

Article ID: 185311  
DOI: 10.5586/aa/185311

**Publication History**  
Received: 2023-09-29  
Accepted: 2024-02-24  
Published: 2024-06-01

**Handling Editor**  
Beata Myśków; West Pomeranian University of Technology, Szczecin, Poland; <https://orcid.org/0000-0001-5062-9841>

**Authors' Contributions**  
NTTP: Research concept and design; NTTP, DTN, TVBT, HHD, BDDT, TVP: Collection and/or assembly of data; NTTP, AHN, BDDT, LTT, QTH, TQDN: Data analysis and interpretation; NTTP, AHN: Writing the article; LTT, QTH, TQDN, PTBT: Critical revision of the article; PTBT: Final approval of the article












**Funding**  
This study was financially supported by the Department of Science and Technology, Thua Thien Hue, Vietnam, under Grant TTH.2018-KC.03, and University of Sciences, Hue University, under Student Research Grant 2020. The authors also acknowledge the partial support of Hue University, under the Core Research Program, Grant No. NCM.DHH.2020.12.

**Competing Interests**  
No competing interests have been declared.

**Copyright Notice**  
© The Author(s) 2024. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits redistribution, commercial and noncommercial, provided that the article is properly cited.

## ORIGINAL RESEARCH

# Molecular characterization of a distinct ginger chemotype from Thua Thien Hue, Vietnam, and the application of PCR-based markers for identifying unknown ginger populations in the region using machine learning

Nguyen Thi Thao Phan <sup>1,2</sup>, Dat Tien Nguyen <sup>1</sup>,  
Thanh Van Bao Tong <sup>1</sup>, Hanh Hong Dang <sup>1</sup>,  
An Hoang Nguyen <sup>1</sup>, Bao Duc Duy Tran <sup>1</sup>, Tri Van Phan <sup>3</sup>,  
Lan Thuy Tran <sup>2</sup>, Quang Tan Hoang <sup>2</sup>,  
Tien Quang Duc Nguyen <sup>1</sup>, Phuong Thi Bich Truong <sup>1\*</sup>

<sup>1</sup>Department of Biology, University of Sciences, Hue University, 77 Nguyen Hue Street, 49000, Hue City, Thua Thien Hue Province, Viet Nam

<sup>2</sup>Laboratory of Gene Technology, Institute of Biotechnology, Hue University, Nguyen Dinh Tu Street, 49000, Hue City, Thua Thien Hue Province, Viet Nam

<sup>3</sup>HUSC High School for Gifted Students, University of Sciences, Hue University, 77 Nguyen Hue Street, 49000, Hue City, Thua Thien Hue Province, Viet Nam

\* To whom correspondence should be addressed. Email: [tbphuong@hueuni.edu.vn](mailto:tbphuong@hueuni.edu.vn)

## Abstract

Ginger (*Zingiber officinale* Roscoe) stands as an esteemed herbaceous spice due to its extensive applications in medical and culinary sectors. The variety of ginger indigenous to Thua Thien Hue, known as Hue's ginger, has long garnered recognition for its distinct aroma and unique oil composition. Regrettably, this ginger variety has intermingled with unidentified ginger types. Thus, the objective of this study is to identify DNA markers that can facilitate the identification of Hue's ginger. Such markers will enable the precise selection and preservation of the authentic ginger chemotype. To substantiate the distinctive genetic attributes of Hue's ginger, we employed two marker techniques: RAPD and *matK* DNA barcoding. The RAPD technique demonstrated its robustness by generating an impressive number of 139 amplicons, with an absolute polymorphic rate of 100%. Among the resulting bands, two region-specific markers, OPA03-480 and OPB01-1150, were delineated for Hue's ginger. These specific markers facilitated the separation of Hue's ginger from other ginger chemotypes, shown by principal coordinates analysis. Furthermore, the alignment of the *matK* gene sequence of Hue's ginger with the reference chloroplast genome substantiated the hypothesis that Hue's ginger possesses distinct genetic characteristics. This alignment revealed three transition variants within the *matK* gene of Hue's ginger. Considering the extensive intermixing of ginger populations in Thua Thien Hue, we constructed an XGBoost machine-learning model using RAPD data to identify the most pivotal markers capable of effectively distinguishing between these populations. Our model identified OPN06-350, OPA03-480, OPD02-500, OPF04-950, and OPN06-300 as the most influential markers for population discrimination. This study not only furnishes molecular markers for the precise identification of a unique Vietnamese ginger chemotype but also advocates for the utilization of machine-learning methodologies employing PCR-based marker data for the identification of pivotal markers, a practice with promising implications for the effective differentiation of plant varieties in future endeavors.

## Keywords

Hue's ginger; RAPD; *matK*; *Zingiber officinale* Roscoe; XGBoost; machine learning

## 1. Introduction

Ginger (*Zingiber officinale* Roscoe) stands as a paramount spice, with its rhizome serving as a source of invaluable tonics (An et al., 2020; Baliga et al., 2011; Engdal et al., 2009; Nicoll & Henein, 2009). The species harbors a plethora of bioactive secondary metabolites, including phenolic compounds and terpenes, each manifesting antiemetic, anti-inflammatory, antioxidant, anti-tumor, anti-cancer, and neuroprotective properties (Mao et al., 2019; Prasad & Tyagi, 2015; Yeh et al., 2014). Its historical role as a folk remedy encompasses diverse therapeutic applications, such as addressing headaches, fever, dyspepsia, nausea, digestive disorders, blood circulation disturbances, and inflammation (Ali et al., 2008; El baroty et al., 2010; Geiger, 2005). Furthermore, the distinctive aroma and traditional medicinal usages of ginger have made it a principal ingredient in the global food processing industry, giving rise to an array of popular processed products, including candied ginger, gingerbread, ginger oil, ginger candy, crystallized ginger, ginger powder, ginger beer, and ginger ale (Govindarajan & Connell, 1982; Nair, 2013; Rajathi et al., 2017; Shukla & Singh, 2007). Ginger's nutraceutical properties have garnered significant attention within the realms of both food processing and pharmaceutical industries (Bag, 2018; Kubra & Jaganmohanrao, 2012; Vasala, 2012).

