Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Authenticity identification of wasabi (Eutrema wasabi) and its analogues by integrated genomic, barcoding-machine learning, and nutritional analysis

Mengdi Zheng

hakuna_matata_mm@qq.com

Xi'an Medical University

Yibo Ma

Shenyang Pharmaceutical University

Yangrui Li

Nanjing Sanhome Pharmaceutical Co., Ltd

Jiayin Guo

Xi'an Medical University

Miaomiao Li

Xi'an Medical University

Yulei Li

Xi'an Medical University

Research Article

Keywords: Eutrema wasabi, Food authenticity, DNA barcoding, Machine learning, Chloroplast genome

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11499665/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Authenticity identification of wasabi (*Eutrema wasabi*) and its analogues by integrated genomic, barcoding-machine learning, and nutritional analysis

Mengdi Zheng[1]*, Yibo Ma[2], Yangrui Li[3], Jiayin Guo[1], Miaomiao Li[1], Yulei Li[1]
(1. Xi'an Medical University, Xi'an, 710021, China; 2. Shenyang Pharmaceutical University, Benxi, 117004, China; 3. Nanjing Sanhome Pharmaceutical Co., Ltd., Nanjing 210018 China)

Abstract: Wasabi's high market value drives its adulteration with taxonomically related analogues, compromising food quality and safety. We developed an integrated approach combining chloroplast genome comparison, DNA barcoding with machine learning classifiers, and nutritional profiling to distinguish wasabi from adulterants. Among seven Brassicaceae species, chloroplast genomes were largely conserved but exhibited species-specific variations in IR boundaries, SSRs, hypervariable regions, and weak A/U-ended codon bias. Of three barcodes (ITS2, *matK*, *rbcL*), ITS2 showed the clearest barcoding gap and best phylogenetic resolution, while machine learning (BLOG and WEKA) identified *matK* as the top marker with 100% accuracy. Nutritional analysis revealed significant differences in flavonoids, polyphenols, and vitamin C across species. This framework offers a robust authenticity solution integrating genomic, molecular, and nutritional data.

Keywords: *Eutrema wasabi*; Food authenticity; DNA barcoding; Machine learning; Chloroplast genome

1. Introduction

Food authenticity has become a global concern, particularly in the condiment industry, where high-value products are frequently adulterated with cheaper substitutes (Wang et al., 2025; Galvin-King et al., 2018). Wasabi (*Eutrema wasabi*) is a characteristic example. Its unique pungent flavor, derived from isothiocyanates such as allyl isothiocyanate (AITC) and 6-(methylsulfinyl)hexyl isothiocyanate (6-MSITC), together with its potential antioxidant and neuroprotective activities, positions it as a premium ingredient in global cuisine (García-Yagüe et al., 2025; Bartkowiak-Wieczorek et al., 2024). However, wasabi cultivation demands stringent conditions—low-temperature running water, shade, and a long growth cycle of 1.5–3 years—resulting in low yield and high market price (Lin, 2008). This economic gap has driven widespread substitution by more accessible species with similar organoleptic properties, notably horseradish (*Armoracia rusticana*) and mustard (*Brassica juncea*) (Cammue et al., 2016).

Although these substitutes may mimic wasabi's pungency to some extent, they differ significantly in flavor profile, bioactive composition, and importantly, allergenic potential—mustard being a declared allergen in the EU (Regulation No. 1169/2011) (Wang et al., 2025a, 2025b). Traditional identification methods based on morphology or sensory evaluation become ineffective after processing into pastes or powders. Chemical methods such as HPLC profiling of isothiocyanates are compromised by heat- or pH-induced degradation during processing (Bartkowiak-Wieczorek et al., 2024). Therefore, rapid and reliable methods to authenticate wasabi and detect its substitutes are urgently needed for quality control and consumer safety.

DNA-based approaches offer species-specific, stable, and traceable markers. DNA barcoding using standardized regions such as ITS2, *matK*, and *rbcL* has been widely applied in food authentication (Martinović et al., 2025; Zheng et al., 2022). ITS2 evolves rapidly and is suitable for distinguishing closely related species, while *matK* and *rbcL* are more conserved and useful for higher-level phylogenetics (Zheng et al., 2022). Recently, machine learning algorithms have been introduced to DNA barcoding analysis, enabling automated extraction of diagnostic nucleotide patterns and improving discrimination of closely related adulterants compared to traditional distance- or tree-based methods (Zheng et al., 2025; Zheng et al., 2026). Meanwhile, the chloroplast genome, with its conserved quadripartite structure and variable regions (e.g., IR boundaries, SSRs, hypervariable sites), serves as a valuable resource for marker development and evolutionary analysis (Zheng et al., 2026; Hejazi et al., 2025). Codon usage bias further reflects molecular evolutionary pressures and can aid in understanding sequence conservation patterns.

Despite existing studies on phylogenetic relationships within Brassicaceae and molecular identification of individual species, systematic comparative research specifically targeting wasabi and its common substitutes remains insufficient. Previous work has largely focused on single-species genome sequencing or evaluation of individual barcode fragments (Chen et al., 2021), lacking comprehensive comparisons from complete chloroplast genomes. Moreover, evaluations combining DNA barcoding with machine learning classification for wasabi authenticity identification are scarce. Given the similarities and differences in taxonomy, flavor, and processing applications among wasabi, horseradish, and mustard, relying solely on traditional barcodes may not meet the rapid identification demands of complex market samples.

To address these gaps, this study presents an integrated three-module framework. The first module compares complete chloroplast genomes and codon usage bias of seven Brassicaceae species to reveal evolutionary background. The second module evaluates three DNA barcodes (ITS2, *matK*, *rbcL*) using both phylogenetic methods and machine learning classifiers (BLOG, WEKA) for species discrimination. Beyond genetic identification, nutritional and functional components are critical for quality evaluation of food products. The third module quantifies total flavonoids, total polyphenols, and vitamin C to compare nutritional profiles of wasabi, horseradish, and mustard. Together, this framework provides molecular markers for authenticity identification and insights into nutritional differences among authentic and substitute products, supporting both food safety and quality evaluation in the condiment industry.

2. Materials and methods

2.1 Sample collection, sequence data, and PCR amplification

Fresh rhizomes of wasabi (*Eutrema wasabi*), horseradish (*Armoracia rusticana*), and mustard (*Brassica juncea*) were collected from Baoshan (Yunnan), Sanmenxia (Henan), and Linyi (Shandong), respectively. Genomic DNA was extracted from 100 mg of each sample (three biological replicates) using a commercial kit (DP305, Tiangen) according to the manufacturer's instructions. PCR primers for ITS2, *matK,* and *rbcL* were used as previously described (Zheng et al., 2022; 2025; 2026).

Chloroplast genomes of seven Brassicaceae species (*E. wasabi, E. heterophyllum, E. yunnanense, A. rusticana, B. juncea, B. nigra, S. alba*) were retrieved from NCBI (accessions: NC028727.1, NC028728.1, OR588868.1, MZ328719.1, NC028272.1, NC030450.1, NC045948.1). CDSs were annotated with CPGAVAS2; redundant and short (<300 bp) or incomplete sequences were excluded, retaining only those with intact start/stop codons for codon usage analysis. Barcode sequences (ITS2, *matK*, *rbcL*) for the same species were downloaded from NCBI; ITS2 was trimmed of flanking 5.8S/28S using the ITS2 Database. All sequences were aligned and trimmed with Clustal Omega, and laboratory-generated sequences from our samples were added. Accession numbers and sample codes are listed in Table S1.

2.2 Chloroplast genome comparison and codon usage bias analysis

IR boundaries were analyzed using IRscope, SSR loci with MISA, and whole-genome sequences with mVISTA. Phylogenetic analysis was conducted in MEGA using the Neighbor-Joining (NJ) method (Kimura 2-parameter model, 1000 bootstrap replicates) (Zheng et al., 2025).

