genetic markers because significant amounts of nucleotide variation can be found in the intergenic regions of the chloroplast genome in spite of the conservancy in the genetic region. Hence, chloroplast intergenic space (CIS) has been a handy tool for systematic studies of diverse plants because the richness of information can be found for differentiation of not only species but also samples within a population in that region [62, 63].

As previously mentioned above, the substitution of *P. quinquefolius* in *P. ginseng* products has been reported, possibly due to its similar morphological appearance and genetically close phylogenetic relationship [64]. Therefore, employing 23 CISs regions (Table 6) proven to be useful for studying the genetic diversity of the Araliaceae family [58], we examined the uniformity of the genetic makeup of 60 samples from Korea and China in comparison with that of *P. quinquefolius* samples. Consequently, none of the 23 CISs regions were polymorphic among the four representative samples (Figure 8), two Korea samples (K24, K27) and two China samples (C01, C27). Notably, primer pgcpir 035 may well resolved the genetic diversity between all the *P. ginseng* samples and *P. quinquefolius* sample, a reported adulterant in ginseng products (Fig. 9). However, all of primers including primer pgcpir 035 failed to show any differences among 60 ginseng samples. This experiment indicates either the seemingly narrow genetic diversity within ginseng samples from different geographical populations or the fact that those primers do not have sufficient intra-genus discrimination power. Therefore, it is suggested to use a set of powerful intra-genus discrimination DNA ginseng markers when it comes to dealing with authentication of samples from dissimilar regions.

Figure 8. PCR results of amplified CIS regions of four *P. ginseng* root samples consisting of two from China (C01, C27) and two from Korea (K24, K27) for the 23 CIS regions.

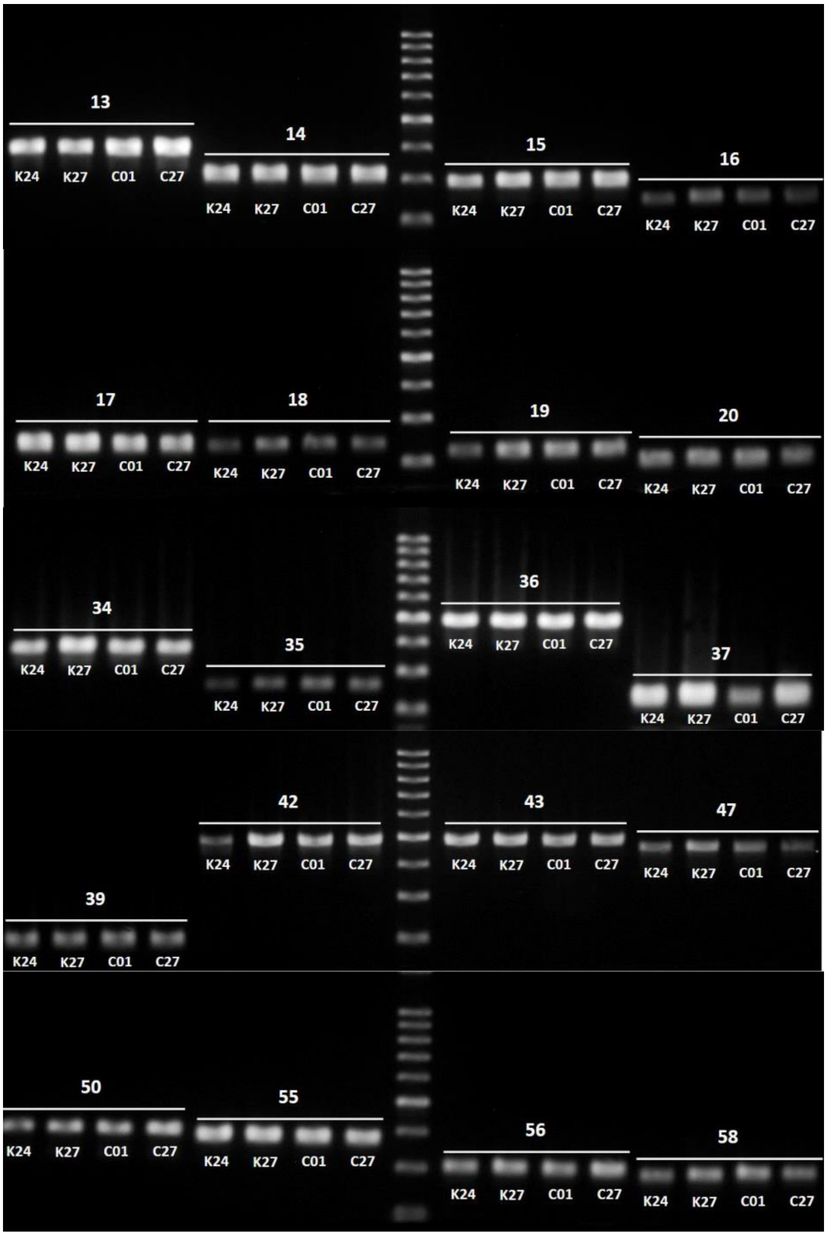
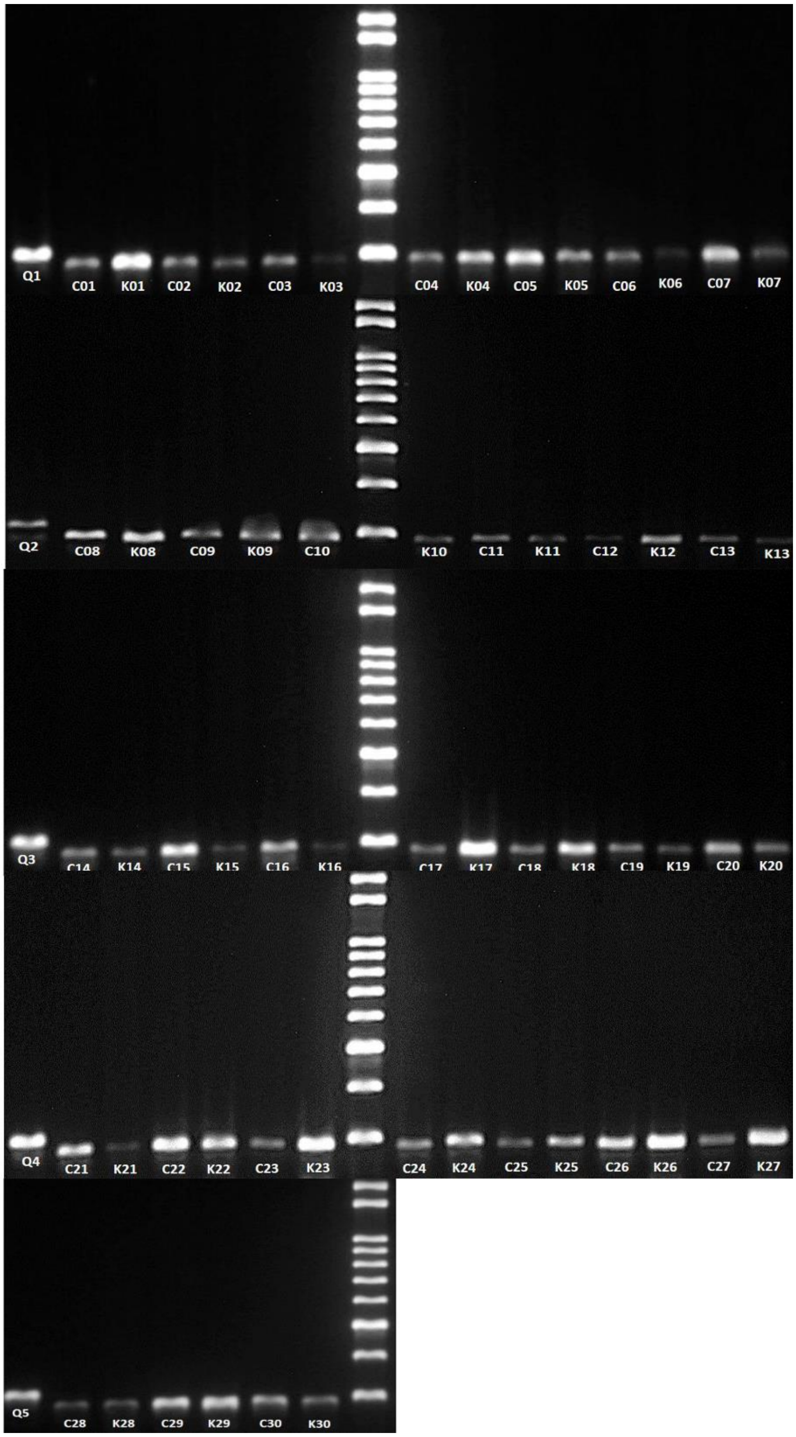


Figure 9. PCR results for the primer 'pgcpir 035' of 60 *P. ginseng* root samples from Korea and China.



3.2. ^{1}H-NMR spectrum inspection and metabolite identification

First, we tried to visually inspect the ^{1}H-NMR spectra of aqueous extracts from Korea and China samples to get a primitive idea about the differences (Fig. 3S). As seen from the spectra, a heavy congestion of high intensity signals could be observed in the carbohydrate region (δ 3.0-4.2 ppm), implying that the samples contain a considerable amount of sugar. There were also some notable signals in the organic region (δ 0.5-3 ppm). The detailed information for the assigned peaks can be found in Table 1 and Figure 10.

Figure 10. Detected peaks in the representative ^{1}H-NMR spectra. Carbohydrate (1): glucose and sucrose. Organic acid: malic acid (2), succinic acid (3), glutaric acid (4), acetic acid (5), 2-oxo- glutaric acid (6), fumaric acid (7) and formic acid (8). Amino acids: tyrosine (9), leucine (10), glutamine (11), choline (12), alanine (13) and arginine (14)

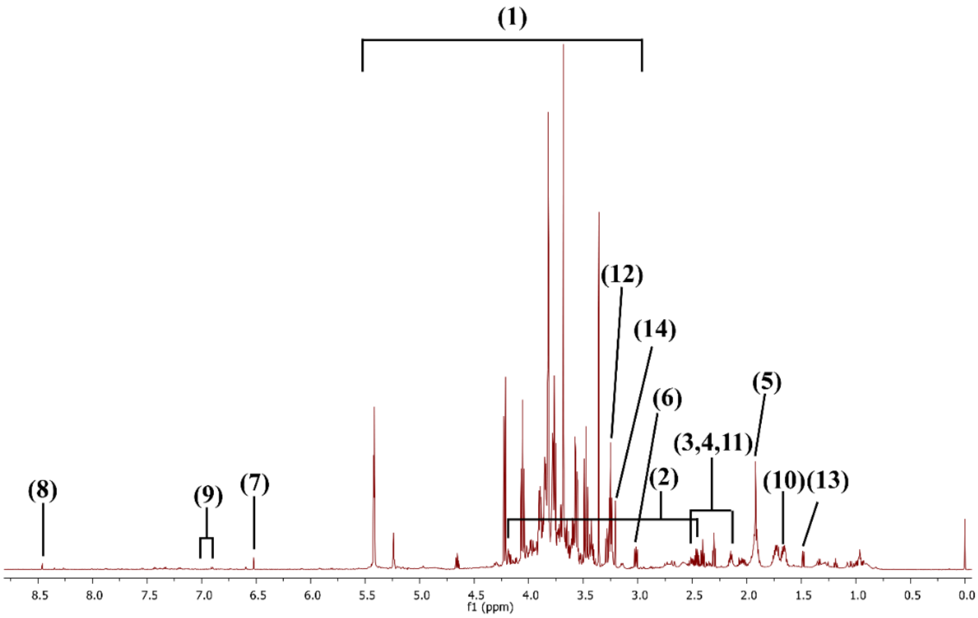
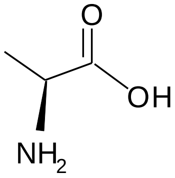
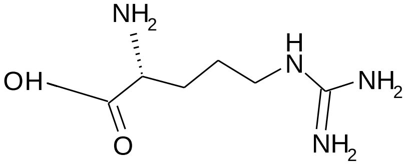
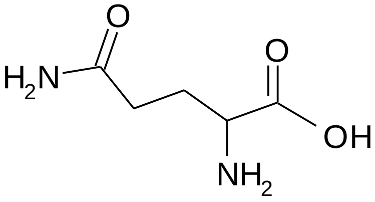
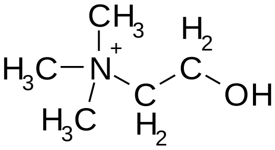
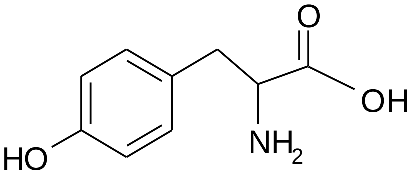
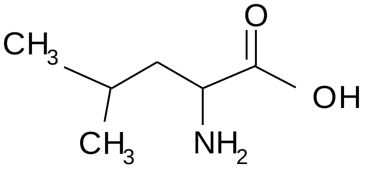
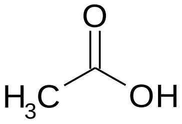
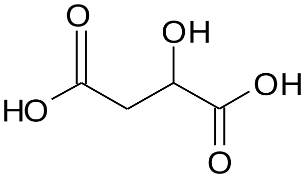
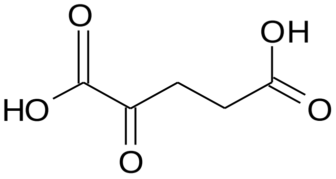
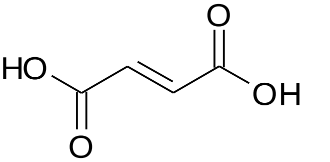
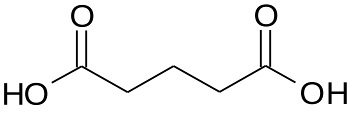
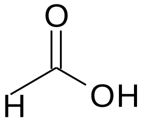
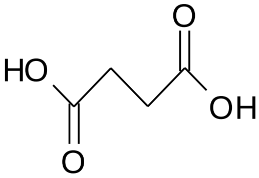
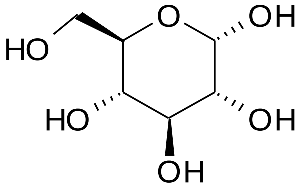
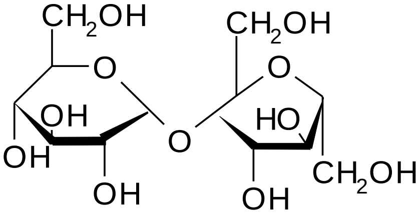


Table 7. ^{1}H-NMR chemical shifts (ppm) and multiplicity of assigned metabolites and coupling constants (in Hertz) of assigned metabolites.

