



Comprehensive classification of USA cannabis samples based on chemical profiles of major cannabinoids and terpenoids

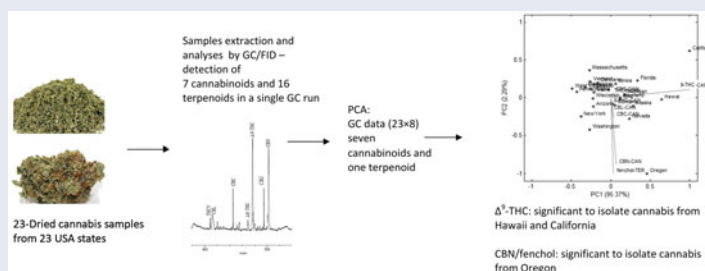
Ramia Z. Albakain^a , Yahya S. Al-Degs^b , James V. Cizdziel^c , and Mahmoud A. Elsohly^{d,e}

^aDepartment of Chemistry, School of Science, The University of Jordan, Amman, Jordan; ^bChemistry Department, The Hashemite University, Zarqa, Jordan; ^cDepartment of Chemistry and Biochemistry, University of Mississippi, University, MS, USA; ^dNational Center for Natural Products Research, University, MS, USA; ^eDepartment of Pharmaceutics and Drug Delivery, School of Pharmacy, University of Mississippi, University, MS, USA

ABSTRACT

Different USA-origin cannabis samples were analyzed by GC-FID to quantify all possible cannabinoids and terpenoids prior to their clustering. Chromatographic analysis confirmed the presence of seven cannabinoids and sixteen terpenoids with variable levels. Among tested cannabinoids, Δ^9 -Tetrahydrocannabinol Δ^9 -THC and cannabinol CBN were available in excess amounts (1.2–8.0 wt%) and (0.22–1.1 wt%), respectively. Fenchol was the most abundant terpenoid with a range of (0.03–1.0 wt%). The measured chemical profile was used to cluster 23 USA states and to group plant samples using different unsupervised multivariate statistical tools. Clustering of plant samples and states was sensitive to the selected cannabinoids/terpenoids. Principal component analysis (PCA) indicated the importance of Δ^9 -THC, CBN, CBG, CBC, THCV, Δ^8 -THC, CBL, and fenchol for samples clustering. Δ^9 -THC was significant to separate California-origin samples while CBN and fenchol were dominant to separate Oregon-origin samples away from the rest of cannabis samples. A special PCA analysis was performed on cannabinoids after excluding Δ^9 -THC (due to its high variability in the same plant) and CBN (as a degradation byproduct for THC). Results indicated that CBL and Δ^8 -THC were necessary to separate Nevada and Washington samples, while, CBC was necessary to isolate Oregon and Illinois plant samples. PCA based on terpenoids content confirmed the significance of caryophyllene, guaiol, limonene, linalool, and fenchol for clustering target. Fenchol played a major role for clustering plant samples that originated from Washington and Nevada. *k*-means method was more flexible than PCA and generated three different classes; samples obtained from Oregon and California in comparison to the rest of other samples were obviously separated alone, which attributed to their unique chemical profile. Finally, both PCA and *k*-means were useful and quick guides for cannabis clustering based on their chemical profile. Thus, less effort, time, and materials will be consumed in addition to decreasing operational conditions for cannabis clustering.

GRAPHICAL ABSTRACT



KEYWORDS

Cannabinoids; cannabis; *k*-means clustering; PCA; terpenoids; unsupervised analysis

Introduction

The size of analytical data in natural products science has increased substantially over the last several years due to the application of advanced instruments and chemometrics. Such development has yielded a better understanding of the chemistry of natural products, particularly for those with medical applications. Recently, a great deal of attention has

been paid to the analysis of natural products using sophisticated devices to identify more of their therapeutic potential. Cannabis is among these materials due to its widespread use and its diverse pharmacological properties.^[1] Cannabis is a chemically complex species containing a large number of active constituents.^[2,3] Herbal cannabis (marijuana), cannabis resin (hashish), and extracts of cannabis resin (hashish

oil) are still the most illicit drugs in the world. More than 8000 tons of cannabis are consumed in the USA every year^[2] where 11.5 million users purchasing around \$10 billion of the drug each year. In addition to the USA, cannabis is very popular in Canada and North America.^[2–4] In 2017, many USA states, including Washington, DC, have legalized the medical use of cannabis, where, 38 licensed producers in Canada are authorized to produce and sell dried marijuana.^[4] Overall, there has been a major increase in domestic production worldwide.

With the aid of advanced chromatographic instruments, a large number of active ingredients in cannabis samples including cannabinoids and terpenoids were identified.^[4,5] Both terpenoids and cannabinoids are known for their variable biological activities.^[6] Terpenoids are of great interest because of their production by plants that are likely to consistently reflect the immediate environment and they are responsible for cannabis' distinctive odor,^[7] whereas, cannabinoids would tend to reveal genetic relationships.^[8]

Today, most nations worldwide regard cannabis as an illegal drug of abuse. Despite the abuse potential of cannabis and its illegal status at the federal level in the USA, research into its chemistry and pharmacology has demonstrated that it also has medicinal properties. Cannabis has a long history of human use as a medicinal plant, intoxicant, and ritual drug.^[9,10] Clinical trials into cannabis, pure cannabinoids, and synthetic analogs have demonstrated some effectiveness as analgesics for chronic neuropathic pain, appetite stimulants for cancer and AIDS patients, multiple sclerosis, pain, inflammation, depression, anxiety, epilepsy, and infection.^[11–15] This has resulted in 16 potential therapeutic uses of cannabinoids, ranging from pain management to neurological disorders.^[16] The increased medical interest in these substances has prompted the development of various cannabis-based medicines such as the oral Δ^9 -THC preparation Marinol[®], a synthetic analog of Δ^8 -THC and an oral mucosal spray containing 1:1 ratio of Δ^8 -THC and cannabidiol (CBD).^[17,18]

A wide variety of analytical techniques have been described for the qualitative and quantitative determination of cannabinoids. Thin Layer Chromatography,^[19] fingerprinting with HPLC,^[20–22] GC and SFC coupled with mass spectrometry, ¹H NMR is useful for studying biochemistry and chemotaxonomy as well as quality control of medicinal plants.^[22] ¹H NMR has been used to fingerprint cannabis aqueous extracts and tinctures^[23] as well as to chemically differentiate cannabis cultivars.^[24] SFC also has been used to analyze cannabis but with limited studies.^[13,25–28] GC, however, is the most commonly used technique for analyzing cannabinoids and terpenoids.^[2,13,29–31] GC has been used to differentiate cannabis from different countries, including Mexico, Colombia, Jamaica, Thailand, and the USA.^[2]

Currently, there are three main classification systems for cannabis. The first is by species based on appearance, THC content, and geographical origins (gene pools) since environmental factors and marijuana cultivated sources can induce different cannabis profiles.^[2,8,19,29,30,32] The second classification is based on the ratio of two major cannabinoids THC and CBD, which is decided by their

corresponding allelic loci.^[33,34] The third is based on both cannabinoids and terpenoids for drug standardization and clinical research purposes.^[34] Novotny et al. reported that data relative to the use of GC analysis of marijuana samples of different origin indicated that the chromatograms appeared to be different, so the correlation between chromatographic data and geographical origin of marijuana samples might be possible.^[35] Hazekamp et al.^[20] reported the impact of changing the environmental conditions on the chemical composition and variability of terpenoids and cannabinoids in 11 cannabis varieties.

Principal component analysis (PCA) and hierarchical cluster analysis (HCA) are useful tools in analytical chemistry used for classifications.^[36–38] These tools were performed in cannabis studies for many purposes;^[35] to identify the compounds most important in distinguishing cannabis varieties, to find the variation on cannabis chemical profiles as a result of growing plants in different batches and with deviations in growth time, to confirm whether the cultivars in the cluster analysis would also be grouped together and to reveal the compounds that were responsible for grouping cultivars between clusters.

There is currently no available systematic clustering for the USA drug-type cannabis samples, which is necessary to explore the similarities/differences, if any, among plant samples. Therefore, this study was carried out to examine the chemical profiles of cannabinoids and terpenoids in cannabis grown in the USA. For the first time, 4 cannabis plants samples obtained from each of the 23 USA States were collected, extracted and analyzed using GC-FID. The plant samples were analyzed to detect all possible cannabinoids and terpenoids, which are necessary for plant samples clustering. The validated method was evaluated for selectivity and precision (i.e., repeatability). Grouping of plant samples from different states was carried out using unsupervised clustering methods, including PCA and *k*-means clustering. Moreover, the significance of cannabinoids and terpenoids for sample clustering has been outlined. To the best of our knowledge, this paper is the first to classify the 23 USA states and to use the PCA and *k*-means clustering together.