Thuy Bieu Ward (Hue City) and the Junction of Tuan (Huong Tho District) represent the primary cultivation regions for one of Vietnam's most well-known ginger chemotypes (Hue's ginger). The Hue's ginger cultivar (characterized by an exceptionally robust flavor) possesses a unique essential oil content, substantiated by the notably high  $\alpha$ -zingiberene amount of 32.52% in its essential oil (Hien et al., 2018). This secondary metabolite exhibits antimicrobial activity against various microorganisms (*Penicillium* spp., *Candida albicans*, *Aspergillus niger*, and *Bacillus subtilis*) and shows potential contributions to antipyretic, antiallergenic, analgesic, antitussive, and chemopreventive effects (El baroty et al., 2010; Sasidharan & Menon, 2010; Sharma et al., 2016). However, uncontrolled trading practices have resulted in the mixing of Hue's ginger with unidentified ginger types, leading to the erosion and dilution of the native variety.

The initial attempt to differentiate Hue's ginger from unknown chemotypes was undertaken based on retrotransposon-based markers (An et al., 2022). Although the study successfully identified distinctive markers for Hue's ginger, the sample sizes from various geographical regions proved insufficient to draw definitive conclusions regarding the genetic attributes of Hue's ginger.

Given the dearth of substantiating genetic evidence for Hue's ginger, we have collected a substantial number of ginger samples and employed different molecular markers to fortify the hypothesis that Hue's ginger genuinely possesses distinct genetic characteristics compared to other chemotypes. Initially, the search for unique markers of Hue's ginger was conducted using the RAPD technique, renowned for its high discriminatory power in elucidating ginger's genetic diversity (Ashraf et al., 2014; Harisaranraj et al., 2009; Ismail et al., 2016; Mia et al., 2014). Subsequently, we sequenced the *matK* gene, a reliable DNA barcode known for its utility in both inter- and intra-specific discrimination (Zhu et al., 2022). Our *matK* sequence was then compared to a publicly available ginger chloroplast genome sequence to identify single nucleotide polymorphisms (SNPs) for prospective Hue's ginger identification.

Simultaneously, considering the pervasive mixing of ginger chemotypes in Vietnam, we aimed to extract a concise set of the most efficient RAPD markers characterized by high discriminatory power and capable of distinguishing ginger chemotypes from various regions. This would facilitate selections for cultivation, breeding, and food/pharmaceutical processing. The identification of these markers was accomplished through the application of the XGBoost (extreme gradient boosting) machine-learning algorithm possessing exceptional accuracy, speed, and efficiency in mitigation of over-fitting (M. Chen et al., 2019). Our study represents the first-time utilization of XGBoost with PCR-based marker data. Furthermore, XGBoost allowed us to identify the most important markers in sample discriminations, which is useful for PCR-based genetic diversity studies, particularly in scenarios involving a substantial number of generated markers.

## 2. Material and methods

### 2.1. Plant materials

One hundred and five (105) ginger leaf samples were collected from nine different localities for the purpose of total genomic DNA extraction (see Table 1). Young ginger leaves were harvested from 2-to-5-month-old plants and subsequently stored in plastic containers at a temperature of  $-20\text{ }^{\circ}\text{C}$  within the Gene Technology Laboratory at the Institute of Biotechnology, Hue University. The samples were further categorized into four distinct subpopulations, as outlined in Table 1. Notably, samples within subpopulations 1 and 3 are presumed to represent Hue's ginger, while the remaining samples are representative of unidentified varieties sourced from various regions within Thua Thien Hue. The specific locations from which the samples were procured are depicted in Figure 1.

### 2.2. RAPD profiling and *matK* sequencing

#### 2.2.1. Total genomic DNA extraction

Genomic DNA was extracted according to the procedure proposed by Doyle and Doyle (Doyle & Doyle, 1987). The RNA in the DNA solutions was digested by  $1\text{ }\mu\text{L}$  of RNase A ( $100\text{ }\mu\text{g}/\mu\text{L}$ ) at the temperature of  $37\text{ }^{\circ}\text{C}$  for 30 minutes. The DNA samples were stored at  $4\text{ }^{\circ}\text{C}$ .

#### 2.2.2. Primer screening

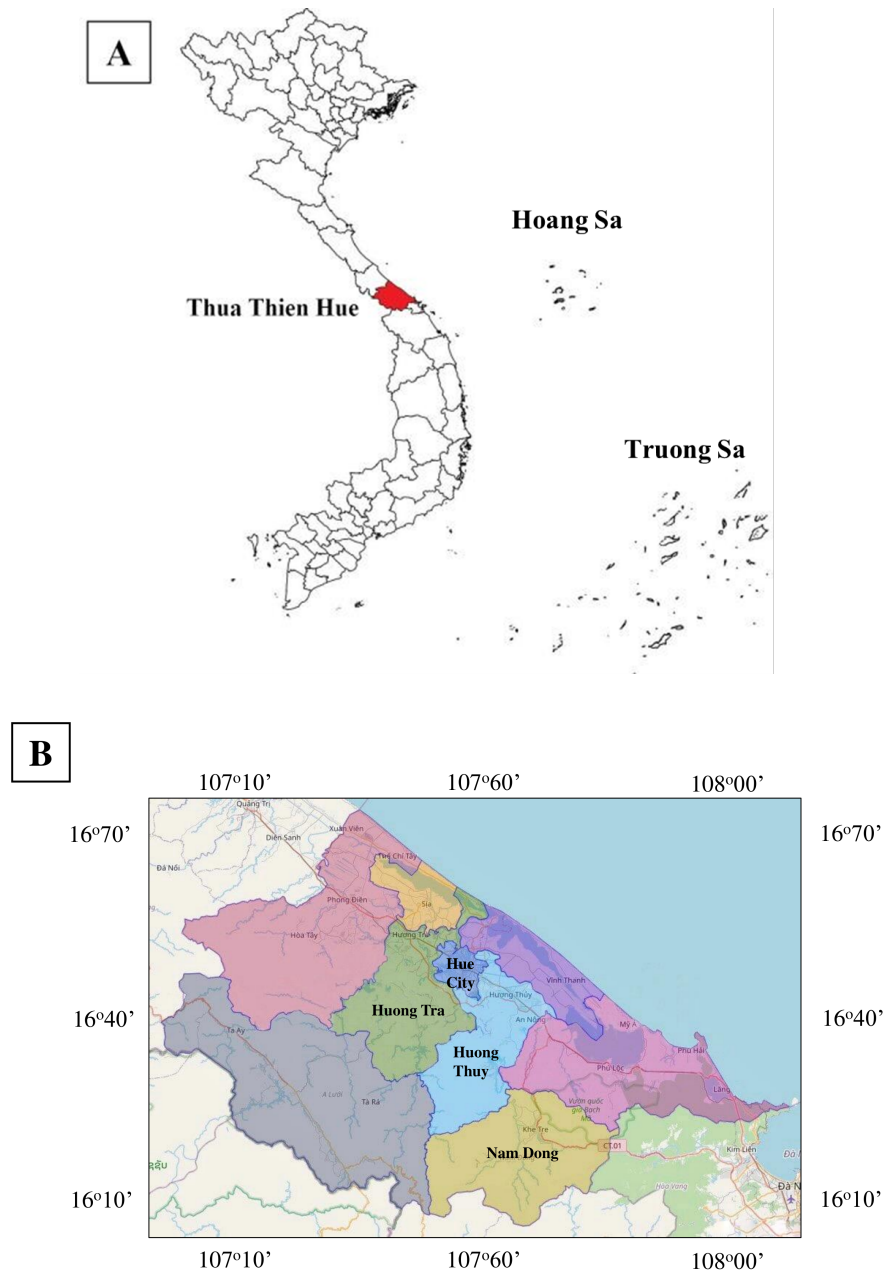
Twenty-two primers were employed in the screening process (Table S1) to select primers that can generate high polymorphic rates. Nine representative DNA samples were amplified by each tested primer.

