

# Discrimination between genetically identical peony roots from different regions of origin based on $^1\text{H}$ -nuclear magnetic resonance spectroscopy-based metabolomics: determination of the geographical origins and estimation of the mixing proportions of blended samples

Jung A Um · Young-Geun Choi · Dong-Kyu Lee · Yun Sun Lee · Chang Ju Lim ·  
Young A Youn · Hwa Dong Lee · Hi Jae Cho · Jeong Hill Park · Young Bae Seo ·  
Hsun-chih Kuo · Johan Lim · Tae-Jin Yang · Sung Won Kwon · Jeongmi Lee

Received: 9 May 2013 / Revised: 23 June 2013 / Accepted: 25 June 2013 / Published online: 16 July 2013  
© Springer-Verlag Berlin Heidelberg 2013

**Abstract** Sixty peony root training samples of the same age were collected from various regions in Korea and China, and their genetic diversity was investigated for 23 chloroplast intergenic space regions. All samples were genetically indistinguishable, indicating that the DNA-based techniques employed were not appropriate for determining the samples' regions of origin. In contrast,  $^1\text{H}$ -nuclear magnetic resonance ( $^1\text{H}$ -NMR) spectroscopy-based metabolomics coupled with multivariate statistical analysis revealed a clear difference between the metabolic profiles of the Korean and Chinese samples. Orthogonal projections on the latent structure-

discrimination analysis allowed the identification of potential metabolite markers, including  $\gamma$ -aminobutyric acid, arginine, alanine, paeoniflorin, and albiflorin, that could be useful for classifying the samples' regions of origin. The validity of the discrimination model was tested using the response permutation test and blind prediction test for internal and external validations, respectively. Metabolomic data of 21 blended samples consisting of Korean and Chinese samples mixed at various proportions were also acquired by  $^1\text{H}$ -NMR analysis. After data preprocessing which was designed to eliminate uncontrolled deviations in the spectral

**Electronic supplementary material** The online version of this article (doi:10.1007/s00216-013-7182-9) contains supplementary material, which is available to authorized users.

J. A. Um · D.-K. Lee · C. J. Lim · J. H. Park · S. W. Kwon (✉)  
College of Pharmacy, Seoul National University,  
Seoul 151-742, South Korea  
e-mail: swkwon@snu.ac.kr

Y. S. Lee · T.-J. Yang  
Department of Plant Science, Plant Genomics and Breeding  
Institute, and Research Institute for Agriculture and Life Sciences,  
College of Agriculture and Life Sciences, Seoul National  
University, Seoul 151-921, South Korea

Y.-G. Choi · J. Lim  
Department of Statistics, Seoul National University,  
Seoul 151-742, South Korea

Y. A. Youn · J. Lee (✉)  
School of Pharmacy, Sungkyunkwan University,  
Suwon 440-746, South Korea  
e-mail: jlee0610@skku.edu

H. D. Lee · H. J. Cho  
Korea Promotion Institute for Traditional  
Medicine Industry, Gyeongsan 712-260,  
South Korea

Y. B. Seo  
Department of Herbalogy, College of Oriental  
Medicine, Daejeon University, Daejeon 300-716,  
South Korea

H.-c. Kuo  
Department of Statistics, National Chengchi  
University, No.64, Sec.2, ZhiNan Road,  
Wenshan District, Taipei 11605, Taiwan

data between the testing and training sets, a new statistical procedure for estimating the mixing proportions of blended samples was established using the constrained least squares method for the first time. The predictive procedure exhibited relatively good predictability (adjusted  $R^2=0.7669$ ), and thus has the potential to be used in the quality control of peony root by providing correct indications for a sample's geographical origins.

**Keywords** Nuclear magnetic resonance spectroscopy · Metabolomics · Geographical origin · Chloroplast intergenic space analysis · High-resolution melting analysis · Constrained least squares method

## Introduction

Peony root, the root of *Paeonia lactiflora* Pall, is one of the most important and widely used traditional Chinese medicines (TCMs) in Asia, especially in Korea and China. It has traditionally been used in the treatment of a variety of diseases including rheumatoid arthritis, hepatitis, muscle cramping, and fever [1]. This root contains a number of bioactive compounds including phenols, triterpenoids, and monoterpene glycosides such as paeoniflorin, albiflorin, benzoylpaeoniflorin, and lactiflorin [2–4] that can possess anti-inflammatory [5], anti-hyperglycemic [6], and anti-hyperlipidemic [7] effects. In Korea, peony roots cultivated in either Korea or China are on the market and they are morphologically very similar regardless of their geographical origins. Considering that the origin is generally regarded as one of the major factors in the grading and the pricing of the TCMs, identification of the region of origin is deemed important in the TCM market to sustain socioeconomic balance. In addition, the geographic origins of medicinal plants may affect medicinal efficacy and quality [8–16]. Nonetheless, the authentication of medicinal plants in the current TCM market is largely carried out by morphological inspection rather than by standardized methods based on scientific measurements. Once the peony roots from different origins are mixed (intentionally or unintentionally) during distribution, morphological inspection is practically impossible for estimating the fractional composition of roots from different origins. Therefore, traditional authentication methods cannot reproducibly assure the quality of TCMs. This shortfall inherently poses the potential dangers of causing unintended pharmacological effects and of yielding incorrect value estimation upon distribution.

As a measure of quality control for the peony root, the Korean Pharmacopoeia [17] suggests that only two monoterpene glycosides, paeoniflorin, and albiflorin, should be quantified. According to Wang et al., however, the major bioactive compounds of the peony root, including paeoniflorin and albiflorin significantly varied depending on the peony root

sample, and it was speculated that the variance was most likely due to “processing procedure and habitat variation” [18]. If the genetic makeup of the TCM products is different, a genetic approach is used as the method of choice for differentiating plants when possible [19–21]. When the plants are genetically similar, however, the discovery of gene markers becomes more challenging and genetic markers do not allow researchers to distinguish between plants of the same species cultivated in different environments.

Metabolomics has been used extensively in plant research to monitor the physiological responses of plants to external stresses and has also been applied to discriminate between plants of similar or identical genotypes [22] as well as to classify medicinal plants of similar species for quality control purposes [23]. Our research group has been developing an  $^1\text{H}$ -nuclear magnetic resonance ( $^1\text{H}$ -NMR) spectroscopy-based metabolomics methods to efficiently distinguish between medicinal herbs of different geographical origins [16] because the sample analysis using this method is simple, reproducible, rapid, and allows for the simultaneous detection of primary and secondary metabolites. Consequently, this method has better predictability and stability than other techniques [23–27].

In this study, peony roots were first collected from various regions in Korea and China and the two DNA-based techniques, chloroplast intergenic space (CIS) analysis, and high-resolution melting (HRM) analysis, revealed they were genetically indistinguishable, i.e., these techniques could not distinguish between samples from different locations. Therefore, as described in this paper, a NMR-based metabolomics approach was developed and used as a simple yet efficient tool for the discrimination between peony root samples of different geographical origins. This paper also describes the first application of a constrained least-squares method for estimating the proportion of blended samples from each of the two geographical origins.

## Materials and methods

### Sample collection and preparation

On the TCM market in Korea, peony root is mostly distributed as the dried roots of *P. lactiflora* Pall cultivated either in Korea or China. In both countries, the peony root is usually harvested between mid-September and late November at the age of 4 years. Thus, in this study, roots of 4-year-old *P. lactiflora* Pall were directly collected from ten cultivation locations in Korea and seven cultivation locations in China during the regular harvesting season to assure the authenticity of the geographical origins of the samples and to minimize any unwanted introduction of variations arising from factors such as the plant age and the harvesting season.