	Metabolite	**Structure**	**Assignment**
Amino acids	Alanine		1.48 (d, J=7.2)
	Arginine		3.24 (t, J=6.9)
	Glutamine		2.46 (m)
	Choline		3.2 (s)
	Tyrosine		7.2 (t, J=8.5) 6.9 (t, J=7.5)

	Leucine		1.7 (m)
Organic acids	Acetic acid		1.91 (s)
	Malic acid		2.68 (dd, J=15.5, 31) 4.3 (dd, J=9.9)
	2-Oxoglutaric acid		3 (t, J=7.4)
	Fumaric acid		6.6 (s)
	Glutaric acid		2.32 (t, J=7.4)

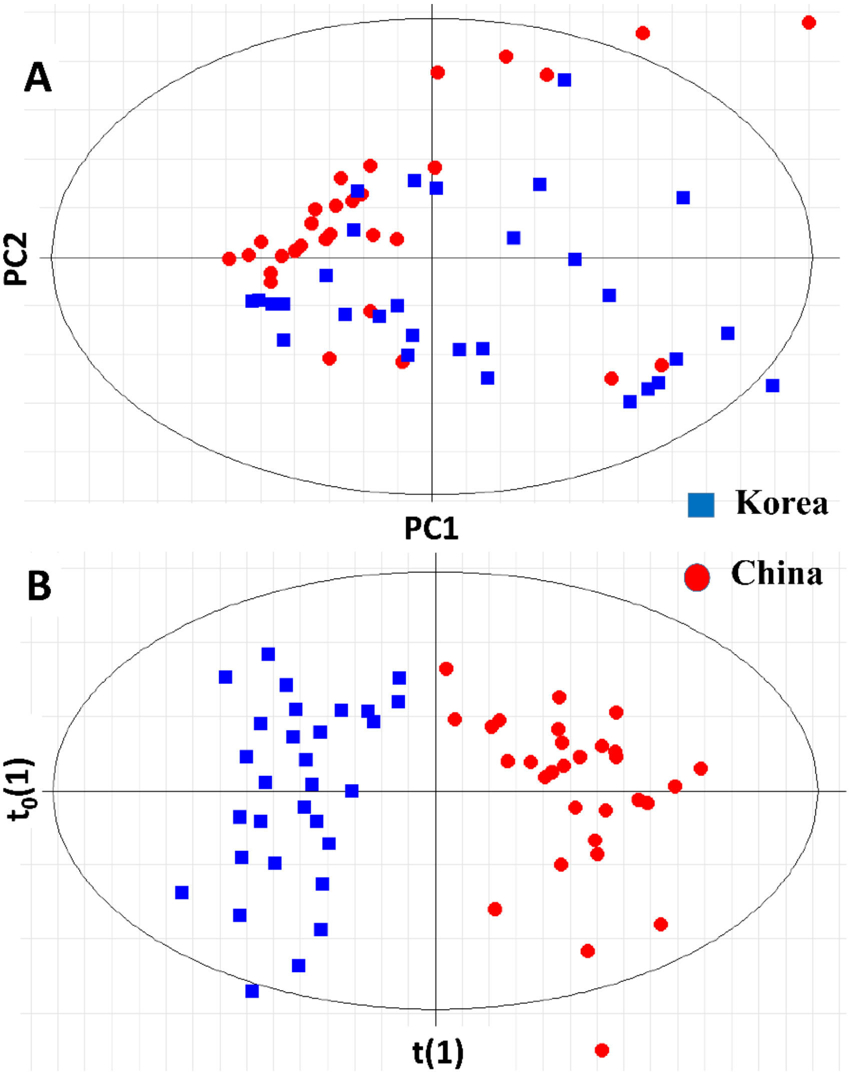
	Formic acid		8.45 (s)
	Succinic acid		2.57 (s)
Carbohydrate	Glucose		5.24 (d, J=3.8)
	Sucrose		5.41 (d, J=3.8)

3.3. Multivariate statistical analysis

Multivariate statistical analysis was used to investigate the differences in ginseng samples from two countries. PCA was first applied to the data without any prior group label (in an unsupervised manner) to identify outliers and visualize the underlying trends as well as showing the variation in the matrix data [55]. Using six principal components with a total variation of 89.2%, the statistical modeling was achieved (Figure 11A). As seen from the PCA plot, there were noticeable overlaps between the two classes of samples, indicating that the samples could not be separated well, despite the high level of fitting being applied. The variation within each set of samples was also considerable, which might arise because the samples were of different ages and were also cultivated in various areas in a single country.

Next, orthogonal projections on latent structure-discriminant analysis (OPLS-DA) was introduced for discrimination and potential markers identification. OPLS-DA is an extended version of the supervised partial least square regression method (PLS-DA) building in an integrated OSC filter, hence allowing better classification and predicting capacity. Furthermore, OPLS-DA can be utilized when PCA fails to discriminate individual classes exhibiting divergence in within-class variation [65]. Indeed, despite a large amount of structural noise, the score plot of the OPLS-DA clearly showed the satisfactory separation between the ginseng samples from two countries (Figure 11B). The model was obtained using one predictive and two orthogonal variations while having an adequate goodness of fit, R_Y^2, of 86% and predictive ability, Q^2, of 89.7%. The total variation of the independent variables defined by the mode in R_X^2 was 0.79%, showing that the OPLS-DA approach is more proper and effective than PCA for discrimination of the cultivation sources.

Figure 11. (A) Score plot of PCA of Korea and China samples. **(B)** Score plot of OPLS-DA of Korea and China samples. *Blue squares* represent the Korea samples; *red dots* represent the China samples.



3.4. Determination of the potential metabolite markers

To further examine how significantly the variable contributed to the discrimination, the loading scores of the variable importance for projection (VIP) (Figure 12) was extracted from the OPLS-DA model. On the *x*-axis, the chemical shifts (ppm) represent the bins, which can be referred to the equivalent to that of the detected metabolites in Table 7. One or more bins belonging to the same metabolites are observable in the spectrum. For example, the bins at 2.12, 2.13, 2.14, 2.15, 2.16 and 2.45, 2.46 ppm were all identified as glutamine. Now those bins having a VIP score more than 1 were derived and further analyzed by univariate statistical analysis. Univariate statistical analysis was introduced to confirm biased distribution, thus removing those samples that were not strongly biased to either one of the groups ($p > 0.05$). This strategy might be conventional, but the strategy is, however, reliable. Finally, we were able to identify seven marker candidates. ^{1}H-^{1}H COSY (Figure 13) were subsequently employed to further confirm the identities. Based on the results, the discriminatory compounds such as glutaric acid, succinic acid, malic acid, choline, glucose and sucrose expressed highly in the Korea samples, while glutamine appeared abundantly in the China samples (Fig. 14).

Sugar metabolism is a especially dynamic process, and metabolic fluxes and sugar concentrations differ greatly in both the development stage and the response to environmental changes. By and large, under low sugar conditions, source activities such as photosynthesis, nutrient mobilization and export are upregulated, while sink activities like growth and storage are upregulated due to carbon sources being plentifully available [66]. In that aspect, as in Korea, the roots are usually harvested in autumn [67], when conditions such as sunlight and rain are much more available, thus allowing accumulation of carbohydrate sources preparing for the coming winter. As in China, the roots are harvested in spring, after a long winter when sunlight and rain are rarely encountered resulting in increased reduction of available carbohydrate sources for energy consumption. Similarly, the organic acids including succinic acid and malic acid are key intermediates in the tricarboxylic acid cycle, so the higher concentration of these metabolites in Korea samples might indicate that energy metabolism of Korean samples was more active than that of China samples. Moreover, the upregulation of amino acid synthesis was reported to be associated with temperature stress and light exposure as well [68, 69]. Therefore, we might conclude that the distinctive environmental stress and the cultivation conditions in each country might influence the differences in the primary metabolites.

Figure 12. VIPs score plot derived from OPLS-DA model.

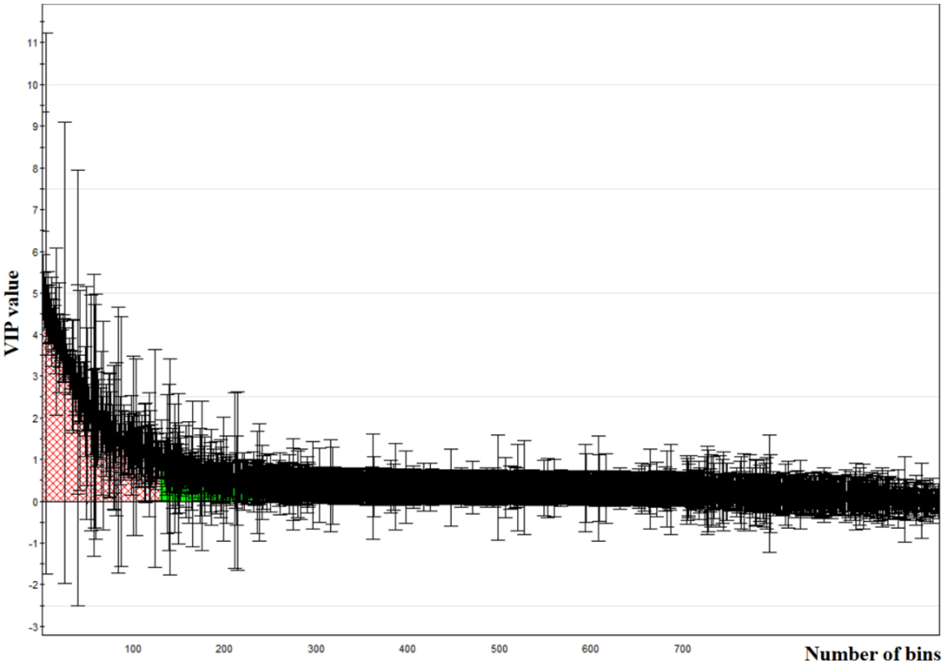
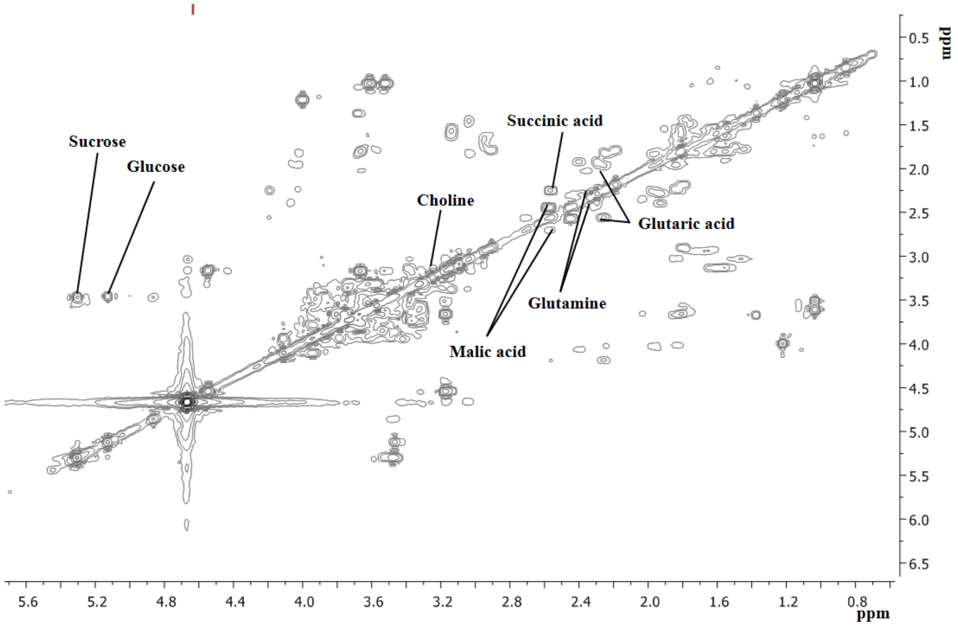


Figure 13. Expanded representative 2D-NMR spectra of *P. ginseng* samples obtained from ^{1}H-^{1}H COSY



3.5. Validation of the model by prediction

Twenty-one sets of processed metabolomics data for seven mixing proportions were applied to the OPLS-DA model samples previously built. As illustrated in Figure 14, the mixed samples were scattered from left to right of the tPS(1) predictive component that was utilized for clean discrimination of the cultivation origins. As we expected, those samples having 100% and 0% Korea ratio were clearly classified into the Korea group and the China group, respectively, thus indicating the robustness of the statistical model. Even though the mixed samples at the ratio of 0% and 10% of Korea somewhat overlapped, they still showed good separation when visualizing the model in 3D score plot (Figure 15) and in this case, only choline and glucose markers contributed to the discrimination, indicating the difficulty in detecting those samples intentionally blended at a tiny proportion.

More interestingly, it is notable that other blended samples were positioned consistently with the degree of mixing. For example, in Korea samples, the 100% mixing ratio mixed well in the clustering while other ratios were gradually located farther away with respect to the level of mixing. Finally,50% Korea mixing samples were located around the main axis of the predictive component tPS(1), implying that the model could not classify the 50% Korea mixing samples to any group because 50% Korea mixing samples expressed the characteristics of both classes of samples. This result not only clearly confirmed the accuracy and efficiency of our method but also encouraged us to build a statistical procedure for estimating mixing proportions of the blended samples.

Figure 14. Score plot of OPLS-DA of mixing samples combining with Korea and China samples. *Blue squares* represent the Korea samples; *red dots* represent the China samples; and *green shapes* represent the mixing ratio.