Codon usage analysis was performed using CodonW1.4.2 and EMBOSS CUSP (Zheng et al., 2026). The following parameters were calculated: base composition at each codon position (A3, T3, C3, G3), GC1, GC2, GC3, GC3s, GCall, GC12, CAI, CBI, Fop, ENC, and RSCU. Neutrality plotting was conducted with GC12 against GC3. ENC-plot (ENC vs. GC3s) included the expected curve: $ENC = 2 + GC3 + 29/[GC3^2 + (1-GC3)^2]$. PR2-plot was performed using G3/(G3+C3) and A3/(A3+T3) as coordinates. Correspondence analysis (COA) was performed on RSCU values using CodonW. Optimal codons were identified based on the top and bottom 10% of genes by ENC value; codons with $\Delta RSCU > 0.08$ and $RSCU > 1$ were considered optimal. PCA was performed on RSCU values of 59 synonymous codons (excluding stop codons, Met, and Trp) using Origin 2021 Pro (covariance matrix; 409 gene - species combinations). Correlation analyses were performed using ChiPlot.

2.3 DNA barcoding and machine learning-based molecular identification

Sequence characteristics (GC content, conserved sites, variable sites, parsimony-informative sites) were calculated using MEGA. Genetic distances (intraspecific and interspecific) were calculated using the Kimura 2-parameter model, and NJ phylogenetic trees were constructed. Secondary structures of ITS2 were predicted using the ITS2 Database and aligned using 4Sale, followed by PNJ tree construction using ProfDistS. Secondary structures of *matK* and *rbcL* were predicted using RNAfold.

For machine learning, the BLOG algorithm randomly divided sequences into training and query sets to extract diagnostic nucleotide logic rules. On the WEKA platform, four classifiers (NaiveBayes, J48, SMO, JRip) were used (Feng, 2021; Zheng et al., 2026). Correct classification rate, error rate, and unclassified rate were used as evaluation metrics.

2.4 Nutritional composition analysis

Fresh samples of wasabi, horseradish, and mustard were cleaned, surface-dried, chopped, and mixed. Total flavonoids and total polyphenols were extracted with 70% ethanol under ultrasonication, while vitamin C was extracted with 60% ethanol. Total flavonoid content was determined by the sodium nitrite-aluminum nitrate-sodium hydroxide colorimetric method at 510 nm using a rutin standard curve. Total polyphenol content was determined by the Folin-Ciocalteu colorimetric method at 760 nm using a gallic acid standard curve. Vitamin C content was determined by the phosphomolybdic acid colorimetric method at 760 nm using an ascorbic acid standard curve. All assays were performed in triplicate in 96-well plates using a microplate reader. Results were expressed as mean ± standard deviation (Singleton et al., 1999; Jia et al., 1999).

3. Results and discussion

3.1 Chloroplast genome comparison and codon usage bias

3.1.1 IR boundary dynamics and junctional plasticity

The chloroplast genomes of all seven species exhibited a typical quadripartite structure, comprising a large single copy (LSC) region, a small single copy (SSC) region, and two inverted repeat (IRa and IRb) regions (Fig. 1A). The total genome lengths ranged from 153,483 bp (*B. juncea*) to 155,136 bp (*A. rusticana*), with *E. wasabi* measuring 153,851 bp. While the overall genomic architecture was highly conserved across species, subtle variations were detected at the four junctions (JLB, JSB, JSA, and JLA). Notably, *A. rusticana* (horseradish), the primary substitute for wasabi, displayed distinct characteristics including the longest total genome length, an expanded SSC region (18,097 bp), and a unique 188-bp intergenic spacer at the JLB junction. In contrast, the three *Eutrema* species (*E. wasabi*, *E. yunnanense*, and *E. heterophyllum*) showed remarkable similarity in genome size and boundary gene organization, necessitating higher-resolution markers for intrageneric discrimination. These junctional micro-variations provide candidate loci for DNA barcoding to authenticate wasabi and distinguish it from common adulterants.

mVISTA comparison of seven Brassicaceae chloroplast genomes, using *E. wasabi* as reference, revealed high overall sequence similarity, with coding regions (e.g., *rpoC1*, *rbcL*, *ycf2*) markedly more conserved than non-coding regions (Fig. 1B). Non-coding intergenic spacers and boundary junctions-including trnK-rps16, rpl22-rps19, rps8-rpl14, petB-petD, and *ycf1* boundaries—exhibited pronounced similarity drops, indicating accumulation of indels and substitutions. Interspecifically, *Eutrema* analogues clustered closest to *E. wasabi*, whereas *A. rusticana*, *Brassica spp.*, and *S. alba* showed greater divergence in these variable loci. In contrast, the ~100–136 kb IR region (containing *rrn*, *trn*, *ndhB*, *rps12*) remained conserved across all species. Collectively, despite overall structural conservation, non-coding and boundary regions harbour sufficient variation to serve as promising targets for developing diagnostic molecular markers, supporting the authentication of wasabi and detection of its common substitutes in processed food products.

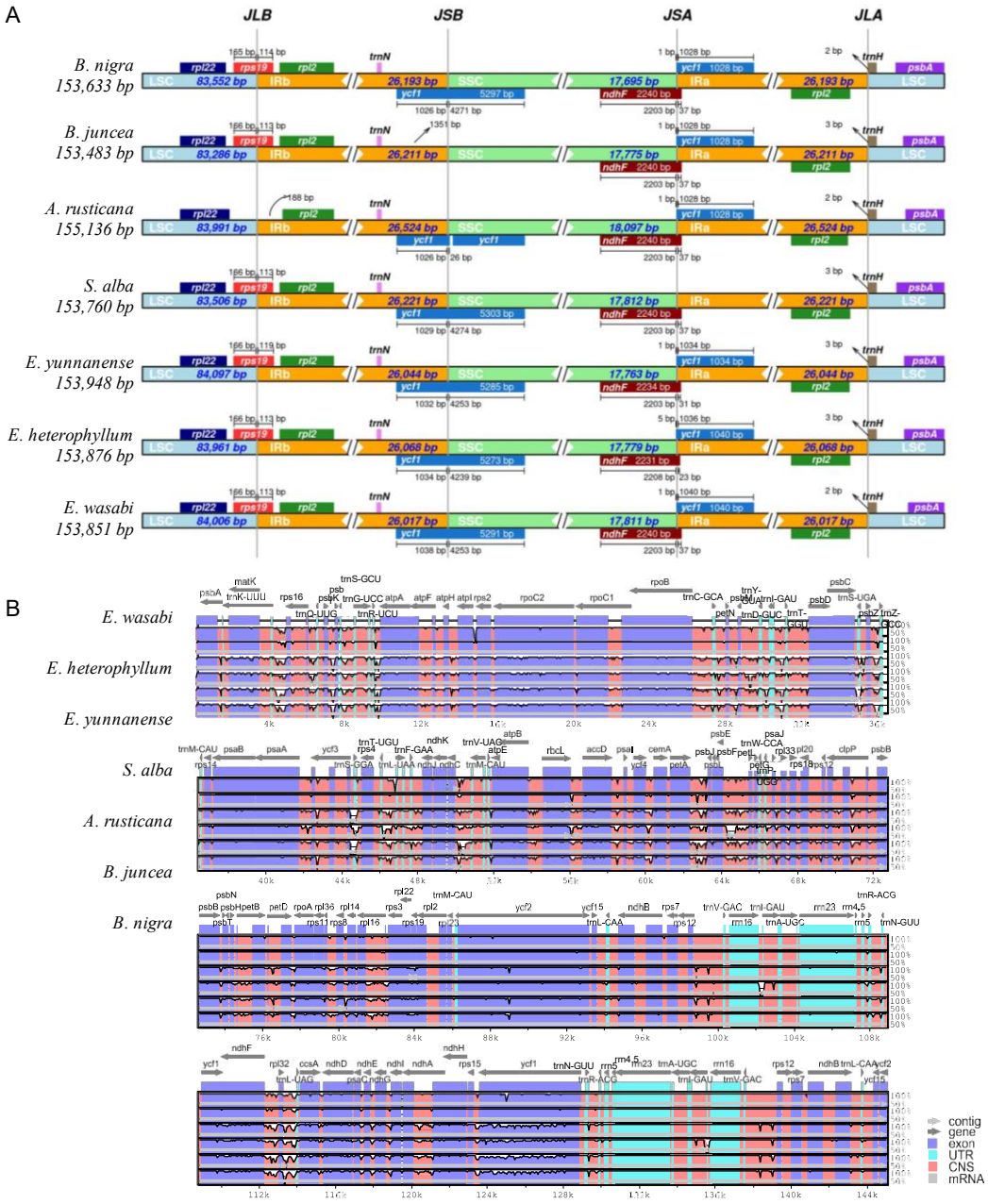


Figure 1. Structural organization and sequence divergence of chloroplast genomes from seven Brassicaceae species. (A) Inverted repeat (IR) boundary analysis showing the four junctions (JLB, JSB, JSA, and JLA) and the total genome sizes (in kb) for each species.