Chosen cultivation locations that were in the same country were relatively far from one another to represent a wide diversity of cultivation environments. Detailed information about the collection locations is provided in Fig. S1 in the Electronic supplementary material (ESM) and Table 1. Thirty samples (K1–30) were collected from the areas of Yeongju, Andong, Uiseong, Yeongcheon, Imsil, Sancheong, Jeongeup, Naju, Muan, and Hwasun in Korea, and 30 samples (C1–30) were collected from the areas of Chongqing, Yuncheng, Anguo, Heze, Neimenggu, Zhumadian, and Bozhou in China. Between 1 and 16 samples were collected from each location. In addition to the 60 training samples described above, four blind samples (TJY01 and TJY03 from Korea; TJY02 and TJY04 from China) were collected separately to assess the predictive capabilities of the discrimination model. Since knowledge of the cultivation origin was crucial in this experimental design, samples were collected in person from the actual cultivation locations indicated in Table 1, and the morphology of each was examined by professionals (Dr. Young-Bae Seo from the College of Oriental Medicine, Daejun University, Daejun, Korea and Dr. Wan-Kyun Hwang from the College of Pharmacy, ChoongAng University, Seoul, Korea) as is the traditional authentication method.

Consistent with the treatment before being sold in the market, the freshly collected samples were dried in the dark. After they were pulverized using an electric blender, powdered samples

were stored in 50 mL conical tubes at  $-20^{\circ}\text{C}$  until use. Powders passing between 125 and 300  $\mu\text{m}$  sieves were used for the DNA analysis and the metabolic profiling.

Korean and Chinese samples were mixed at various proportions to produce seven blended samples. Before blending, all of the 30 powdered samples from Korea were mixed in the same proportion to produce a representative Korean sample, and the same procedure was performed for the Chinese samples. Then, the blended samples were prepared in triplicate by mixing the two representative Korean and Chinese samples at seven different ratios: 0 (sample Nos. 1–3), 10 (sample Nos. 4–6), 25 (sample Nos. 7–9), 50 (sample Nos. 10–12), 75 (sample Nos. 13–15), 90 (sample Nos. 16–18), and 100 % (sample Nos. 19–21) of the Korean sample.

#### DNA extraction and PCR analysis

Total DNA was extracted from the powdered samples using the modified cetyl trimethylammonium bromide method as previously described [28]. A total of 48 primer pairs [29] were applied to amplify the intergenic spaces of the chloroplast genome (Table S1 in the ESM). Polymerase chain reaction (PCR) was performed in a 20- $\mu\text{L}$  volume containing 20 ng DNA, 20 pmol of primer pairs, 2.5 mM of dNTPs, and 0.4 unit of *Taq* polymerase (Vivagen, Seongnam, Korea). PCR amplification was performed as follows: 5 min at  $94^{\circ}\text{C}$ ; 38 cycles of  $95^{\circ}\text{C}$  for 30 s,  $55^{\circ}\text{C}$  for 30 s, and  $72^{\circ}\text{C}$  for 20 s and a final extension at  $72^{\circ}\text{C}$  for 5 min. The PCR products were separated by electrophoresis on a 1 % agarose gel. Polyacrylamide gel electrophoresis (PAGE) on a 9 % gel was used for identifying InDel polymorphisms. HRM analysis was carried out using a LightCycler 480 (Roche Applied Science, Mannheim, Germany) to identify nucleotide sequence variation between collections.

#### Metabolite extraction

One hundred milligrams of powdered material was vortexed in 1.5 mL of a 1:1 mixture of  $\text{CD}_3\text{OD}$  and a pH 6.1 phosphate buffer composed of 4 mM sodium phosphate monobasic and 2 mM sodium phosphate dibasic in  $\text{D}_2\text{O}$  and was then extracted by ultrasonication at room temperature for 15 min. 3-(trimethylsilyl)propionic-2,2,3,3- $\text{d}_4$  acid (TMSP; 0.025 % ( $w/v$ )) was used as the internal chemical shift standard. The extracted sample was centrifuged at  $13,000\times g$  for 10 min, followed by filtration of the supernatant through a cellulose membrane (0.45  $\mu\text{m}$ ). Six hundred microliters of the filtrate was transferred to a 5-mm NMR tube for NMR analysis.

#### NMR spectroscopy

Spectrum measurement of the 60 training samples and the 4 blind samples was performed at room temperature using a

**Table 1** Geographical origins of *Paeonia lactiflora* collected in Korea and China

| Country | Location (City, Province) | Number of samples (sample name)   |
|---------|---------------------------|---|
| Korea   | Yeongju, Gyeongbuk        | 4 (K1, K2, K3, and K4)  |
|         | Andong, Gyeongbuk         | 3 (K5, K6, and K10)   |
|         | Uiseong, Gyeongbuk        | 3 (K7, K8, and K9)  |
|         | Yeongcheon, Gyeongbuk     | 3 (K11, K12, and K13)   |
|         | Imsil, Jeonbuk            | 3 (K14, K15, and K16)   |
|         | Sancheong, Gyeongnam      | 3 (K17, K18, and K19)   |
|         | Jeongeup, Jeonbuk         | 3 (K20, K21, and K22)   |
|         | Naju, Jeonnam             | 2 (K23 and K24)   |
|         | Muan, Jeonnam             | 3 (K25, K26, and K27)   |
|         | Hwasun, Jeonnam           | 3 (K28, K29, and K30)   |
| China   | Chongqing                 | 4 (C1, C2, C3, and C4)  |
|         | Yuncheng, Shanxi Sheng    | 6 (C5, C6, C10, C11, C12, and C13)  |
|         | Anguo, Hebei Sheng        | 1 (C7)  |
|         | Heze, Shandong Sheng      | 1 (C8)  |
|         | Neimenggu                 | 1 (C9)  |
|         | Zhumadian, Henan Sheng    | 1 (C14)   |
|         | Bozhou, Anhui Sheng       | 16 (C15, C16, C17, C18, C19, C20, C21, C22, C23, C24, C25, C26, C27, C28, C29, and C30) |
|         |                           |   |
|         |                           |   |
|         |                           |   |

JEOL NMR ECA 500 spectrometer, equipped with a TH5 probe (JEOL, Tokyo, Japan) to establish a model for origin discrimination. An AVANCE 500 FT-NMR spectrometer (Bruker biospin, Rheinstetten, Germany) was utilized for the 21 blended samples. Acquisition parameters were a 5.7- $\mu$ s (45°) pulse, 9,384.0-Hz spectral width, number of scans equal to 8, and 5-s relaxation delay with 64 transients collected into 32 k data points. Undesired signal caused by the residual water was removed by presaturation during the relaxation delay. The data were Fourier transformed using a line-broadening factor of 0.2 Hz, phase and baseline were corrected, and the data were then calibrated by shifting the TMSP signal to 0.0 ppm using Delta software version 4.3.6 (JEOL, Tokyo, Japan). One-dimensional NMR spectral peaks were assigned using Chenomx NMR software (version 7.1, Chenomx Inc.) by referring to the chemical shifts and the coupling constants in previously reported papers [30–33] or in a free database [34]. Identities of selected peaks were further confirmed by two-dimensional correlation methods including total correlation spectroscopy (TOCSY), heteronuclear single quantum coherence (HSQC), and heteronuclear multiple bond coherence (HMBC). Two-dimensional NMR spectra were measured on the same spectrometer that was used for one-dimensional measurements.