Material and method

Cannabis samples

Representative cannabis samples were obtained from the supply of materials provided from seized samples by The Drug Enforcement Administration (DEA) and submitted to the National Institute on Drug Abuse (NIDA) for analysis under a national potency monitoring program. The samples were arrived in sealed plastic bags and stored in a dry cool storage facility in the Coy Waller Complex at the University of Mississippi prior to analysis. The samples were selected from 23 states that have enacted Medical Marijuana laws, including: Alaska, Arizona, California, Colorado, Delaware, Florida, Hawaii, Illinois, Maine, Maryland, Massachusetts, Michigan, Montana, Nevada, New York, Ohio, Oregon, Pennsylvania, Vermont, Washington, West Virginia, Wisconsin, and Mississippi.

Reagents

Standards of the most common cannabinoids (cannabigerol CBG, cannabichromene CBC, Δ^9 -(trans)-tetrahydrocannabinol THC, cannabicyclol CBL, cannabinol CBN, Δ^9 -tetrahydrocannabinol Δ^9 -THC, Δ^8 -tetrahydrocannabinol Δ^8 -THC) and terpenoids (α -pinene, β -pinene, α -humulene, β -caryophyllene, α -terpinol, myrcene, limonene, caryophyllene oxide, fenchol, linalool, sabinene, carveol, terpinolene, cineol, guaiol, α -bisabolol) were purchased from Sigma-Aldrich® (St. Louis, MO). Structural formulae of compounds are provided in Table 1.

All standards were of analytical grade. The internal standard, phenanthrene (99% purity) was supplied from Sigma-Aldrich®. All solvents used for extraction and other preparations were of HPLC ultra-grade: acetone and ethyl acetate ($\geq 99.7\%$), hexane ($\geq 98.5\%$), ethanol ($>98\%$), and methanol ($\geq 99.8\%$) were purchased from Sigma-Aldrich®. Chloroform ($\geq 99.8\%$) was provided by Fischer Scientific (Bridgewater, NJ). Ultrapure water ($18\text{ M}\Omega\text{ cm}^{-1}$) generated by Milli-Q Plus water purification system (Millipore, Billerica, MA) was used to prepare aqueous solutions and dilutions.

Cannabinoids and terpenoids extraction from the plant

Dried cannabis samples were manicured so that the material has a homogenous mixture of leaf particles with no seeds or stems. A 100 mg portion was transferred to a test tube and 3.0 mL of extraction solution (methanol-chloroform 9:1 v/v spiked with 0.2 mg/mL phenanthrene) was added. Phenanthrene is an ideal internal standard (IS) as it is not present in cannabis and also does not interfere with GC analysis. Phenanthrene was used as both a retention time marker (R_t between the terpenoids and cannabinoids) and as a reference to calculate quantities of the solute of interest. The extraction tube was then placed in an ultrasonic water bath for 15 min to break down the plant tissue and to allow soluble cannabis and terpenes to dissolve in the extraction solution. The samples were then centrifuged for 30 s at 2000 rpm. Finally, the extract was filtered using the Acrodisc syringe filter (PAU-Gelman Lab, 0.45 μm , 25 mm diameter) and collected in a screw-capped amber vial. Samples were stored in a freezer (-10°C) until analysis time. Duplicate extractions and injections were made for each cannabis sample.

Gas chromatography - flame ionization detector

Twenty-three standard stock solutions of the tested cannabinoids and terpenoids were prepared at the concentration of 100 $\mu\text{g}/\text{mL}$ in pure MeOH. Each solution was injected separately to identify the retention time of each component. The GC profiles of cannabis extracts were all generated in the splitless mode using an Agilent GC 6890 series system equipped with a 7683B autosampler. The GC column was an Agilent, DB-5, 30 m length, 0.25 mm internal diameter, film thickness 0.25 μm , (J&W Scientific Inc., Folsom, CA).

The injector and detector temperatures were set at 250 and 300 $^\circ\text{C}$, respectively. Helium was used as the carrier gas at a flow rate of 25 cm/s, the airflow rate was set at 300 mL/min. Hydrogen at 30 mL/min as detector gas and helium at 30 mL/min as makeup gas were used. Solute detection was achieved using the Flame Ionization Detector FID. The oven temperature was programmed in a linear generic gradient mode from 70 to 250 $^\circ\text{C}$ at a rate of 5.0 $^\circ\text{C}/\text{min}$. The final temperature was set to 250 $^\circ\text{C}$ and was held for 19 min, followed by 8 min to reset the oven to the starting temperature. The total overall runtime was 63 min/sample. The injection volume was 1.0 μL . The GC-FID was controlled by GC Chemstation software version B.04.01.

Chromatographic quantification of extracted compounds

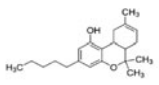
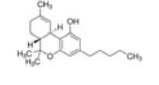
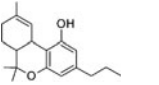
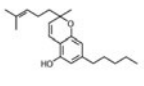
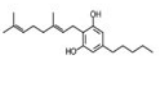
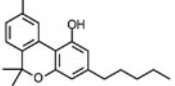
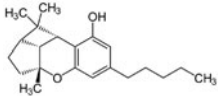
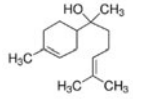
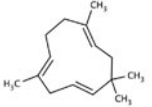

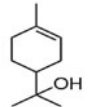
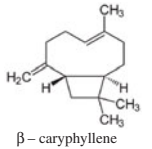
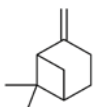
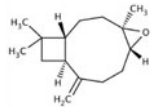
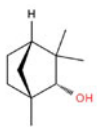
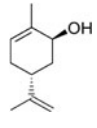
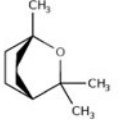
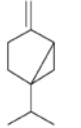
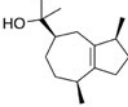
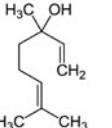
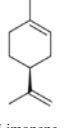
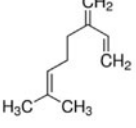
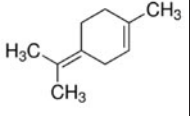
Selectivity was determined by injecting solvent blank to confirm that there were no false signal peaks at the targeted retention time. Each cannabinoid and terpenoid standard were individually injected to determine the retention time. Intraday reproducibility was determined by injecting an aliquot of a reference sample of β -pinene (100 $\mu\text{g}/\text{mL}$) five times from the same vial in a single day ($n=5$). The peak area ratio was calculated for each solute (Peak ratio = Peak area of solute/Peak area of IS) and further used to quantify the content of the detected solute. The contents of cannabinoids and terpenoids were reported as an average of two identical injections.

Unsupervised clustering of cannabis samples

PCA and HCA: Initially, the GC data was arranged in a data matrix $X_{n \times b}$, where n is the number of samples (states) and b is the number of measured variables (cannabinoids/terpenoids). For the current system, matrix X has a size of 23×23 (23 cannabis \times 23 cannabinoids/terpenoids). Matrix X was subjected to different clustering methods as will be discussed below. None of the collected samples was assigned to a class membership; hence, unsupervised clustering methods were applied.^[39,40]

The main adopted unsupervised methodologies for grouping/clustering objects are principal component analysis PCA and hierarchical clustering HCA.^[39,40] In PCA, matrix X is decomposed into two matrices; T (score matrix) and L (loading matrix) using singular value decomposition or non-linear iterative partial least-squares.^[40] Hence, $X_{23 \times 23}$ is decomposed to $T_{23 \times h}$ and $L_{h \times 23}$ where h is the number of optimum factors needed for matrix decomposition.^[39,40] The value of h is often estimated by leave-one-out cross-validation mythology.^[40] Once h is computed, scores and loading vectors are viewed to assess the clusters and the significance of variables for clustering samples. In HCA, GC data is displayed in a certain way to emphasize their natural clusters and patterns in a two-dimensional space. The results are often presented in the form of a dendrogram which allows quick visualization of clusters and correlations among tested samples.^[39]

Table 1. Structural formulae of tested cannabinoids and terpenoids.