The total reaction volume comprised  $20\text{ }\mu\text{L}$  and included the following components: random primer ( $20\text{ }\mu\text{M}$ ), GoTaq Green Master Mix 2X (Promega, USA), genomic DNA ( $50\text{ ng}/\mu\text{L}$ ), and nuclease-free double-distilled water (ddw).

The PCR reaction protocol involved an initial denaturation step at  $95\text{ }^{\circ}\text{C}$  for 3 minutes, followed by 42 amplification cycles (comprising denaturation at  $92\text{ }^{\circ}\text{C}$  for one minute, primer annealing at  $36\text{ }^{\circ}\text{C}$  for one minute, and primer extension at  $72\text{ }^{\circ}\text{C}$  for one minute), and a final extension step at  $72\text{ }^{\circ}\text{C}$  for 10 minutes. Subsequently, electrophoresis of the amplicons was performed on a 1.4% agarose gel supplemented with ABM's SafeView<sup>TM</sup> Classic dye, utilizing an applied voltage of 40 V for 70 minutes. Electrophoresis images were captured using the Ultra Slim LED Illuminator (Miulab, wavelength: 470 nm), and the lengths of the amplicons were determined through the utilization of GeneRuler 1 kb DNA Ladder (ThermoScientific, # SM0313). Primers exhibiting high polymorphic rates were chosen for RAPD analysis.

**Table 1** Localities of Thua Thien Hue province where ginger leaves were collected.

District/ City	Ward	Sample codes	Number of samples	Sub-population	Coordinates
Hue City	Thuy Bieu	B	17	Pop1	16.4446° N, 107.5511° E
	Thuy Xuan	X	10	Pop2	16.4261° N, 107.5789° E
	An Tay	H	8		16.4304° N, 107.6045° E
Huong Tra	Huong Tho	T	10	Pop3	16.3936° N, 107.5747° E
Huong Thuy	Thuy Bang	A	15	Pop4	16.3979° N, 107.5974° E
	Thuy Phu	U	10		16.3806° N, 107.7153° E
	Thuy Phuong	P	15		16.4327° N, 107.6333° E
	Thuy Duong	D	15		16.4419° N, 107.6168° E
Nam Dong	Thuong Long	N	5		16.0755° N, 107.6275° E



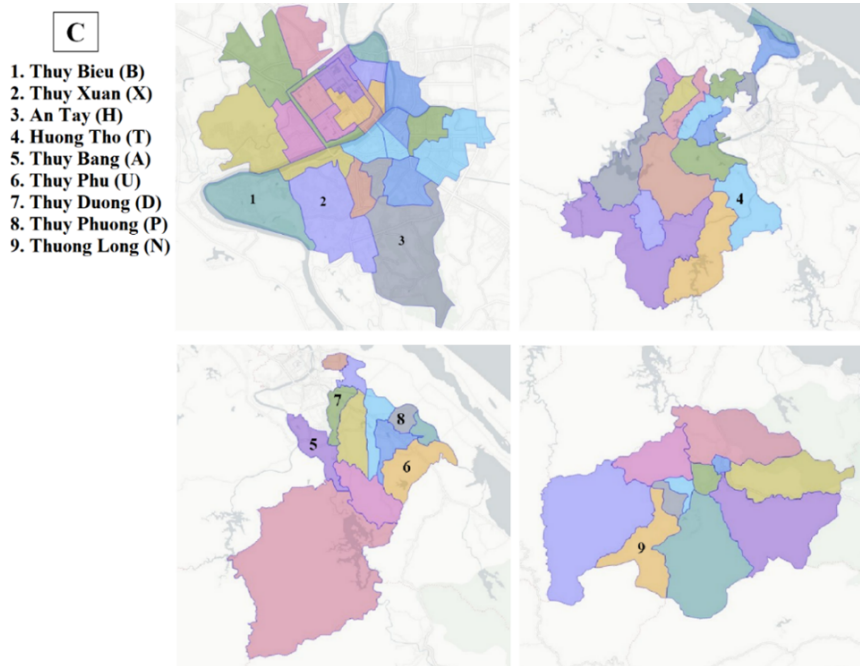
**Figure 1** Continued on next page.

### 2.2.3. RAPD amplification

The 105 genomic DNA samples were first diluted to a concentration of 50 ng/μL and subsequently subjected to amplification using a set of selected primers (OPA03, OPA04, OPA07, OPA09, OPB01, OPB18, OPD02, OPF04, OPN03, and OPN06). The total reaction volume, PCR reaction protocol, agarose gel electrophoresis, and amplicon visualization are described in Subsection 2.2.2.

### 2.2.4. *matK* sequencing

To establish a DNA barcode for Hue's ginger, an analysis of the *matK* gene sequences was carried out using two representative samples, B1 and B2. The initial step involved the isolation of the *matK* gene from the DNA samples through PCR, with the following reaction components: GoTaq Green Master Mix 2X (Promega, USA), 3F\_Kim F primer (CGTACAGTACTTTTGTGTTTACGAG) (10 μM), 1R\_Kim



**Figure 1** Thua Thien Hue Province (A), regions in Thua Thien Hue (B), and Wards of each region (C) from which ginger leaves were collected.