#### NMR data processing and statistical analysis

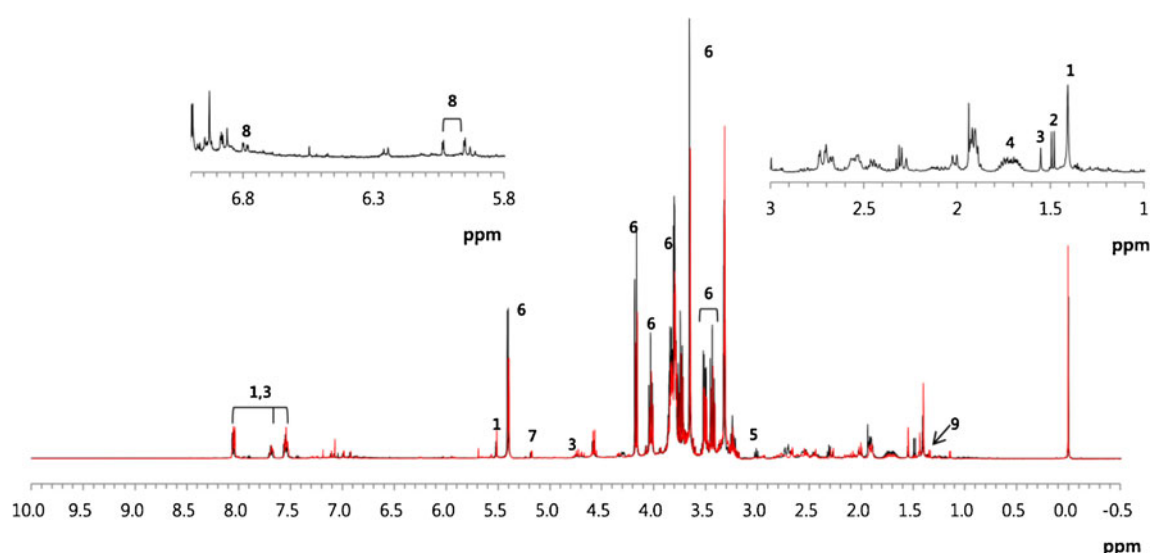
$^1\text{H}$ -NMR spectra were converted to the ASCII format for data processing and were segmented into 241 bins between 0.2 and 10 ppm with a bin width of 0.04 ppm. Bins corresponding to the residual methanol peaks between 3.29 and 3.37 ppm were removed, producing a total of 239 bins. The integral values were normalized against the TMSP signal. The resulting data sets were imported into SIMCA-P software (version 12.0,

Umetrics, Umea, Sweden) for multivariate statistical calculations and plotting. Univariate statistical analysis was conducted using SPSS (version 12.0, SPSS, Chicago, IL).

## Results and discussion

### DNA profiling of the collected peony samples

Random amplified polymorphic DNA analysis was carried out to classify peony cultivars of similar morphologies which may or may not be of the same species [35]. Microsatellite markers were developed to investigate the genetic diversity of the peony species but were not designed to determine the location of cultivation [36]. Nuclear genome-based markers, in which a number of variations can be found, are a weak tool for classification because too many variations exist within populations [29]. In contrast, chloroplast genomes are generally highly conserved among plants but a considerable amount of nucleotide variation was identified in the intergenic regions in the chloroplast genomes despite the conservation of genic regions. Therefore, CISs have been used for the systematic study of various plants. These regions provide solid information for distinguishing between not only different species but also different samplings within the same species [37, 38]. Consequently, 48 CISs were examined to identify polymorphisms among the 60 samples and a total of 23 CISs were successfully amplified. None of the 23 CIS regions were polymorphic among the four representative samples, two Korean samples (K12 and K21) and two Chinese samples (C15 and C20) (Fig. S2a in the ESM), and the CIS region of *rpl23-trnI*(CAU) did not show any diversity among the 60 samples (Fig. S2b in the ESM).

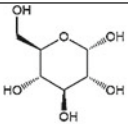
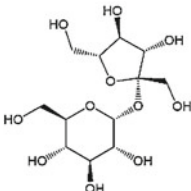
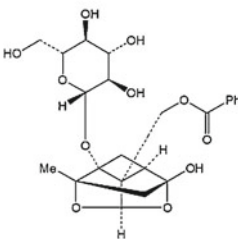
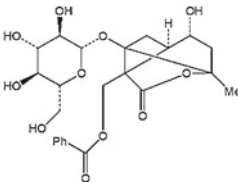
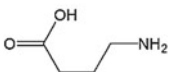
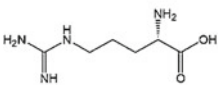
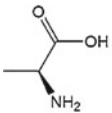
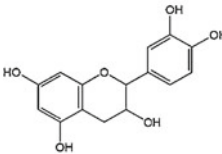
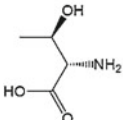


**Fig. 1** Overlaid representative  $^1\text{H}$ -NMR spectra of the samples collected from Korea (black K18) and China (red C26). 1 Peoniflorin, 2 alanine, 3 albiflorin, 4 arginine, 5 GABA, 6 sucrose, 7 glucose, 8 (+)-catechin, 9 threonine

These findings indicated very narrow genetic diversity within plants of the same species of *P. lactiflora* from different geographic locations.

HRM analysis has been applied to detect SNPs and small InDels based on the melting curves of PCR amplicons from nuclear genomes in plant species [39, 40]. In the current

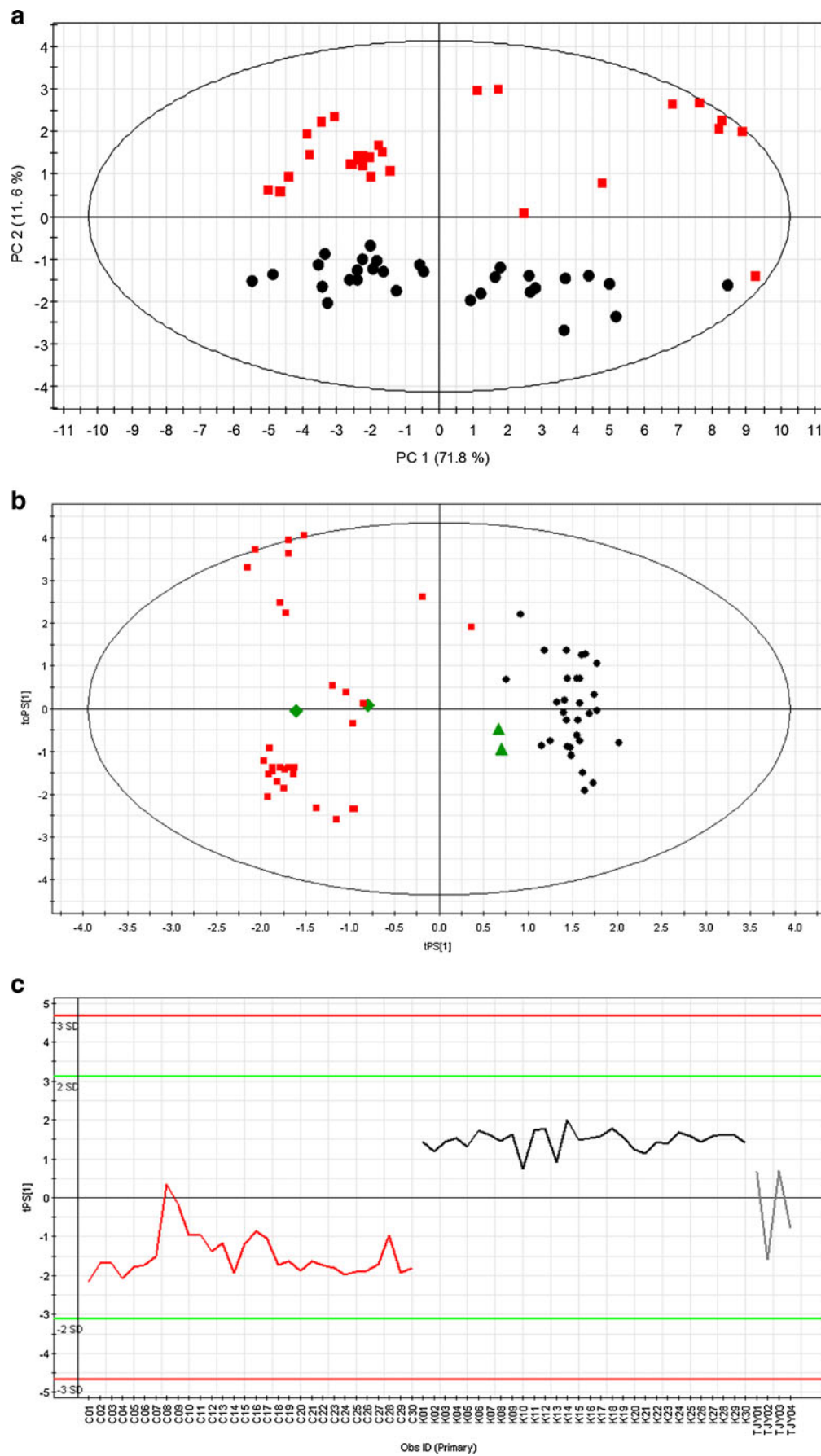
**Table 2**  $^1\text{H}$ -NMR chemical shifts (in parts per million), multiplicity, and coupling constants (in Hertz) of assigned metabolites