(B) mVISTA-based pairwise alignment of the complete chloroplast genomes, using *Eutrema wasabi* as the reference. Gray arrows indicate gene orientation and transcription direction. Colored regions denote: purple-blue for exons/encoding sequences (CDS), light blue for untranslated regions (UTRs), and red for conserved non-coding sequences (CNS). The black similarity curves represent the sequence identity relative to the reference, with the vertical scale ranging from 50% to 100%. The seven species analyzed (from top to bottom) are *E. wasabi* (reference), *E. heterophyllum, E. yunnanense, Sinapis alba, Armoracia rusticana, Brassica juncea*, and *B. nigra*.

3.1.2 SSR profiling and species-specific repeat signatures

As shown in Figure S1, a total of 64 to 79 SSR loci meeting the screening criteria were identified across the seven chloroplast genomes. Three types of SSRs were detected:

mononucleotide, dinucleotide, and trinucleotide repeats, with each species distinguishable by the quantity and composition of these repeat types. *Armoracia rusticana* (horseradish) harbored the highest number of SSR loci (79), whereas *Eutrema wasabi* (wasabi) contained the fewest (64). Among all seven species, mononucleotide repeats were the most abundant, predominantly composed of A/T motifs, accounting for 82.35% to 91.14% of the total SSRs. This pronounced AT-bias reflects the characteristic nucleotide composition of chloroplast genomes, resulting in the majority of SSR loci being constituted by polyA or polyT repeat sequences.

3.1.3 Complete chloroplast genome phylogeny and taxonomic authentication

The NJ phylogenetic tree based on complete chloroplast genomes resolved all seven Brassicaceae species with 100% bootstrap support (Fig. S2). The three *Eutrema* species formed a monophyletic clade, with *Armoracia rusticana* as their sister group; *Brassica juncea*, *B. nigra* and *Sinapis alba* constituted another well‑supported lineage, consistent with traditional taxonomy. The robust resolution confirms that complete chloroplast genome data reliably distinguish *E. wasabi* from its common adulterants (*A. rusticana*, *Brassica spp.*), providing a solid phylogenetic basis for molecular authentication of wasabi products.

3.1.4 Codon base composition and usage bias parameters

Base composition analysis (Table S2) revealed highly similar patterns across all seven species: GC1 (44.74–45.07%) > GC2 (37.33–37.54%) > GC3 (28.81–29.20%), with GC3s ranging from 28.90% to 29.10%. The markedly lower GC3/GC3s values indicate a strong A/T (U) preference at the third synonymous codon position. Codon usage metrics—CAI (0.1661–0.1679), CBI (−0.1098 to −0.0966), Fop (0.3511–0.3554), and ENC (47.42–47.55)—showed minimal interspecific variation and consistently pointed to weak overall codon usage bias. The low CAI and high ENC values suggest relaxed translational selection pressure, typical of plant chloroplast genomes(Wang et al., 2026a, 2026b). Taken together, these parameters confirm that chloroplast coding sequences are evolutionarily conserved among the seven species, with limited discriminatory power for species-level authentication. Nevertheless, these findings establish a foundational baseline for distinguishing mutation--driven versus selection--driven evolutionary forces, and support the integration of codon usage data with other variable markers in the multi--tiered wasabi authentication framework.

3.1.5 Correlations among codon usage indices

Correlation analysis among nine codon usage indices (Fig. S3) revealed that GC_all positively correlated with GC1, GC2 and GC3, with GC1‑GC2 showing stronger mutual correlation than their correlations with GC3, reflecting the compositional independence of the wobble position. CAI correlated positively with CBI and Fop, confirming their concordance in quantifying translational selection. GC3s exhibited significant positive correlations with ENC, indicating that third‑position synonymous GC content is closely linked to overall codon bias variation. Importantly, GC_all and ENC showed no pronounced correlation, suggesting that total GC content alone cannot explain codon usage shifts. These patterns imply that mutation pressure and natural selection differentially affect overall nucleotide composition versus codon‑level

preferences, justifying the subsequent neutrality, ENC‑ and PR2‑plot analyses. In summary, these correlations confirm the conserved codon usage landscape among the seven species, with limited inter‑species discriminatory value, supporting their role as auxiliary rather than primary markers in the authentication framework.

3.1.6 Neutrality plot reveals combined effects of mutation and selection on codon usage

Neutrality plot analysis (Fig. S4) across seven Brassicaceae species yielded regression slopes of 0.15‑0.35, deviating from both 0 and 1, indicating that both mutation pressure and natural selection synergistically shape codon usage, with the shallow slope suggesting a predominant role of selection at the first two codon positions. *Eutrema wasabi* and its analogues showed nearly overlapping scatter distributions and comparable slopes, reflecting shared evolutionary history, whereas Armoracia rusticana and Brassica species exhibited slightly steeper slopes and more dispersed GC3 values, implying relaxed constraints or distinct mutational profiles. The clustering of Eutrema versus the more dispersed adulterants supports codon usage parameters as auxiliary discriminative features for machine‑learning authentication, though their limited interspecific divergence necessitates integration with other high‑resolution markers.

3.1.7 ENC-plot reveals significant natural selection beyond mutational bias

ENC‑plot analysis (Fig. S5) showed that most chloroplast genes in all seven species fell below the expected curve, indicating significant natural selection beyond mutational bias, particularly at intermediate GC3s (~25‑35%). Eutrema wasabi and its analogues exhibited near‑identical profiles, while *Armoracia rusticana* and *Brassica* species showed slightly broader dispersion, implying relaxed constraints. Combined with neutrality plots, these results confirm that both forces shape codon usage, yet the limited interspecific divergence reinforces that codon metrics alone are insufficient and should be integrated with higher‑resolution markers for robust authentication.

3.1.8 PR2-plot reveals A/U-ending bias and limited interspecific divergence in codon usage

PR2‑plot analysis (Fig. S6) revealed that chloroplast genes of all seven species deviated from the central null expectation, clustering predominantly in the lower‑left quadrant (A3/(A3+T3) < 0.5, G3/(G3+C3) < 0.5), reflecting a characteristic A/U‑ending preference. *Eutrema wasabi* and its analogues exhibited tightly clustered, near‑identical distributions, underscoring their shared mutational and selective regimes. In contrast, *Armoracia rusticana* and *Brassica species* showed slightly more dispersed patterns, suggesting relaxed constraints. Collectively, the consistent A/U bias across all species confirms that codon usage metrics alone offer limited species‑discriminatory power, supporting their role as auxiliary rather than primary markers within the integrated genomic‑nutritional authentication framework.

3.1.9 RSCU-based correspondence analysis of codon usage variation

Correspondence analysis (COA) based on RSCU values (Fig. S7) showed that, in all seven species, the majority of chloroplast genes clustered near the origin with limited dispersion,

indicating codon usage patterns close to the global mean. *Eutrema wasabi* and its analogues exhibited nearly overlapping distributions, reflecting close phylogenetic affinity and shared constraints, whereas *Armoracia rusticana* and *Brassica* species displayed slightly broader dispersion with a few outliers, suggesting subtle lineage‐specific heterogeneity. Importantly, COA primarily visualizes RSCU variation landscape without implying causality. Integrated with correlation, neutrality, ENC‐ and PR2‐plot analyses, these results confirm that codon usage bias is shaped by multiple interacting factors—with natural selection playing a prominent role—yet offers limited species‐discriminatory power, reinforcing its auxiliary role in the multi‐tiered authentication framework.

3.1.10 Optimal codon identification reveals conserved A/U-ending preference and limited interspecific divergence

Optimal codons were identified using the top and bottom 10% of ENC‐ranked genes, with ΔRSCU > 0.08 and RSCU > 1 as criteria (Fig. 2). Each species harboured 8‐9 optimal codons; the three Eutrema species shared nine (GCA, GGA, CAU, UUG, AAU, AGA, UCA, ACA, UAU), while *Armoracia rusticana*, *Brassica juncea* and *B. nigra* each had eight, and *Sinapis alba* differed by only one codon. Seven optimal codons (GCA, GGA, UUG, AAU, AGA, UCA, ACA) were universally conserved, six of which end in A/U (UUG being the only exception), confirming the A/U‐ending preference. The near‐identical optimal codon repertoires between *E. wasabi* and its analogues, and the minimal divergence from substitutes, indicate that optimal codon composition alone is insufficient for species discrimination. Nevertheless, it provides a foundational reference for understanding translational selection within the integrated authentication framework.

