Characterization and Discrimination of Volatile Compounds in Different Types of Tobacco Extracts Using GC-MS

Wang Long, Li Yuangen, Wei Shen, Zhang Yanfang, He Aimin, Jiehui Li, Wang Xiaoru, Yan Jin, Chen Weihua*

Hebei China Tobacco Industry Co., Ltd., Shijiazhuang 050051, China

Abstract: Flavor quality is a core attribute of tobacco and is closely related to the composition of volatile organic compounds (VOCs), which are influenced by genetic background, ecological conditions, and curing. To characterize and discriminate the volatile profiles of tobacco from different origins, tobacco samples from Guangdong, Zimbabwe, and Yunnan were analyzed by gas chromatography-mass spectrometry (GC-MS) combined with multivariate statistical analysis. A total of 197 VOCs were identified, including alcohols, ketones, acids, esters, aldehydes, heterocycles, phenols, and alkanes. The number and total content of VOCs differed markedly among origins. Guangdong contained 106 compounds with the highest total content (771.69 mg/kg), dominated by alkenes and acids; Zimbabwe contained 110 compounds (665.42 mg/kg), with ketones and phenols as the most abundant classes; Yunnan contained the largest number of compounds (136) but the lowest total content (576.21 mg/kg) and was richest in alcohols, esters, and heterocycles. Fifty-seven compounds were shared by all three origins, forming a core volatile skeleton, while Yunnan and Zimbabwe shared 94 compounds and Yunnan possessed 34 unique compounds. Principal component analysis and partial least squares discriminant analysis clearly separated the three origins, and VIP > 1.0 screening identified 82 differential markers. Terpenes such as p-cymene, ocimene, and pinene were enriched in Yunnan; norisoprenoid ketones and aromatic esters such as β-damascone and benzyl acetate accumulated in Zimbabwe; and short-chain acids such as propanoic acid were higher in Guangdong. Aroma activity evaluation further indicated a shared core aroma skeleton, including β-damascenone, nonanal, and linalool, together with origin-specific aroma notes. These results suggest three ecological chemotypes: Yunnan (terpene/heterocycle-enriched), Zimbabwe (ketone/ester-synergistic), and Guangdong (acid/alcohol-based). p-Cymene, β-damascone, propanoic acid, and maltol are proposed as candidate discriminative indicators for origin identification and flavor-oriented regulation.

Keywords: Tobacco; Volatile Organic Compounds; GC-MS; Multivariate Statistical Analysis; Origin Discrimination; PLS-DA; Aroma-active Compounds; Biomarkers

DOI: 10.57237/j.cse.2026.03.002

Funding: The “Research and Application Project on Low-Temperature Continuous Phase Change Extraction of Volatile Tar Oil from Natural Tobacco Leaves” (Project No. HBZY2024D001)

*Corresponding author: Chen Weihua, 604631656@qq.com

Received: September 30, 2026; Accepted: October 24, 2026; Published Online: November 2, 2026 http://www.agrforestry.com

1 Introduction

Flavor quality is a core attribute of tobacco quality [1]. It not only directly shapes consumers' sensory experience but also underpins product innovation and brand competitiveness [2]. Tobacco aroma is governed by hundreds of volatile organic compounds acting in concert; these compounds span multiple chemical classes, including alcohols, ketones, acids, esters, aldehydes, phenols, and terpenes [3]. Their composition and relative abundance are jointly regulated by genetic background, ecological conditions, and curing processes. Variations in variety, geographical environment, climatic factors, and agricultural practices are sufficient to induce pronounced differentiation in chemical composition and aroma profiles, thereby shaping markedly distinct tobacco leaf style characteristics [4].