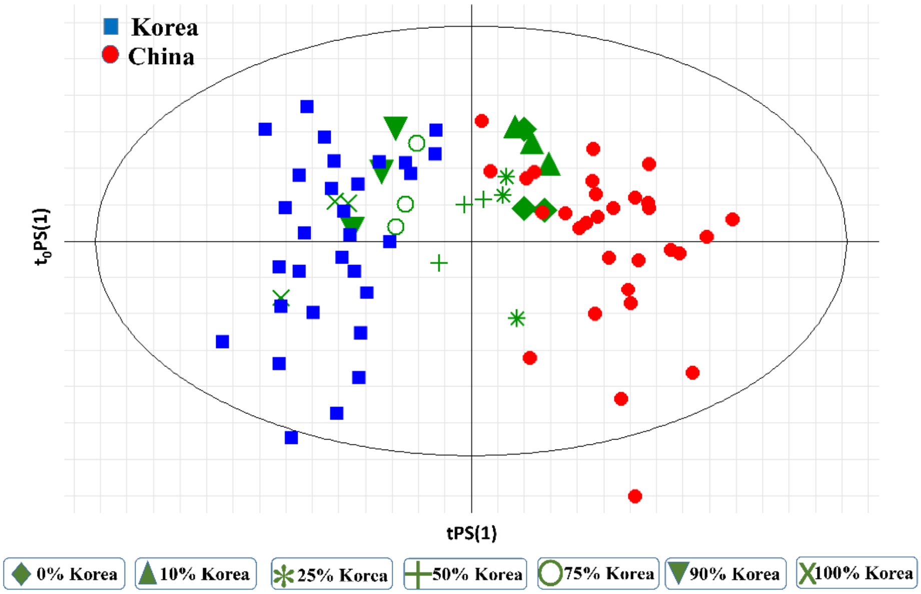
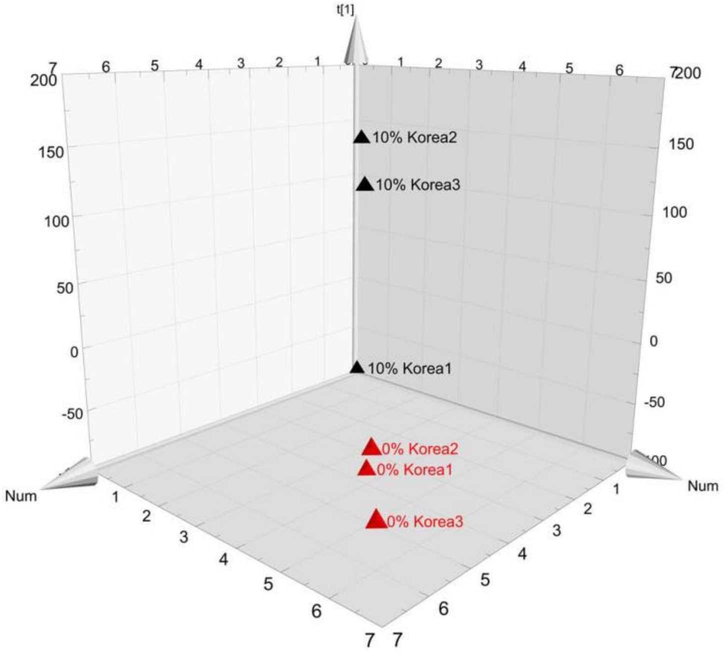


Figure 15. 3D OPLS-DA score plot of two 0% and 10% Korea groups. Black triangles represent 10% Korea group while red triangles represent 0% Korea groups. Only choline and glucose markers contributed to the discrimination of these two groups.



3.6. Setting up a statistical procedure for assessing the mixing ratio of blended samples from dissimilar origins

As a practical aspect, it could be of great importance to predict the unknown mixing proportion of Korea and China samples whenever they are intentionally blended. We constructed the aggregated constraint least squares method, a statistical approach to estimate the mixing proportion from ^{1}H-NMR spectra.

We obtained 21 ^{1}H-NMR spectra of blended samples together with the information of the true mixing proportions. The information is expressed as a vector and scalar pair, where $(\boldsymbol{y}^{(i)}, \pi^{(i)})$ $(i = 1, 2, \ldots, 21)$, where $\boldsymbol{y}^{(i)} = \left(y_{0.20}^{(i)}, y_{0.21}^{(i)}, \ldots, y_{6.00}^{(i)}\right)$ is a vector of intensities, and $\pi^{(i)}$ is the true mixing proportion of Korea sample in the i^{th} sample. The subscription 0.20, 0.21, … , and 6.00 means ppm (binning of 0.01 ppm), for example, $y_{0.20}^{(i)}$ indicates the ^{1}H-NMR intensity at 0.20 ppm of i^{th} sample. The values of $\pi^{(i)}$ are $\pi^{(i)} = 0$ (pure China) for $i = 1, 2, 3$, $\pi^{(i)} = 0.1$ for $i = 4, 5, 6$, and the rest is 0.25, 0.5, 0.75, 0.9 and 1 (pure Korea), in the same manner. The parameters $\pi^{(i)}$ are assumed to be unknown during the analysis except for the pure China ($i = 1, 2, 3$) and the pure Korea ($i = 19, 20, 21$). In addition, $\hat{\boldsymbol{v}}^{CN}$ and $\hat{\boldsymbol{v}}^{KR}$ were set as the mean vectors of pure China and pure Korea, respectively, i.e., $\hat{\boldsymbol{v}}^{CN} = \frac{1}{3}\sum_{i=1}^{3} \boldsymbol{y}^{(i)}$ and $\hat{\boldsymbol{v}}^{KR} = \frac{1}{3}\sum_{i=19}^{21} \boldsymbol{y}^{(i)}$.

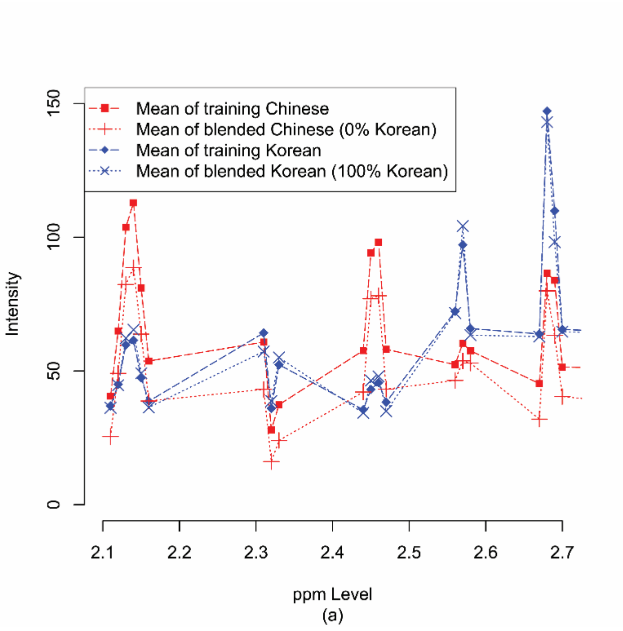
The key idea of estimation is as follows. If a blended sample $\boldsymbol{y}$ has a mixture proportion of π, the spectral intensity at k-ppm, say y_k, is expected to have a mean of $\pi\hat{\mu}_k{}^{KR} + (1 - \pi)\hat{\mu}_k{}^{CN}$, where $\hat{\mu}_k{}^{CN}$ and $\hat{\mu}_k{}^{KR}$ are the intensity at k ppm of the mean vectors of 30 training samples from China and Korea, respectively. We did not use all bins of spectral intensities because many bins in the spectra were baseline noise. Instead, we chose a subset of bins corresponding to 0.01 ppm-nearest neighborhoods of choline (3.21-3.23 ppm), glucose (5.23-5.26 ppm), glutaric acid (2.31-2.33 ppm), glutamine (2.11-2.17 ppm and 2.43-2.47 ppm), malic acid (2.67-2.70 ppm and 4.29-4.32 ppm), succinic acid (2.56-2.58 ppm) and sucrose (5.39-5.42 ppm). The bins selected were discovered from the t-test procedures of the training samples in the previous section, which ensured that the mean intensities $\hat{\mu}_k{}^{CN}$ and $\hat{\mu}_k{}^{KR}$ were statistically well separated. The consideration of the 0.01 ppm nearest neighbor is for reducing the unexpected error from the horizontal shift.

Note that there could be an uncontrolled deviation between the blended samples and the training samples due to the differences in the experiment conditions. As Figure 16A shows, the mean of the representative blended samples and ($\hat{\boldsymbol{\nu}}^{CN}$ and $\hat{\boldsymbol{\nu}}^{KR}$) in the selected bins was shifted from the means of the training samples ($\hat{\boldsymbol{\mu}}^{CN}$ and $\hat{\boldsymbol{\mu}}^{KR}$). To reduce the shift, we further pre-processed the data of the blended samples utilizing a pointwise linear matching, $\hat{\mu}_k^{CN} = a_k \hat{\nu}_k^{CN} + b_k$ and $\hat{\mu}_k^{KR} = a_k \hat{\nu}_k^{KR} + b_k$ for each $k \in K$. The parameters a_k and b_k were calculated directly by solving the system of linear equations. Then, the blended sample $\boldsymbol{y}^{(i)}$ was redefined by $y_k^{(i)} \coloneqq a_k y_k^{(i)} + b_k$ for every $i = 1,2, \dots ,21$. The shift of the mean spectra between blended and training samples was removed after the transformation (Figure 16B). This transformation step also allows the removal of the deviation in the mean spectra of samples acquired on a different machine and under different experimental conditions, thus making it altogether functional in the constrained least squares model.

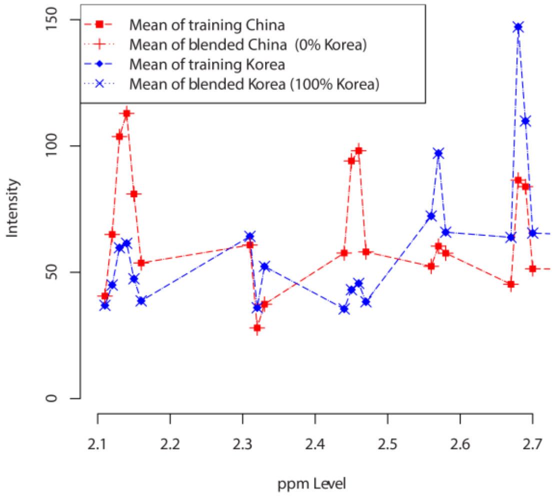
Then, the aggregated constraint least squares method was utilized to estimate the mixture proportions. The details are explained here for completeness. The set of all 37 bins corresponding to seven biomarkers are denoted by K. From the choice of bins, $\hat{\nu}_k^{KR}$ and $\hat{\nu}_k^{CN}$ have sufficient dispersion to distinguish two groups for each $k \in K$. The estimation of π has two steps, bin-wise estimation and aggregation. First, fix each $k \in K$ and find π_k minimizing $(y_k - \pi_k \hat{\mu}_k^{KR} - (1 - \pi_k)\hat{\mu}_k^{CN})^2$ subject to $0 \leq \pi_k \leq 1$. The solution of the problem, namely, $\hat{\pi}_k$ is explicitly $\hat{\pi}_k = \min\left(\max\left(\frac{y_k - \hat{\mu}_k^{CN}}{\hat{\mu}_k^{KR} - \hat{\mu}_k^{CN}}, 0\right), 1\right)$. As a second step, π was estimated as the median of $\hat{\pi}_k$, $\hat{\pi} \coloneqq \text{median}\ (\hat{\pi}_k : k = 1,2, \dots , K)$. We used median because median is more robust for the unexpected outlier than the mean. We applied this procedure to 21 blended samples, respectively, to calculate $\hat{\pi}^{(i)}$ $(i = 1, 2, \ldots, 21)$, the estimate of the mixing proportion. Figure 17 illustrates the true mixing proportion ($\pi^{(i)}$) against the estimated mixing proportion ($\hat{\pi}^{(i)}$) for $i = 1, 2, \ldots, 21$. The linear trend was quantified as the adjusted R^2 of 0.8343, thus indicating that the model also has a good predictability for estimating the mixing proportions of blended ginseng

Figure 16. Mean spectra at 2.10 - 2.70 ppm, Dashed lines represent the mean ($\hat{\boldsymbol{\mu}}$) of the spectra of the training data set; dotted lines represent the mean ($\hat{\boldsymbol{\nu}}$) of the spectra of the pure samples from the blended samples; blue lines represent Korea samples; and red lines represent China samples **(A)**. After pre-processing, the spectra of pure samples from the blended samples (dotted lines) are synchronized to those of the training samples (dashed lines). The shift of the mean spectra between blended and training samples was removed after the transformation **(B)**.

A



B



(b)

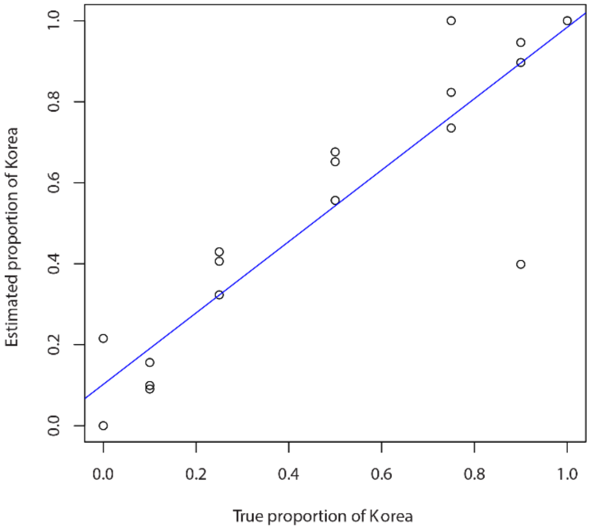


Figure 17. Plot for the true proportion ($\boldsymbol{\pi}^{(i)}$) versus the estimated proportion ($\hat{\boldsymbol{\pi}}^{(i)}$) for $i = 1, 2, \dots, 21$. $\hat{\boldsymbol{\pi}}^{(i)}$ is from the predictive model constructed using the aggregation of the constrained least squares method. The adjusted R^2 was 0.8343

.

4. Conclusion

In this study, 60 ginseng samples from different geographical areas, namely Korea and China, were found to be indistinguishable using DNA-based approach due to the very narrow genetic diversity among those samples. However, ^{1}H-NMR-based metabolomics with OPLS-DA statistical models clearly clustered the samples according to the geographical origins. Several metabolites contributing to the discrimination were also found to be potential markers. Samples of different mixing proportions were applied to the newly built OPLS-DA model, being separated well according to the ratios of mixing. Consequently, we believe that the ease and transferability of our approach as well as its applicability to other products could contribute to the establishment of a better quality control method for ginseng particularly and other herbal medicine, and thus promote a safer market and greater consumer confidence by preventing origin counterfeiting.

References

[1] J. Sullivan, P. Joyce, Model Selection in Phylogenetics, Annual Review of Ecology, Evolution, and Systematics 36 (2005) 445-466.

[2] F. Ronquist, Dispersal-Vicariance Analysis: A New Approach to the Quantification of Historical Biogeography, Systematic Biology 46(1) (1997) 195-203.