R primer (ACCCAGTCCATCTGGAAATCTTGTTTC) (10  $\mu$ M), genomic DNA (100 ng/ $\mu$ L), and nuclease-free ddw.

Subsequently, PCR amplification was carried out using the following thermal program: an initial denaturation step at 95 °C for 10 minutes, followed by 30 amplification cycles (consisting of denaturation at 95 °C for one minute, annealing at 48 °C for one minute, and extension at 72 °C for one minute), and a final extension step at 72 °C for 10 minutes.

The resulting amplicons were subjected to visualization through the aforementioned electrophoresis procedure and subsequently purified using the GeneJET Gel Extraction Kit (ThermoFisher Scientific, USA) in accordance with the manufacturer's purification protocol. The purified PCR products were then forwarded to 1st BASE (Apical Scientific Sdn Bhd, Malaysia) for Sanger sequencing.

### 2.3. Data analysis

#### 2.3.1. Band scoring

PCR-RAPD products were systematically assigned numerical identifiers according to the presence or absence of bands, thereby generating a binary matrix. Concretely, bands that manifested were assigned the numerical value "1", whereas those absent were assigned "0". Bands that were unsighted and failed to meet the criteria for evaluation were also assigned "0". Amplicons that were exclusive to a specific geographical region and detected in multiple samples were classified as "region-specific markers," while PCR products observed in only one sample were designated as "unique markers".

#### 2.3.2. Evaluation of RAPD discriminatory power and cluster analysis

To evaluate if the RAPD technique had generated significant differences among the four subpopulations (comprising 105 samples), AMOVA (analysis of molecular variance) was used, with the permutation value of 999 ( $p < 0.01$ ). AMOVA was performed using GenAlEx v6.51 (Peakall & Smouse, 2006).

To observe the genetic relationships of all the samples, PCoA (principal coordinates analysis) was performed using the distance matrix generated from the initial binary matrix. PCoA was also performed using GenALEx v6.51.

### 2.3.3. Identification of the most powerful markers for the discrimination of the subpopulations

Of all the RAPD markers generated, an XGBoost machine-learning model was used to extract the most useful markers with the highest capacity to distinguish the subpopulations using the aforementioned binary matrix as input data. The appearance of markers in the matrix was used as features, and the subpopulation names were used as labels for model building and prediction. The method for important markers identification was based on a reported procedure with some modifications (Nguyen-Hoang et al., 2024) as follows.

The model was generated employing the `XGBClassifier` class sourced from the `xgboost` module (T. Chen & Guestrin, 2016). The model utilized the training dataset, encompassing 80% of the total input data. Data partitioning was conducted using the `train_test_split()` function from the `sklearn` module (Pedregosa et al., 2011). Model optimization was carried out through the fine-tuning of parameters, including "learning\_rate", "max\_depth", "min\_child\_weight", "gamma", and "colsample\_bytree", employing the `RandomizedSearchCV` class from the `sklearn` module (Pedregosa et al., 2011). Subsequently, the optimized model was saved using the `joblib.dump()` function (*Joblib: Computing with Python Functions.*, n.d.).

To evaluate the model's efficacy in distinguishing subpopulations, the input data was divided into training and testing datasets in an 8:2 ratio, employing multiple random seeds spanning the range from 1 to 100. The optimized model was then fitted to each training dataset, and prediction was made for the corresponding testing dataset. Assessment of the model's accuracy was carried out by means of the `accuracy_score()` function, available within the `sklearn` module (Pedregosa et al., 2011). By iteratively conducting the fitting and prediction process, a collection of accuracy scores was obtained. These scores were subsequently leveraged to compute the mean accuracy across the various times of model fitting.

For the identification of the most salient RAPD markers for subpopulation discrimination, the input data was partitioned in an 8:2 ratio, utilizing random seeds ranging from 1 to 100. The model was then fitted to these 100 training datasets. Following each fitting iteration, the command

```
"model_name.get_booster().get_score(importance_type="gain")"
```

was executed to compile lists enumerating markers alongside their respective gain scores, signifying their discriminatory potential. Subsequently, the 100 lists of markers and their gain scores were amalgamated into a comprehensive data frame. The markers exerting the greatest influence were ascertained by calculating the mean gain scores across the lists.