Cannabinoids		
 Δ ⁸ -THC/Δ ⁸ -Tetrahydrocannabinol	 Δ ⁹ -THC /Δ ⁹ -Tetrahydrocannabinol	 THCV/Δ ⁹ -Tetrahydrocannabinol
 CBC/Cannabichromene	 CBG/Cannabigerol	 CBN/Cannabinol
 CBL/ Cannabicyclol		
Terpenoids		
 α-bisabolol	 α-humulene	 α-pinene
 α-terpinol	 β-caryophyllene	 β-pinene
 Caryophylleneoxide	 Fenchol	 Carveol
 Cineol	 Sabinene	 Guaiol
 Linalool	 Limonene	 Myrcene
 Terpinolene		

***k*-Means**

k-means clustering is an advanced unsupervised clustering method that is applied for a large number of samples.^[41-43]

The aim of *k*-means is to find clusters in the cannabis samples with the maximum number of groups given by the variable *k*. *k*-means can estimate the centroid of a given cluster,

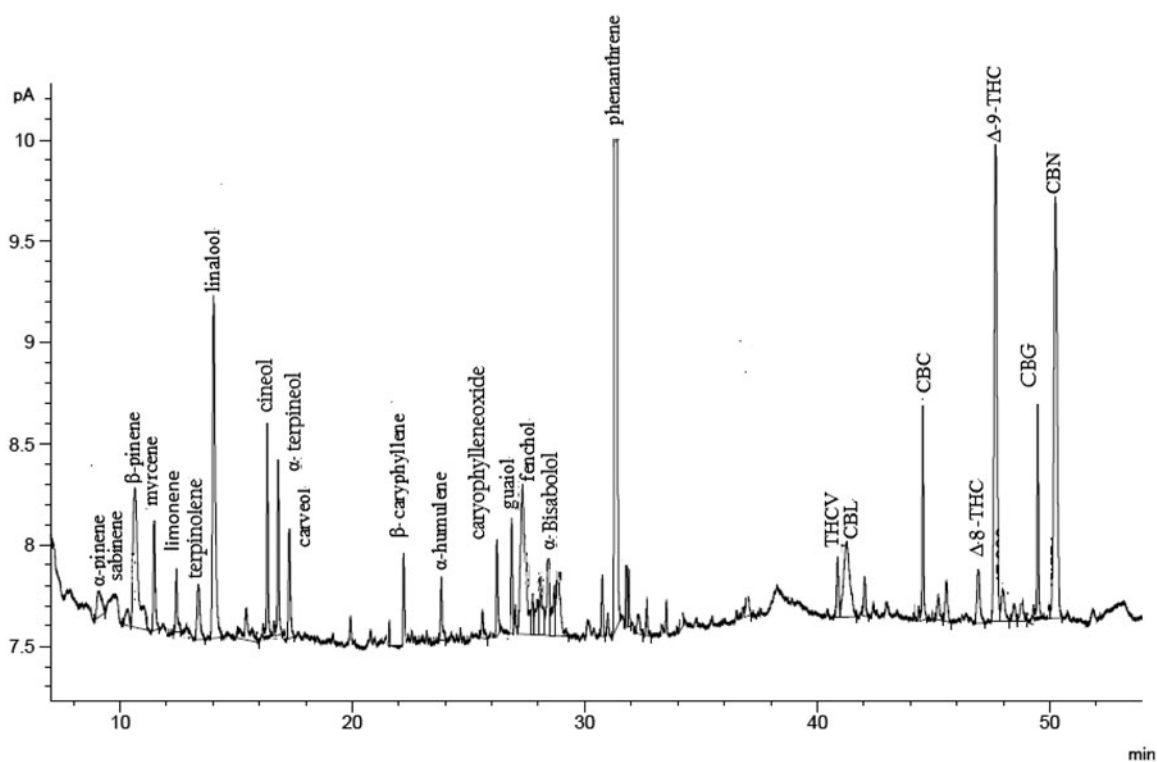


Figure 1. GC-FID chromatogram of cannabinoids and terpenoids in Alaska-origin cannabis.

which can then be used to assign a new sample into defined k -clusters. The inputs for k -means are X and the number of clusters is k .^[41–43]

Statistical software

The statistical analysis including principal component analysis PCA and hierarchical cluster analysis HCA were performed using Chemoface 1.61 software^[44] which runs under Matlab[®] (Mathworks, 8.6, USA) and k -means clustering was performed using XLSTAT software (Excel, Microsoft[®]).

Results and discussion

GC analysis and variability of cannabinoids and terpenoids in the samples

Among known active ingredients in cannabis, seven cannabinoids and sixteen terpenoids were detected in the plant samples over the 23 USA states. The detected cannabinoids and terpenoids have variable proportions but comparable to those reported for Canadian cannabis.^[4] All solutes were separated in 55 min. As a model example, Figure 1 depicts the GC chromatogram of Alaska-origin cannabis.

As shown in Figure 1, all solutes along with IS were fairly separated within 55 min. GC analysis indicated that the degradation of cannabinoids and terpenoids was not encountered. The sharp and intense GC peak positioned at 31.30 min was for phenanthrene (IS) and used to estimate the contents of other components. The position of phenanthrene encountered very small variations in both intensity and retention time during all injections. The proposed GC

method was stable and convenient to quantify all the 23 solutes. In general, cannabinoid solutes like CBN, Δ^9 -THC, CBG, and CBC have higher peak intensity and eluted at longer retention times compared to terpenoids, which would be attributed to the higher polarity of cannabinoids. Interday reproducibility was determined by injecting the same reference sample 12 times using fresh aliquots on each day ($n=12$). The intra and inter-day precisions (RSD%) were 0.37 and 0.32%, respectively. The method was precise in terms of repeatability and intermediate precision. Instrumental precision (RSD), defined as the variation in the peak area of the IS to all solutes was found to be 1.22%. The contents (provided as wt%) of cannabis samples are provided in Table 2.

As indicated in Table 2, seven cannabinoids were detected in the samples. The detected cannabinoids were related to six different classes: Δ^8 -Tetrahydrocannabinol (Δ^8 -THC), Δ^9 -Tetrahydrocannabinol (Δ^9 -THC and THCV), Cannabichromene (CBC), Cannabigerol (CBG), Cannabicyclol (CBL), and Cannabinol (CBN).^[7] It is known that cannabinoids would be available in neutral and acidic forms and quantification of both forms will require silylation/methylation of the acidic ones before GC analysis.^[12] Hence, the provided data in Table 2 gave the total contents of neutral and acidic forms of cannabinoids as no silylation of the acidic groups was carried out.

Moreover, most cannabinoids are available in their neutral form, for example, 10 isolated forms are known for Δ^9 -Tetrahydrocannabinol and only two of these are in acidic form.^[7] Among detected cannabinoids, the contents of Δ^9 -THC (1.2–8.0%) and CBN (0.22–1.1%) were notably higher than the rest of other ingredients. It is known that the