| Metabolite                               | Chemical Structure  | Assignment  |
|--|---|---|
| $\alpha$ -Glucose <sup>a</sup>           |    | 5.41 (d, $J=3.7$ , H-1)   |
| Sucrose <sup>a</sup>                     |    | 5.19 (d, $J=3.7$ , H-1)   |
| Paeoniflorin <sup>b</sup>                |   | 7.44 (t, $J=8.0$ , H-3'', 5''), 7.68 (tt, $J=1.7$ , 7.5, H-4''), 8.05 (dd, $J=1.4$ , 8.3, H-2'', 6''), 1.42 (s, H-10) |
| Albiflorin <sup>b</sup>                  |  | 7.44 (t, $J=8.0$ , H-3'', 5''), 7.68 (tt, $J=1.7$ , 7.5, H-4''), 8.05 (dd, $J=1.4$ , 8.3, H-2'', 6''), 1.54 (s, H-10) |
| $\gamma$ -Aminobutyric acid <sup>a</sup> |  | 2.32 (d, $J=7.5$ , H-2)   |
| Arginine <sup>a</sup>                    |  | 1.91 (m, $J=2.9$ , H-16, 17)  |
| Alanine <sup>a</sup>                     |  | 1.48 (d, $J=7.2$ , H-3)   |
| (+)-Catechin <sup>b</sup>                |  | 6.05 (d, $J=2.4$ , H-8)   |
| Threonine <sup>a</sup>                   |  | 1.32 (d, $J=6.6$ , H-5)   |

<sup>a</sup> Assigned using standard data in BioMagResBank (<http://www.bmrb.wisc.edu/>)

<sup>b</sup> Assigned by comparison with published data





**Fig. 2** Score plots of PCA (a) and OPLS-DA (b) for the 60 samples and the origin prediction (c). Black circles for the Korean samples; red squares for the Chinese samples; green triangles for the blind test samples from Korea; green diamonds for the blind test samples from China; black line for the Korean samples; red line for the Chinese samples; gray line for the blind test samples (TJY01 and TJY03 from Korea; TJY02 and TJY04 from China)

study, HRM analysis was performed on the amplified CIS regions for *rpl23-trnI*(CAU) (Fig. S2c in the ESM). The HRM profiles of the 60 samples revealed no variations, however, indicating a lack of an association in sequence variations. These results suggest that discriminating between samples of peony root grown in different geographic locations is not feasible using DNA-based methods.

#### <sup>1</sup>H-NMR spectrum measurement and metabolite identification

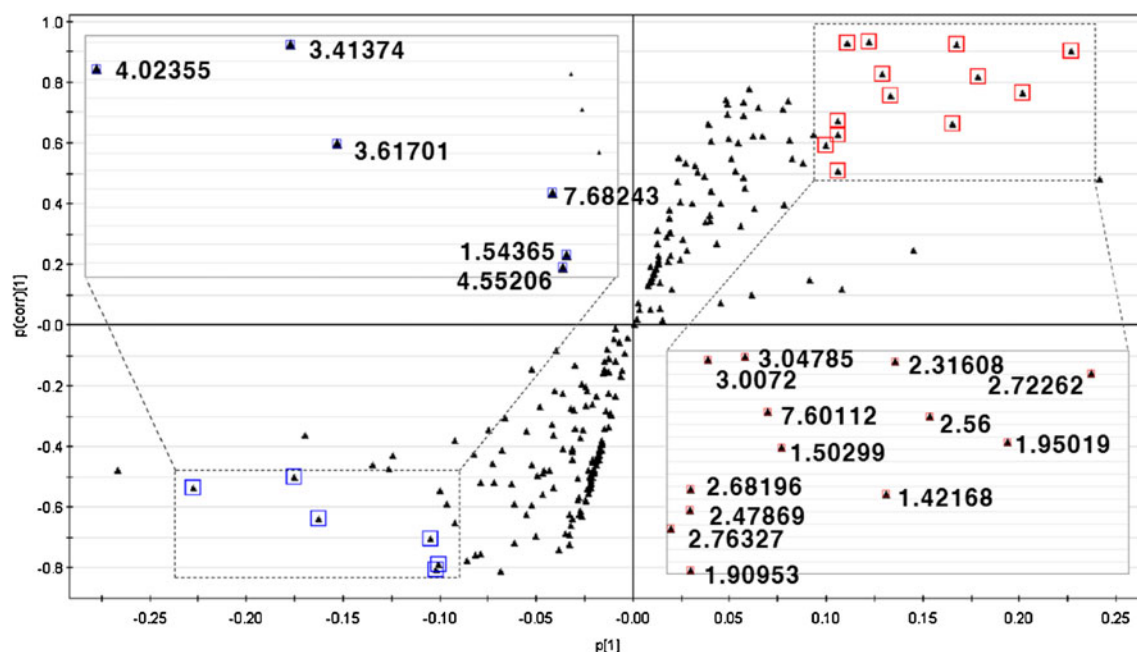
Metabolic profiles of the aqueous methanol extracts of the peony root samples were collected using <sup>1</sup>H-NMR analysis. Representative one-dimensional <sup>1</sup>H-NMR spectra of the peony root samples from Korea and China are shown in Fig. 1. They display dominant signals in the carbohydrate (3.0–5.0 ppm), the aliphatic/organic acid (0.5–3.0 ppm), and the aromatic (6.0–8.0 ppm) regions. The assignment of the identified metabolites is shown in Table 2. Many peaks originating from carbohydrates were observed in the region between 3.1 and 4.3 ppm, and the characteristic signals corresponding to sucrose and α-glucose appeared at 5.19 (d, J=3.7 Hz, H-1) and 5.41 (d, J=3.7 Hz, H-1) ppm, respectively. Identified aliphatic/organic acids included γ-aminobutyric acid (GABA;

2.32 ppm), arginine (1.91 ppm), alanine (1.48 ppm), and threonine (1.32 ppm). In the aromatic region, peaks of (+)-catechin (6.05 ppm) were found and the peaks corresponding to paeoniflorin and albiflorin, which are the two representative metabolites of *P. lactiflora*, were observed at 7.44 (t, J=8.0 Hz, H-3", 5"), 7.68 (tt, J=1.7, 7.5 Hz, H-4"), and 8.05 (dd, J=1.4, 8.3 Hz, H-2", 6") ppm. The CH<sub>3</sub> singlet signal of paeoniflorin was found at 1.42 ppm, while that of albiflorin was shifted downfield (1.54 ppm), and the distinctive acetal proton signal of paeoniflorin was detected at 5.53 ppm.