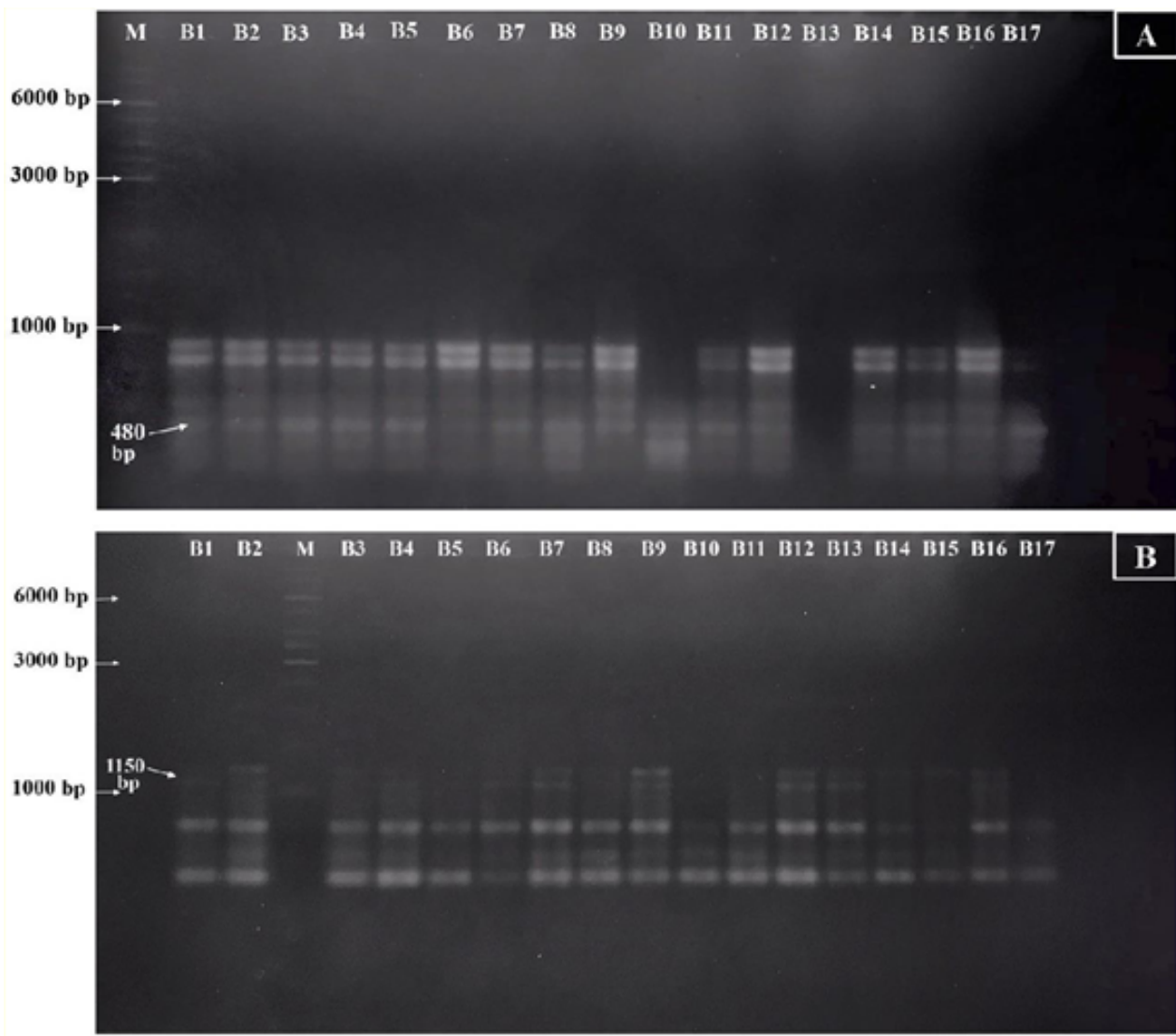
### 2.3.4. Hue's ginger DNA barcoding

The *matK* sequences obtained from samples B1 and B2 underwent a preprocessing step to eliminate regions characterized by low-quality base calling. Subsequently, the sequences were aligned against the reference ginger chloroplast genome (NC\_044775) to identify any Single Nucleotide Polymorphisms (SNPs) present within the *matK* gene of these samples. These analytical procedures were executed utilizing Bioedit (v7.2.5) (Hall, 1999).

## 3. Result

### 3.1. Primer screening

Among the 22 RAPD primers assessed, 10 primers, namely OPA03, OPA04, OPA07, OPA09, OPB01, OPB18, OPD02, OPF04, OPN03, and OPN06, demonstrated a poly-



**Figure 2** Observations of region-specific markers of Hue's ginger in Thuy Bieu, namely OPA03-480 (A) and OPB01-1150 (B). The illustrated codes (B1–B17) on each lane represent 17 different samples from Thuy Bieu Ward (Hue City). The lane with the character code “M” is for GeneRuler™ 1kb DNA Ladder.

morphic rate of 100%. The other 12 primers exhibited low amplicon numbers and polymorphic rates (Figure S1). Therefore, these primers were not used in further analyses.

The amplicon numbers detected in the nine representative samples were consistent when the selected primers were used to amplify all 105 samples in the main analysis described in Subsection 3.2 below (Figure S2).

### 3.2. RAPD profiling and identification of unique RAPD markers for Hue's ginger

All 10 RAPD primers exhibited a 100% polymorphic rate. The number of amplified bands and their respective lengths ranged from 10 to 19 bands and 200 to 2,400 bp (see Table 2 and Figure S2). All 139 generated PCR products displayed polymorphism, with 14 unique markers and 16 region-specific markers. Notably, the OPA03 and OPB01 primers yielded the highest number of both region-specific and unique bands. Conversely, no unique bands were observed for the OPA04, OPA07, OPB18, and OPF04 primers (Table 2).

Three markers specific to Hue's ginger were identified through RAPD profiling. Specifically, two region-specific markers for ginger samples from Thuy Bieu ward (OPA03-480 and OPB01-1150) were detected in 94.12% and 52.94% of Thuy Bieu

**Table 2** RAPD amplifications of the 105 ginger genomic DNA samples.

No.	Primer	Number of bands	Number of polymorphic bands	Amplicon lengths (bp)	Polymorphic rate (%)	Region-specific band (bp)	Unique band (bp)	Samples
1	OPA03	13	13	375–1500	100		1500	T8
						1450		H3, H4, H8
						1350		H3, H4, H8
						1000		H3–H5, H7–H8
							625	H3
						460		N3–N5
						480		B1–B12, B14–B17
						400		N3–N5
								H3–H5
								X4, X5
2	OPA04	16	16	450–2000	100	1340		P5, P7, P10, P12–P15
3	OPA07	15	15	200–1650	100	1500		T5–T6, T9–T10
						1450		
						1050		
4	OPA09	19	19	250–1750	100		1500	D11
							1395	H2
							1175	H7
5	OPB01	14	14	400–2400	100		2400	H4
							2000	H4
							1750	H4
						1150		B1–B2, B7–B9, B12–B14, B17
						550		N1–N5
6	OPB18	10	10	250–1700	100			
7	OPD02	12	12	300–1950	100		1950	H3
							1400	H8
						450		N1–N5
8	OPF04	11	11	400–1400	100	1400		H6, H8
9	OPN03	14	14	250–1600	100		1200	A7
							275	X3
10	OPN06	15	15	250–1700	100	1500		H7, H8
							1125	H8
							1000	H8
						900		X7, X9



**Table 3** Analysis of the molecular variance of the four ginger subpopulations.

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	Percentage of total variance (%)	<i>PhiPT</i>	<i>p</i>
Among populations	3	313.25	104.42	22	0.22	0.001**
Within populations	101	1508.26	14.93	78		
Total	104	1821.51		100		

*df* - degree of freedom; *SS* - Sum of Squares; *MS* - Mean of Squares.

samples, respectively (see Table 2 and Figure 2). Additionally, the OPA07-1050 marker was exclusively present in samples from Huong Tho, another region of Hue's ginger cultivation, albeit in only 40% of the Huong Tho samples (Table 2).

### 3.3. Assessment of RAPD discrimination power and genetic relationship of the ginger samples

The RAPD markers demonstrated significant discriminatory power in elucidating genetic variations among the subpopulations. Specifically, AMOVA revealed that 78% and 22% of the total genetic diversity existed within and among subpopulations, respectively ( $p < 0.01$ ) (Table 3).