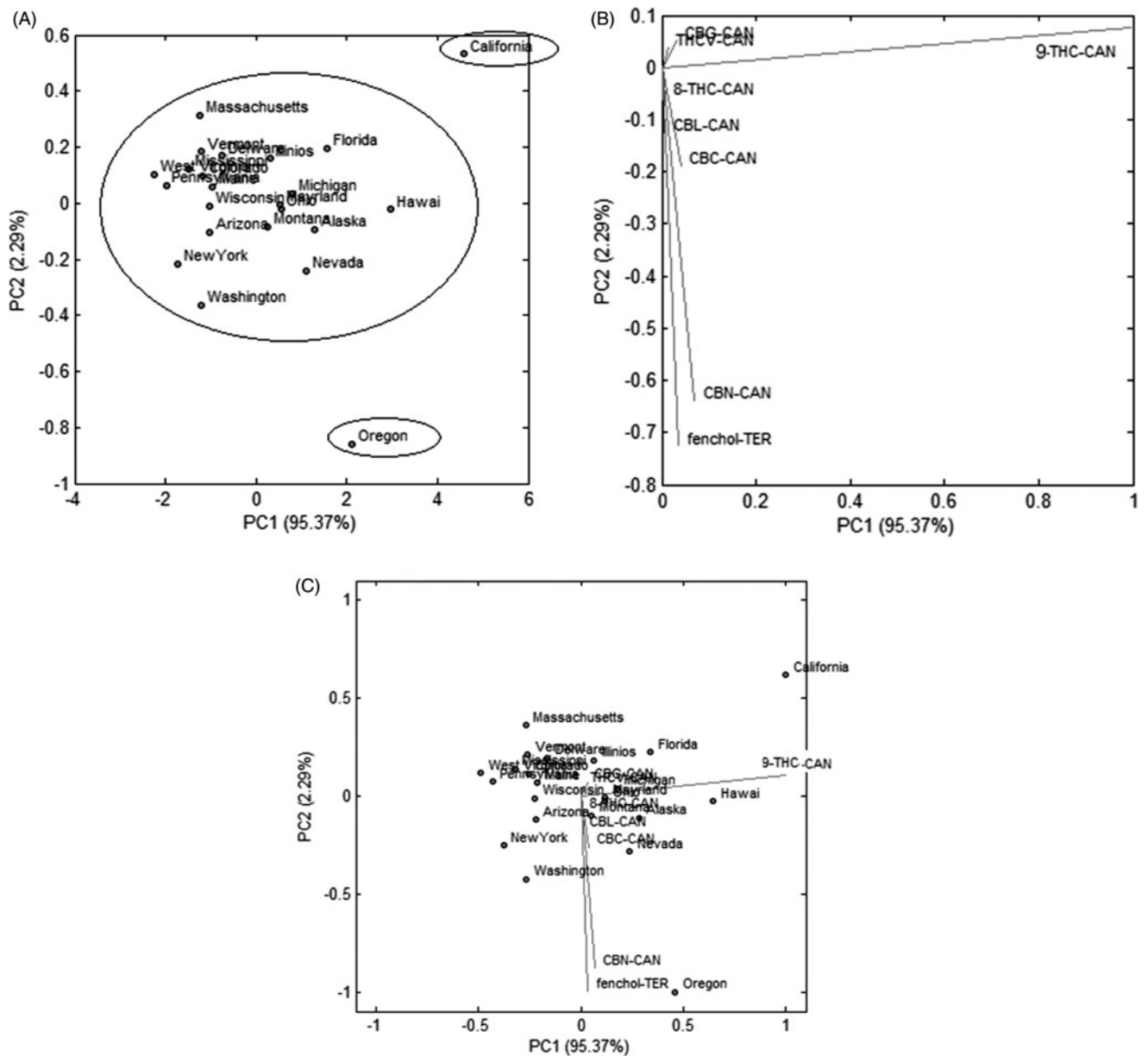


Figure 2. PCA outputs, (A) score plot, (B) loading plot, and (C) bi-plot obtained for pre-selected cannabinoids and terpenoids components.

97.66%. Hence, loading, score, and bi-plots can be viewed using two factors only.

As indicated in Figure 2A, the score plot indicated three main clusters collecting a different number of states. The three clusters were: California, Oregon, and (Alaska, West Virginia, Wisconsin, Washington, Vermont, Pennsylvania, Montana, Ohio, Mississippi, Nevada, New York, Massachusetts, Michigan, Maryland, Maine, Illinois, Florida, Colorado, Arizona, Hawaii, and Delaware). Accordingly, cannabis samples obtained from California and Oregon are significantly different from the rest of samples from other states. Again, this clustering was obtained when seven cannabinoids and one terpenoid included in the analysis. Figure 2B showed the most significant solutes for sample clustering. Both CBN and fenchol were positively correlated as they appeared in the same direction as shown in Figure 2B. Δ^9 -

THC was not correlated with other cannabinoids and more significant for sample clustering (appeared alone in the plot, Figure 2B). Most of the variables (CBC, CBL, 8-THC, CBG, and THCV) were positioned close to the center and this indicating their limited usage for sample clustering. It was interesting to notice the limited performance of some important cannabinoids (CBC and CBG) for cannabis clustering compared to fenchol. In summary, fenchol, CBN and Δ^9 -THC seem to be the most significant variables for sample clustering. As depicted in Figure 2C, Δ^9 -THC was of high efficiency to separate California from the rest of the states (Δ^9 -THC and California were in the same direction). To a less extent, Δ^9 -THC was effective to separate Hawaii from the rest of the states. On the other hand, CBN and fenchol were dominant to separate Oregon away from the rest of the cannabis samples obtained from other states.

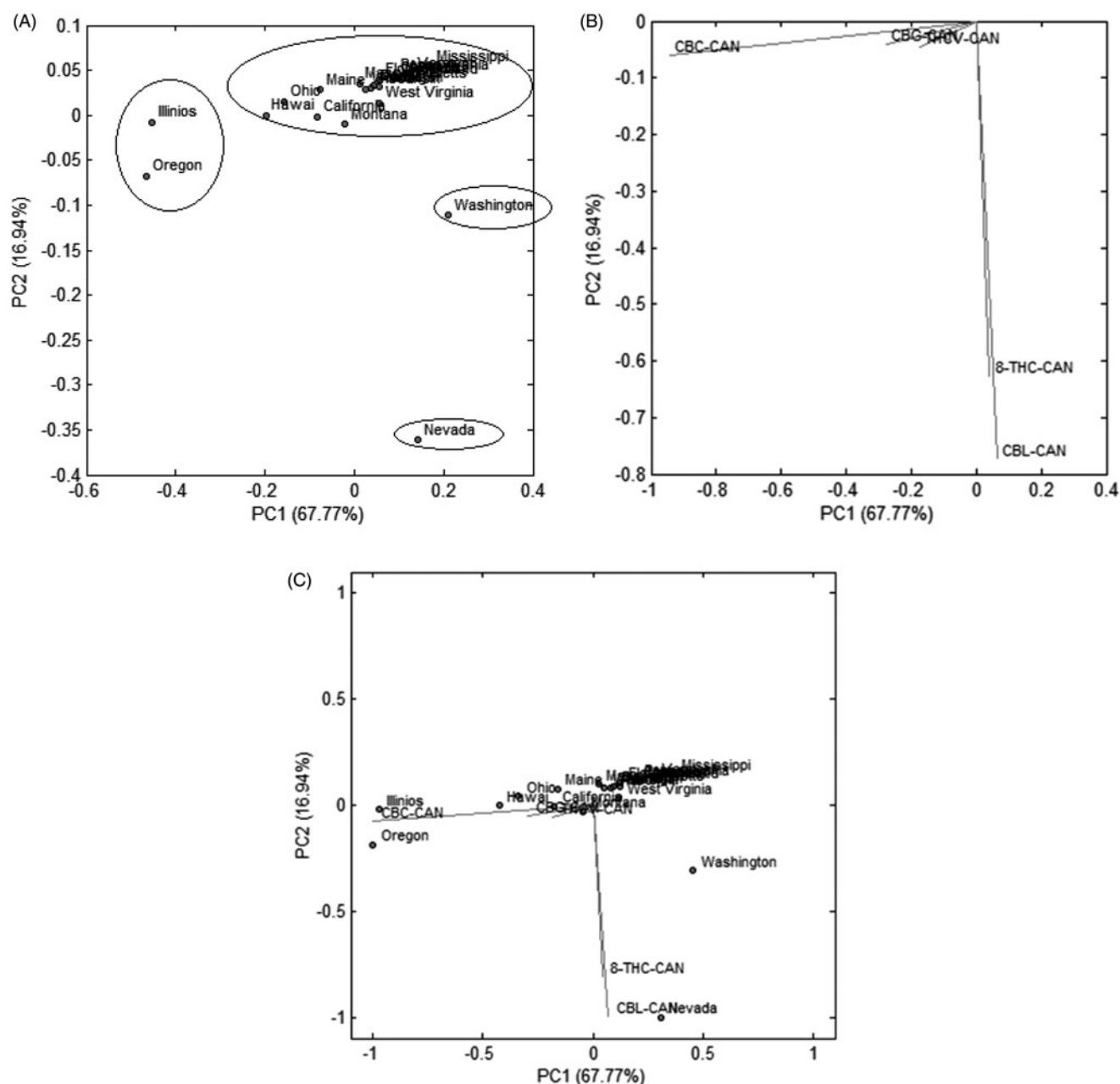


Figure 3. PCA outputs: (A) score plot, (B) loading plot, and (C) bi-plot based on a limited number of cannabinoids.

Cannabinoids-based-PCA analysis

Compared to terpenoids, the cannabinoids group is more significant due to their: (a) therapeutic uses including pain management and neurological disorders,^[4,5,7,11] and (b) large abundance in cannabis.^[7] As shown in Table 2, Δ^9 -THC, CBN, CBG, and CBC were available in large excess compared to the rest of the compounds. In fact, Δ^9 -THC is a common constituent with levels varying even within the same sample depending on the composition of the sample (i.e., leaves vs. bud, vs. mixture and the ratio of small leaves to large leaves). Hence, the variation in THC content among samples is expected and cannot be a distinguishing factor. At the same time, CBN is a degradation product of THC and only reflects the age of the sample or storage

environments and not the sample's origin. If all samples were harvested at the same time and kept under the same storage conditions, then, CBN would be included in the clustering job. In this section, PCA was performed by taking the following five cannabinoids: Δ^8 -THC, THCV, CBC, CBG, and CBL. A data matrix $X_{23 \times 5}$ containing the GC chromatographic data was subjected to PCA and the results are provided in Figure 3.