In terms of analytical approaches, headspace solid-phase microextraction coupled with gas chromatography–mass spectrometry (HS-SPME-GC-MS) has become a mainstream platform for tobacco volatile metabolome analysis, because it requires no organic solvents, offers high enrichment efficiency, and causes little damage to thermolabile analytes [5]. In recent years, this technique has been widely applied to flavor characterization of tobacco from different origins [6]. Previous studies combining headspace gas chromatography–ion mobility spectrometry (HS-GC-IMS) and HS-SPME-GC-MS detected 93 and 80 volatile compounds, respectively, from six flue-cured tobacco varieties and, using orthogonal partial least squares discriminant analysis (OPLS-DA), identified 19 key differential compounds capable of distinguishing different aroma types [7]. For Guizhou honey-sweet aroma-type flue-cured tobacco, HS-SPME-GC-MS combined with machine learning was applied to leaf samples from 14 origins, identifying 27 discriminatory volatiles; odor activity value (OAV) analysis further pinpointed nonanal and β-cyclocitral as the most critical aroma-active substances [3].

In terms of chemometric strategies, tobacco-related studies have established a flue-cured tobacco aroma-type classification model based on gas chromatography mass spectrometry combined with principal component analysis (PCA) and discriminant analysis, demonstrating the robustness of aroma component data for variety and origin discrimination [7]. Moreover, for origin authentication of tobacco extracts, chemometrics has been shown to effectively screen characteristic components that differentiate tobacco extracts from Yunnan and Zimbabwe [8]. Collectively, these findings indicate that chemometric analysis of volatile components provides a sound methodological basis for accurate cross-origin discrimination.

However, existing studies have mostly focused on flavor comparisons involving a single origin or a few origins, and systematic comparisons of volatile metabolomes across ecological regions and multiple origins remain relatively limited. To address this gap, the present study selected tobacco from three production regions with typical ecological differences (Guangdong, Zimbabwe, and Yunnan) and used HS-SPME-GC-MS to systematically characterize their volatile metabolomes. PCA and partial least squares discriminant analysis (PLS-DA) models were employed to assess the feasibility of origin discrimination, and differential markers were screened using variable importance in projection (VIP) > 1.0 as the threshold. This study aims to provide a quantifiable marker panel for tobacco origin traceability and a metabolic basis for the targeted regulation of tobacco flavor from different production regions.

2 Materials and Methods

2.1 Materials

Tobacco leaves were supplied by Hebei China Tobacco Industry Co., Ltd. After drying under controlled environmental conditions, the samples were pulverized with a grinder and sieved to obtain the required particle-size fractions. The resulting tobacco powders were stored in sealed containers at room temperature until extraction. For extraction, butane (purity ≥ 99.99%; Guangzhou Shenyuan Gas Co., Ltd., Guangzhou, China) was used as the solvent for low-temperature continuous phase-transition extraction. All chemicals and reagents used in chemical analysis were of analytical or chromatographic grade. Guangdong tobacco was defined as Group A, Zimbabwe tobacco as Group B, and Yunnan tobacco as Group C

2.2 The Extraction of Tobacco Oil

Low-temperature continuous phase-transition extraction (LTCP) was performed using a custom-built continuous phase-transition extraction system. In brief, 400 g of pulverized tobacco powder was weighed and loaded into a 3 L extraction vessel. Butane was liquefied by compression

and continuously circulated through the extraction vessel at a flow rate of 60 L/h. During extraction, the liquefied solvent penetrated the tobacco matrix and dissolved the target components. The extract solution was then transferred to a separation vessel, where the solvent underwent phase transition via heating and depressurization, enabling the separation and collection of the tobacco oil. The evaporated butane was subsequently recompressed into a liquid state and returned to the extraction vessel for continuous recycling.