The discerning capabilities of RAPD markers resulted in the clear separation of Hue's ginger samples from Thuy Bieu from other ginger chemotypes. Notably, 17 Hue's ginger samples from Thuy Bieu formed a distinct cluster in the 1–2 and 2–3 PCoA plots. Conversely, the samples from Huong Tho clustered with others, indicating potential hybridization or mixing with unidentified ginger types from different regions (Figure 3).

### 3.4. *matK* barcoding of Hue's ginger samples from Thuy Bieu ward

The PCR reactions for *matK* gene isolation were successful, yielding a single discernible amplicon within the length range of 750 to 900 bp for the B1 and B2 samples. The alignment of the ginger reference chloroplast genome with the processed *matK* gene sequences of the B1 and B2 samples identified three SNPs that may serve as barcodes for future Hue's ginger identification. These substitutions are exclusively transitions, namely A to G at position 3, G to A at position 12, and G to A at position 849 (Figure 4). The *matK* gene sequences of Hue's ginger samples have been deposited in the Genbank database, with accession number MZ202362 and MZ202363.

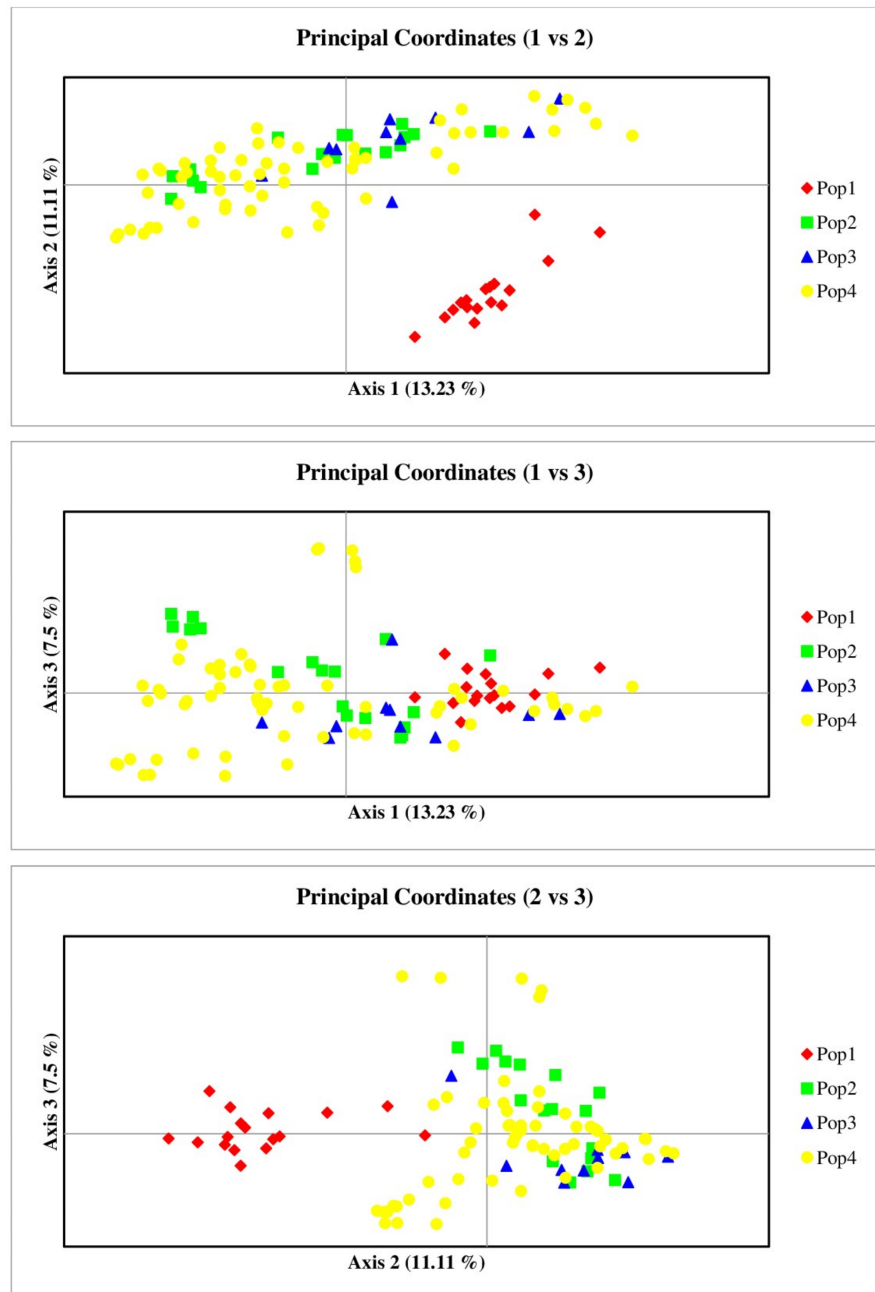
### 3.5. Identification of RAPD markers with the highest discriminatory power

The application of the XGBoost model for discriminating the four subpopulations yielded a successful outcome, with a mean accuracy score of  $91.63 \pm 5.20\%$ . Among all generated markers, OPN06-350, OPA03-480, OPD02-500, OPF04-950, and OPN06-300 exhibited the greatest contributions to subpopulation differentiation, with the mean gain values of 8.31, 6.80, 6.40, 4.40, and 4.20, respectively (Figure 5).

Remarkably, 39.57% (55/139) of the generated markers were identified by the XGBoost model as having no contribution to subpopulation identifications, characterized by a mean gain of zero. Notably, 93.33% (28) of the unique and region-specific markers fell within this category, as their discriminatory power was limited to individual and regional detections (refer to Figure 5).