The chromatographic data were presented by two PC factors only with an accumulative variance of 84.71%, which allowed for sample clustering. From the score plot (Figure 3A), four clusters were identified: (Washington), (Nevada), and (Illinois/Oregon), and one large cluster containing the rest of other states. Removing Δ^9 -THC, CBN,

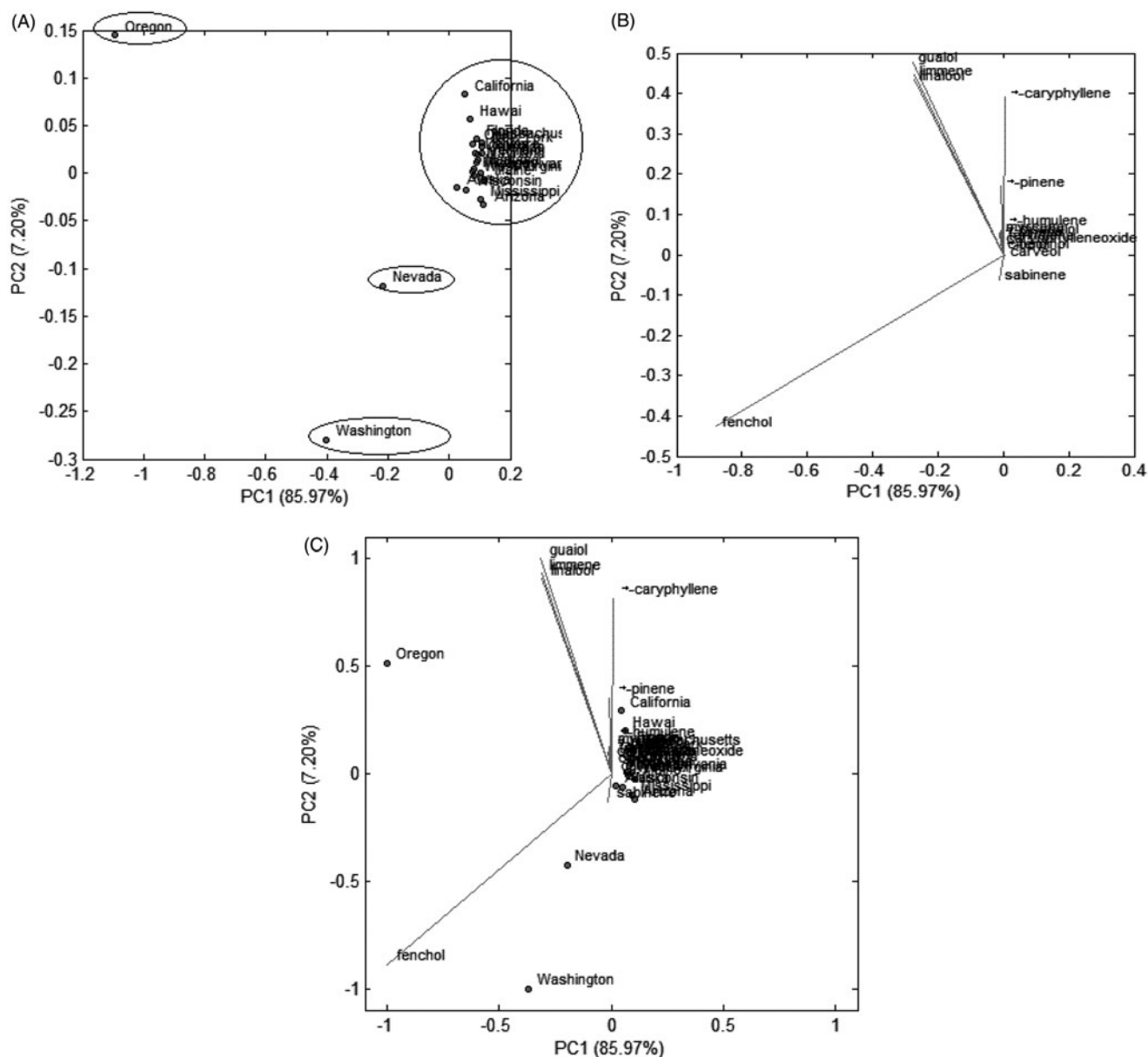


Figure 4. PCA outputs: (A) score plot, (B) loading plot, and (C) bi-plot based on terpenoids contents only.

and fenchol from the analysis has been separated Nevada, Illinois, Oregon, and Washington from the rest of the states as shown in Figure 3A. Interestingly, California-origin cannabis was clustered with other states and this would indicate the importance of Δ^9 -THC, CBN, and fenchol for this state (Compare Figure 2A and Figure 3A). In the meantime, the removal of Δ^9 -THC, CBN, and fenchol has improved the separation of Ohio, Hawaii, and Montana from the rest of other states. The position of other states was not affected upon excluding terpenoids from PCA (Compare Figure 2A with Figure 3A). Figure 3B indicated that CBC, CBL, and Δ^8 -THC were the most significant variables needed for sample clustering, CBL and Δ^8 -THC have the same influence and both not highly correlated with CBC (angle 90°). The rest of the variables were accumulated in the center indicating their limited applicability for sample clustering. Bi-plot (Figure 3C), indicated that CBL and Δ^8 -THC were necessary to separate Nevada and Washington, while, CBC was

necessary to isolate Oregon and Illinois from the rest of the states. At this stage, it is clear that sample clustering is sensitive to the selected cannabinoids.

Terpenoids-based-PCA analysis

Terpenoids are known for their distinctive odor and synergistic interactions with cannabinoids in the treatment of pain, inflammation, depression, anxiety, addiction, epilepsy, and cancer.^[7,8] Accordingly, the classification of cannabis samples based on terpenoids only deserved further investigation. A data matrix $X_{23 \times 16}$ containing GC data of all terpenoids was created and subjected to PCA, the results are provided in Figure 4.

As was the case in the earlier PCA, the data matrix was presented by two factors with an accumulative variance of 93.17%. The score plot (Figure 4A) exhibited four different groups: Oregon, Nevada, Washington, and one large cluster

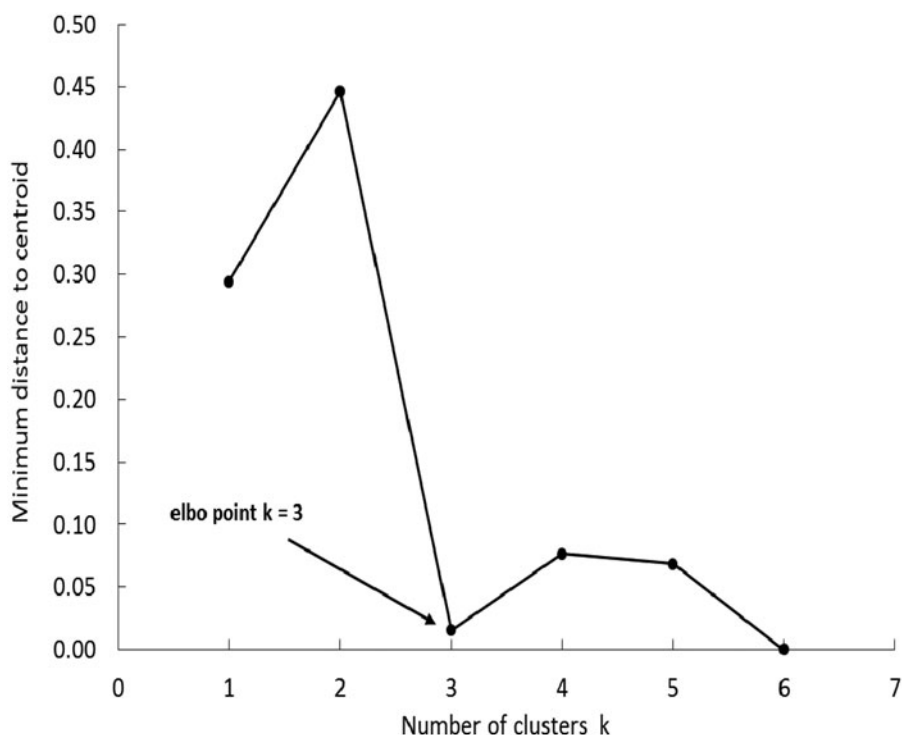


Figure 5. Determination of the optimum number of clusters by k -means algorithm.

containing the rest of the states. The variation in terpenoids contents was the reason behind the separation of cannabis samples obtained from Washington, Nevada, and Oregon. Separation of other samples was not possible. Among variables, caryophyllene, guaiol, limonene, linalool, and fenchol were the most significant for clustering especially for Oregon, Nevada, and Washington (Figure 4B).