#### Multivariate statistical analysis for the differentiation between geographical origins

The metabolomics data obtained by NMR analysis were first subjected to PCA. The PCA score plot and loading plot from the <sup>1</sup>H-NMR spectra of *P. lactiflora* are shown in Fig. S3 in the ESM and Fig. 2a, respectively. Although PCA was used to identify outliers and to indicate data patterns and trends as an unsupervised classification method, the score plot projected into two dimensions showed a clear separation between Korean and Chinese samples by PC2 (11.6 %). The one exception to this trend was sample C8, which was collected from China but segregated with the Korean samples (Fig. 2a). Based on the PCA loading plot of PC2 (Fig. S3 in the ESM), aromatic compounds and sugars that were present at high levels in the Chinese samples and amino acids, which showed higher intensities in the Korean samples, facilitated the differentiation between the samples.

PCA is not considered a proper mechanism for identifying potential markers or predicting unknowns because of the



**Fig. 3** S-plot constructed from the OPLS-DA model. Cutoff value of  $p \geq |0.1|$  and correlation value of  $p(\text{corr}) \geq |0.5|$  are marked

inherently undefined dependent variables. Therefore, orthogonal projections on latent structure-discriminant analysis (OPLS-DA) was performed after the PCA analysis. In OPLS-DA, important variables contributing to group separation based on the dependent variables are maximized, allowing for better discrimination and elucidation of marker candidates [41]. Consistent with the results of the PCA analysis, the score plot of the OPLS-DA displayed a clear separation pattern between the samples of Chinese and Korean origins with the exception of sample C8 (Fig. 2b). The model was established using one predictive and two orthogonal variations and had an acceptable goodness of fit,  $R_Y^2$ , of 92.2 % and predictive ability,  $Q^2$ , of 89.7 %. The overall variation of the independent variables explained by the model in  $R_X^2$  was 87.5 %. Only 10.7 % of that variation was correlated to the dependent variables, whereas the residual 76.8 % was systemically structured and uncorrelated to response. Overall, OPLS-DA was more appropriate than PCA for discriminating between peony root samples of different geographic origins.

Figure 2b shows a clear separation of the samples into two groups according to the cultivation origin, with the exception of C8. Korean samples were relatively concentrated compared with Chinese samples, which were dispersed over a wide area. When the individual samples were labeled in the score plots based on a smaller regional scale such as city level (Fig. S4 in the ESM), samples within the same province or nearby provinces in the same country showed a tendency to cluster. The sample C8 collected from Heze, Shandong Sheng was found far from the other Chinese samples in both the PCA and the OPLS-DA score plots (Fig. 2). Instead, the sample C8 plotted close to the Korean samples. The misclassification of C8 suggests that its metabolomic profile may be more similar to those of the Korean samples instead of the Chinese samples. Indeed, Heze is close to Korea in terms of geographical latitude compared with the districts from which the other Chinese samples were collected (Fig. S1 in the ESM). Thus, the cultivation environment in Shandong Sheng is likely to be similar to that of Korea. Metabolomic analysis of more specimens from the Shandong Sheng area and comparison of the cultivation environments is necessary to test this hypothesis.

#### Validation of the discrimination model

After establishing the discrimination model, the response permutation test (Y-scrambling) was performed for internal validation. In Y-scrambling,  $y$  variables were randomly shuffled, and the models were rebuilt. These new models were compared with the original models to test the possibility that the original model arose by chance. Both the  $Q^2$  and  $R^2$  parameters decreased substantially after 200 rounds of permutation, and the  $y$ -intercept of the  $Q^2$  regression line was  $-0.111$ , indicating that the model was not over-fit (Fig. S5 in the ESM).

Of the 60 training samples, C8 behaved differently from the other Chinese samples in the predicted plots with a prediction score above 0, indicating it was predicted to be of Korean origin (Fig. 2c). The overall classification rate was 98.33 %. In addition to the statistical validation, four blind samples (TJY01 and TJY03 from Korea; TJY02 and TJY04 from China) were analyzed to further assess the predictive capabilities of the model. As displayed in Fig. 2c, the geographical origins of the blind samples were correctly predicted.

#### Determination of potential marker metabolites

Putative marker metabolites, which might contribute to the differentiation between samples of different geographical origins, were extracted from the S-plot constructed from the OPLS-DA model (Fig. 3). In the S-plot, the covariance  $p$  and correlation  $p(\text{corr})$  variables are displayed, and the variables with higher absolute values on both the horizontal and vertical axes contribute significantly to group separation with a high reliability. In this study, 22 variables showing  $p \geq |0.1|$  and  $p(\text{corr}) \geq |0.5|$ , which are marked in blue or red squares in the dotted rectangles in Fig. S6 in the ESM, were selected as marker candidates. Their identities were confirmed by 2D NMR analysis including TOCSY, HSQC, and HMBC as shown in Fig. S6 in the ESM.

The marker candidates were further analyzed by the univariate statistical analysis,  $t$  test with  $\alpha=0.0002$ , which was obtained by the Bonferroni correction ( $0.05/239$ ) for the multiple comparisons. This strategy made the statistical analysis more conservative but more reliable. It was found that each marker candidate was significantly different between the groups ( $p < 0.0002$ ). Based on the identified markers, Korean samples were characterized by peaks between 1 and 3 ppm, including the signals from GABA, arginine, alanine, and the  $\text{CH}_3$  singlet signal of paeoniflorin (1.42 ppm). In contrast,

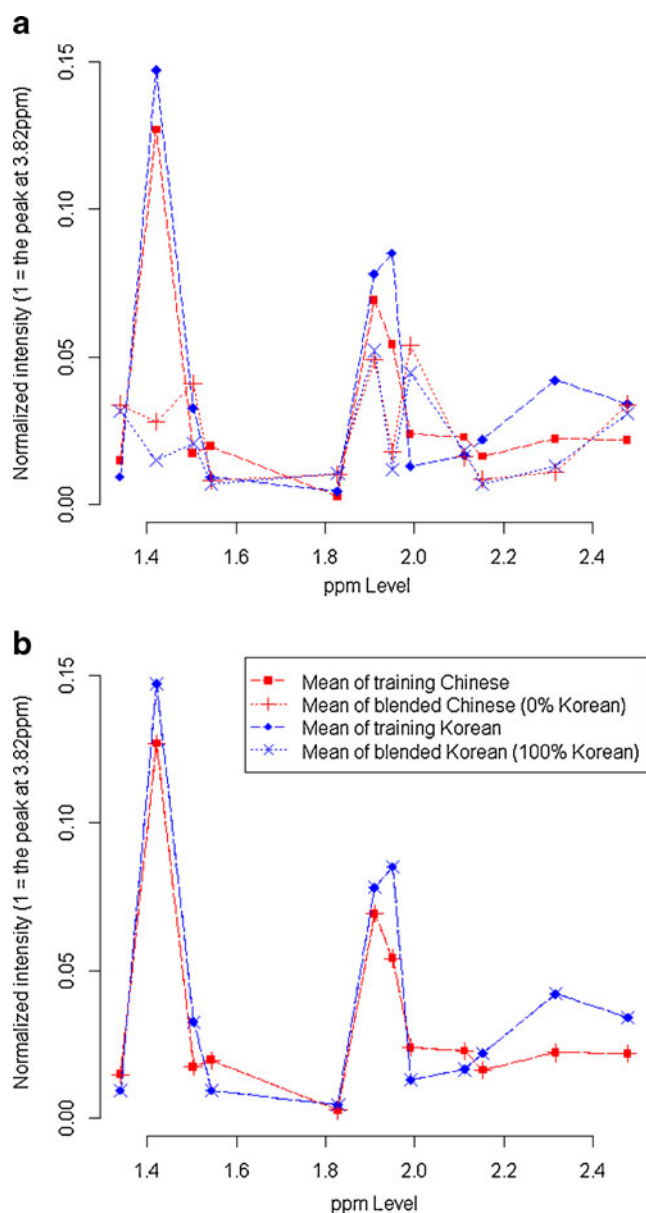
**Table 3** Relative quantification of five marker metabolites based on their peak areas in NMR spectra