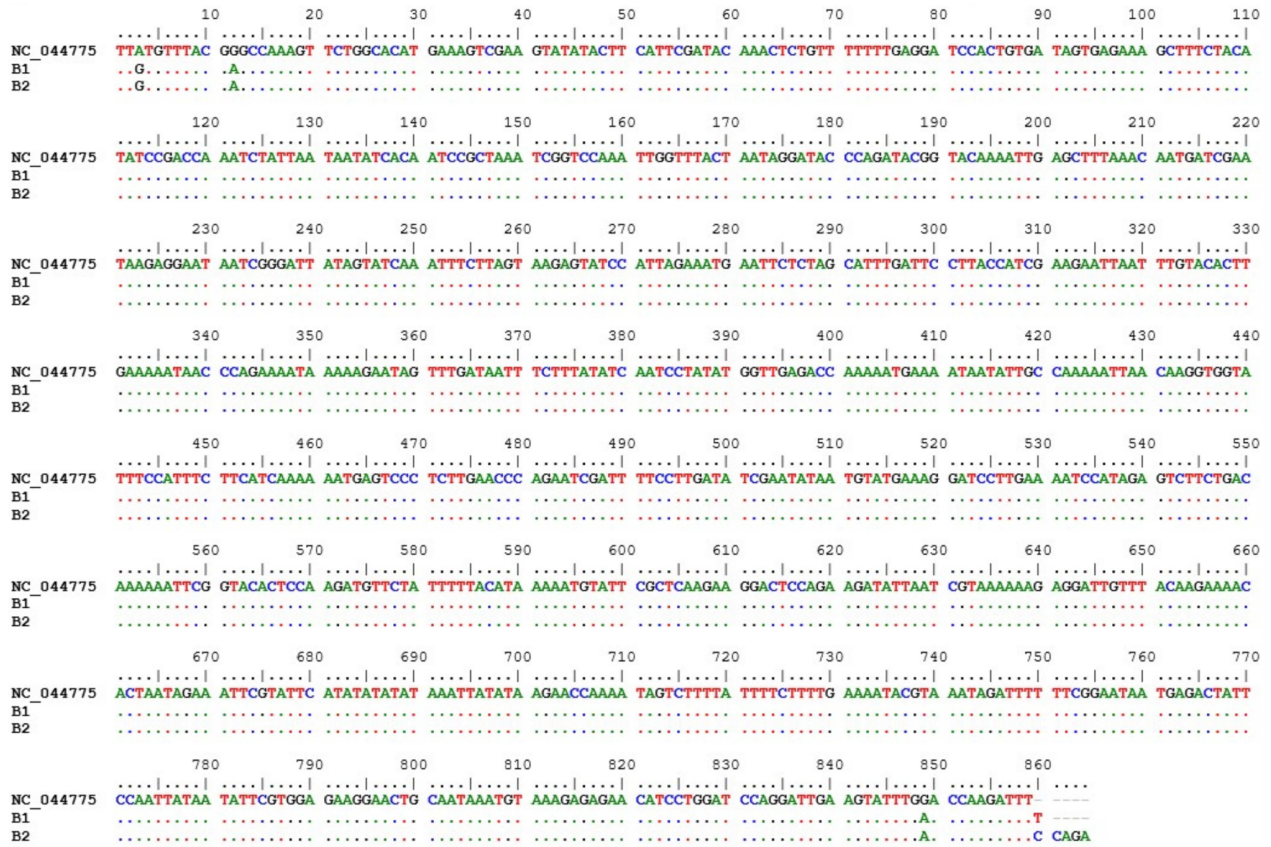
## 4. Discussion

In this study, we employed RAPD as our chosen method to observe genetic variations and differentiate ginger samples from various regions. This selection is based on a substantial body of evidence from previous studies, which has consistently identified RAPD as one of the most efficient marker systems for genetic diversity analysis (Ardiyani et al., 2021; Baruah et al., 2019; Blanco et al., 2016; Jatoi et al., 2006; Mohanty et al., 2014; Motlagh et al., 2023; Nayak et al., 2005; Zheng et al., 2015). However, our

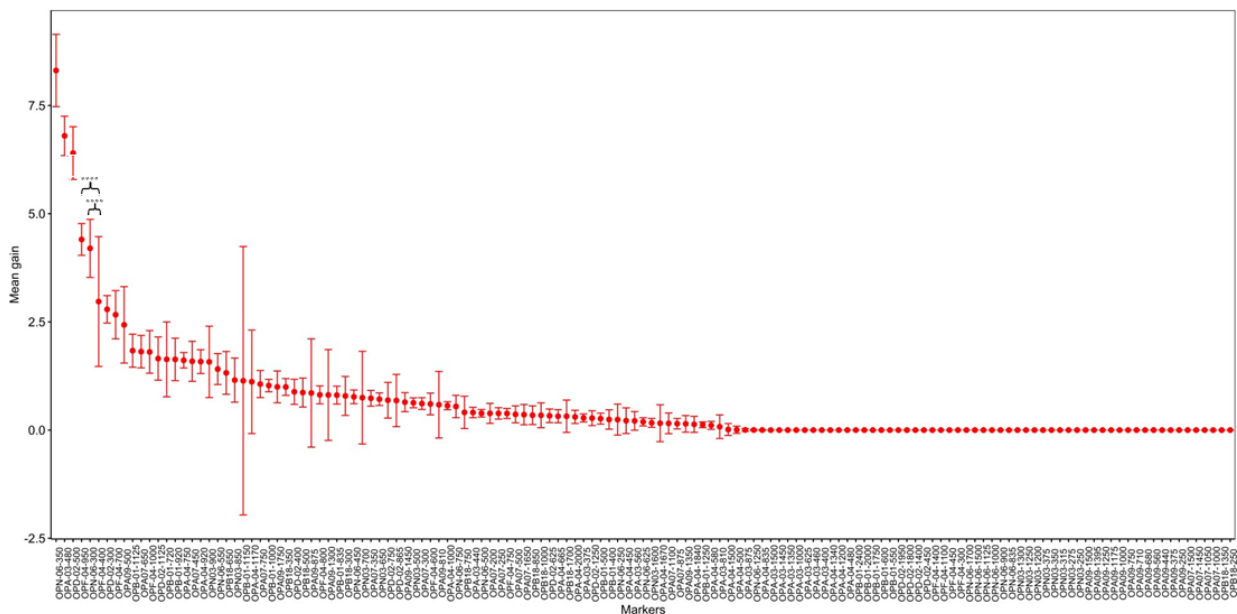


**Figure 3** Employing PCoA to delineate groupings among the ginger samples. The PCoA analysis utilizing RAPD markers resulted in the formation of two well-defined clusters. The red points denoting Hue's ginger samples were consolidated within a single cluster, while the remaining samples constituted a separate cluster. This discernible clustering underscores the genetic distinctiveness of ginger originating from Thuy Bieu Ward (Pop1) in comparison to specimens from the other geographical regions. Axes 1, 2, and 3 represent the three principal coordinates that account for the largest proportions of sample variation.