As shown in Figure 4C, fenchol was efficient to separate Washington and Nevada from the rest of the states while guaiol, limonene, and linalool were helpful for separating Oregon from the rest of other states. In summary, PCA indicated that cannabis samples obtained from Oregon, Washington, and Nevada were distinguished from the rest of other plant samples.

***k*-Means clustering: including all variables for clustering**

For the first time in Cannabis study, k -means clustering is used. The benefit of k -means is that it can include all variables in the clustering job. In k -means clustering, choosing k is necessary to get the optimum number of centroid or clusters. To estimate the correct number of clusters in data matrix X , the algorithm is tested for a range of k values and the results are compared. In fact, there is no procedure for determining the correct k value. However, monitoring the mean distance among states and their cluster centroid is often adopted to estimate the optimum k value.^[41–43] For the current system, the minimum distance to the centroid is plotted against k as shown in Figure 5.

As shown in Figure 5, the distance to centroid was notably decreased at $k = 3$ which is also known as elbow point, represented the proper number of clusters for the current system. Accordingly, the 92 cannabis samples that have been

Table 3. Outputs of k -means clustering for samples based on cannabinoids/terpenoids contents.

Sum of weights	8	12	3
Within-class variance	0.289	0.271	2.182
Minimum distance to centroid	0.058	0.167	0.482
Average distance to centroid	0.451	0.450	1.117
Maximum distance to centroid	0.796	0.922	1.519
Class	1	2	3
No. of States	8	12	3
	Alaska	West Virginia	Oregon
	Montana	Wisconsin	Hawaii
	Ohio	Washington	California
	Nevada	Vermont	
	Michigan	Pennsylvania	
	Maryland	Mississippi	
	Illinois	New York	
	Florida	Massachusetts	
		Maine	
		Colorado	
		Arizona	
		Delaware	
Central state of class	Michigan	Colorado	Hawaii

collected across 23 USA states in this study can be clustered into three groups and this classification was totally based on their cannabinoids/terpenoids content. Besides the outlined method, cross-validation, theoretic jump, and silhouette methods^[42] were also presented to estimate k value. The characteristic results of k -means for clustering samples are summarized in Table 3.

As indicated in Table 3, all states were classified into three classes, each containing a different number of states. Class 1, Class 2, and Class 3 grouped 35, 52, and 13% of states (i.e., cannabis) samples, respectively. Accordingly, 52% of collected cannabis samples have comparable features or comparable chemical composition of terpenoids/cannabinoids. Another main finding from k -nearest is the unique

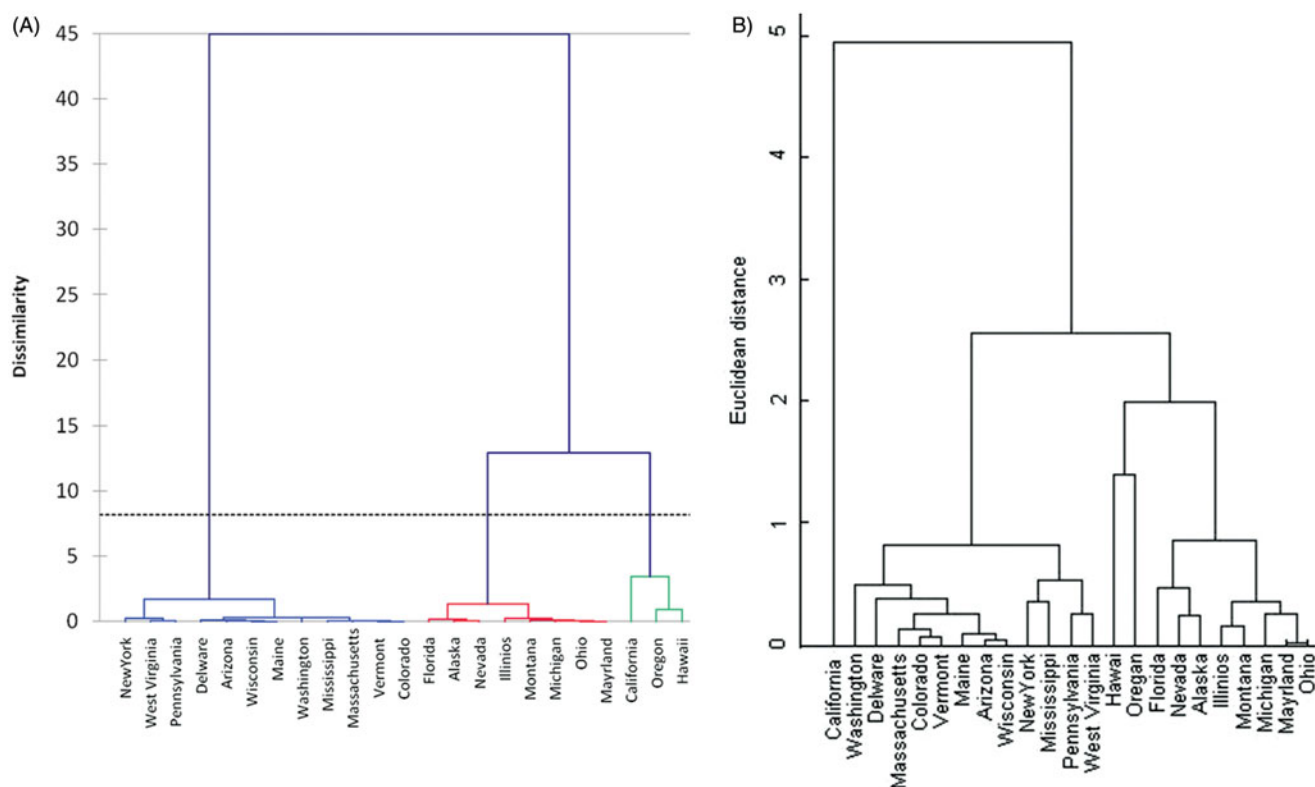


Figure 6. Dendrograms generated by (A) *k*-means-HCA and (B) PCA-HCA.

clustering of samples obtained from Oregon, Hawaii, and California, which may deserve more investigation in the future. Another interesting point regarding Class 3 was the higher distance of samples from the centroid (1.117). States grouped in Class 1 and 2 were of comparable distances to their centroids (0.451–0.450). For the three classes, the central states were Michigan, Colorado and Hawaii, for class 1, 2, and 3, respectively. Moreover, results indicated that distances between the central states were 2.0, 2.1, and 4.2 between Michigan–Colorado, Michigan–Hawaii, and Colorado–Hawaii, respectively. The earlier distances indicated that the distance between Colorado–Hawaii is rather high. This was attributed to the lower contents of CBN and Δ^9 -THC in Colorado-origin cannabis compared to those obtained from Hawaii. Another option of *k*-means is running HCA analysis based on the estimated centroids. The obtained dendrograms which estimated using Euclidean distance is presented in Figure 6.

As indicated in Figure 6A, three main clusters were identified and the cluster (Oregon, Hawaii and California) was obviously closer to the cluster (Alaska, Montana, Ohio, Nevada, Michigan, Maryland, Illinois, and Florida) and this agrees with the earlier results where the distance between Hawaii and Michigan (the central states in their class) is closer when compared to Colorado (the central state in the last class observed in the dendrogram). It is interesting to mention that the outputs of any adopted clustering method are highly sensitive to the input variables. Figure 6B displayed the dendrogram obtained by PCA including all variables. In fact, both PCA and *k*-means generated comparable plots with few exceptions. For example, in both methods,

the cannabis obtained from Washington, Delaware, Massachusetts, Colorado, Vermont, Maine, Arizona, Wisconsin, New York, Mississippi, Pennsylvania, West Virginia were clustered together. In addition, cannabis samples obtained from Florida, Nevada, Alaska, Illinois, Montana, Michigan, Maryland, and Ohio were also obviously clustered together by both methods.

The main difference between PCA and *k*-means was observed for California, Oregon, and Hawaii. As shown in Figure 6B, Oregon and Hawaii were grouped together while California was clustered alone by PCA, however, the three states were clustered in one group by *k*-means. One explanation for clustering Hawaii, California, and Oregon in the same cluster (using *k*-means) was attributed to the high total contents of cannabinoids and terpenoids 9.14, 10.39, and 10.56%, respectively, compared to the rest of other samples.