| Metabolite                  | Relative quantification |             |                        |
|-----------------------------|-------------------------|-------------|------------------------|
|                             | China                   | Korea       | $p$ value <sup>a</sup> |
| Paeoniflorin <sup>b</sup>   | 0.820±0.079             | 0.997±0.029 | 9.98 E-07              |
| Albiflorin <sup>b</sup>     | 0.118±0.061             | 0.060±0.114 | 8.95 E-11              |
| $\gamma$ -Aminobutyric acid | 0.147±0.041             | 0.287±0.016 | 4.80 E-20              |
| Arginine                    | 0.345±0.042             | 0.586±0.033 | 1.23 E-10              |
| Alanine                     | 0.115±0.049             | 0.227±0.052 | 1.93 E-11              |

<sup>a</sup> Determined by  $t$  test

<sup>b</sup> Quantified using the areas of non-overlapped peaks (paeoniflorin, 1.42 ppm; albiflorin, 1.54 ppm)

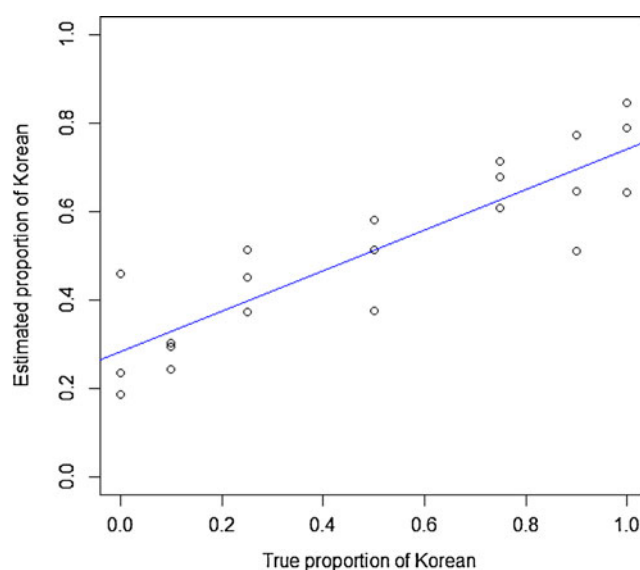




**Fig. 4** Mean spectra before (a) and after (b) data preprocessing. Dashed lines for the mean ( $\bar{\mu}$ ) of the spectra of the training data set; dotted lines for the mean ( $\bar{v}$ ) of the spectra of blended (pure) sample; blue lines for Korean samples; red lines for Chinese samples

Chinese samples were characterized by peaks in the  $\text{CH}_3$  singlet peak of albiflorin (1.54 ppm). Their relative quantification is presented in Table 3.

In plants, GABA, which is produced by decarboxylation of glutamate, is regarded not only as a signaling molecule but also as a metabolite that plays a role in the response to biotic or abiotic stresses and in the carbon-nitrogen balance [42]. Elevated levels of GABA, as found in Korean samples in the present study, have been found to be associated with defense against insects and protection against oxidative and osmotic stresses [42], some of which might have occurred more



**Fig. 5** Predictive model constructed using the constrained least squares method displaying the true proportion ( $\pi^{(i)}$ ) versus the estimated proportion ( $\hat{\pi}^{(i)}$ ) for  $i=1,2,\dots,21$ . Adjusted  $R^2$  was 0.7669

prevalently in Korea than China. Levels of the two amino acid markers, arginine and alanine were significantly higher in Korean samples than in those of Chinese origin. Previously, exposure to more light [43] and to temperature stress (heat shock or cold shock) [44] were reported to be associated with upregulation of amino acid synthesis, implying that the differences in the amount of sunshine and temperatures between Korea and China might have caused the different levels of the amino acid marker metabolites. While the concentration of paeoniflorin, which exhibits a neuroprotective effect [45], was higher in the Korean samples, the concentration of albiflorin, which possesses anti-oxidative activity [46], was higher in the Chinese samples. Although no specific reports are available on the correlation of these secondary metabolites with the environmental conditions, production of secondary metabolites is known to be influenced by a number of abiotic stresses, including humidity, light intensity, and water supply [47], and by pesticide usage [48].

Construction of a new statistical procedure for estimating the mixing proportions of blended samples of different origins

Metabolomics data from the 21 blended samples were used to establish a new statistical procedure for estimating the mixing proportions of the blended samples. The 21 NMR spectra contained vector and scalar pairs,  $(y^{(i)}, \pi^{(i)})$  ( $i=1,2,\dots,21$ ), where  $y^{(i)}=(y_1^{(i)}, y_2^{(i)}, \dots, y_{239}^{(i)})$  was a vector of intensity and  $\pi^{(i)}$  was the true mixing proportion of the Korean sample in the  $i$ th sample. The  $\pi^{(i)}$  was assumed to be unknown during analysis except for the first three ( $i=1,2,3$ ;  $\pi^{(i)}=0$  for 0 %) and the last three ( $i=19,20,21$ ;  $\pi^{(i)}=1$  for 100 %)

samples, which were known to be pure Chinese and Korean, respectively. Then,  $\pi$  was estimated by testing the blended sample's  $y$ . In addition,  $\hat{v}^{CN}$  and  $\hat{v}^{KR}$  were set as the mean vectors of the representative blended Chinese (0 % Korean) and Korean (100 % Korean) samples, i.e.,  $\hat{v}^{CN} = \frac{1}{3} \sum_{t=1}^3 y^{(t)}$  and  $\hat{v}^{KR} = \frac{1}{3} \sum_{t=19}^{21} y^{(t)}$ .

It is expected that  $y_k$ , the spectral intensity of the  $k$ th bin, has a mean of  $\pi \hat{\mu}_k^{KR} + (1-\pi) \hat{\mu}_k^{CN}$ , where  $\hat{\mu}_k^{KR}$  and  $\hat{\mu}_k^{CN}$  are the  $k$ -th intensity of the mean vectors of 30 training samples from Korea and China, respectively. A subset of bins were selected to be used for estimating the mixing proportion based on the premise that the two means  $\hat{\mu}_k^{KR}$  and  $\hat{\mu}_k^{CN}$  are statistically separated from each other if the  $k$ th bins of the spectra show disparity between Korean and Chinese samples. As a result, 12 bins were selected, i.e.,  $K = \{28, 30, 32, 33, 40, 42, 43, 44, 47, 48, 52, 56\}$ , corresponding to the spectral region between 1.34 and 2.48 ppm. These bins include the peaks at 1.42 ppm for paeoniflorin, 1.48 ppm for alanine, 1.54 ppm for albiflorin, 1.91 ppm for arginine, and 2.32 ppm for GABA, which were identified as the marker variables for origin discrimination by the OPLS-DA analysis and the t-test.