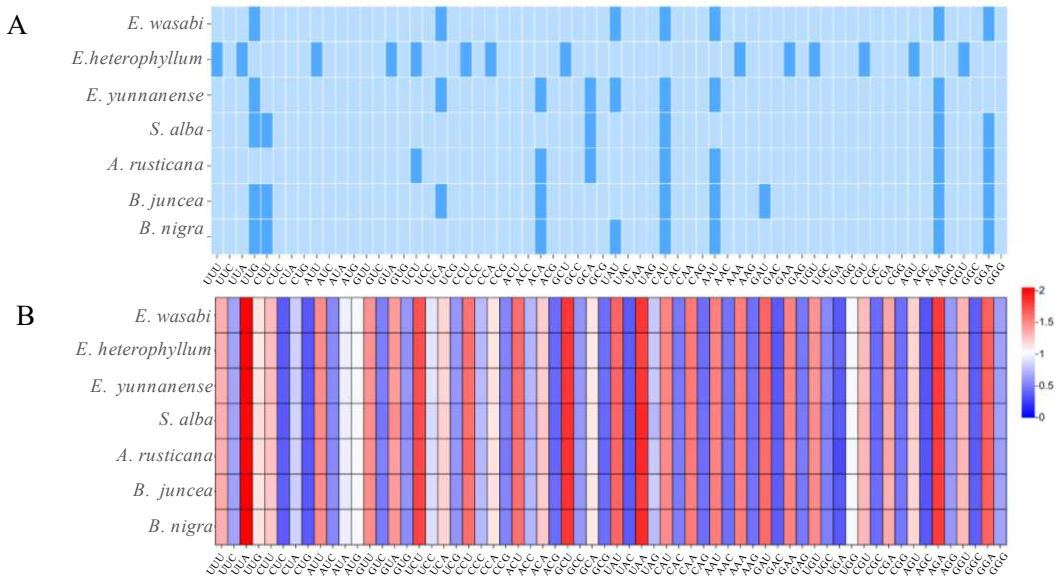


Figure 2. Optimal codon identification and RSCU patterns across seven Brassicaceae species. (A) Dark blue: optimal codons in each species (RSCU > 1 and △RSCU ≥ 0.08); light blue: codons with RSCU > 1 only. (B) RSCU heatmap showing relative synonymous codon usage values, with red indicating higher RSCU and blue indicating lower RSCU. The three *Eutrema* species share an identical set of nine optimal codons, while *Armoracia rusticana*, *Brassica juncea*, and *B. nigra* possess eight, and *Sinapis alba* differs by one codon. Seven optimal codons (GCA, GGA, UUG, AAU, AGA, UCA, ACA) are universally conserved, predominantly ending with A/U.

3.1.11 PCA reveals dispersed RSCU variation without species-specific clustering

As shown in Figure 3, principal component analysis (PCA) was performed on 59 synonymous codon RSCU values across 409 gene-species combinations. The first four principal components—PC1 (9.16%), PC2 (7.86%), PC3 (6.68%), and PC4 (6.35%)—collectively explained only ~30.05% of total variance (Fig. 3A), indicating that RSCU differences are dispersed across numerous codon variables rather than concentrated in major axes. In all six 2D projections, sample points from all seven species clustered near the origin with substantial overlap, lacking species-specific groupings; *Eutrema wasabi*, *E. heterophyllum*, and *E. yunnanense* were extensively intermingled, while *Armoracia rusticana*, *B. juncea*, *B. nigra*, and *S. alba* remained largely overlapping with the *Eutrema* cluster. The 3D PCA (Fig. 3B) corroborated these findings, with most samples occupying a shared spatial envelope without species-stratified layering. Combined with the conserved A/U-ending optimal codons (Fig. 2), these results confirm that chloroplast RSCU patterns are highly similar across all seven species with only dispersed, minor differences, underscoring the need to integrate RSCU data with other genomic and nutritional markers for robust authentication.

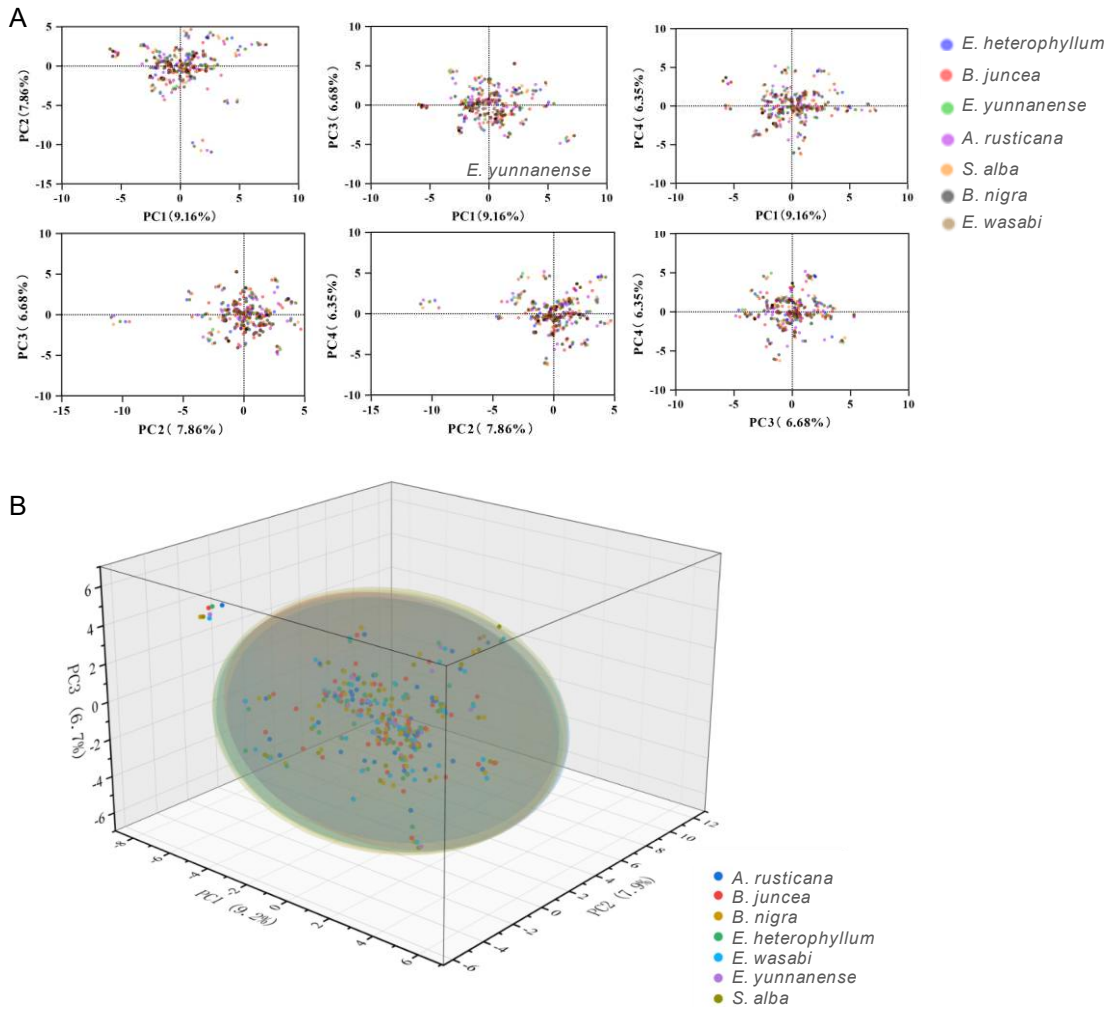


Figure 3. Principal component analysis (PCA) of relative synonymous codon usage (RSCU) across seven Brassicaceae species. (A) Two-dimensional projections of PC1-PC2, PC1-PC3, PC1-PC4, PC2-PC3, PC2-PC4,

and PC3-PC4. PC1, PC2, PC3, and PC4 explain 9.16%, 7.86%, 6.68%, and 6.35% of variance, respectively. All species (color-coded) cluster near the origin with extensive overlap, lacking species-specific separation. (B) Three-dimensional projection (PC1, PC2, PC3) revealing a shared spatial distribution without clear species-stratified layering. The translucent ellipsoid encompasses the majority of samples across all species, highlighting the conserved codon usage landscape.