[3] M. Nei, PHYLOGENETIC ANALYSIS IN MOLECULAR EVOLUTIONARY GENETICS, Annual Review of Genetics 30(1) (1996) 371-403.

[4] W.J. Murphy, E. Eizirik, W.E. Johnson, Y.P. Zhang, O.A. Ryder, S.J. O'Brien, Molecular phylogenetics and the origins of placental mammals, Nature 409(6820) (2001) 614-618.

[5] M. Wink, Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective, Phytochemistry 64(1) (2003) 3-19.

[6] N. Hounsome, B. Hounsome, D. Tomos, G. Edwards-Jones, Plant Metabolites and Nutritional Quality of Vegetables, Journal of Food Science 73(4) (2008) R48-R65.

[7] F. Bourgaud, A. Gravot, S. Milesi, E. Gontier, Production of plant secondary metabolites: a historical perspective, Plant Science 161(5) (2001) 839-851.

[8] A.C. Kushalappa, R. Gunnaiah, Metabolo-proteomics to discover plant biotic stress resistance genes, Trends in Plant Science 18(9) (2013) 522-531.

[9] M. Wink, L. Witte, Evidence for a wide-spread occurrence of the genes of quinolizidine alkaloid biosynthesis, FEBS Letters 159(1-2) (1983) 196-200.

[10] P. Tetenyi, A Chemotaxonomic Classification of the Solanaceae, Annals of the Missouri Botanical Garden 74(3) (1987) 600-608.

[11] S.-O. Yang, S.W. Lee, Y.O. Kim, S.-H. Sohn, Y.C. Kim, D.Y. Hyun, Y.P. Hong, Y.S. Shin, HPLC-based metabolic profiling and quality control of leaves of different Panax species, J. Ginseng Res. 37(2) (2013) 248-253.

[12] P.H. Chan, K.Y. Zheng, K.W. Tsim, H. Lam, Metabonomic analysis of water extracts from Chinese and American ginsengs by 1H nuclear magnetic resonance: identification of chemical profile for quality control, Chinese Medicine 7(1) (2012) 25.

[13] G.X. Xie, Y. Ni, M.M. Su, Y.Y. Zhang, A.H. Zhao, X.F. Gao, Z. Liu, P.G. Xiao, W. Jia, Application of ultra-performance LC-TOF MS metabolite profiling techniques to the analysis of medicinal Panax herbs, Metabolomics 4(3) (2008) 248-260.

[14] S. Zhu, K. Zou, H. Fushimi, S. Cai, K. Komatsu, Comparative study on triterpene saponins of ginseng drugs, Planta Med. 70(7) (2004) 666-677.

[15] E. Nocerino, M. Amato, A.A. Izzo, The aphrodisiac and adaptogenic properties of ginseng, Fitoterapia 71, Supplement 1 (2000) S1-S5.

[16] J. Wu, H.K. Jeong, S.E. Bulin, S.W. Kwon, J.H. Park, I. Bezprozvanny, Ginsenosides protect striatal neurons in a cellular model of Huntington's disease, Journal of Neuroscience Research 87(8) (2009) 1904-1912.

[17] N. Angelova, H.-W. Kong, R. van der Heijden, S.-Y. Yang, Y.H. Choi, H.K. Kim, M. Wang, T. Hankemeier, J. van der Greef, G. Xu, R. Verpoorte, Recent methodology in the phytochemical analysis of ginseng, Phytochemical Analysis 19(1) (2008) 2-16.

[18] S. Jia, J. Li, N. Yunusova, J.H. Park, S.W. Kwon, J. Lee, A New Application of Charged Aerosol Detection in Liquid Chromatography for the Simultaneous Determination of Polar and Less Polar Ginsenosides in Ginseng Products, Phytochemical Analysis 24(4) (2013) 374-380.

[19] G. Theodoridis, H.G. Gika, I.D. Wilson, LC-MS-based methodology for global metabolite profiling in metabonomics/metabolomics, TrAC 27(3) (2008) 251-260.

[20] D.-K. Lee, M.H. Yoon, Y.P. Kang, J. Yu, J.H. Park, J. Lee, S.W. Kwon, Comparison of primary and secondary metabolites for suitability to discriminate the

origins of Schisandra chinensis by GC/MS and LC/MS, Food Chemistry 141(4) (2013) 3931-3937.

[21] G. Xie, Y. Ni, M. Su, Y. Zhang, A. Zhao, X. Gao, Z. Liu, P. Xiao, W. Jia, Application of ultra-performance LC-TOF MS metabolite profiling techniques to the analysis of medicinal Panax herbs, Metabolomics 4(3) (2008) 248-260.

[22] N. Kim, K. Kim, B.Y. Choi, D. Lee, Y.-S. Shin, K.-H. Bang, S.-W. Cha, J.W. Lee, H.-K. Choi, D.S. Jang, D. Lee, Metabolomic Approach for Age Discrimination of Panax ginseng Using UPLC-Q-Tof MS, J. Agric. Food. Chem. 59(19) (2011) 10435-10441.

[23] N. Kim, K. Kim, D. Lee, Y.-S. Shin, K.-H. Bang, S.-W. Cha, J.W. Lee, H.-K. Choi, B.Y. Hwang, D. Lee, Nontargeted Metabolomics Approach for Age Differentiation and Structure Interpretation of Age-Dependent Key Constituents in Hairy Roots of Panax ginseng, J. Nat. Prod. 75(10) (2012) 1777-1784.

[24] M. Dan, M. Su, X. Gao, T. Zhao, A. Zhao, G. Xie, Y. Qiu, M. Zhou, Z. Liu, W. Jia, Metabolite profiling of Panax notoginseng using UPLC-ESI-MS, Phytochemistry 69(11) (2008) 2237-44.

[25] H.-J. Kim, C.-W. Cho, J.-T. Hwang, N. Son, J.H. Choi, G.-S. Shim, C.-K. Han, LC-MS-based metabolomic analysis of serum and livers from red ginseng-fed rats, J. Ginseng Res. 37(3) (2013) 371-378.

[26] Y. Chen, Z. Zhao, H. Chen, T. Yi, M. Qin, Z. Liang, Chemical Differentiation and Quality Evaluation of Commercial Asian and American Ginsengs based on a UHPLC–QTOF/MS/MS Metabolomics Approach, Phytochem. Anal. 26(2) (2015) 145-160.

[27] S. Zhu, H. Fushimi, S. Cai, K. Komatsu, Phylogenetic relationship in the genus Panax: inferred from chloroplast trnK gene and nuclear 18S rRNA gene sequences, Planta Med. 69(7) (2003) 647-53.

[28] Y. Sasaki, H. Fushimi, H. Cao, S.-Q. Cai, K. Komatsu, Sequence Analysis of Chinese and Japanese <i>Curcuma</i> Drugs on the 18S rRNA Gene and <i>trn</i>K Gene and the Application of Amplification-Refractory Mutation System Analysis for Their Authentication, Biol. Pharm. Bull. 25(12) (2002) 1593-1599.

[29] D. Charif, J. Thioulouse, J.R. Lobry, G. Perrière, Online synonymous codon usage analyses with the ade4 and seqinR packages, Bioinformatics 21(4) (2005) 545-547.

[30] J. Chakerian, S. Holmes, Computational Tools for Evaluating Phylogenetic and Hierarchical Clustering Trees, J. Comput. Graph. Stat. 21(3) (2012) 581-599.

[31] P.W.a.D. Che*, Constructing phylogenetic trees using interacting pathways, Bioinformation 9(7) (2013) 363–367.

[32] V. Makarenkov, D. Kevorkov, P. Legendre, Phylogenetic network construction approaches, in: R.M.B. Dilip K. Arora, B.S. Gautam (Eds.), Applied Mycology and Biotechnology, Elsevier2006, pp. 61-97.

[33] R. van den Berg, H. Hoefsloot, J. Westerhuis, A. Smilde, M. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC Genomics 7(1) (2006) 142.

[34] M.C. Peel, B.L. Finlayson, T.A. Mcmahon, Updated world map of the Köppen-Geiger climate classification, Hydrology and Earth System Sciences Discussions 4(2) (2007) 439-473.

[35] M. Wink, L. Witte, Evidence for a wide-spread occurrence of the genes of quinolizidine alkaloid biosynthesis: Induction of alkaloid accumulation in cell suspension cultures of alkaloid-‘free’ species, FEBS Lett. 159(1–2) (1983) 196-200.

[36] Y.-Z. Xiang, H.-C. Shang, X.-M. Gao, B.-L. Zhang, A Comparison of the ancient use of ginseng in traditional Chinese medicine with modern pharmacological experiments and clinical trials, Phytother. Res. 22(7) (2008) 851-858.

[37] B.-K. Shin, S.W. Kwon, J.H. Park, Chemical diversity of ginseng saponins from Panax ginseng, J. Ginseng Res. 39(4) (2015) 287-298.

[38] K. Metori, M. Furutsu, S. Takahashi, The preventive effect of ginseng with du-zhong leaf on protein metabolism in aging, Biol. Pharm. Bull 20(3) (1997) 237-242.

[39] Y. Jang, W.-J. Lee, G.-S. Hong, W.-S. Shim, Red ginseng extract blocks histamine-dependent itch by inhibition of H1R/TRPV1 pathway in sensory neurons, J. Ginseng Res. 39(3) (2015) 257-264.

[40] W.Y. Kim, J.M. Kim, S.B. Han, S.K. Lee, N.D. Kim, M.K. Park, C.K. Kim, J.H. Park, Steaming of Ginseng at High Temperature Enhances Biological Activity, J. Nat. Prod. 63(12) (2000) 1702-1704.

[41] D.O. Kennedy, A.B. Scholey, Ginseng: potential for the enhancement of cognitive performance and mood, Pharmacol. Biochem. Behav. 75(3) (2003) 687-700.

[42] J. Sharma, P.K. Goyal, Chemoprevention of chemical-induced skin cancer by Panax ginseng root extract, J. Ginseng Res. 39(3) (2015) 265-273.

[43] L.J. Wallace, S.M.A.L. Boilard, S.H.C. Eagle, J.L. Spall, S. Shokralla, M. Hajibabaei, DNA barcodes for everyday life: Routine authentication of Natural Health Products, Food Res. Int. 49(1) (2012) 446-452.

[44] N. Techen, S.L. Crockett, I.A. Khan, B.E. Scheffler, Authentication of Medicinal Plants Using Molecular Biology Techniques to Compliment Conventional Methods, Curr. Med. Chem. 11(11) (2004) 1391-1401.

[45] J. Song, H. Yao, Y. Li, X. Li, Y. Lin, C. Liu, J. Han, C. Xie, S. Chen, Authentication of the family Polygonaceae in Chinese pharmacopoeia by DNA barcoding technique, J. Ethnopharmacol. 124(3) (2009) 434-439.

[46] R. Srirama, U. Senthilkumar, N. Sreejayan, G. Ravikanth, B.R. Gurumurthy, M.B. Shivanna, M. Sanjappa, K.N. Ganeshaiah, R. Uma Shaanker, Assessing species admixtures in raw drug trade of Phyllanthus, a hepato-protective plant using molecular tools, J. Ethnopharmacol. 130(2) (2010) 208-215.

[47] L. Jaakola, M. Suokas, H. Häggman, Novel approaches based on DNA barcoding and high-resolution melting of amplicons for authenticity analyses of berry species, Food Chem. 123(2) (2010) 494-500.

[48] I. Bruni, F. De Mattia, A. Galimberti, G. Galasso, E. Banfi, M. Casiraghi, M. Labra, Identification of poisonous plants by DNA barcoding approach, Int. J. Legal Med. 124(6) (2010) 595-603.

[49] S. Moco, J. Vervoort, S. Moco, R.J. Bino, R.C.H. De Vos, R. Bino, Metabolomics technologies and metabolite identification, TrAC-Trend. Anal. Chem. 26(9) (2007) 855-866.

[50] L. Nyadong, G.A. Harris, S. Balayssac, A.S. Galhena, M. Malet-Martino, R. Martino, R.M. Parry, M.D. Wang, F.M. Fernández, V. Gilard, Combining Two-Dimensional Diffusion-Ordered Nuclear Magnetic Resonance Spectroscopy, Imaging Desorption Electrospray Ionization Mass Spectrometry, and Direct Analysis in Real-Time Mass Spectrometry for the Integral Investigation of Counterfeit Pharmaceuticals, Anal. Chem. 81(12) (2009) 4803-4812.

[51] J. Kim, Y. Jung, B. Song, Y.-S. Bong, D.H. Ryu, K.-S. Lee, G.-S. Hwang, Discrimination of cabbage (Brassica rapa ssp. pekinensis) cultivars grown in different geographical areas using 1H NMR-based metabolomics, Food Chem. 137(1–4) (2013) 68-75.

[52] D.M.A.M. Luykx, S.M. van Ruth, An overview of analytical methods for determining the geographical origin of food products, Food Chem. 107(2) (2008) 897-911.

[53] L. Li, C. Zhao, Y. Chang, X. Lu, J. Zhang, Y. Zhao, J. Zhao, G. Xu, Metabolomics study of cured tobacco using liquid chromatography with mass spectrometry: Method development and its application in investigating the chemical differences of tobacco from three growing regions, J. Sep. Sci. 37(9-10) (2014) 1067-1074.

[54] M. Lv, J. Chen, Y. Gao, J. Sun, Q. Zhang, M. Zhang, F. Xu, Z. Zhang, Metabolomics based on liquid chromatography with mass spectrometry reveals the chemical difference in the stems and roots derived from Ephedra sinica, J. Sep. Sci. 38(19) (2015) 3331-3336.

[55] H.K. Kim, Y.H. Choi, R. Verpoorte, NMR-based metabolomic analysis of plants, Nat. Protoc. 5(3) (2010) 536-549.

[56] H.K. Kim, Y.H. Choi, R. Verpoorte, NMR-based plant metabolomics: where do we stand, where do we go?, Trends Biotechnol. 29(6) (2011) 267-275.

[57] G.C. Allen, M.A. Flores-Vergara, S. Krasynanski, S. Kumar, W.F. Thompson, A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide, Nat. Protoc. 1(5) (2006) 2320-2325.

[58] J. Kim, J.-Y. Jung, H.-I. Choi, N.-H. Kim, J. Park, Y. Lee, T.-J. Yang, Diversity and evolution of major Panax species revealed by scanning the entire chloroplast intergenic spacer sequences, Genet. Resour. Crop Ev. 60(2) (2013) 413-425.