The mean spectral intensities of the representative blended samples ( $\hat{v}^{CN}$  and  $\hat{v}^{KR}$ ) in the selected region deviated greatly from those of the training samples ( $\hat{\mu}^{CN}$  and  $\hat{\mu}^{KR}$ ) (Fig. 4a) most likely because of random measurement errors. To reduce these uncontrolled deviations, the data from the blended samples were preprocessed utilizing a point-wise linear matching,  $\hat{\mu}_k^{CN} = a_k \hat{v}_k^{CN} + b_k$  and  $\hat{\mu}_k^{KR} = a_k \hat{v}_k^{KR} + b_k$  for each  $k \in K$ . Determining the values for  $a_k$  and  $b_k$  was directly achieved by solving a system of linear equations. Then, for every  $i = 1, 2, \dots, 21$ ,  $y^{(i)}$  was redefined as  $y_k^{(i)} = a_k y_k^{(i)} + b_k$ . Using this transformation, the deviation in the mean spectra between blended and training samples was removed (Fig. 4b).

The preprocessed data from the 12 bins above was used in the constrained least-squares method to establish a procedure for the estimation of the mixing proportions. The set of bins to be used was denoted as  $K$ , consisting of bins such that, for each  $k \in K$ ,  $\hat{\mu}_k^{KR}$  and  $\hat{\mu}_k^{CN}$  have enough disparity to distinguish between the two groups. To estimate  $\pi$ , we propose a two-step procedure. First, for each  $k \in K$ ,  $\pi_k$  was found to minimize  $(y_k - \pi_k \hat{\mu}_k^{KR} - (1-\pi_k) \hat{\mu}_k^{CN})^2$  subject to  $0 \leq \pi_k \leq 1$ . The minimum of the above was at  $\hat{\pi}_k = \min \left( \max \left( \frac{y_k - \hat{\mu}_k^{CN}}{\hat{\mu}_k^{KR} - \hat{\mu}_k^{CN}}, 0 \right), 1 \right)$ . Then,  $\pi$  was estimated as  $\hat{\pi} := \frac{1}{\#(K)} \sum_{k \in K} \hat{\pi}_k$ . Finally, a model for estimating the mixing proportions of blended samples was established. Figure 5 displays the plot of the true mixing proportion ( $\pi^{(i)}$ ) against the estimated mixing proportion ( $\hat{\pi}^{(i)}$ ) for  $i = 1, 2, \dots, 21$ , which was calculated from the established model for the seven groups of blended samples.

An adjusted  $R^2$  of 0.7669 for the plot indicates that the model to estimate the mixing proportions of blended peony roots had a relatively good predictive capability.

Peony roots contain a variety of bioactive components including the marker metabolites paeoniflorin and albiflorin identified in the present study, and their compositions may vary according to their geographical origins. Accordingly, the blended samples composed of Korean and Chinese peony roots may have different pharmacological effects than the pure Korean or the pure Chinese samples depending on the mixing ratio. The known pharmacological effects of peony root, including anti-inflammatory and immunomodulatory effects [1], may be evaluated on the blended samples at various mixing ratios to address the question above, although peony root is generally prepared as "Tang (concoction)" in which other herbal medicines are mixed according to traditional oriental formulas. However, discrimination of the geographical origins of peony root and the estimation of the mixing proportions in blended samples has more value in terms of cultural and economic effects than in therapeutic effects. Considering that the origin of the peony root is considered essential for determining pricing in Korea, the roots' geographical origins are always subject to counterfeiting. Correct indications of the geographical origins could help promote local socio-economic development and develop consumer trust and loyalty.

## Conclusions

In this study, two DNA-based techniques, CIS analysis and HRM analysis were applied in parallel with metabolomics analysis and were found to be inadequate to differentiate peony root samples of the same species, *P. lactiflora*, according to their cultivation areas. On the other hand, their metabolic profiles analyzed by  $^1\text{H-NMR}$  spectroscopy showed clear differences based on their geographical origins. OPLS-DA enabled clear classification of the samples according to their origin, and several metabolites were identified as potential markers. Internal and external validation by the response permutation test and blind prediction test, respectively, suggested the established model was practical for use in discriminating between peony root samples based on their geographical origin. Finally, a new statistical procedure using a constrained least-squares method allowed for the estimation of the mixing proportions of blended samples. Consequently, we suggest that the current study based on the  $^1\text{H-NMR}$  metabolomics could be applied for the quality control of TCMs. Because the geographical origins are regarded as essential in the grading and the pricing of the peony root, correct identification of the origin by the suggested method may help promote local socio-economic development and build consumer trust and loyalty by preventing origin counterfeiting.

**Acknowledgments** This work was supported by the following grants: the Next-Generation BioGreen 21 Program (No. PJ008202) from the Rural Development Administration, Republic of Korea, the Medicinal Herbs Discrimination Project from the Ministry of Health and Welfare, Republic of Korea, the Yujeonja-Donguibogam project based on Traditional herbs (No. 2012M3A9C4048796), Republic of Korea, and the Basic Science Research Program (No. 2011-0024225) of the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (MEST), Republic of Korea.