3.2 Molecular identification based on DNA barcodes

3.2.1 Sequence characteristics and divergence of DNA barcode regions

Three core DNA barcodes (ITS2, *rbcL*, and *matK*) were characterized in *Eutrema wasabi* and its related or substituted species (Table S3). ITS2 (35 sequences, 255 bp, GC 58.15%) contained 40 variable sites (15.69%, 5 singletons, 35 parsimony-informative), indicating substantial interspecific divergence. *rbcL* (36 sequences, 555 bp, GC 44.38%) had only 21 variable sites (3.78%, 0 singletons, 21 parsimony-informative), reflecting strong conservation that limits species-level discrimination. *matK* (40 sequences, 736 bp, GC 30.83%) showed the highest variation with 68 variable sites (9.24%, 1 singleton, 67 parsimony-informative), offering the richest informative content. Collectively, ITS2 provides a useful barcoding gap for closely related species, *matK* delivers the most abundant informative variation, and *rbcL* is more suitable for higher-level phylogenetic inference, establishing a robust multi-locus framework for authenticating *E. wasabi* and detecting its substitutes.

3.2.2 Multi-locus barcoding and RNA secondary structure analysis for wasabi authentication

Phylogenetic trees (NJ and PNJ) based on ITS2, *matK* and *rbcL* (Fig. 4) showed that ITS2 provided clear intergeneric separation (98% bootstrap) but limited intrageneric resolution (35‑52%). In contrast, *matK* gave the strongest intrageneric support (78‑100%), while *rbcL* offered moderate phylogenetic stability (29‑100%). Genetic distance analysis (Table S4) revealed that *matK* had the widest barcoding gap (intraspecific 0.0000‑0.0008; interspecific 0.0061‑0.0616), ITS2 showed moderate divergence (0.0000‑0.0102 vs 0.0154‑0.0906), and *rbcL* displayed overlapping ranges (0.0000‑0.0013 vs 0.0000‑0.0275), compromising its discriminatory power. Distances from *E. wasabi* to other species confirmed *matK* as the most informative single locus. RNA secondary structures (Fig. 5) further distinguished *E. wasabi* from all other species through unique loop and helix arrangements. Collectively, the multi‑locus integration—ITS2 for intergeneric separation, *matK* for rich variable sites and an optimal barcoding gap, and *rbcL* for phylogenetic background—combined with species‑specific structural diagnostics, provides a robust framework for authenticating wasabi and detecting its primary adulterant, *A. rusticana*.

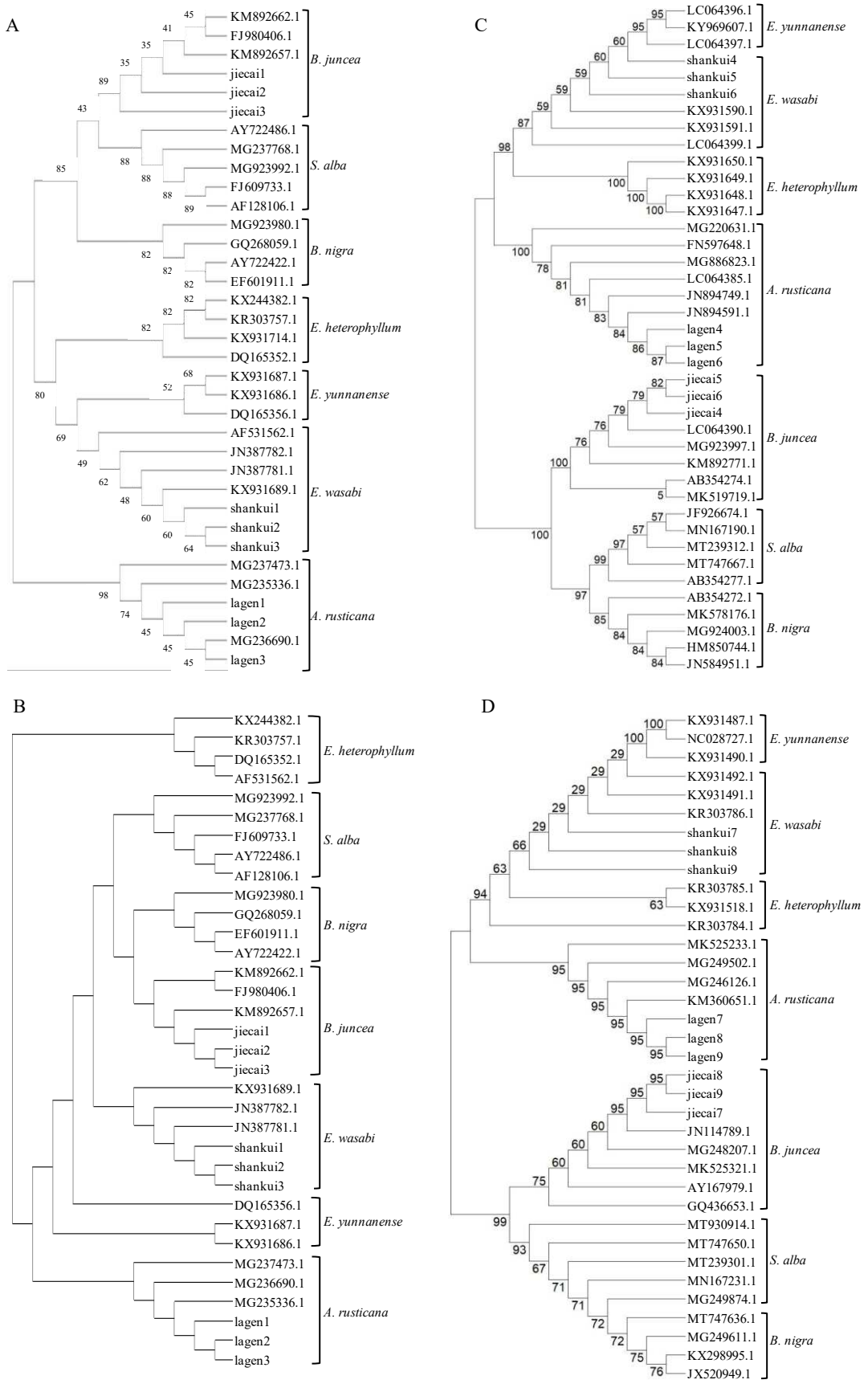

A
KM892662.1
FJ980406.1
KM892657.1
jiecai1
jiecai2
jiecai3
B. juncea
AY722486.1
MG237768.1
MG923992.1
FJ609733.1
AF128106.1
S. alba
MG923980.1
GQ268059.1
AY722422.1
EF601911.1
B. nigra
KX244382.1
KR303757.1
KX931714.1
DQ165352.1
E. heterophyllum
KX931687.1
KX931686.1
DQ165356.1
E. yunnanense
AF531562.1
JN387782.1
JN387781.1
KX931689.1
shankui1
shankui2
shankui3
E. wasabi
MG237473.1
MG235336.1
lagen1
lagen2
MG236690.1
lagen3
A. rusticana
B
KX244382.1
KR303757.1
DQ165352.1
AF531562.1
E. heterophyllum
MG923992.1
MG237768.1
FJ609733.1
AY722486.1
AF128106.1
S. alba
MG923980.1
GQ268059.1
EF601911.1
AY722422.1
B. nigra
KM892662.1
FJ980406.1
KM892657.1
jiecai1
jiecai2
jiecai3
B. juncea
KX931689.1
JN387782.1
JN387781.1
shankui1
shankui2
shankui3
E. wasabi
DQ165356.1
KX931687.1
KX931686.1
E. yunnanense
MG237473.1
MG236690.1
MG235336.1
lagen1
lagen2
lagen3
A. rusticana
C
LC064396.1
KY969607.1
LC064397.1
E. yunnanense
shankui4
shankui5
shankui6
KX931590.1
KX931591.1
LC064399.1
E. wasabi
KX931650.1
KX931649.1
KX931648.1
KX931647.1
E. heterophyllum
MG220631.1
FN597648.1
MG886823.1
LC064385.1
JN894749.1
JN894591.1
lagen4
lagen5
lagen6
A. rusticana
jiecai5
jiecai6
jiecai4
LC064390.1
MG923997.1
KM892771.1
AB354274.1
MK519719.1
B. juncea
JF926674.1
MN167190.1
MT239312.1
MT747667.1
AB354277.1
S. alba
AB354272.1
MK578176.1
MG924003.1
HM850744.1
JN584951.1
B. nigra
D
KX931487.1
NC028727.1
KX931490.1
E. yunnanense
KX931492.1
KX931491.1
KR303786.1
shankui7
shankui8
shankui9
E. wasabi
KR303785.1
KX931518.1
KR303784.1
E. heterophyllum
MK525233.1
MG249502.1
MG246126.1
KM360651.1
lagen7
lagen8
lagen9
A. rusticana
jiecai8
jiecai9
jiecai7
JN114789.1
MG248207.1
MK525321.1
AY167979.1
GQ436653.1
B. juncea
MT930914.1
MT747650.1
MT239301.1
MN167231.1
MG249874.1
S. alba
MT747636.1
MG249611.1
KX298995.1
JX520949.1
B. nigra