[59] E. Ulrich, H. Akutsu, J. Doreleijers, Y. Harano, Y. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. Schulte, D. Tolmie, R. Kent Wenger, H. Yao, J. Markley, BioMagResBank, Nucleic Acids Res. 36 (2008) D402 - D408.

[60] D. Wishart, D. Tzur, C. Knox, R. Eisner, A. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. Duggan, G. Macinnis, A. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. Sykes, H. Vogel, L. Querengesser, HMDB: the human metabolome database, Nucleic Acids Res. 35 (2007) D521 - D526.

[61] J. Xia, R. Mandal, I.V. Sinelnikov, D. Broadhurst, D.S. Wishart, MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis, Nucleic Acids Res. 40(W1) (2012) W127-W133.

[62] R.J. Britten, L. Rowen, J. Williams, R.A. Cameron, Majority of divergence between closely related DNA samples is due to indels, Proc. Natl. Acad. Sci. U.S.A. 100(8) (2003) 4661-4665.

[63] D.E. McCauley, The use of chloroplast DNA polymorphism in studies of gene flow in plants, Trends Ecol. Evol. 10(5) (1995) 198-202.

[64] K. Komatsu, S. Zhu, H. Fushimi, T.K. Qui, S. Cai, S. Kadota, Phylogenetic Analysis Based on 18S rRNA Gene and matK Gene Sequences of Panax vietnamensis and Five Related Species, Planta Med. 67(05) (2001) 461-465.

[65] M. Bylesjö, M. Rantalainen, O. Cloarec, J.K. Nicholson, E. Holmes, J. Trygg, OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification, J. Chemometr. 20(8-10) (2006) 341-351.

[66] F. Rolland, E. Baena-Gonzalez, J. Sheen, SUGAR SENSING AND SIGNALING IN PLANTS: Conserved and Novel Mechanisms, Annu. Rev. Plant Biol. 57(1) (2006) 675-709.

[67] M. Horacek, J.-S. Min, S.-C. Heo, G. Soja, Discrimination between ginseng from Korea and China by light stable isotope analysis, Analytica Chimica Acta 682(1–2) (2010) 77-81.

[68] N. Bouché, H. Fromm, GABA in plants: just a metabolite?, Trends Plant Sci. 9(3) (2004) 110-115.

[69] G. Noctor, A.-C.M. Arisi, L. Jouanin, C.H. Foyer, Manipulation of Glutathione and Amino Acid Biosynthesis in the Chloroplast, Plant Physiol. 118(2) (1998) 471-482.

Publication

- H.T. Nguyen, D.-K. Lee, W.J. Lee, G. Lee, S.J. Yoon, B.-k. Shin, M.D. Nguyen, J.H. Park, J. Lee, S.W. Kwon, UPLC-QTOFMS based metabolomics followed by stepwise partial least square-discriminant analysis (PLS-DA) explore the possible relation between the variations in secondary metabolites and the phylogenetic divergences of the genus Panax, Journal of Chromatography B 1012–1013 (2016) 61-6
- H.T. Nguyen, D.-K. Lee, Y.-G. Choi, J.-E. Min, S.J. Yoon, Y.-H. Yu, J. Lim, J. Lee, S.W. Kwon, J.H. Park, A 1H NMR-based metabolomics approach to evaluate the geographical authenticity of herbal medicine and its application in building a model effectively assessing the mixing proportion of intentional admixtures: A case study of Panax ginseng: Metabolomics for the authenticity of herbal medicine, Journal of Pharmaceutical and Biomedical Analysis 124 (2016) 120-128

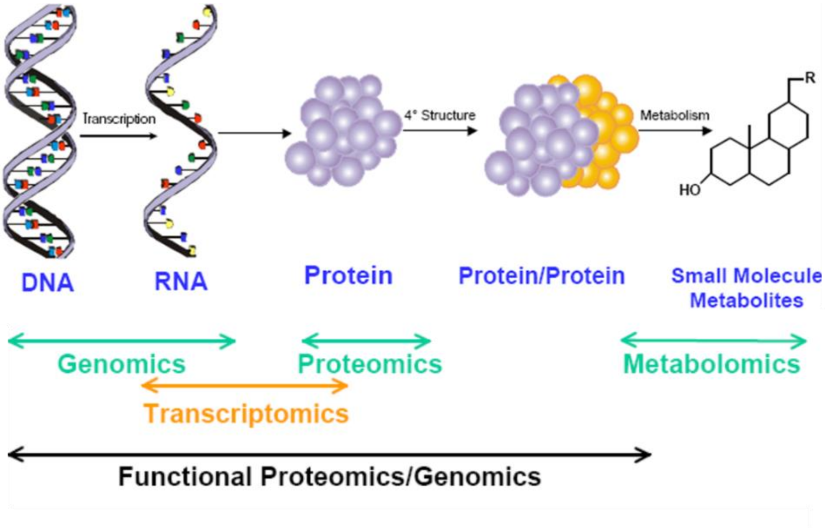
Index

1. Metabolomics approach

Metabolomics approach is scientific study focusing on the small molecule metabolites. Metabolites are compounds synthesized by plants for both essential functions. There are 2 types of metabolites: primary and secondary metabolites. Primary metabolites essentially play their role in growth and development while secondary metabolites are known for their specific functions such as pollinator attraction or defense against herbivory.

Metabolomics, the downstream product of genomics, transcriptomics and proteomics, is an recently emerging approach of system biology that has provided often unexpected and unique insights into various biological processes. Unlike the genome or proteome, changes in the metabolome are rapid and represent the final response of an organism to both internal and external stimuli. Hence, metabolomics is particularly conducive to identifying pathophysiologically affected processes and moreover elucidating novel physiological and pathological mechanisms.

The things that make metabolomics approach different from the conventional analytical chemistry is that metabolomics approach embraces high throughput analysis and therefore creates a very complex data matrix. Multivariate statistical analysis is often employed to reduce the dimension of such complex data matrix and thus help interpret the meaning of the data set.

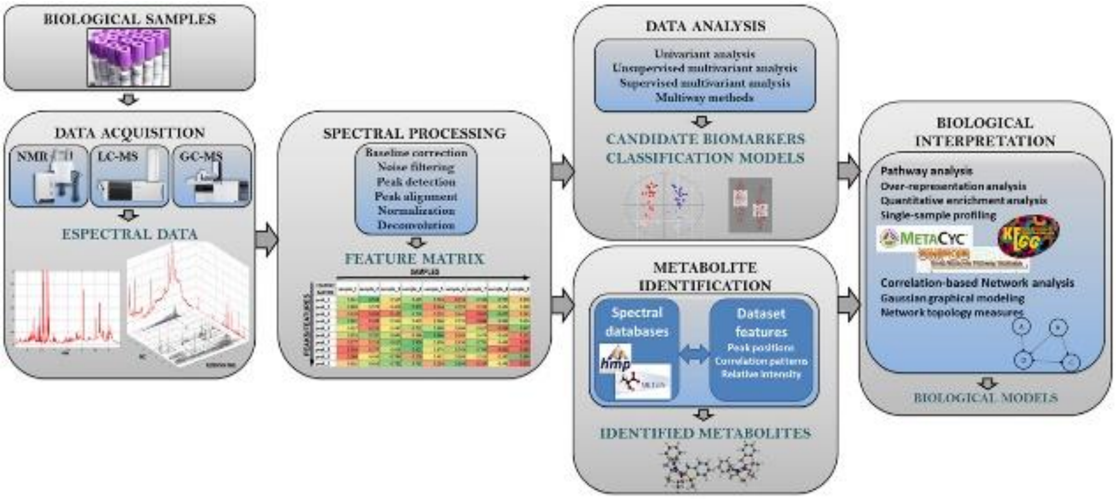


(hos.ufl.edu/meteng/PCB%205530%202012/Metabolomics)

2. Metabolomics analytical methods

It should be noted that the choice of analytical methods is simplified by the fact that the many substances in a living metabolism are interlinked in synthesis and function, with each substance providing information about some of the others. When a large subset of these substances are quantitatively analyzed, the metabolites measured can be chosen by analytical convenience and economy rather than maximum information content per metabolite.

Those analytical platforms usually employed in metabolomics approach are Liquid chromatography couple with mass detector (LC-MS), Gas chromatography couple with mass detector (GC-MS) and Nuclear magnetic resonance (NMR).

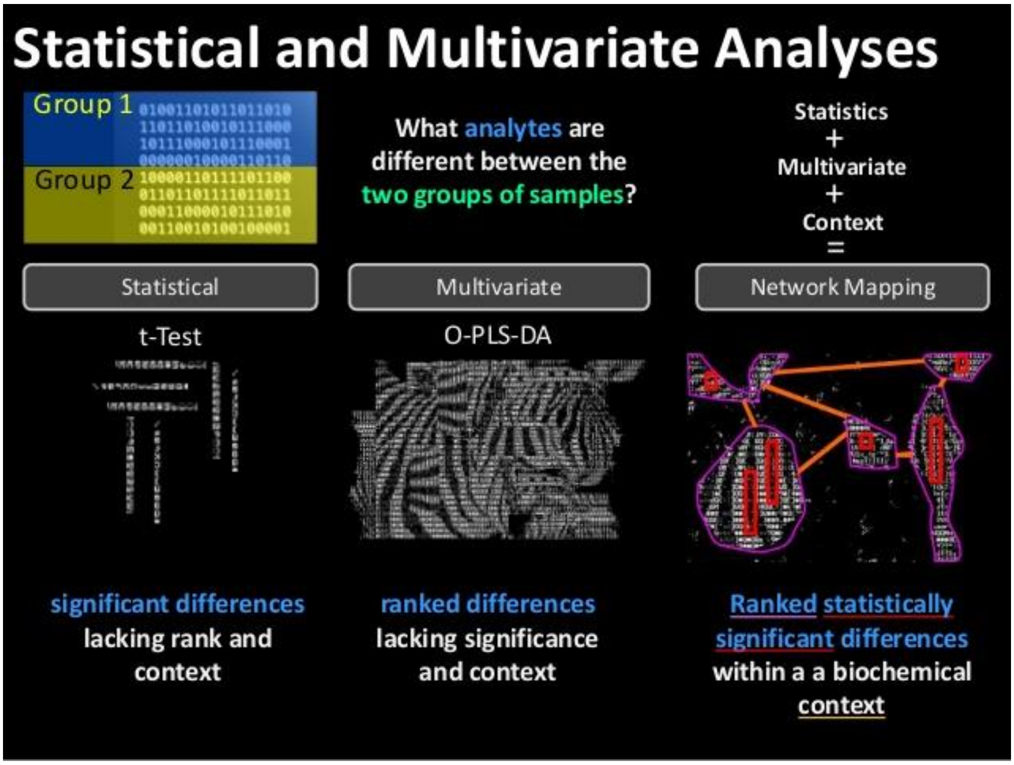


(http://journal.frontiersin.org/article/10.3389/fbioe.2015.00023/full)

3. Multivariate statistical analysis

Multivariate analysis is the area of statistics that deals with observations made on many variables. The main objective is to study how the variables are related to one another, and how they work in combination to distinguish between the cases on which the observations are made.

The analysis of multivariate data permeates every research discipline: biology, medicine, environmental science, sociology, economics, education, linguistics, archaeology, anthropology, psychology and behavioural science, to name a few, and has even been applied in philosophy. All natural and physical processes are essentially multivariate in nature—the challenge is to understand the process in a multivariate way, where variables are connected and their relationships understood, as opposed to a bunch of univariate processes, single variables at a time, isolated from one another.



Statistical and Multivariate Analyses

What analytes are different between the two groups of samples?

Statistics + Multivariate + Context =

Statistical	Multivariate	Network Mapping
t-Test	O-PLS-DA	
significant differences lacking rank and context	ranked differences lacking significance and context	Ranked statistically significant differences within a a biochemical context

(http://imdevsoftware.wordpress.com/2014/10/11/2014-)

4. Statistical Platforms

MetaboAnalys-is a set of online tools for metabolomic data analysis and interpretation, created by members of the Wishart Research Group at the University of Alberta. It was first released in May 2009 and version 2.0 was released in January 2012. MetaboAnalyst provides a variety of analysis methods that have been tailored for metabolomic data. These methods include metabolomic data processing, normalization, multivariate statistical analysis, and data annotation. The current version is focused on biomarker discovery and classification.

MetaboAnalyst supports a wide variety of data input types commonly generated by metabolomic studies including GC/LC-MS raw spectra, MS/NMR peak lists, NMR/MS peak intensity table, NMR/MS spectral bins, and metabolite concentrations.

MetaboAnalyst has four modules:

- Data processing
- Statistical analysis (one-factor, two-factor, and time-series data)
- Functional enrichment analysis
- Metabolic pathway analysis

MetaboAnalyst is part of a suite of metabolomics databases that also includes Human Metabolome Database (HMDB), DrugBank, Toxin and Toxin-Target Database, and The Small Molecule Pathway Database. The HMDB has over 7900 human metabolites and roughly 7200 associated DNA and protein sequences, that are linked to these metabolite entries. While DrugBank includes information on 6707 drugs and 4228 non-redundant drug targets, enzymes, transporters, and carriers, T3DB houses over 2900 common toxins and environmental pollutants. The suite is rounded out by SMPDB with its pathway diagrams for more than 350 human metabolic and disease pathways.

(http://www.metaboanalyst.ca/)

5. Metabolome online library

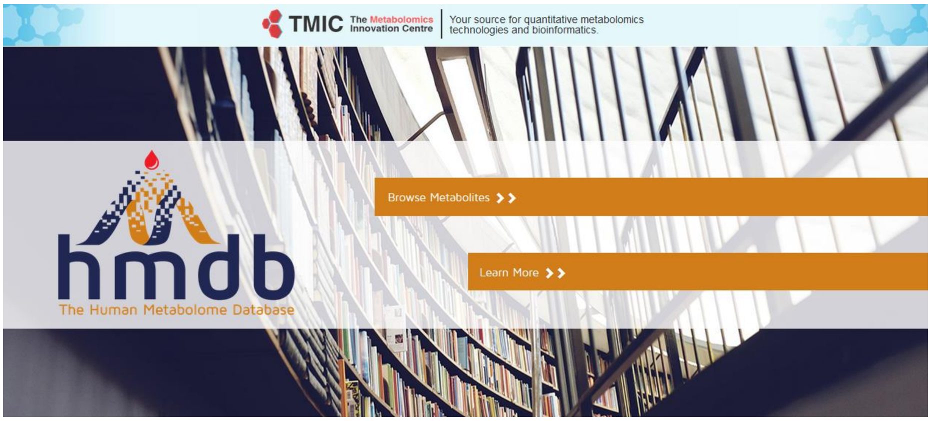
The Human Metabolome Database (HMDB) is a freely available electronic database containing detailed information about small molecule metabolites found in the human body. It is intended to be used for applications in metabolomics, clinical chemistry, biomarker discovery and general education. The database is designed to contain or link three kinds of data

1) chemical data

2) clinical data

3) molecular biology/biochemistry data

The database contains 42,003 metabolite entries including both water-soluble and lipid soluble metabolites as well as metabolites that would be regarded as either abundant (> 1 uM) or relatively rare (< 1 nM). Additionally, 5,701 protein sequences are linked to these metabolite entries. Each MetaboCard entry contains more than 110 data fields with 2/3 of the information being devoted to chemical/clinical data and the other 1/3 devoted to enzymatic or biochemical data. Many data fields are hyperlinked to other databases (KEGG, PubChem, MetaCyc, ChEBI, PDB, UniProt, and GenBank) and a variety of structure and pathway viewing applets. The HMDB database supports extensive text, sequence, chemical structure and relational query searches.