2.3 GC-MS Analysis of Tobacco Oil

The chemical compositions of the tobacco oils obtained by the different extraction methods were analyzed using a GCMS-QP2020 NX system. Before analysis, the samples were diluted with chromatographic-grade solvent and filtered through 0.22-μm membranes. Separation was performed on an SH-Polar WAX capillary column (60 m × 0.25 mm i.d., 0.25-μm film thickness), using high-purity helium (>99.999%) as the carrier gas at a flow rate of 1 mL/min. The oven temperature was initially held at 50°C for 1 min, increased to 120°C at 5°C/min, then to 180°C at 4°C/min and held for 7 min, and finally to 240°C at 4°C/min and held for 8 min. The injector temperature was set at 250°C, while the transfer line, ion source, and quadrupole temperatures were 240°C, 230°C, and 150°C, respectively. A 1-μL aliquot of each sample was injected in split mode at a split ratio of 20:1. Mass spectra were acquired in full-scan mode over the m/z range of 28–450 at an ionization voltage of 70 eV. The relative contents of individual compounds were calculated using the peak area normalization method.

2.4 Statistical Analysis

All data were analyzed using SPSS Statistics (version 13) and are presented as mean ± standard deviation (SD). Statistical comparisons were performed using one-way analysis of variance (ANOVA), followed by Duncan's post-hoc test. A p-value < 0.05 was considered statistically significant. Principal component analysis (PCA), hierarchical clustering analysis (HCA), and heatmap visualization were employed to evaluate differences in the chemical profiles of the tobacco oils.

3 Results

3.1 Overall Characteristics of Volatile Organic Compounds and TIC Profiles

The total ion chromatograms (TICs) of tobacco samples from Guangdong (A), Zimbabwe (B), and Yunnan (C) are shown in Figure 1. The TICs displayed complex and origin-specific peak patterns across the early-eluting regions I (RT: 8.13-18.35min). The Guangdong group (A) exhibited relatively intense peaks in the regions associated with alkenes, consistent with its highest total volatile content (771.69 mg/kg) and the dominance of these compound classes. Common peaks were also clearly observed among the three TICs, corresponding to the shared compounds that form the core volatile skeleton of tobacco across the regions II. Consistent with the higher chemical similarity between Yunnan and Zimbabwe, their TICs showed a relatively closer overall peak distribution than that of Guangdong. The chromatographic features provided an intuitive fingerprint of origin-related volatile differences.

A total of 197 volatile organic compounds (VOCs) were identified across the three origins. By chemical class, these included alcohols (42), ketones (37), acids (23), esters (21), aldehydes (15), heterocycles (19), phenols (14), alkanes (12), and others (14). The number of compounds and total content differed significantly among the groups in Figure 2. The Guangdong group (A) contained 106 compounds with a total content of 771.69 mg/kg, dominated by alkenes and acids. The Zimbabwe group (B) contained 110 compounds with a total content of 665.42 mg/kg, with ketones and phenols being the most abundant classes. The Yunnan group (C) contained the largest number of compounds (136) but the lowest total content (576.21 mg/kg), and was richest in alcohols, esters, and heterocycles. Fifty-seven compounds were shared by all three groups, constituting the core volatile skeleton of tobacco. Yunnan and Zimbabwe shared 94 compounds, indicating relatively high chemical similarity, while Yunnan possessed 34 unique compounds, suggesting potential flavor complexity.