Figure 4. Phylogenetic trees based on three DNA barcodes for *Eutrema wasabi* and related or substituted species. (A) ITS2 neighbor-joining (NJ) tree; (B) ITS2 parsimony neighbor-joining (PNJ) tree; (C) *matK* NJ tree; (D) *rbcL* NJ tree. Bootstrap support values (%) are indicated at nodes.

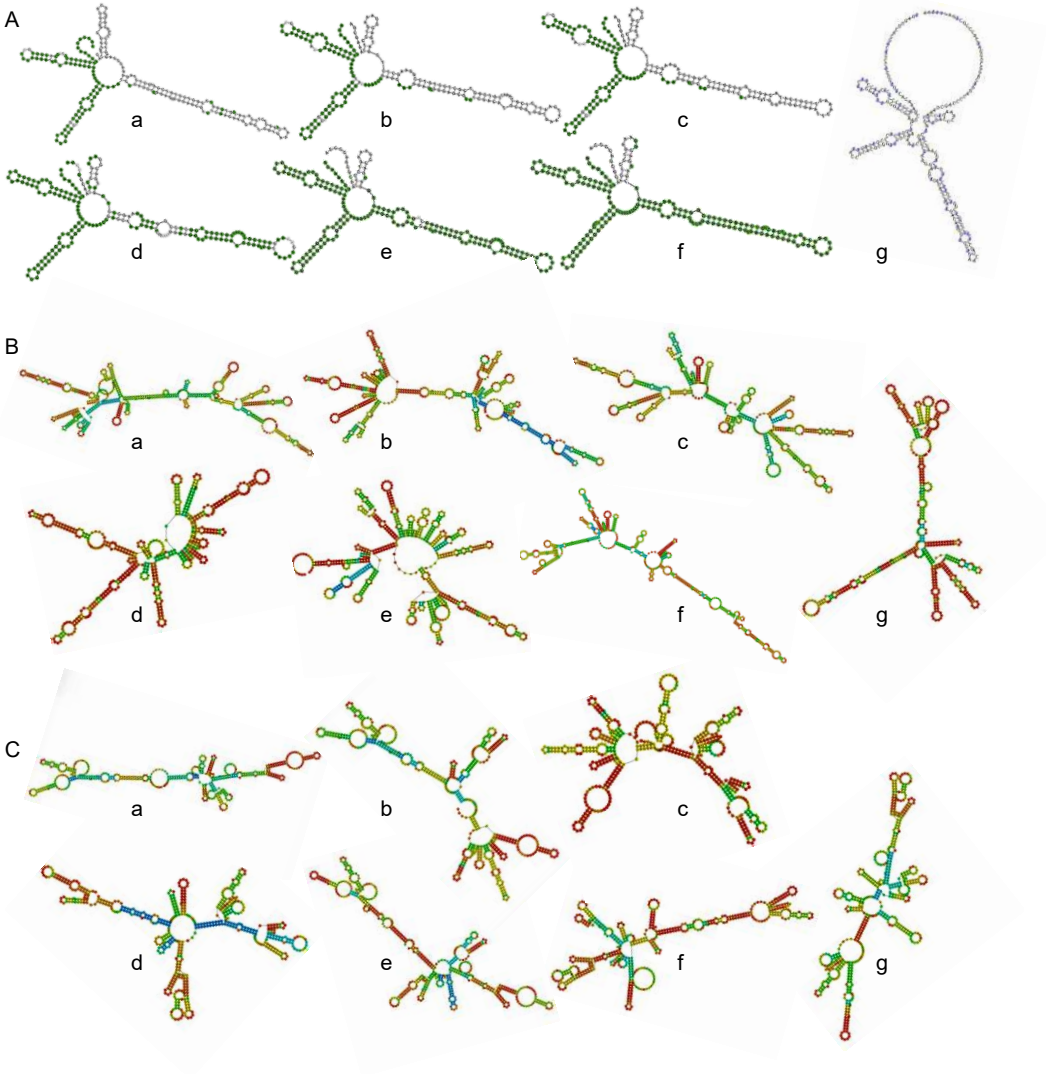


Figure 5. Predicted secondary structures of ITS2 (A), *matK* (B), and *rbcL* (C) across seven Brassicaceae species. Species a–g are as follows: *Eutrema wasabi*, *E. heterophyllum*, *E. yunnanense*, *Sinapis alba*, *Armoracia rusticana*, *Brassica juncea*, and *B. nigra*.

3.3 Machine learning-based barcode classification for wasabi authentication

Machine‑learning classification using BLOG and WEKA (NaiveBayes, J48, SMO, JRip) was applied to ITS2, *rbcL* and *matK* (Fig. 6). BLOG achieved 100% correct classification for both *rbcL* and *matK*, but only 88.99% for ITS2, indicating that *rbcL* and *matK* harbour more stable diagnostic positions. In WEKA, *matK* consistently outperformed the others across all classifiers (correct rates: 97.37% for NaiveBayes, J48 and SMO; 89.47% for JRip). *rbcL* scored 86.11% (NaiveBayes, J48, SMO) and 77.78% (JRip), while ITS2 gave the lowest rates (75.76%, 69.70%, 57.58%). Overall, BLOG showed the highest average accuracy, followed by NaiveBayes and SMO, with JRip being least stable. These observations support the view that *matK* demonstrated

the most robust performance, establishing it as the preferred barcode for machine‑learning authentication. Notably, ITS2, despite good phylogenetic resolution, performed poorly in machine‑learning models—likely because traditional methods leverage global evolutionary signals, whereas classifiers rely on discrete diagnostic sites, which are more abundant in *matK*. Thus, *matK* is recommended for intelligent authentication, complementing ITS2 for rapid screening.

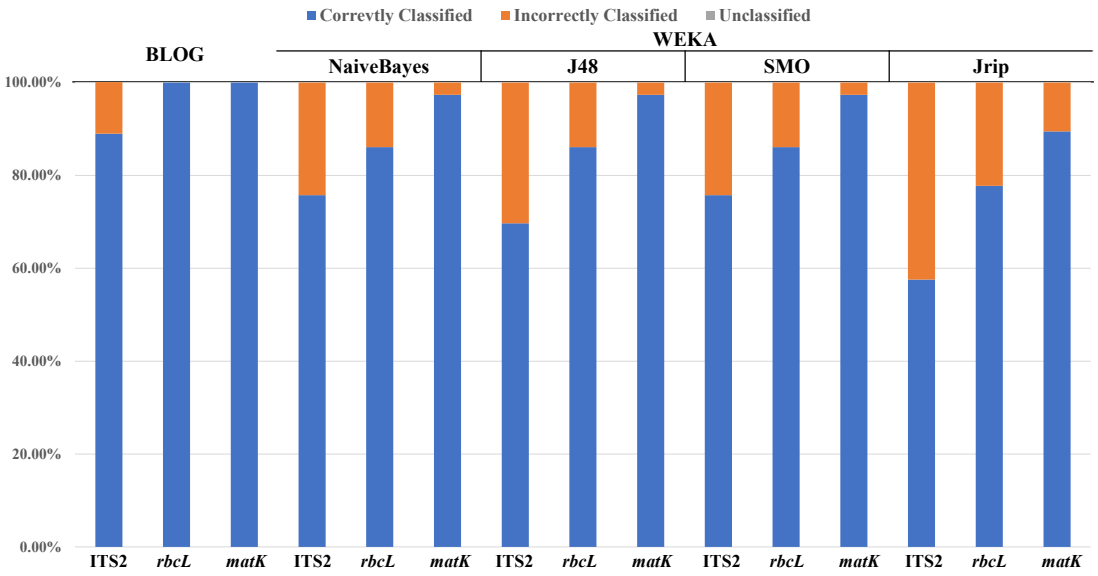


Figure 6. Machine learning classification accuracy of three DNA barcodes (ITS2, *rbcL*, *matK*) for authenticating *Eutrema wasabi*. Stacked bars represent correct classification rates (%) across BLOG algorithm and four WEKA classifiers (NaiveBayes, J48, SMO, JRip).