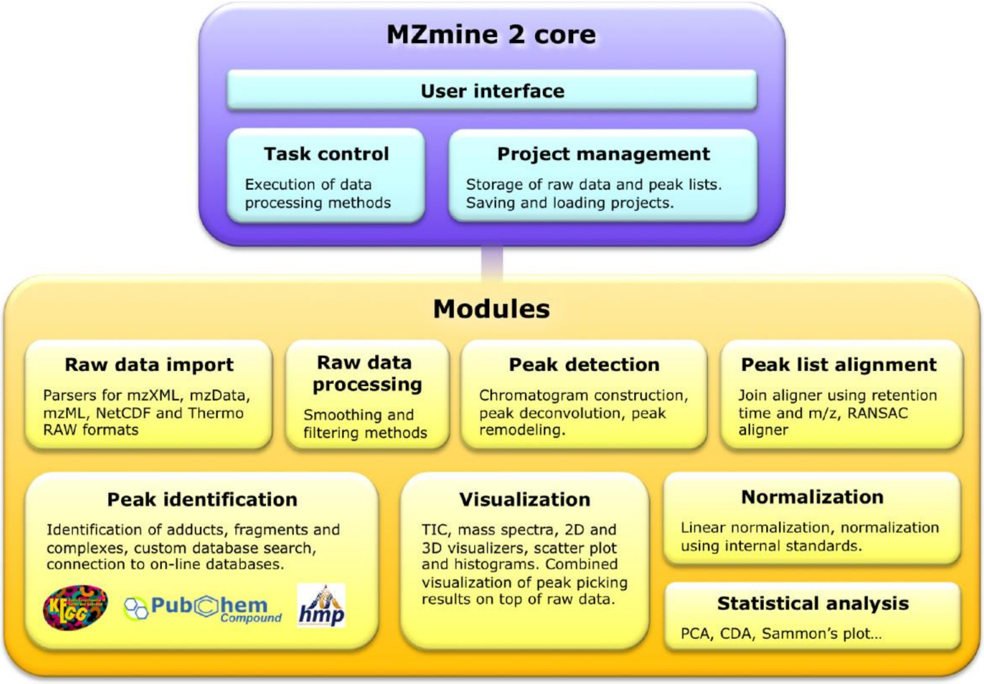

TMIC The Metabolomics Innovation Centre
Your source for quantitative metabolomics technologies and bioinformatics.
hmdb
The Human Metabolome Database
Browse Metabolites ❯❯
Learn More ❯❯

(http://www.hmdb.ca/)

6. Framework for processing metabolomics profile data

Mzmine is modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. Mzmine is distributed free of charge and could be downloaded online (http://mzmine.github.io/)

The software could be run on the Windows, Mac OS X, and Linux platforms.


MZmine 2 core
User interface
Task control
Execution of data processing methods
Project management
Storage of raw data and peak lists. Saving and loading projects.
Modules
Raw data import
Parsers for mzXML, mzData, mzML, NetCDF and Thermo RAW formats
Raw data processing
Smoothing and filtering methods
Peak detection
Chromatogram construction, peak deconvolution, peak remodeling.
Peak list alignment
Join aligner using retention time and m/z, RANSAC aligner
Peak identification
Identification of adducts, fragments and complexes, custom database search, connection to on-line databases.
KEGG
PubChem Compound
hmp
Visualization
TIC, mass spectra, 2D and 3D visualizers, scatter plot and histograms. Combined visualization of peak picking results on top of raw data.
Normalization
Linear normalization, normalization using internal standards.
Statistical analysis
PCA, CDA, Sammon's plot...

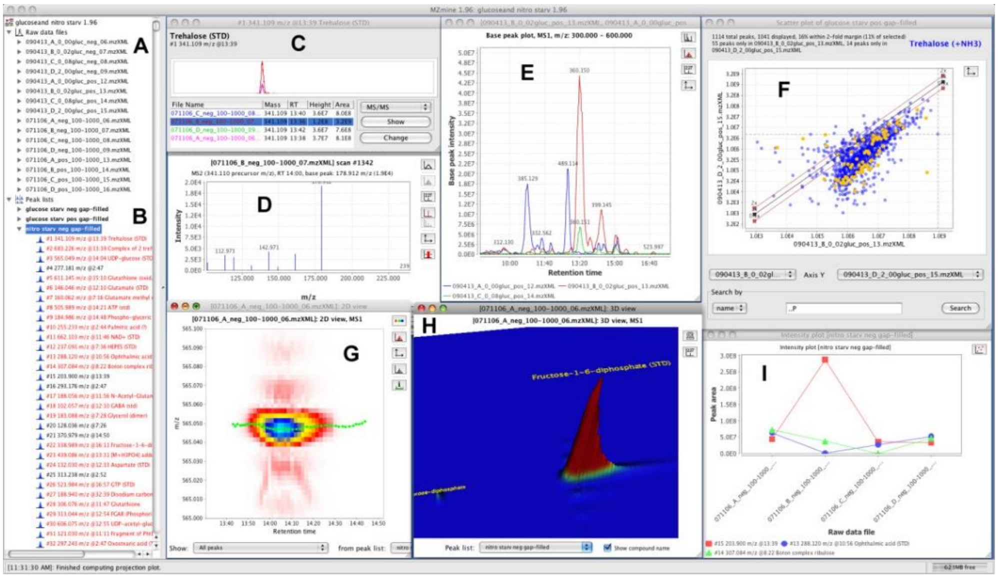
(http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-395)

6.1. Raw data file format support

Mzmine can process both unit mass resolution and accurate mass resolution MS data in both continuous and centroid modes, including fragmentation (MS^n) scans. The currently supported file formats are mzML, mzXML, mzData, NetCDF, and RAW format used natively by Thermo Fisher Scientific instruments (requires installation of Thermo Xcalibur). If other mass spectra from other instrument is used, it is required to convert the file to CDF format prior to processing by MZmine.

6.2. Data visualization

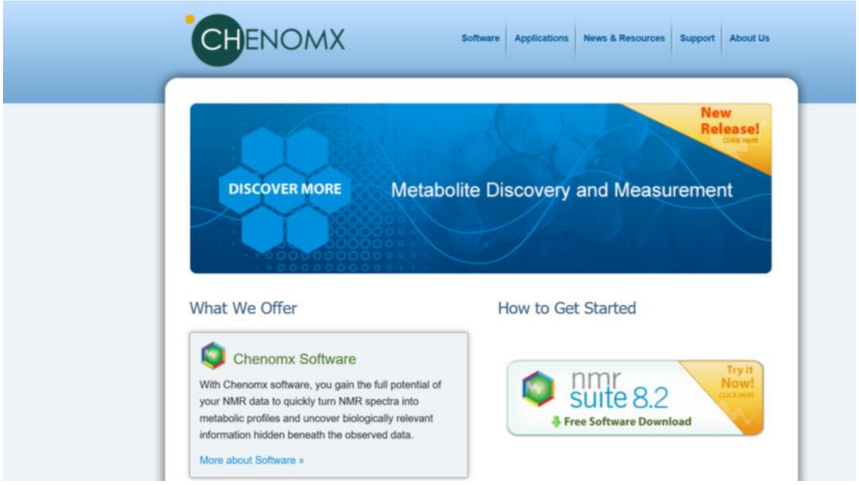
MZmine 2 includes several of visualization modules. Following the goal of providing the user with an intuitive interface, the visualizers automatically annotate raw data with the obtained peak picking and identification results, allowing for quick orientation when large amounts of data are being processed.



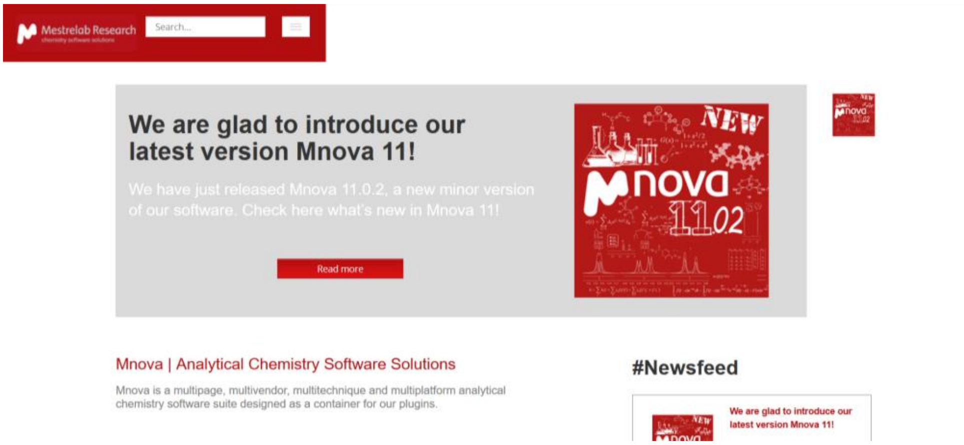
(A) imported samples, (B) peak lists including single peak list contents, (C) peak shapes for an identified metabolite across multiple samples, (D) MS/MS spectrum of a metabolite, (E) combined base peak plot for multiple samples, (F) scatter plot of peak areas across two samples, (G) 2D plot of a detected peak, mass-to-charge ratio vs. retention time, (H) 3D view of a detected peak, and (I) intensity plot for specific peaks across multiple samples.

7. Nuclear Magnetic Resonance (NMR) software

There are several computer software that are developed for handling the raw spectra of NMR data. Among which, MestReNova and Chenomex have been the usual method of choice. MestReNova could be downloaded at http://mestrelab.com, while Chenomex could be found at http://www.chenomx.com.



With Chenomx software, you gain the full potential of your NMR data to quickly turn NMR spectra into metabolic profiles and uncover biologically relevant information hidden beneath the observed data.



We are glad to introduce our latest version Mnova 11!

We have just released Mnova 11.0.2, a new minor version of our software. Check here what's new in Mnova 11!

Mnova | Analytical Chemistry Software Solutions

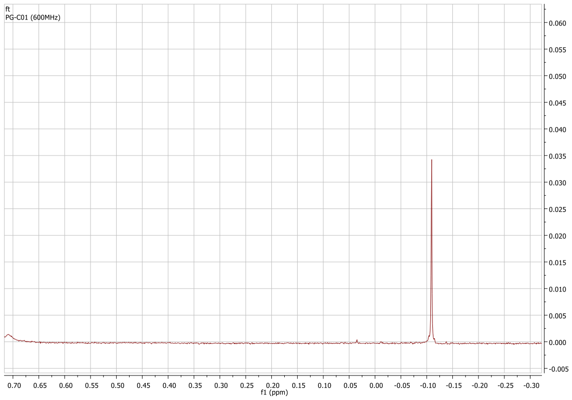
Mnova is a multipage, multivendor, multitechnique and multiplatform analytical chemistry software suite designed as a container for our plugins.

We are glad to introduce our latest version Mnova 11!

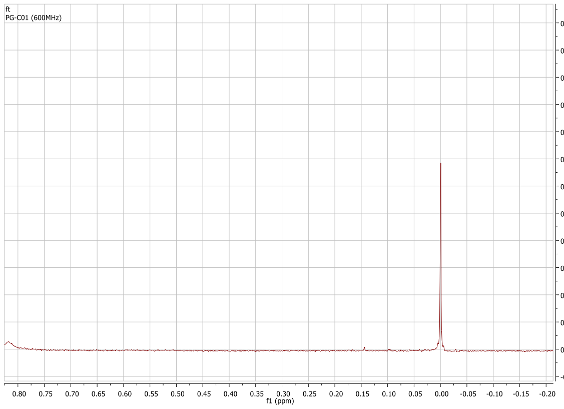
8. ^{1}H-NMR metabolomics using MestReNova

- ✓ Measure the sample. Note that at this step, the water suppression must be acquired (the operator should do this step).
- ✓ Collect the raw file of the measured samples.
- ✓ Open the 1H-NMR raw file of sample. Use the **JDF** file in the folder received from the operator.
- ✓ Adjust the ppm of TMS internal standard to 0 ppm.

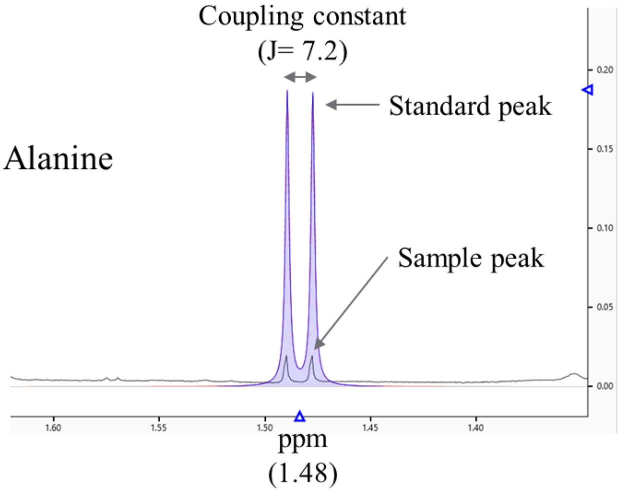
Before adjustment



After adjustment

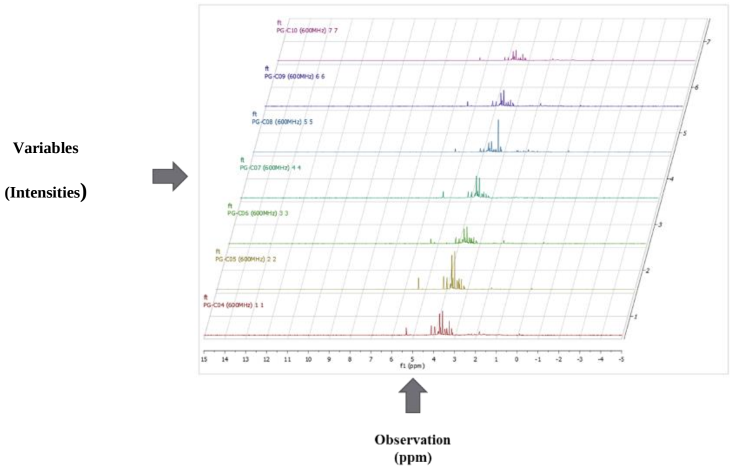


- ✓ The profiling of the metabolites is done by comparing the chemical shift (ppm) and coupling constant (Hzt) of the standards to that of the peak in the spectra. The standard library could be download at online library such as HMDB, BMRB.