## References

- He DY, Dai SM (2011) Anti-inflammatory and immunomodulatory effects of *Paeonia lactiflora* Pall., a traditional Chinese herbal medicine. *Front Pharmacol* 2:10
- Guo D, Ye G, Guo H (2006) A new phenolic glycoside from *Paeonia lactiflora*. *Fitoterapia* 77(7–8):613–614
- Ikeda N, Fukuda T, Jyo H, Shimada Y, Murakami N (1996) Quality evaluation on *Paeoniae radix*. I. Quantitative analysis of monoterpene glycosides constituents of *Paeoniae radix* by means of high performance liquid chromatography. Comparative characterization of the external figures, processing method and the cultivated areas. *J Pharm Soc Jpn* 116(2):138–147
- Kamiya K, Yoshioka K, Saiki Y, Ikuta A, Satake T (1997) Triterpenoids and flavonoids from *Paeonia lactiflora*. *Phytochemistry* 44(1):141–144
- Jiang D, Chen Y, Hou X, Xu J, Mu X, Chen W (2011) Influence of *Paeonia lactiflora* roots extract on cAMP-phosphodiesterase activity and related anti-inflammatory action. *J Ethnopharmacol* 137(1):914–920
- Hsu F-L, Lai C-W, Cheng J-T (1997) Antihyperglycemic effects of paeoniflorin and 8-debenzoylpaeoniflorin, glucosides from the root of *Paeonia lactiflora*. *Planta Med* 63(04):323,325
- Yang HO, Ko WK, Kim JY, Ro HS (2004) Paeoniflorin: an antihyperlipidemic agent from *Paeonia lactiflora*. *Fitoterapia* 75(1):45–49
- Kang J, Lee S, Kang S, Kwon HN, Park JH, Kwon SW, Park S (2008) NMR-based metabolomics approach for the differentiation of ginseng (*Panax ginseng*) roots from different origins. *Arch Pharm Res* 31(3):330–336
- Kim SH, Kim DH, Park J-H, Choi EJ, Park S, Lee KY, Jeon M-J, Kim YC, Sung SH (2010) Discrimination of *Scrophularia* spp. according to geographic origin with HPLC-DAD combined with multivariate analysis. *Microchem J* 94(2):118–124
- Kuhnen S, Bernardi Ogliairi J, Dias PF, da Silva Santos M, Ferreira AG, Bonham CC, Wood KV, Maraschin M (2010) Metabolic Fingerprint of Brazilian maize landraces silk (stigma/styles) using NMR spectroscopy and chemometric methods. *J Agric Food Chem* 58(4):2194–2200
- Lima MR, Felgueiras ML, Graca G, Rodrigues JE, Barros A, Gil AM, Dias AC (2010) NMR metabolomics of esca disease-affected *Vitis vinifera* cv. Alvarinho leaves. *J Exp Bot* 61(14):4033–4042
- Pereira GE, Gaudillere J-P, Leeuwen C, Hilbert G, Maucourt M, Deborde C, Moing A, Rolin D (2006) 1H NMR metabolite fingerprints of grape berry: comparison of vintage and soil effects in Bordeaux grapevine growing areas. *Anal Chim Acta* 563(1–2):346–352
- Staneva J, Denkova P, Todorova M, Evstatieva L (2011) Quantitative analysis of sesquiterpene lactones in extract of *Arnica montana* L. by <sup>1</sup>H NMR spectroscopy. *J Pharm Biomed Anal* 54(1):94–99
- Wen H, Jeon B, Moon S, Song Y, Kang S, Yang H-J, Song Y, Park S (2010) Differentiation of antlers from deer on different feeds using an NMR-based metabolomics approach. *Arch Pharm Res* 33(8):1227–1234
- van Leeuwen C, Friant P, Choné X, Tregoat O, Koundouras S, Dubourdieu D (2004) Influence of climate, soil, and cultivar on terroir. *Am J Enol Vitic* 55(3):207–217
- Kang J, Choi MY, Kang S, Kwon HN, Wen H, Lee CH, Park M, Wiklund S, Kim HJ, Kwon SW, Park S (2008) Application of a <sup>1</sup>H nuclear magnetic resonance (NMR) metabolomics approach combined with orthogonal projections to latent structure-discriminant analysis as an efficient tool for discriminating between Korean and Chinese herbal medicines. *J Agric Food Chem* 56(24):11589–11595
- Korean Pharmacopoeia (2007). Korea
- Wang Q, Liu R, Gua H, Ye M, Huo C, Bi K, Guo D (2005) Simultaneous LC determination of major constituents in red and white peony root. *Chromatographia* 62:581–588
- Arumugasundaram S, Ghosh M, Veerasamy S, Ramasamy Y (2011) Species discrimination, population structure and linkage disequilibrium in *Eucalyptus camaldulensis* and *Eucalyptus tereticornis* using SSR markers. *PLoS One* 6(12):e28252
- Zhang CY, Wang FY, Yan HF, Hao G, Hu CM, Ge XJ (2012) Testing DNA barcoding in closely related groups of *Lysimachia* L. (Myrsinaceae). *Mol Ecol Resour* 12(1):98–108
- Ebihara A, Nitta JH, Ito M (2010) Molecular species identification with rich floristic sampling: DNA barcoding the pteridophyte flora of Japan. *PLoS One* 5(12):e15136
- Kim HK, Choi YH, Verpoorte R (2010) NMR-based metabolomic analysis of plants. *Nat Protoc* 5(3):536–549
- Gilard V, Balayssac S, Malet-Martino M, Martino R (2010) Quality control of herbal medicines assessed by NMR. *Curr Pharm Anal* 6(4):234–245
- Kim HK, Choi YH, Verpoorte R (2010) NMR-based metabolomic analysis of plants. *Nat Protocols* 5(3):536–549
- Krishnan P, Kruger NJ, Ratcliffe RG (2005) Metabolite fingerprinting and profiling in plants using NMR. *J Exp Bot* 56(410):255–265
- Verpoorte R, Choi Y, Kim H (2007) NMR-based metabolomics at work in phytochemistry. *Phytochem Rev* 6(1):3–14
- Viant MR, Rosenblum ES, Tieerdema RS (2003) NMR-based metabolomics: a powerful approach for characterizing the effects of environmental stressors on organism health. *Environ Sci Technol* 37(21):4982–4989
- Doyle J, Doyle J (1987) Genomic plant DNA preparation from fresh tissue—CTAB method. *Phytochem Bull* 19(11)
- Kim JH, Jung JY, Choi HI, Kim NH, Park JY, Lee Y, Yang TJ (2012) Diversity and evolution of major *Panax* species revealed by scanning the entire chloroplast intergenic spacer sequences. *Genet Resour Crop Evol* 60(2):413–425
- Kim JS, Kim YJ, Lee JY, Kang SS (2008) Phytochemical studies on *Paeoniae radix* (2)—phenolic and related compounds. *Kor J Pharmacogn* 39(1):28–36
- Kim JS, Kim YJ, Lee SY, Kang SS (2008) Phytochemical studies on *Paeoniae radix* (3)—triterpenoids. *Kor J Pharmacogn* 39(1):37–42
- Yean MH, LEE JY, Kim JS, Kang SS (2008) Studies on *Paeoniae radix* (1)—monoterpene glucosides. *Kor J Pharmacogn* 39(1):19–27
- Yoo JS, Song MC, Ahn EM, Lee YH, Rho YD, Baek NI (2006) Quantitative analysis of paeoniflorin from *Paeonia lactiflora* using H-NMR. *Nat Prod Sci* 12(4):237–240
- BioMagResBank (2008) [http://nar.oxfordjournals.org/content/36/suppl\\_1/D402.abstract](http://nar.oxfordjournals.org/content/36/suppl_1/D402.abstract)
- Hosoki T, Nagasako T, Kimura D, Nishimoto K, Hasegawa R, Ohta K, Sugiyama M, Haruki K (1997) Classification of herbaceous peony [*Paeonia lactiflora*] cultivars by random amplified polymorphic DNA (RAPD) analysis. *J Jpn Soc Hortic Sci* 65(4):843–849

36. Li L, Cheng FY, Zhang QX (2011) Microsatellite markers for the Chinese herbaceous peony *Paeonia lactiflora* (Paeoniaceae). *Am J Bot* 98(2):e16–e18
37. Britten RJ, Rowen L, Williams J, Cameron RA (2003) Majority of divergence between closely related DNA samples is due to indels. *Proc Natl Acad Sci USA* 100(8):4661–4665
38. McCauley DE (1995) The use of chloroplast DNA polymorphism in studies of gene flow in plants. *Trends Ecol Evol* 10(5):198–202
39. Lehmensiek A, Sutherland MW, McNamara RB (2008) The use of high resolution melting (HRM) to map single nucleotide polymorphism markers linked to a covered smut resistance gene in barley. *Theor Appl Genet* 117(5):721–728
40. Montgomery J, Wittwer CT, Palais R, Zhou L (2007) Simultaneous mutation scanning and genotyping by high-resolution DNA melting analysis. *Nat Protoc* 2(1):59–66
41. Bylesjö M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J (2006) OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J Chemometr* 20(8–10):341–351
42. Bouché N, Fromm H (2004) GABA in plants: just a metabolite? *Trends Plant Sci* 9(3):110–115
43. Noctor G, Arisi A-CM, Jouanin L, Foyer CH (1998) Manipulation of glutathione and amino acid biosynthesis in the chloroplast. *Plant Physiol* 118(2):471–482
44. Kaplan F, Kopka J, Haskell DW, Zhao W, Schiller KC, Gatzke N, Sung DY, Guy CL (2004) Exploring the temperature-stress metabolome of *Arabidopsis*. *Plant Physiol* 136(4):4159–4168
45. Wang D, Tan Q-R, Zhang Z-J (2013) Neuroprotective effects of paeoniflorin, but not the isomer albiflorin, are associated with the suppression of intracellular calcium and calcium/calmodulin protein kinase II in PC12 cells. *J Mol Neurosci* (in press)
46. Suh KS, Choi EM, Lee YS, Kim YS (2013) Protective effect of albiflorin against oxidative-stress-mediated toxicity in osteoblast-like MC3T3-E1 cells. *Fitoterapia* (in press)
47. Ramakrishna A, Ravishankar GA (2011) Influence of abiotic stress signals on secondary metabolites in plants. *Plant Signal Behav* 6(11):1720–1731
48. Lydon J, Duke SO (1989) Pesticide effects on secondary metabolism of higher plants. *Pestic Sci* 25(4):361–373