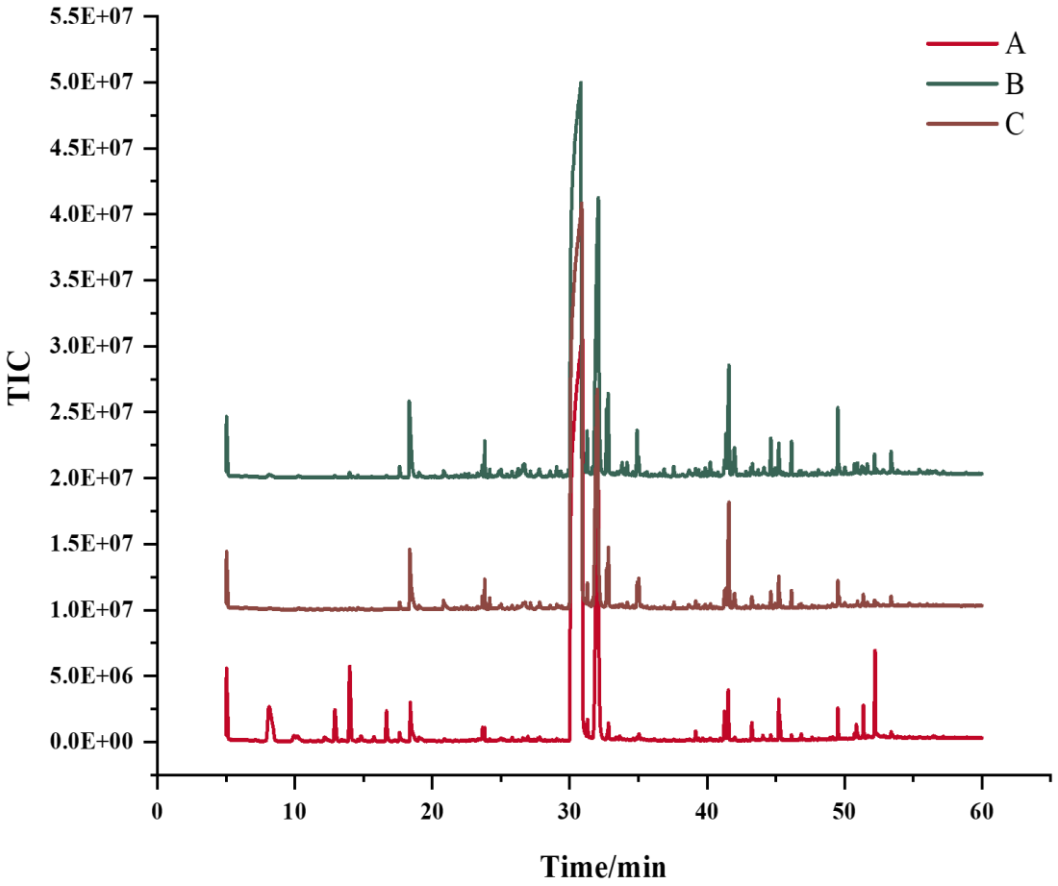


Figure 1 Total ion current chromatogram of volatile components from tobacco samples

3.2 Differential Analysis of Major Compound Classes

As shown in Figure 2, among the major compound classes, acids, alcohols, ketones, phenols, and aldehydes exhibited pronounced origin-dependent differences. Acetic acid was the overwhelmingly dominant acid, with the highest content in Zimbabwe (31.28 mg/kg), followed by similar levels in Guangdong (13.09 mg/kg) and Yunnan (13.51 mg/kg); notably, nonanoic acid was detected only in Guangdong tobacco (8.73 mg/kg) and, owing to its musty and pungent odor and possible origin from the oxidative metabolism of nonanal, may serve as a potential marker for Guangdong tobacco [9]. 2-Phenylethanol was highest in Guangdong (18.97 mg/kg), contributing rose-like and honey-like notes, whereas cis-linalool oxide was detected only in Yunnan (14.40 mg/kg), exhibiting woody and floral characteristics and acting as a characteristic marker of Yunnan tobacco [10]; several unique alcohols, including α-terpineol and cedrol, were also found in Yunnan. For ketones, 6,8-nonadien-2-one was exceptionally high in Zimbabwe (24.13 mg/kg) and can enhance the fullness of tobacco aroma, while β-ionone and β-damascone were significantly enriched in Zimbabwe, and megastigmatrienone and isophorone were rich in Yunnan [11]. In phenols, maltol content was similar in Yunnan (5.50 mg/kg) and Zimbabwe (5.25 mg/kg) and much higher than in Guangdong (1.91 mg/kg), imparting caramel-like sweetness, whereas guaiacol was detected only in Yunnan (1.82 mg/kg), exhibiting smoky and spicy characteristics and serving as a key marker of Yunnan tobacco [3]. Finally, nonanal was highest in Zimbabwe (2.72 mg/kg) with fatty and rose-like notes, and decanal was highest in Guangdong (2.99 mg/kg) with a sweet orange aroma; both are important aroma contributors [12].