Conclusions

The outputs of 23 USA states grouping and cannabis plant samples clustering based on cannabis contents by PCA were sensitive to the selected of the measured chemical profile, that is, cannabinoids and terpenoids. Fenchol (terpenoid), CBN (cannabinoid) and Δ^9 -THC (cannabinoid) were significant input variables for cannabis clustering; Δ^9 -THC was essential to isolate California-origin cannabis from the rest of 23 samples. Samples clustering based on cannabinoid content revealed that CBL and Δ^8 -THC were dominant to isolate samples from Nevada and Washington. In the meantime, CBC was the controlling variable to separate Oregon and Illinois products from the rest of the samples. Samples

obtained from Oregon, Nevada, and Washington were clustered alone by PCA based on terpenoid content only. Both PCA and *k*-means outputs confirmed that the number of cannabis samples from the same group of clustered states could be minimized since they have the same content of cannabinoids and terpenoids. Medically, since different cannabis from different states showed similarities in their content, hence, any of them could be chosen. Thus, fewer efforts, time, and materials will be consumed in addition to decreasing operational conditions. Finally, the use of PCA and *k*-means was a useful and quick guide for samples classification based on the cannabis chemical profiles contents.

Acknowledgments

Dr. Ramia Albakain gratefully acknowledges the financial support for her Fulbright postdoctorate year 2017/2018 at The University of Mississippi, USA, provided by (1) The Binational Fulbright Commission in Jordan, and (2) the University of Jordan. Indeed, great appreciation goes to Dr. Omar Marzouk for his experimental help and technical support. Many thanks also go to Dr. Mohamed Radwan, Elsayed Ibrahim, Avery Claire Jones, and Chandrani Gon Majumdar for their assistance.

Disclosure statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

ORCID

Ramia Z. Albakain  <http://orcid.org/0000-0003-0419-0828>
 Yahya S. Al-Degs  <http://orcid.org/0000-0002-9555-7594>
 James V. Cizdziel  <http://orcid.org/0000-0002-4107-0190>
 Mahmoud A. Elsohly  <http://orcid.org/0000-0002-0019-2001>

References

- [1] Lee, D. C.; Schlien, N. J.; Peters, E. N.; Dworkin, R. H.; Turk, D. C.; Strain, E. C.; Vandrey, R. Systematic Review of Outcome Domains and Measures Used in Psychosocial and Pharmacological Treatment Trials for Cannabis Use Disorder. *Drug. Alcohol. Depend.* **2019**, *194*, 500–517. DOI: [10.1016/j.drugalcdep.2018.10.020](https://doi.org/10.1016/j.drugalcdep.2018.10.020).
- [2] Brenneisen, R.; Elsohly, M. Chromatographic and Spectroscopic Profiles Cannabis of Different Origins: Part 1. *J. Forensic Sci.* **1988**, *33*, 12583J–121404. DOI: [10.1520/JFS12583J](https://doi.org/10.1520/JFS12583J).
- [3] ProCon.org. Legal Medical Marijuana States and DC 2019. <http://medicalmarijuana.procon.org/view.resource.php?resourceID=000881>.
- [4] Jin, D.; Jin, S.; Yu, Y.; Lee, C.; Chen, J. Classification of Cannabis Cultivars Marketed in Canada for Medical Purposes by Quantification of Cannabinoids and Terpenes Using HPLC-DAD and GC-MS. *J. Anal. Bioanal. Tech.* **2017**, *8*, 1–9. DOI: [10.4172/2155-9872.1000349](https://doi.org/10.4172/2155-9872.1000349).
- [5] Elsohly, M. Constituents of Cannabis Sativa. In *Handbook of Cannabis*; Elsohly, M., Gul, W., Eds.; Oxford Univ. Press: Oxford, UK, **2014**.
- [6] McPartland, J. M.; Russo, E. B. Cannabis and Cannabis Extracts. *J. Cannabis Ther.* **2001**, *1*, 103–132. DOI: [10.1300/J175v01n03_08](https://doi.org/10.1300/J175v01n03_08).
- [7] Brenneisen, R. *Chemistry and Analysis of Phytocannabinoids and Other Cannabis Constituents. Forensic Science and Medicine: Marijuana and the Cannabinoids*; Humana Press Inc.: Totowa, NJ, **2007**.
- [8] ElSohly, M. A.; Stanford, D. F.; Murphy, T. P. Chemical Fingerprinting of Cannabis as a Means of Source Identification. In *Marijuana and the Cannabinoids*; Elsohly, M., Ed.; Humana Press: New York, NY, **2007**; pp 51–66.
- [9] Russo, E. B. History of Cannabis and Its Preparations in Saga, Science, and Sobriquet. *Chem. Biodiv.* **2007**, *4*, 1614–1648. DOI: [10.1002/cbdv.200790144](https://doi.org/10.1002/cbdv.200790144).
- [10] Pertwee, R. G. *Handbook of Cannabis*; Oxford University Press: Oxford, UK, **2014**; pp 33–22.
- [11] Russo, E. B. Taming THC: Potential Cannabis Synergy and Phytocannabinoid-Terpenoid Entourage Effects. *Br. J. Pharmacol.* **2011**, *163*, 1344–1364. DOI: [10.1111/j.1476-5381.2011.01238.x](https://doi.org/10.1111/j.1476-5381.2011.01238.x).
- [12] Wang, M.; Wang, Y.-H.; Avula, B.; Radwan, M. M.; Wanas, A. S.; Mehmedic, Z.; van Antwerp, J.; ElSohly, M. A.; Khan, I. A. Quantitative Determination of Cannabinoids in Cannabis and Cannabis Products Using Ultra-High-Performance Supercritical Fluid Chromatography and Diode Array/Mass Spectrometric Detection. *J. Forensic Sci.* **2017**, *62*, 602–611. DOI: [10.1111/1556-4029.13341](https://doi.org/10.1111/1556-4029.13341).
- [13] Baron, E. Comprehensive Review of Medicinal Marijuana, Cannabinoids, and Therapeutic Implications in Medicine and Headache: What a Long Strange Trip It's Been *Headache.* **2015**, *55*, 885–916. DOI: [10.1111/head.12570](https://doi.org/10.1111/head.12570).
- [14] Schier, A. R. d M.; Ribeiro, N. P. d O.; e Silva, A. C. d O.; Hallak, J. E. C.; Crippa, J. A. S.; Nardi, A. E.; Zuardi, A. W. Cannabidiol, a Cannabis Sativa Constituent, as an Anxiolytic Drug. *Revista Brasileira de Psiquiatria* **2012**, *34*, S104–S117. DOI: [10.1016/S1516-4446\(12\)70057-0](https://doi.org/10.1016/S1516-4446(12)70057-0).
- [15] Porter, B.; Jacobson, C. Report of a Parent Survey of Cannabidiol-Enriched Cannabis use in Pediatric Treatment-Resistant Epilepsy. *Epilepsy Behav.* **2013**, *29*, 574–577.
- [16] <http://www.hc-sc.gc.ca/dhp-mps/marihuana/med/infoprof-eng.php>.
- [17] Amar, M. B. Cannabinoids in Medicine: A Review of Their Therapeutic Potential. *J. Ethnopharmacol.* **2006**, *105*, 1–25.
- [18] Hazekamp, A.; Grotenhermen, F. Review on Clinical Studies with Cannabis and Cannabinoids 2005-2009. *Cannabinoids* **2010**, *5*, 1–21.
- [19] Sherma, J.; F, R. Thin Layer Chromatography in the Analysis of Cannabis and Its Components and Synthetic Cannabinoids. *J. Liq. Chrom. Relat. Tech.* **2019**, *42*, 613–628.
- [20] Hazekamp, A.; Fishedick, J. T. Cannabis - from Cultivar to Chemovar. *Drug Test. Anal.* **2012**, *4*, 660–667. DOI: [10.1002/dta.407](https://doi.org/10.1002/dta.407).
- [21] Hazekamp, A.; Peltenburg, A.; Verpoorte, R.; Giroud, C. Chromatographic and Spectroscopic Data of Cannabinoids from *Cannabis sativa* L. *J. Liq. Chrom. Relat. Tech.* **2005**, *28*, 2361–2382. DOI: [10.1080/10826070500187558](https://doi.org/10.1080/10826070500187558).
- [22] Van der Kooy, F.; Maltese, F.; Choi, Y. H.; Kim, H. K.; Verpoorte, R. Quality Control of Herbal Material and Phytopharmaceuticals with MS and NMR Based Metabolic Fingerprinting. *Planta Med.* **2009**, *75*, 763–775. DOI: [10.1055/s-0029-1185450](https://doi.org/10.1055/s-0029-1185450).
- [23] Politi, M.; Peschel, W.; Wilson, N.; Zloh, M.; Prieto, J. M.; Heinrich, M. Direct NMR Analysis of Cannabis Water Extracts and Tinctures and Semiquantitative Data on D9-THC and D9-THC-Acid. *Phytochemistry* **2008**, *69*, 562–570. DOI: [10.1016/j.phytochem.2007.07.018](https://doi.org/10.1016/j.phytochem.2007.07.018).
- [24] Choi, Y. H.; Kim, H. K.; Hazekamp, A.; Erkelens, C.; Lefeber, A. W. M.; Verpoorte, R. Metabolomic Differentiation of Cannabis Sativa Cultivars Using ¹H NMR Spectroscopy and Principal Component Analysis. *J. Nat. Prod.* **2004**, *67*, 953–957. DOI: [10.1021/np049919c](https://doi.org/10.1021/np049919c).
- [25] Breitenbach, S.; Rowe, W. F.; McCord, B.; Lurie, I. S. Assessment of Ultra High Performance Supercritical Fluid Chromatography as a Separation Technique for the Analysis of Seized Drugs: Applicability to Synthetic Cannabinoids.