3.4 Comparative nutritional quality of wasabi, horseradish, and mustard

The contents of three antioxidant-related components—vitamin C, total flavonoids, and total polyphenols—were compared among *Eutrema wasabi* (wasabi), *Armoracia rusticana* (horseradish), and *Brassica juncea* (mustard) to assess their nutritional quality differences (Table 1). *E. wasabi* exhibited the highest total flavonoid content (2.539 ± 0.072 mg RE/g FW), significantly exceeding *A. rusticana* (1.022 ± 0.035) and *B. juncea* (0.912 ± 0.029). Similarly, total polyphenols followed the same ranking: *E. wasabi* (1.900 ± 0.055 mg GAE/g FW) > *A. rusticana* (1.735 ± 0.049) > *B. juncea* (1.274 ± 0.041). In contrast, vitamin C content showed minimal interspecific variation, with *E. wasabi* (2.154 ± 0.061 mg/g FW) marginally higher than *A. rusticana* (2.148 ± 0.057 mg/g FW) and *B. juncea* (2.045 ± 0.053 mg/g FW). These results demonstrate that total flavonoids and total polyphenols serve as robust discriminative nutritional markers for distinguishing authentic wasabi from its common substitutes, whereas vitamin C is relatively conserved across species. In summary, these antioxidant-related quality metrics provide a "species origin‑functional component" analytical basis that complements DNA barcoding and chloroplast genomic authentication, following the multi-marker logic recently applied to evaluate substitute sources of high-value medicinal materials (He et al., 2026).

Table 1. Contents of antioxidant-related components in wasabi (*Eutrema wasabi*), horseradish (*Armoracia rusticana*), and mustard (*Brassica juncea*). Values are presented as mean ± standard deviation. RE, rutin equivalent; GAE, gallic acid equivalent; FW, fresh weight. *E. wasabi* exhibits significantly higher total flavonoids and total polyphenols than *A. rusticana* and *B. juncea*, while vitamin C content shows minimal interspecific

variation. Standard curve equations: vitamin C, A = 0.0063C + 0.3822 (R^2 = 0.9929); total flavonoids, A = 0.0009C + 0.0915 (R^2 = 0.9911); total polyphenols, A = 0.0056C + 0.1439 (R^2 = 0.9966).

Parameter	Wasabi (*E. wasabi*)	Horseradish (*A. rusticana*)	Mustard (*B. juncea*)
Vitamin C content (mg/g)	2.154±0.061	2.148±0.057	2.045±0.053
Total flavonoid content (mg RE/g FW)	2.539±0.072	1.022±0.035	0.912±0.029
Total polyphenol content (mg GAE/g FW)	1.900±0.055	1.735±0.049	1.274±0.041

4. Discussion

This study demonstrates that the chloroplast genomes of seven Brassicaceae species share a conserved overall architecture yet harbour discernible differences in IR boundaries, hypervariable intergenic spacers, SSR loci and DNA barcode sequences, sufficient to distinguish *Eutrema wasabi* from its relatives and common substitutes. These findings support the view that authentication should not rely on a single "universal" marker, but rather on a hierarchical framework combining whole-genome candidate screening, rapid barcode validation, and multi-marker cross-verification.

Our results are consistent with earlier studies on *Eutrema* and Brassicaceae (Zhang, 2024; Haga et al., 2019; Li et al., 2021), confirming that coding regions are highly conserved due to strong functional constraints, while non-coding and boundary regions accumulate more insertions, deletions and substitutions. The diagnostic value of IR boundaries, SSRs and hypervariable spacers likely reflects differential selective pressures: functionally essential genes (*rpoC1, rbcL, ycf2, rrn* clusters) remain conserved, whereas regions such as trnK-rps16, rpl22-rps19, petB-petD and *ycf1* boundaries show greater variability (Hong et al., 2026). These variable loci are promising candidates for species-specific marker development.

At the DNA barcode level, ITS2 provided clear intergeneric separation (98% bootstrap support) but limited intrageneric resolution (35–52%), while matK gave the strongest intrageneric support (78–100%) and the widest barcoding gap (intraspecific 0.0000–0.0008; interspecific 0.0061–0.0616). Genetic distance analysis confirmed matK as the most informative single locus for distinguishing *E. wasabi* from its substitutes, whereas *rbcL* showed overlapping intra-/interspecific distances, compromising its discriminatory power. RNA secondary structures further corroborated these differences, with *E. wasabi* displaying unique loop and helix arrangements distinct from all other species.

Machine-learning classification using BLOG and WEKA (NaiveBayes, J48, SMO, JRip) revealed that *matK* consistently outperformed ITS2 and *rbcL*, achieving 100% accuracy in BLOG and 89.5–97.4% in WEKA, whereas ITS2 performed poorly in ML models (down to 57.6%) despite good phylogenetic resolution. This discrepancy arises because traditional phylogenetic methods exploit global sequence variation, whereas ML classifiers depend on discrete diagnostic nucleotide positions—which are more abundant in matK. Therefore, *matK* is recommended as the

preferred core marker for intelligent authentication, while ITS2 remains useful for rapid screening, and *rbcL* provides taxonomic background.

Codon usage analysis showed that all seven species exhibit weak A/U-ending bias, low CAI and high ENC values, with natural selection predominating over mutation pressure (neutrality, ENC- and PR2-plots). However, interspecific differences in codon metrics were minimal, and PCA showed extensive overlap among species, confirming that codon usage alone is not diagnostic for species-level authentication. Thus, these parameters serve as auxiliary evolutionary context rather than primary discriminative tools.

Nutritional profiling revealed that *E. wasabi* contains significantly higher total flavonoids (2.539 mg RE/g FW) and total polyphenols (1.900 mg GAE/g FW) than horseradish and mustard, while vitamin C is conserved across species. These chemical markers provide robust, complementary criteria for quality-based authentication and can be combined with genomic data in an integrated authenticity workflow.

Our integrated framework—chloroplast genomic comparison, ML-enhanced barcoding, and nutritional quality assessment—addresses the critical issue of wasabi adulteration, particularly in processed products where morphology and sensory evaluation fail. For practical deployment, we propose a tiered strategy: ITS2 for initial screening, matK (or hypervariable chloroplast spacers) for confirmatory identification, and—for disputed or deeply processed samples—whole-genome based assays combined with flavonoid/polyphenol quantification. This approach enhances detection efficiency and scientific rigor for label authenticity, traceability and regulatory control. Future work should expand sequence databases, validate ML models on commercial samples, and incorporate targeted metabolomics (e.g., isothiocyanates by LC-MS/MS) to further strengthen the authentication system.

5. Conclusions

This study establishes an integrated three--module framework for wasabi authenticity control, combining chloroplast genomic comparison, DNA barcoding with machine--learning classification, and nutritional metabolite profiling. Chloroplast genomes of seven Brassicaceae species are structurally conserved but harbour species--specific variations in IR boundaries, SSRs and hypervariable regions, providing candidate markers for differentiating Eutrema wasabi from its substitutes. Among barcodes, ITS2 excels in traditional phylogenetic screening, while matK achieves the highest and most stable accuracy across machine--learning classifiers (BLOG: 100%; WEKA: 89.5–97.4%), making it the preferred core marker for intelligent authentication. Critically, nutritional profiling reveals that wasabi contains significantly higher total flavonoids (2.539 mg RE/g FW) and total polyphenols (1.900 mg GAE/g FW) than horseradish and mustard, whereas vitamin C is conserved and offers limited discriminatory power.