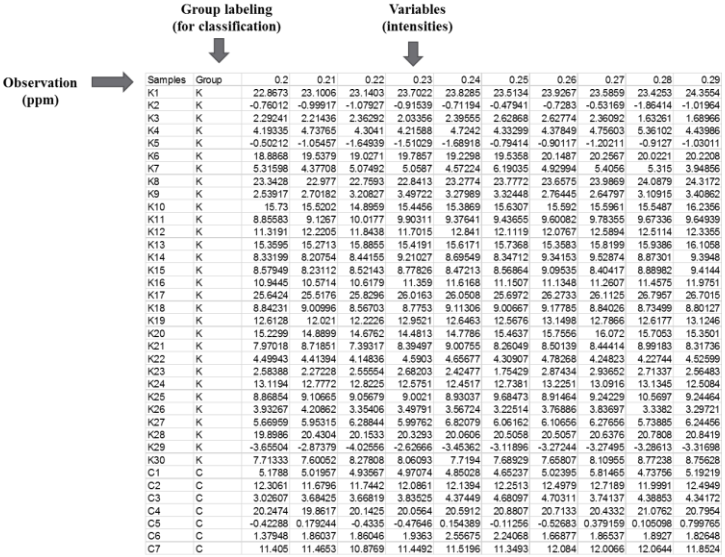


- ✓ After metabolite profiling, the data could be confirmed again by comparing the chemical shift (ppm) and coupling constant (Hzt) of the putatively identified compounds to standards prepared in our lab. All the detected metabolites should be listed in a table along with the characteristic chemical shift and coupling constant.

- ✓ Open all the raw spectra of the acquired samples and overlay all the spectra in one spectra. From this overlaid spectra, the data would be extracted and be ready for multivariate statistical analysis.



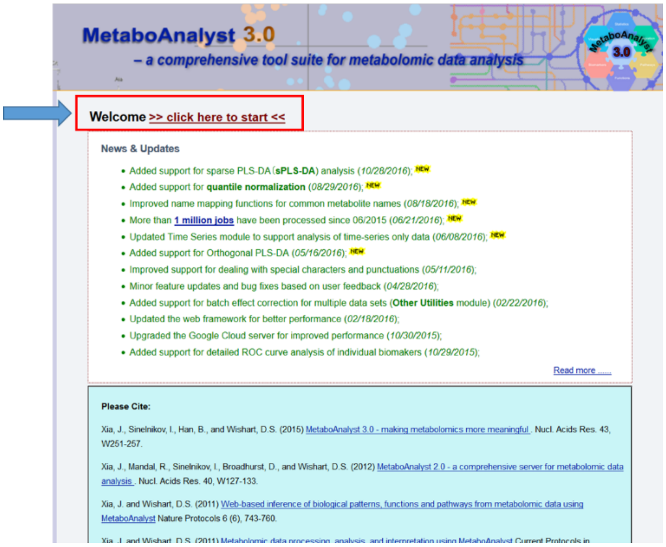
- ✓ For multivariate statistical analysis, data are extracted and formatted as follows


Group labeling
(for classification)
Variables
(intensities)
Observation
(ppm)

Samples	Group	0.2	0.21	0.22	0.23	0.24	0.25	0.26	0.27	0.28	0.29
K1	K	22.8673	23.1006	23.1403	23.7022	23.8285	23.5134	23.9267	23.5859	23.4253	24.3554
K2	K	-0.76012	-0.99917	-1.07927	-0.91539	-0.71194	-0.47941	-0.7283	-0.53169	-1.86414	-1.01964
K3	K	2.29241	2.21436	2.36292	2.03356	2.39555	2.62868	2.62774	2.36092	1.63261	1.68966
K4	K	4.19335	4.73765	4.3041	4.21588	4.7242	4.33299	4.37849	4.75603	5.36102	4.43986
K5	K	-0.50212	-1.05457	-1.64939	-1.51029	-1.68918	-0.79414	-0.90117	-1.20211	-0.9127	-1.03011
K6	K	18.8868	19.5379	19.0271	19.7857	19.2298	19.5358	20.1487	20.2567	20.0221	20.2208
K7	K	5.31598	4.37708	5.07492	5.0587	4.57224	6.19035	4.92994	5.4056	5.315	3.94856
K8	K	23.3428	22.977	22.7593	22.8413	23.2774	23.7772	23.6575	23.9869	24.0879	24.3172
K9	K	2.53917	2.70182	3.20827	3.49722	3.27989	3.32448	2.76445	2.64797	3.10915	3.40862
K10	K	15.73	15.5202	14.8959	15.4456	15.3869	15.6307	15.592	15.5961	15.5487	16.2356
K11	K	8.85583	9.1267	10.0177	9.90311	9.37641	9.43655	9.60082	9.78355	9.67336	9.64939
K12	K	11.3191	12.2205	11.8438	11.7015	12.841	12.1119	12.0767	12.5894	12.5114	12.3355
K13	K	15.3595	15.2713	15.8855	15.4191	15.6171	15.7368	15.3583	15.8199	15.9386	16.1058
K14	K	8.33199	8.20754	8.44155	9.21027	8.69549	8.34712	9.34153	9.52874	8.87301	9.3948
K15	K	8.57949	8.23112	8.52143	8.77826	8.47213	8.56864	9.09535	8.40417	8.88982	9.4144
K16	K	10.9445	10.5714	10.6179	11.359	11.6168	11.1507	11.1348	11.2607	11.4575	11.9751
K17	K	25.6424	25.5176	25.8296	26.0163	26.0508	25.6972	26.2733	26.1125	26.7957	26.7015
K18	K	8.84231	9.00996	8.56703	8.7753	9.11306	9.00667	9.17785	8.84026	8.73499	8.80127
K19	K	12.6128	12.021	12.2226	12.9521	12.6463	12.5676	13.1498	12.7866	12.6177	13.1246
K20	K	15.2299	14.8899	14.6762	14.4813	14.7786	15.4637	15.7556	16.072	15.7053	15.3501
K21	K	7.97018	8.71851	7.39317	8.39497	9.00755	8.26049	8.50139	8.44414	8.99183	8.31736
K22	K	4.49943	4.41394	4.14836	4.5903	4.65677	4.30907	4.78268	4.24823	4.22744	4.52599
K23	K	2.58388	2.27228	2.55554	2.68203	2.42477	1.75429	2.87434	2.93652	2.71337	2.56483
K24	K	13.1194	12.7772	12.8225	12.5751	12.4517	12.7381	13.2251	13.0916	13.1345	12.5084
K25	K	8.86854	9.10665	9.05679	9.0021	8.93037	9.68473	8.91464	9.24229	10.5697	9.24464
K26	K	3.93267	4.20862	3.35406	3.49791	3.56724	3.22514	3.76886	3.83697	3.3382	3.29721
K27	K	5.66959	5.95315	6.28844	5.99762	6.82079	6.06162	6.10656	6.27656	5.73885	6.24456
K28	K	19.8986	20.4304	20.1533	20.3293	20.0606	20.5058	20.5057	20.6376	20.7808	20.8419
K29	K	-3.65504	-2.87379	-4.02556	-2.62666	-3.45362	-3.11896	-3.27244	-3.27495	-3.28613	-3.31698
K30	K	7.71333	7.60052	8.27808	8.06093	7.7194	7.68929	7.65807	8.10955	8.77238	8.75628
C1	C	5.1788	5.01957	4.93567	4.97074	4.85028	4.65237	5.02395	5.81465	4.73756	5.19219
C2	C	12.3061	11.6796	11.7442	12.0861	12.1394	12.2513	12.4979	12.7189	11.9991	12.4949
C3	C	3.02607	3.68425	3.66819	3.83525	4.37449	4.68097	4.70311	3.74137	4.38853	4.34172
C4	C	20.2474	19.8617	20.1425	20.0564	20.5912	20.8807	20.7133	20.4332	21.0762	20.7954
C5	C	-0.42288	0.179244	-0.4335	-0.47646	0.154389	-0.11256	-0.52683	0.379159	0.105098	0.799765
C6	C	1.37948	1.86037	1.86046	1.9363	2.55675	2.24068	1.66877	1.86537	1.8927	1.82648
C7	C	11.405	11.4653	10.8769	11.4492	11.5196	11.3493	12.084	12.0066	12.0644	11.8524

9. Multivariate statistical analysis procedure

- ✓ Prepare the file with correct format and go the metaboanalyst website.



MetaboAnalyst 3.0

– a comprehensive tool suite for metabolomic data analysis

Welcome >> click here to start <<

News & Updates

- Added support for sparse PLS-DA (**sPLS-DA**) analysis (*10/28/2016*); NEW
- Added support for **quantile normalization** (*08/29/2016*); NEW
- Improved name mapping functions for common metabolite names (*08/18/2016*); NEW
- More than **1 million jobs** have been processed since 06/2015 (*06/21/2016*); NEW
- Updated Time Series module to support analysis of time-series only data (*06/08/2016*); NEW
- Added support for Orthogonal PLS-DA (*05/16/2016*); NEW
- Improved support for dealing with special characters and punctuations (*05/11/2016*);
- Minor feature updates and bug fixes based on user feedback (*04/28/2016*);
- Added support for batch effect correction for multiple data sets (**Other Utilities** module) (*02/22/2016*);
- Updated the web framework for better performance (*02/18/2016*);
- Upgraded the Google Cloud server for improved performance (*10/30/2015*);
- Added support for detailed ROC curve analysis of individual biomakers (*10/29/2015*);

Read more ......

Please Cite:

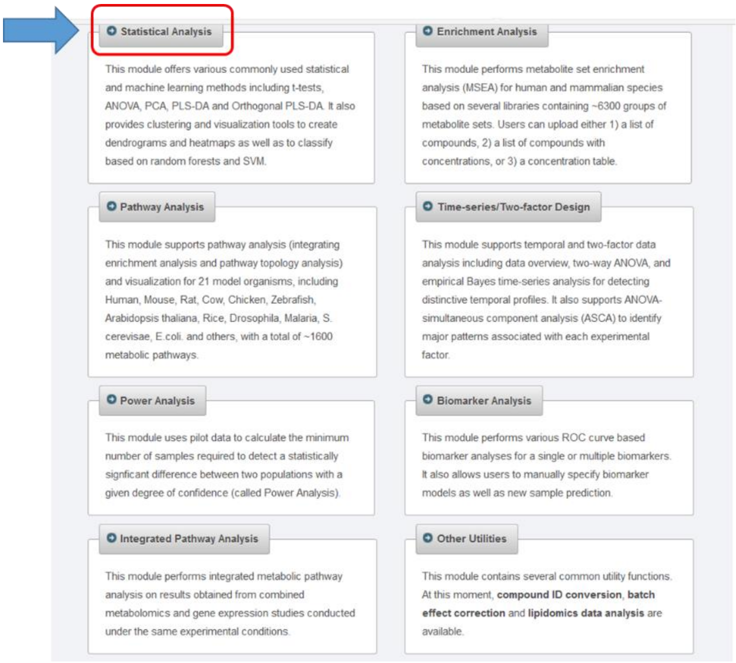
Xia, J., Sinelnikov, I., Han, B., and Wishart, D.S. (2015) MetaboAnalyst 3.0 - making metabolomics more meaningful . Nucl. Acids Res. 43, W251-257.

Xia, J., Mandal, R., Sinelnikov, I., Broadhurst, D., and Wishart, D.S. (2012) MetaboAnalyst 2.0 - a comprehensive server for metabolomic data analysis . Nucl. Acids Res. 40, W127-133.

Xia, J. and Wishart, D.S. (2011) Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst Nature Protocols 6 (6), 743-760.

Xia, J. and Wishart, D.S. (2011) Metabolomic data processing, analysis, and interpretation using MetaboAnalyst Current Protocols in

- ✓ Select Statistical analysis options



Statistical Analysis

This module offers various commonly used statistical and machine learning methods including t-tests, ANOVA, PCA, PLS-DA and Orthogonal PLS-DA. It also provides clustering and visualization tools to create dendrograms and heatmaps as well as to classify based on random forests and SVM.

Enrichment Analysis

This module performs metabolite set enrichment analysis (MSEA) for human and mammalian species based on several libraries containing ~6300 groups of metabolite sets. Users can upload either 1) a list of compounds, 2) a list of compounds with concentrations, or 3) a concentration table.

Pathway Analysis

This module supports pathway analysis (integrating enrichment analysis and pathway topology analysis) and visualization for 21 model organisms, including Human, Mouse, Rat, Cow, Chicken, Zebrafish, Arabidopsis thaliana, Rice, Drosophila, Malaria, S. cerevisae, E.coli. and others, with a total of ~1600 metabolic pathways.

Time-series/Two-factor Design

This module supports temporal and two-factor data analysis including data overview, two-way ANOVA, and empirical Bayes time-series analysis for detecting distinctive temporal profiles. It also supports ANOVA-simultaneous component analysis (ASCA) to identify major patterns associated with each experimental factor.

Power Analysis

This module uses pilot data to calculate the minimum number of samples required to detect a statistically signficant difference between two populations with a given degree of confidence (called Power Analysis).

Biomarker Analysis

This module performs various ROC curve based biomarker analyses for a single or multiple biomarkers. It also allows users to manually specify biomarker models as well as new sample prediction.