- J. Chrom. A.* **2016**, *1440*, 201–211. DOI: [10.1016/j.chroma.2016.02.047](https://doi.org/10.1016/j.chroma.2016.02.047).
- [26] Toyo'oka, T.; Kikura-Hanajiri, R. A Reliable Method for the Separation and Detection of Synthetic Cannabinoids by Supercritical Fluid Chromatography with Mass Spectrometry, and Its Application to Plant Products. *Chem. Pharm. Bull.* **2015**, *63*, 762–769. DOI: [10.1248/cpb.c15-00170](https://doi.org/10.1248/cpb.c15-00170).
- [27] Bäckström, B.; Cole, M. D.; Carrott, M. J.; Jones, D. C.; Davidson, G.; Coleman, K. A Preliminary Study of the Analysis of Cannabis by Supercritical Fluid Chromatography with Atmospheric Pressure Chemical Ionisation Mass Spectroscopic Detection. *Sci. Just.* **1997**, *37*, 91–97. DOI: [10.1016/S1355-0306\(97\)72153-1](https://doi.org/10.1016/S1355-0306(97)72153-1).
- [28] Later, D. W.; Richter, B. E.; Knowles, D. E.; Andersen, M. R. Analysis of Various Classes of Drugs by Capillary Supercritical Fluid Chromatography. *J. Chromatogr. Sci.* **1986**, *24*, 249–253. DOI: [10.1093/chromsci/24.6.249](https://doi.org/10.1093/chromsci/24.6.249).
- [29] Hillig, K. W.; Mahlberg, P. G. A Chemotaxonomic Analysis of Cannabinoid Variation in Cannabis (Cannabaceae). *Am. J. Bot.* **2004**, *91*, 966–975. DOI: [10.3732/ajb.91.6.966](https://doi.org/10.3732/ajb.91.6.966).
- [30] Hillig, K. W. A Chemotaxonomic Analysis of Terpenoid Variation in Cannabis. *Biochem. Syst. Ecol.* **2004**, *32*, 875–891. DOI: [10.1016/j.bse.2004.04.004](https://doi.org/10.1016/j.bse.2004.04.004).
- [31] Lehmann, T.; Brenneisen, R. High Performance Liquid Chromatographic Profiling of Cannabis Products. *J. Liq. Chrom.* **1995**, *18*, 689–700. DOI: [10.1080/10826079508009265](https://doi.org/10.1080/10826079508009265).
- [32] Hillig, K. Genetic Evidence for Speciation in Cannabis (Cannabaceae). *Genet. Resour. Crop Evol.* **2005**, *52*, 161–180. DOI: [10.1007/s10722-003-4452-y](https://doi.org/10.1007/s10722-003-4452-y).
- [33] Zuardi, A. W.; Hallak, J. E. C.; Crippa, J. A. S. Interaction between Cannabidiol (CBD) and Δ^9 -Tetrahydrocannabinol (THC): Influence of Administration Interval and Dose Ratio between the Cannabinoids. *Psychopharmacology* **2012**, *219*, 247–249. DOI: [10.1007/s00213-011-2495-x](https://doi.org/10.1007/s00213-011-2495-x).
- [34] Stromberg, L. Minor Components of Cannabis Resin III: Comparative Gas Chromatographic Analysis of Hashish. *J. Chrom.* **1972**, *68*, 253–258.
- [35] Novotny, M.; Lee, M. L.; Low, C.-E.; Raymond, A. Analysis of Marijuana Samples from Different Origins by High-Resolution Gas-Liquid Chromatography for Forensic Application. *Anal. Chem.* **1976**, *48*, 24–29. DOI: [10.1021/ac60365a039](https://doi.org/10.1021/ac60365a039).
- [36] Al Bakain, R.; Rivals, I.; Sassiati, P.; Thiébaud, D.; Hennion, M.-C.; Euvrard, G.; Vial, J. Comparison of Different Statistical Approaches to Evaluate the Orthogonality of Chromatographic Separations: Application to Reverse Phase Systems. *J. Chrom. A.* **2011**, *1218*, 2963–2975. DOI: [10.1016/j.chroma.2011.03.031](https://doi.org/10.1016/j.chroma.2011.03.031).
- [37] Al Bakain, R.; Rivals, I.; Sassiati, P.; Thiébaud, D.; Hennion, M.-C.; Euvrard, G.; Vial, J. Impact of the Probe Solutes Set on Orthogonality Evaluation in Reverse Phase Chromatographic Systems. *J. Chrom. A.* **2012**, *1232*, 231–241. DOI: [10.1016/j.chroma.2011.12.056](https://doi.org/10.1016/j.chroma.2011.12.056).
- [38] Al Bakain, R. Z.; Al-Degs, Y.; Andri, B.; Thiébaud, D.; Vial, J.; Rivals, I. Supercritical Fluid Chromatography of Drugs: Parallel Factor Analysis for Column Testing in a Wide Range of Operational Conditions. *J. Anal. Meth. Chem.* **2017**, *2017*, 1–13. DOI: [10.1155/2017/5340601](https://doi.org/10.1155/2017/5340601).
- [39] Otto, M. *Chemometrics: Statistics and Computer Application in Analytical Chemistry*, 3rd ed.; Wiley-VCH: New York, NY, **2016**. ISBN 978-3-527-34097-2.
- [40] Breerton, R. G. *Applied Chemometrics for Scientists*; John Wiley & Sons: Chichester, UK, 2007.
- [41] MacQueen, J. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*; University of California Press: Berkeley, CA, 1967; pp 281–297.
- [42] Meng, Y.; Liang, J.; Cao, F.; He, Y. A New Distance with Derivative Information for Functional k -Means Clustering Algorithm. *Inform. Sci.* **2018**, *463–464*, 166–185. DOI: [10.1016/j.ins.2018.06.035](https://doi.org/10.1016/j.ins.2018.06.035).
- [43] Heil, J.; Häring, V.; Marschner, B.; Stumpe, B. Advantages of Fuzzy k -Means over k -Means Clustering in the Classification of Diffuse Reflectance Soil Spectra: A Case Study with West African Soils. *Geoderma* **2019**, *337*, 11–21. DOI: [10.1016/j.geoderma.2018.09.004](https://doi.org/10.1016/j.geoderma.2018.09.004).
- [44] Nunes, C. A.; Freitas, M. P.; Pinheiro, A. C. M.; Bastos, S. C. Chemoface: A Novel Free User-Friendly Interface for Chemometrics. *J. Braz. Chem. Soc.* **2012**, *23*, 2003–2010. DOI: [10.1590/S0103-50532012005000073](https://doi.org/10.1590/S0103-50532012005000073).
- [45] Novak, J.; Zitterl-Eglseer, K.; Deans, S. G.; Franz, C. M. Essential Oils of Different Cultivars of *Cannabis sativa* L. and Their Antimicrobial Activity. *Flavour Fragr. J.* **2001**, *16*, 259–262. DOI: [10.1002/ffj.993](https://doi.org/10.1002/ffj.993).
- [46] <http://www.maximummyield.com/definition/3932/fenchol>