Figure 2 Clustering heatmap of volatile organic compounds in tobacco leaf extracts from the same source

3.3 Multivariate Statistical Analysis

To intuitively assess overall metabolic differences among tobacco samples from different origins, unsupervised PCA was first applied to the volatile data. In the PCA score plot (Figure 3), PC1 and PC2 explained 39.21% and 23.71% of the variance, respectively, with a cumulative explained variance of 62.92%. Although this cumulative contribution was moderate, consistent with the dispersed nature of volatile components in complex biological matrices, samples from Guangdong (A), Zimbabwe (B), and Yunnan (C) separated clearly along PC1 and PC2, showing tight within-group clustering and distinct between-group boundaries. This indicates that the three origins possess intrinsically different volatile metabolic profiles, rather than differences driven by individual outliers.

To improve classification and screen differential markers, a supervised PLS-DA model was constructed in Figure 4 and 5. PLS-DA substantially enhanced between-group discrimination. Cross-validation (200-fold) yielded R^2Y=0.7641, Q^2=-1.306, demonstrating good model fit, predictive ability, and robustness. These results support the reliability of VIP-based differential marker screening and provide a solid statistical basis for identifying specific markers.

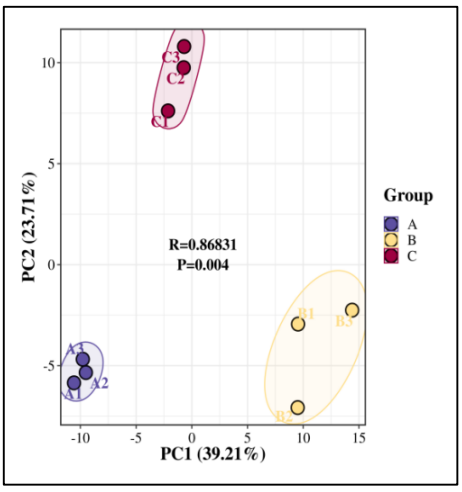


Figure 3 principal component analysis(PCA)

To further identify the key chemical markers responsible for metabolic differences among the three origins, VIP > 1.0 in the PLS-DA model was used as the threshold, and 82 differential compounds were screened (VIP 1.08–1.96). Terpenes, heterocycles, and short-chain organic acids were dominant, together accounting for more than 68% of the differential metabolites. This indicates fundamental differences among origins in terpene biosynthesis, Maillard reactions, and primary energy metabolism [4].