These chemical metrics serve as robust, complementary quality discriminators, supporting both authenticity verification and nutritional quality evaluation. The multi--layered approach—integrating evolutionary genomic insights, AI--enhanced molecular identification, and functional component analysis—offers a practical, scalable solution for combating wasabi adulteration in processed products, with transferable potential to other high--value food commodities. Future work should expand sequence databases and incorporate targeted metabolomics (e.g., isothiocyanates by LC--MS/MS) to further strengthen this authentication system.

Declarations

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Xi'an Medical University 2024 Annual Special Program for Scientific and Technological Capability Enhancement (Grant No. 2024NLTS122).

Authors' contributions

Mengdi Zheng: conceptualization, methodology, supervision, funding acquisition, writing—original draft, writing—review and editing, project administration; Yibo Ma: investigation, data curation, visualization; Yangrui Li: resources, investigation, formal analysis; Jiayin Guo: formal analysis; Miaomiao Li: visualization; Yulei Li: data curation.

Acknowledgements

Not applicable.

Data availability

The authors confirm that the data supporting the findings of this study are available within the article and as its supplementary materials.

References

Bartkowiak-Wieczorek, J., Malesza, M., Malesza, I., et al. (2024). Methylsulfinyl hexyl isothiocyanate (6-MSITC) from wasabi is a promising candidate for the treatment of cancer, Alzheimer's disease, and obesity. *Nutrients*, *16*(15), 2509. doi:10.3390/nu16152509.

Cammue, B. P. A., De Loose, M., Van Droogenbroeck, B., et al. (2016). Duplex real-time PCR for the determination of wasabi (*Eutrema wasabi*) contents in horseradish (*Armoracia rusticana*) products applying the ΔΔCt method. *European Food Research and Technology*, *242*(7), 1079–1088. doi:10.1007/s00217-015-2615-7.

Chen, X., Li, Z., Wu, X., et al. (2021). Identification of Components of *Eutrema wasabi*, *Armoracia rusticana* and Mustard in Pungent Condiment by Internal Transcribed Spacer 2 (ITS2). *Journal of Food Science and Biotechnology*, *40*(9), 94–100. doi:10.3969/j.issn.1673-1689.2021.09.012.

Feng, T. (2021). Molecular identification of DNA barcoding of three medicinal plants based on machine learning approaches [Master's thesis, Guangdong Pharmaceutical University]. doi:10.27690/d.cnki.ggdyk.2021.000096.

Galvin-King, P., Haughey, S. A., & Elliott, C. T. (2018). Herb and spice fraud: The drivers, challenges and detection. *Food Control*, *88*, 85–97. doi:10.1016/j.foodcont.2017.12.031.

García-Yagüe, Á. J., Cueto-Díaz, E. J., Escoll, M., et al. (2025). Dual targeting of Keap1 and Gsk-3 by hexaraphane in the regulation of transcription factor Nrf2. *Free Radical Biology and Medicine*, *239*, 579–593. doi:10.1016/j.freeradbiomed.2025.07.051.

Haga, N., Kobayashi, M., Michiki, N., et al. (2019). Complete chloroplast genome sequence and phylogenetic analysis of wasabi (*Eutrema japonicum*) and its relatives. *Scientific Reports*, *9*(1), 14377. doi:10.1038/s41598-019-49667-z.

He, L., Wang, P., Wang, Z., et al. (2026). Comparative analysis of morphological, molecular, and physicochemical markers to evaluate *Trollius ledebouri* Rchb. as a potential alternative source to *Trollius chinensis* Bunge for high-quality Flos Trollii supplements. *Biology*, *15*(4), 332. doi:10.3390/biology15040332.

Hejazi, F. A., Mohammadi, P., & Soorni, A. (2025). Comparative chloroplast genomics of *Teucrium* species reveals genome evolution, phylogenetic relationships, and candidate molecular markers. *Scientific Reports*, *15*(1), 44318. doi:10.1038/s41598-025-29339-x.

Hong, S., Du, J., Zhang, Z., et al. (2026). Structural Characteristics and Phylogenetic Study of Chloroplast Genome in Baiyu Bean (*Phaseolus lunatus* L.). *Journal of Sichuan Agricultural University*, *44*(1), 148–164. doi:10.16036/j.issn.1000-2650.202510689.

Jia, Z., Tang, M., & Wu, J. (1999). The determination of flavonoid contents in mulberry and their scavenging effects on superoxide radicals. *Food Chemistry*, *64*(4), 555–559. doi:10.1016/S0308-8146(98)00102-2.

Li, M., Zhang, R., Li, J., et al. (2021). Analyses of chloroplast genome of *Eutrema japonicum* provide new insights into the evolution of *Eutrema* species. *Agronomy*, *11*(12), 2546. doi:10.3390/agronomy11122546.

Lin, M. (2008). Study on Low Temperature Fresh-keeping and Drying Technology of Wasabi. *Journal of Chinese Institute of Food Science and Technology*, *8*(1), 83–88. doi:10.16429/j.1009-7848.2008.01.013.

Martinović, A., Ruppitsch, W., et al. (2025). From gene to table—DNA barcoding the backbone of next-generation food integrity and safety. *Meat Technology*, *66*(3), 25–33. doi:10.18485/meattech.2025.66.3.4.

Singleton, V. L., Orthofer, R., & Lamuela-Raventós, R. M. (1999). Analysis of total phenols and other oxidation substrates and antioxidants by means of Folin-Ciocalteu reagent. *Methods in Enzymology*, *299*, 152–178. doi:10.1016/S0076-6879(99)99017-1.

Wang, H., Wang, Z., Liu, Z., Smagghe, G., Zhao, X., Li, D., & Gai, Y. (2026a). Integrative chloroplast genomics of Polygonaceae: Evolutionary dynamics, codon optimization, and phylogenetic resolution. *BMC Plant Biology*, *26*(1), 888. doi:10.1186/s12870-026-08560-4.

Wang, W., Fan, C., Zhang, Y., et al. (2025a). Research progress and enlightenments on food fraud at home and abroad. *Journal of Food Safety and Quality*, *16*(12), 25–33. doi:10.19812/j.cnki.jfsq11-5956/ts.20241218007.

Wang, Z., Guo, Y., Chen, Y., Gao, F., Xue, Y., Niu, Q., Li, D., Smagghe, G., & Gai, Y. (2026b). Chloroplast genome sequencing of *Poa pratensis* and comparative chloroplast genomics analysis with other 29 grass species. *Functional & Integrative Genomics*, *26*(1), 44. doi:10.1007/s10142-025-01814-8.

Wang, Z., Song, X., Zhang, P., et al. (2025b). Analysis of label issues and research on countermeasures for imported prepackaged foods. *Journal of Food Safety and Quality*, *16*(8), 288–295. doi:10.19812/j.cnki.jfsq11-5956/ts.20250221002.

Zhang, Z. (2024). Isothiocyanates, Phylogenetics, and Taxonomy of tuber mustard (*Eutrema japonicum*) and its Relatives [Master's thesis, Sichuan University of Science & Engineering]. doi:10.27703/d.cnki.gsclg.2024.000180.

Zheng, M., Gao, M., Zhang, Z., et al. (2026). Integrating DNA barcoding and machine learning for species identification: Comparative genomics and codon usage bias of chloroplasts in *Gentiana* sect. *Cruciata*. *Journal of Plant Physiology*, *319*, 154730. doi:10.1016/j.jplph.2026.154730.

Zheng, M., Zhang, C., Ma, R., et al. (2022). Identification of moslae herba and its adulterants based on *matK*, ITS2 and its secondary structure. *Chinese Journal of Modern Applied Pharmacy*, *39*(17), 2222–2228. doi:10.13748/j.cnki.issn1007-7693.2022.17.008.

Zheng, M., Zhao, Y., Gao, M., et al. (2025). Chloroplast structure, codon usage bias, and machine learning-based molecular identification using DNA barcoding of Sophorae Tonkinensis Radix et Rhizoma (Shan Dou Gen) and its analogues. *Fitoterapia*, *188*, 107005. doi:10.1016/j.fitote.2025.107005.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 0725SupplementaryInformationrevised.docx