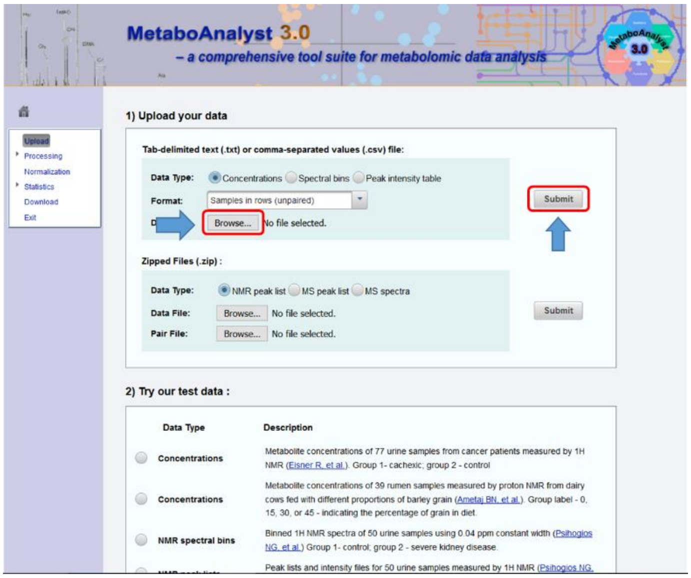
Integrated Pathway Analysis

This module performs integrated metabolic pathway analysis on results obtained from combined metabolomics and gene expression studies conducted under the same experimental conditions.

Other Utilities

This module contains several common utility functions. At this moment, **compound ID conversion**, **batch effect correction** and **lipidomics data analysis** are available.

- ✓ Click Browse to select the designated file and click submit to upload the file.



1) Upload your data

Tab-delimited text (.txt) or comma-separated values (.csv) file:

Data Type: ● Concentrations ○ Spectral bins ○ Peak intensity table

Format: Samples in rows (unpaired)

D Browse... No file selected.

Submit

Zipped Files (.zip) :

Data Type: ● NMR peak list ○ MS peak list ○ MS spectra

Data File: Browse... No file selected.

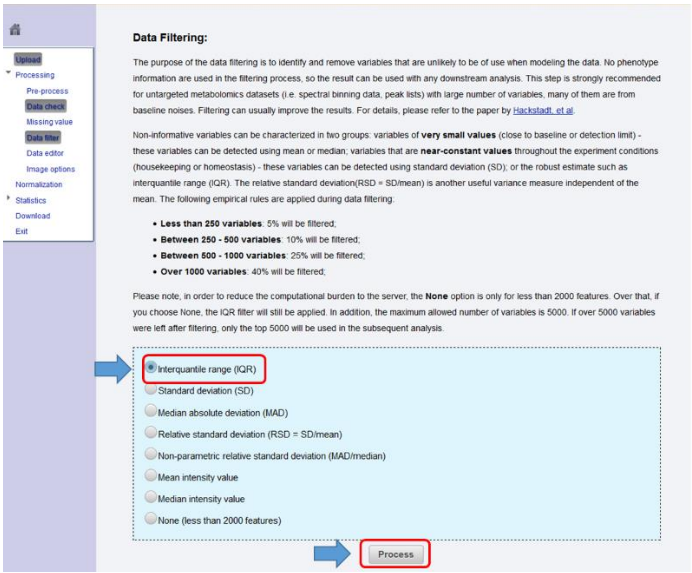
Pair File: Browse... No file selected.

Submit

2) Try our test data :

	Data Type	Description
○	Concentrations	Metabolite concentrations of 77 urine samples from cancer patients measured by 1H NMR (Eisner R. et al.). Group 1- cachexic; group 2 - control
○	Concentrations	Metabolite concentrations of 39 rumen samples measured by proton NMR from dairy cows fed with different proportions of barley grain (Ametaj BN. et al.). Group label - 0, 15, 30, or 45 - indicating the percentage of grain in diet.
○	NMR spectral bins	Binned 1H NMR spectra of 50 urine samples using 0.04 ppm constant width (Psihogios NG. et al.) Group 1- control; group 2 - severe kidney disease.
○		Peak lists and intensity files for 50 urine samples measured by 1H NMR (Psihogios NG.

- ✓ After successfully uploading the file, the first step is data filtering, the usual choice of data filtering would be Interquantile range (IQR). Click process to continue



Data Filtering:

The purpose of the data filtering is to identify and remove variables that are unlikely to be of use when modeling the data. No phenotype information are used in the filtering process, so the result can be used with any downstream analysis. This step is strongly recommended for untargeted metabolomics datasets (i.e. spectral binning data, peak lists) with large number of variables, many of them are from baseline noises. Filtering can usually improve the results. For details, please refer to the paper by Hackstadt, et al.

Non-informative variables can be characterized in two groups: variables of **very small values** (close to baseline or detection limit) - these variables can be detected using mean or median; variables that are **near-constant values** throughout the experiment conditions (housekeeping or homeostasis) - these variables can be detected using standard deviation (SD); or the robust estimate such as interquantile range (IQR). The relative standard deviation(RSD = SD/mean) is another useful variance measure independent of the mean. The following empirical rules are applied during data filtering:

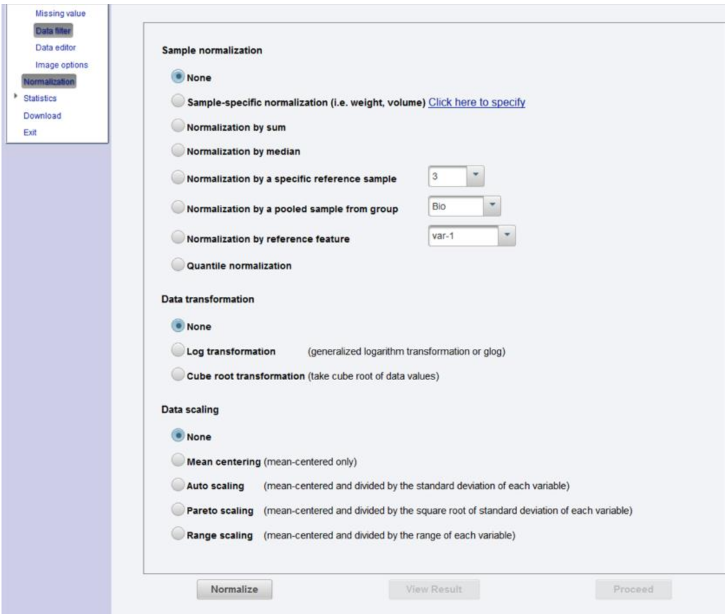
- **Less than 250 variables**: 5% will be filtered;
- **Between 250 - 500 variables**: 10% will be filtered;
- **Between 500 - 1000 variables**: 25% will be filtered;
- **Over 1000 variables**: 40% will be filtered;

Please note, in order to reduce the computational burden to the server, the **None** option is only for less than 2000 features. Over that, if you choose None, the IQR filter will still be applied. In addition, the maximum allowed number of variables is 5000. If over 5000 variables were left after filtering, only the top 5000 will be used in the subsequent analysis.

● Interquantile range (IQR)
○ Standard deviation (SD)
○ Median absolute deviation (MAD)
○ Relative standard deviation (RSD = SD/mean)
○ Non-parametric relative standard deviation (MAD/median)
○ Mean intensity value
○ Median intensity value
○ None (less than 2000 features)

Process

- ✓ Next is the step involving sample normalization, data transformation and data scaling. The choice in this step are essentially based on the types of samples being analyzed. After selecting all the necessary features, click proceed to continue.



Sample normalization

● None

○ Sample-specific normalization (i.e. weight, volume) Click here to specify

○ Normalization by sum

○ Normalization by median

○ Normalization by a specific reference sample 3

○ Normalization by a pooled sample from group Bio

○ Normalization by reference feature var-1

○ Quantile normalization

Data transformation

● None

○ Log transformation (generalized logarithm transformation or glog)

○ Cube root transformation (take cube root of data values)

Data scaling

● None

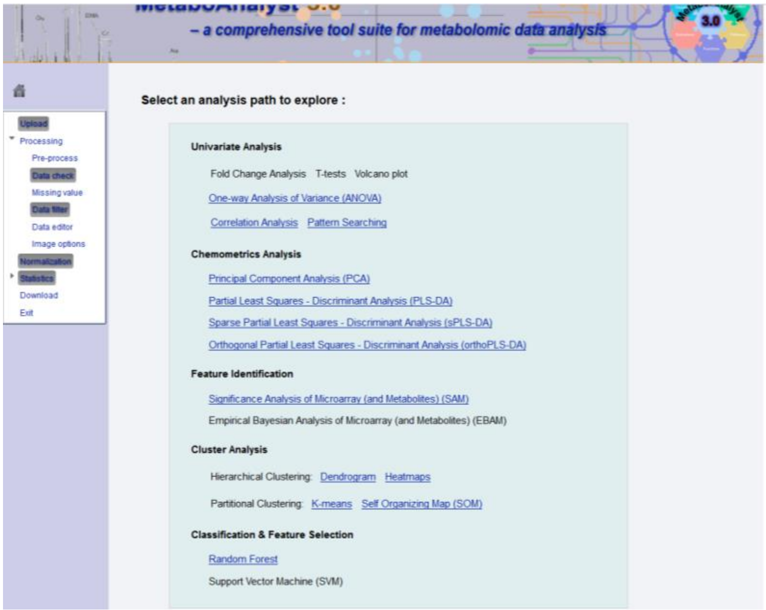
○ Mean centering (mean-centered only)

○ Auto scaling (mean-centered and divided by the standard deviation of each variable)

○ Pareto scaling (mean-centered and divided by the square root of standard deviation of each variable)

○ Range scaling (mean-centered and divided by the range of each variable)

- ✓ At this step, click on any feature of choice to continue.



Select an analysis path to explore :

Univariate Analysis

Fold Change Analysis T-tests Volcano plot

One-way Analysis of Variance (ANOVA)

Correlation Analysis Pattern Searching

Chemometrics Analysis

Principal Component Analysis (PCA)

Partial Least Squares - Discriminant Analysis (PLS-DA)

Sparse Partial Least Squares - Discriminant Analysis (sPLS-DA)

Orthogonal Partial Least Squares - Discriminant Analysis (orthoPLS-DA)

Feature Identification

Significance Analysis of Microarray (and Metabolites) (SAM)

Empirical Bayesian Analysis of Microarray (and Metabolites) (EBAM)

Cluster Analysis

Hierarchical Clustering: Dendrogram Heatmaps

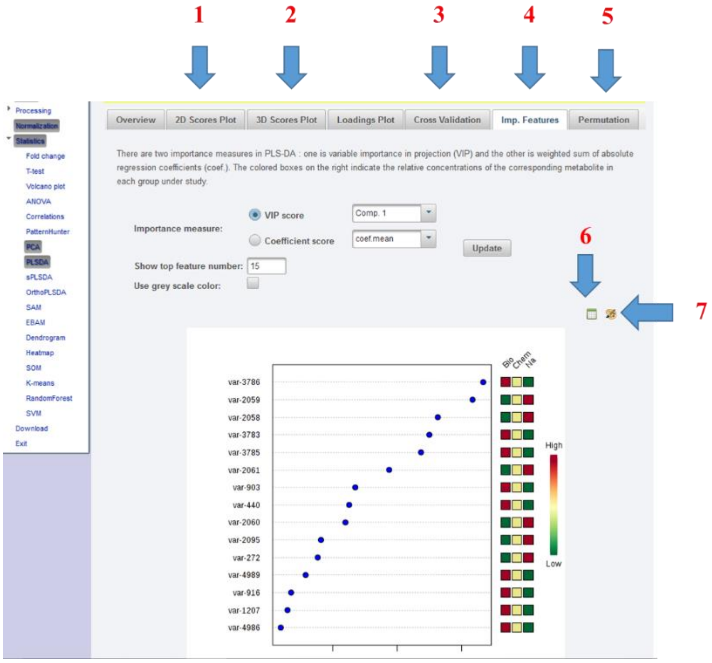
Partitional Clustering: K-means Self Organizing Map (SOM)

Classification & Feature Selection

Random Forest

Support Vector Machine (SVM)

For classification purpose, click PCA or PLS-DA depending on the efficiency of the model. After selecting the desired feature, for the important features that contribute significantly to separating groups on the model could be derived by.



There are two importance measures in PLS-DA : one is variable importance in projection (VIP) and the other is weighted sum of absolute regression coefficients (coef.). The colored boxes on the right indicate the relative concentrations of the corresponding metabolite in each group under study.

Importance measure: ● VIP score Comp. 1
○ Coefficient score coef.mean

Show top feature number: 15

Use grey scale color: □

1. Illustrating the model in two dimension
2. Illustrating the model in three dimension
3. Finding the optimal components employed in building the meaningful model
4. Examining which feature contributing significantly to building the meaning model
5. Validate the model
6. Extracting the table containing the important features.
7. Extracting the figures.

국문초록

천연물은 전통적으로나 역사적으로나 다양한 치료 목적을 가지고 사용되었다. 그러나, 천연물의 사용이 급증하면서 시장은 저급 재료를 섞은 불순품등으로부터 피해를 입고 있으며, 실제로 경제적인 측면 등으로 인하여 그런 사례가 늘고 있다. 따라서 천연물의 성분을 기반으로 한 정확한 감별법의 개발은 필수적이다. 이 연구에서는, multi-platform metabolomics 를 천연물의 품질관리에 적용하는 방법을 연구하였다. 먼저, UPLC-QTOFMS 에 기반한 대사체학을 이용하여 인삼으로써 가장 많이 사용되는 Panax 속 식물종인 *Panax ginseng*, *Panax vietnamensis*, *Panax notoginseng*, *Panax quinquefolius* 를 판별하였다. 더 나아가 ^{1}H-NMR 기반 대사체학을 이용하여 원산지가 다른 *Panax ginseng* 을 감별하였다. 원산지 감별법은 또한 원산지 혼입률이 각자 다른 시료의 원산지 구성 비율을 확인하는데 유용하다는 것을 확인하였고, 그를 판별하는 통계모델을 구축하였다. 이러한 대사체학에 기반한 종 및 원산지 판별법은 phylogenetic variation 에 결합될 수 있다는 것을 마지막으로 확인하였다. 결론적으로, 대사체학을 기반으로 한 한약재의 품질관리법은 적용가능성이 높아 실제로 사용가능하며, 시장에서의 불순품 혼입 우려를 덜어 실용적으로 사용가능하다는 것을 밝혔다.

주요어: 대사체학; 인삼; 진위감별; 계통학; QTOF-MS; ^{1}H-NMR

학번: 2013-22583