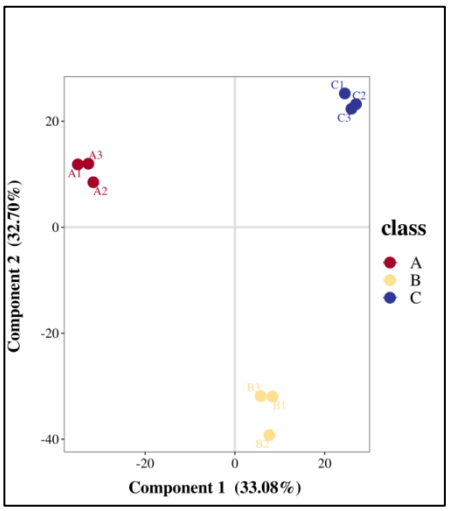


Figure 4 Partial Least Squares Discriminant Analysis

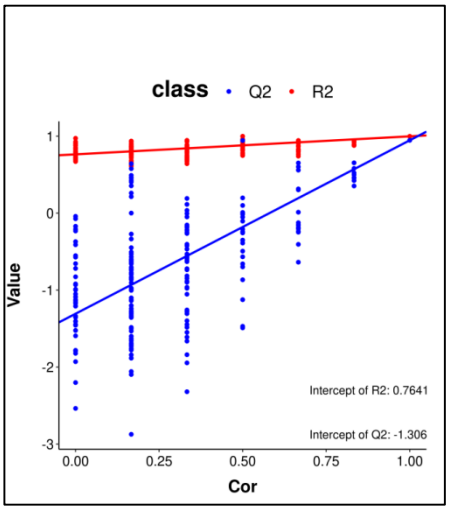


Figure 5 Permutation test of the PLS-DA model (200 permutations)

(1) VIP > 1.8: core discriminant factors. This tier included p-cymene (VIP = 1.96), 2,4,6-octatriene (1.92), bicyclo[3.1.0]hex-2-ene (1.90), (Z)-β-ocimene (1.86), and 5-(hydroxymethyl)dihydrofuran-2(3H)-one (1.82). Terpenoid markers such as p-cymene, ocimene, and bicyclic alkenes varied strongly among origins and were highly enriched in Yunnan. This likely reflects upregulation of the MEP pathway under Yunnan's high-altitude, high-UV conditions [8]. p-Cymene, a precursor of carvacrol and thymol, is a preferred quantitative marker for distinguishing Yun-

nan from low-altitude origins such as Guangdong and Zimbabwe. Heterocyclic/lactone compounds, including 5-(hydroxymethyl)dihydrofuran-2(3H)-one and 2-piperidone (1.81), are Maillard/Strecker degradation products and were higher in Yunnan and Zimbabwe than in Guangdong, indicating differences in protein and sugar degradation during curing [13].

(2) 1.5 ≤ VIP < 1.8. This tier included β-damascone (1.77), limonene (1.77), β-pinene (1.79), benzyl acetate (1.64), 2-cyclopenten-1-one (1.74), and propanoic acid (1.80). β-Damascone and benzyl acetate had VIP > 1.6 and very low sensory thresholds, supporting the floral, fruity, sweet, sour aroma style of Zimbabwe [9]. β-Damascone also showed an extremely high OAV in Zimbabwe (453,849), indicating that it is both a flavor-active compound and a chemical biomarker for Zimbabwe origin [7]. Propanoic acid was the only short-chain fatty acid with VIP > 1.8 and was much higher in Guangdong and Zimbabwe than in Yunnan, suggesting that acid accumulation is linked to fatty acid β-oxidation under specific temperature–humidity conditions [14].

(3) 1.0 < VIP < 1.5. This tier included maltol (1.30), vanillin (1.25), furfural (1.24), and long-chain fatty acid methyl esters such as methyl 9,12-octadecadienoate (1.16). Although individually less discriminatory than terpenes, these compounds showed origin-specific or extreme content differences. For example, the maltol peak area in Yunnan was 2-3 times that in Guangdong. In PLS-DA or OPLS-DA models, they can reduce residuals and improve Q2. Maltol and vanillin, despite VIP ≈ 1.3, are key contributors to caramel-sweet notes.

4 Discussion

This study compared the volatile metabolomes of tobacco from Guangdong (A), Zimbabwe (B), and Yunnan (C) using HS-SPME-GC-MS combined with multivariate statistics and aroma activity evaluation. The three origins differed markedly in volatile composition. A total of 197 compounds were identified. Yunnan had the most compounds (136) but the lowest total content, whereas Guangdong had the fewest compounds (106) but the highest total content. Alcohols and ketones were dominant classes, while acids, terpenes, and heterocycles were strongly origin-dependent. These patterns are consistent with the combined effects of genetic background, ecological conditions, and curing. The high diversity but low abundance of volatiles in Yunnan suggests a broad but less accumulated volatile pool, whereas the high total content in Guangdong indicates stronger accumulation of a narrower set of compounds, possibly giving a simpler but more intense aroma profile.

Multivariate models achieved clear origin discrimination. PCA explained 62.92% of the variance in the first two principal components, with samples clustering naturally by origin. PLS-DA showed significant discrimination ($R = 0.8683$, $P = 0.004$) and was validated as robust and not overfitted, confirming that volatile fingerprints can serve as reliable chemical fingerprints for origin authentication. VIP analysis identified 82 differential markers, which can be grouped into three core discriminant clusters. Terpenes such as p-cymene, ocimene, and pinene (VIP > 1.85) were highly enriched in Yunnan, likely reflecting activation of the MEP pathway under high-altitude, high-UV conditions; p-cymene is a promising quantitative marker for Yunnan origin. Norisoprenoid ketones and aromatic esters, such as β-damascone and benzyl acetate (VIP ≈ 1.70), accumulated in Zimbabwe. Given their low odor thresholds and the extremely high OAV of β-damascone, these compounds form the basis of Zimbabwe's intense floral–fruity style and are robust biomarkers. Short-chain acids and alkanes, such as propanoic acid and 3-methylvaleric acid (VIP ≈ 1.50), were higher in Guangdong, suggesting active lipid metabolism and a fresh, sweet, mild smoke profile. Future work should integrate sensory evaluation, transcriptomics, and targeted quantification to validate these markers across larger sample sets and multiple harvest years. The influence of curing and processing should also be decoupled from origin effects to support robust, simplified discriminant models for practical origin traceability and flavor regulation.

5 Conclusions

In summary, HS-SPME-GC-MS metabolomics combined with multivariate statistics and OAV analysis revealed origin-dependent volatile profiles among Guangdong, Zimbabwe, and Yunnan tobacco: 197 compounds were identified, with Yunnan richest in compound number (136) but lowest in total content, Guangdong fewest (106) but highest total content, and alcohols and ketones as dominant classes. PCA and PLS-DA achieved clear origin discrimination, and VIP > 1.0 screening yielded 82 differential markers. Terpenes such as p-cymene, ocimene, and pinene were enriched in Yunnan; norisoprenoid ketones and aromatic esters such as β-damascone and benzyl acetate accumulated in Zimbabwe; and short-chain acids and alkanes

such as propanoic acid and 3-methylvaleric acid were higher in Guangdong. OAV analysis further revealed a shared core aroma skeleton (β-damascenone, nonanal, and linalool; OAV > 1000) together with distinct origin-specific notes, supporting three ecological chemotypes: Yunnan (terpene/heterocycle-enriched), Zimbabwe (ketone/ester-synergistic), and Guangdong (acid/alcohol-based). Accordingly, p-cymene, β-damascone, propanoic acid, and maltol are recommended as core discriminant indicators for a simplified, accurate origin-identification model and for flavor-directed regulation.

References

[1] Chen L, Zhang J, Su G, Xie X, Shi X, Zhang L, et al. Deciphering the mechanism of light quality regulating the quality of sun-cured yellow tobacco based on GC-MS non-targeted metabolomics. Front Plant Sci. 2026, 17, 1770112. https://doi.org/10.3389/fpls.2026.1770112

[2] Wei K, Tang J, Qiu S, Chen S, Wu S, Zhao Y, et al. Integrative metabolomics and transcriptomics profiling reveals enhanced aroma precursor biosynthesis in tobacco leaves at the mature harvest stage. Ind Crop Prod. 2026, 249, 123647. https://doi.org/10.1016/j.indcrop.2026.123647

[3] Zhao Y, Li H, Zhong R, Feng L, Wang S, Lin Y, et al. Aromatic compounds analysis of Guizhou honey-sweet tobacco. Front Plant Sci. 2026, 17, 1878964. https://doi.org/10.3389/fpls.2026.1878964

[4] Wu Y, Zhao G, Li Y, Li G, Wang W, Yang L, et al. Metabolomic analysis of Yunnan cigar tobacco leaves: impact of geography and climate on flavor characteristics and machine learning-based origin traceability. Front Plant Sci. 2026, 16, 1703429. https://doi.org/10.3389/fpls.2025.1703429

[5] Chen C, Liu Y, Luo Y, Li S, Tang K, Xue X, et al. Comprehensive Characterization of Aroma Compounds in Four Different Varieties of Cigar Tobacco Leaves from China Using HS-SPME-GC × GC-TOF-MS Combined with Flavor Profiling. ACS Omega. 2025, 10(6), 5819-5828. https://doi.org/10.1021/acsomega.4c09407

[6] Jing Y, Chen W, Qiu X, Qin S, Gao W, Li C, et al. Exploring Metabolic Characteristics in Different Geographical Locations and Yields of *Nicotiana tabacum L*. Using Gas Chromatography-Mass Spectrometry Pseudotargeted Metabolomics Combined with Chemometrics. Metabolites. 2024, 14(4), 176. https://doi.org/10.3390/metabo14040176

[7] Chen X, Chen C, Liu Y, Tang K, Chen M, Qiu J, et al. Flavor Characteristics of Different Varieties of Flue-Cured Tobacco Based on Headspace Gas Chromatography-Ion Mobility Spectrometry and Headspace Solid-Phase Microextraction-Gas Chromatography-Mass Spectrometry. ACS Omega. 2025, 10(20), 20382-20401. https://doi.org/10.1021/acsomega.5c00343

[8] Zhao G, Kong G, Li M, Wu Y, Yao H, Li W, et al. Comparative metabolomic of cigar tobacco leaves: A preliminary investigation of potential discriminatory metabolites between Yunnan and Foreign. Biochemistry and Biophysics Reports. 2026, 45, 102461. https://doi.org/10.1016/j.bbrep.2026.102461

[9] Wu X, Liu S, Guo Z, Zhang F, Liu T, Dou Y. A comprehensive map of key aroma-active compounds in cigar tobacco via GC-IMS and GC-O-MS. Front Plant Sci. 2026, 17, 1749662. https://doi.org/10.3389/fpls.2026.1749662

[10] Zhao L, Li Y, Cai J, Guo Z, Zhang S, Shang S, et al. Volatile Profiles of Different Tobacco Cultivars and their Correlation with Sensory Quality of Heated Tobacco. Curr Anal Chem. 2025, 21(4), 345-355. https://doi.org/10.2174/0115734110350070241108054934

[11] Zhang M, Guo D, Wu G, Han P, Shi Y, Zheng T, et al. Analysis of volatile compound metabolic profiles during the fermentation of filler tobacco leaves through integrated E-nose, GC-MS, GC-IMS, and sensory evaluation. Journal of Chromatography. A. 2024, 1737, 465472. https://doi.org/10.1016/j.chroma.2024.465472

[12] Zhao M, Jiang Y, Hu B, Li R, He J, Wang Z, et al. Unraveling the key chemical components transformation and aroma formation mechanisms during tobacco curing-drying process: A comprehensive review. Dry Technol. 2026, 44(8), 1064-1093. https://doi.org/10.1080/07373937.2026.2642377

[13] Shen X, Wang H, Yao L, Sun M, Yu C, Kang W, et al. Microbial communities and carotenoid metabolism associated with aroma-related volatile profiles in fermented cigar tobacco leaves. Industrial Crops and Products. 2026, 245, 123286. https://doi.org/10.1016/j.indcrop.2026.123286

[14] Zhao L, Shang S, Tian Y, Gao Y, Song Z, Peng L, et al. Integrative analysis of sensory evaluation and non-targeted metabolomics to unravel tobacco leaf metabolites associated with sensory quality of heated tobacco. Front Plant Sci. 2023, 14, 1123100. https://doi.org/10.3389/fpls.2023.1123100