primer screening process revealed that not all RAPD primers could generate sufficient polymorphic marker numbers. Therefore, to maximize the possibilities of Hue's ginger identification and discrimination of other heavily mixed ginger sub-populations, we discarded 12/22 RAPD primers with low detected polymorphisms and employed only 10 primers (showing the polymorphic rate of 100%) for subsequent analyses. The impressive polymorphic rate of 100% remained unchanged when the 10 selected primers were utilized to amplify 105 ginger samples, surpassing the rates observed by Sajeev et al. (2011) and Baruah et al. (2019) (92.66% and 84.1%, respectively). The remarkable discriminatory power of RAPD is further exemplified in our study, where the markers accounted for a substantial 22% of the genetic variation among



**Figure 4** Aligning the *matK* sequences of B1 and B2 to the ginger reference chloroplast genome reveals substitutional variants at positions 3, 12, and 849.



**Figure 5** Assessment of marker significance in subpopulation discrimination utilizing the XGBoost model. The relative contributions of markers in delineating subpopulations are depicted through mean gain values. Among the 139 RAPD markers generated, OPN06-350, OPA03-480, OPD02-500, OPF04-950, and OPN06-300 emerge as the markers exerting the most pronounced impact on discrimination accuracy. Asterisks within the plot highlight statistically significant disparities between the mean gains of OPF04-950/OPN06-300 and OPF04-400 ( $p < 0.0001$ ). All values presented in the plot are represented as means  $\pm$  standard deviations.

subpopulations (Table 3), despite the fact that ginger chemotypes in Thua Thien Hue exhibit extensive mixing, and ginger cultivation predominantly relies on vegetative propagation (Zahid et al., 2021).

In addition to its robust discriminatory power, we opted for RAPD due to its capacity to yield a substantial number of amplicons, a characteristic exemplified by Baruah et al. (2019), who reported the observation of 196 bands. Our study further underscores RAPD's suitability as a marker, as it generated a total of 139 bands, reaffirming its effectiveness. Significantly, we identified RAPD markers specific to Hue's ginger in Thuy Bieu (OPA03-480 and OPB01-1150) and Huong Tho (OPA07-1050). These specific markers played a pivotal role in distinguishing Thuy Bieu samples from others, as evident in the PCoA plots (Figure 3). Notably, this differentiation of Thuy Bieu samples aligns with the findings reported by An et al. (2022). The presence of distinct genetic features in Thuy Bieu ginger is further substantiated by the identification of three substitution variants in the *matK* gene (Figure 4). Collectively, these diverse molecular pieces of evidence solidify the conclusion that Thuy Bieu represents the sole region in Thua Thien Hue where Hue's ginger is cultivated without genetic admixture. Conversely, samples from Huong Tho have unfortunately intermingled with unidentified ginger types, rendering them indistinguishable via PCoA, with the region-specific band of Huong Tho appearing in only 40% of the samples from this region (Table 2).

The utilization of machine-learning algorithms to differentiate plant accessions through PCR-based markers remains relatively limited. To the best of our knowledge, only three studies have delved into this subject (Beiki et al., 2012; Costa et al., 2019; Vásquez et al., 2010). Most notably, single nucleotide polymorphisms (SNPs) constitute the predominant marker type employed by the majority of studies implementing machine learning in plant research. In particular, machine learning algorithms are primarily harnessed for predicting plant traits based on SNP-based genotypic data, often generated through microarray or sequencing technologies (Kang et al., 2023; Sirsat et al., 2022; Zhang et al., 2023). However, when it comes to discriminating plant accessions or predicting plant phenotypes, the utilization of SNPs for constructing machine learning models presents significant drawbacks. Notably, the microarray technology is not universally applicable across all plant species, as it necessitates pre-existing genomic information. Furthermore, sequencing entire genomes or transcriptomes for a substantial number of plant accessions incurs substantial costs (Friel et al., 2021; You et al., 2018).

Hence, for the task of distinguishing plant accessions using genotype data, cost-effective, straightforward, universally applicable, and expeditious PCR-based marker techniques like RAPD (Babu et al., 2021) offer promising alternatives to SNPs. The efficacy of PCR-based markers in machine learning model development is underscored by the high accuracy scores achieved in our study (91.63%) and by Beiki et al. (2012) (100%). Moreover, in Costa et al. (2019) work, where PCR-based markers served as input genotypic data for the construction of a deep learning model, successful discrimination between two grapevine cultivars (420-A and Kober 5BB) with identical genetic origins and similar morphological characteristics was achieved. Beyond their high predictive accuracy, our XGBoost model also identified the most pivotal markers for subpopulation classification. These essential markers have the potential to significantly facilitate future ginger classification efforts in Thua Thien Hue. Specifically, future investigations could focus solely on RAPD primers that yield the most effective markers for classification, thereby streamlining the primer selection process from a pool of numerous RAPD candidates.

## 5. Supplementary material

The following supplementary materials are available for this article:

**Figure S1.** Amplifications of nine representative DNA samples by 12 unselected RAPD primers.

**Figure S2.** Electrophoresis of amplified DNA samples by ten selected RAPD primers.

**Table S1.** List of primers used for the screening process.

## References

- Ali, B. H., Blunden, G., Tanira, M. O., & Nemmar, A. (2008). Some phytochemical, pharmacological and toxicological properties of ginger (*Zingiber officinale* Roscoe): A review of recent research. *Food and Chemical Toxicology: An International Journal Published for the British Industrial Biological Research Association*, 46(2), 409–420. <https://doi.org/10.1016/j.fct.2007.09.085>
- An, N. H., Chien, T. T. M., Nhi, H. T. H., Nga, N. T. M., Phuc, T. T., Thuy, L. T. N., Thanh, T. V. B., Nguyen, P. T. T., & Phuong, T. T. B. (2020). The effects of sucrose, silver nitrate, plant growth regulators, and ammonium nitrate on microrhizome induction in perennially-cultivated ginger (*Zingiber officinale* Roscoe) from Hue, Vietnam. *Acta Agrobotanica*, 73(2), Article 7329. <https://doi.org/10.5586/aa.7329>
- An, N. H., Nguyen, P. T. T., Lan, T. T., Quang, H. T., & Phuong, T. T. B. (2022). Genetic diversity analysis based on retrotransposon microsatellite amplification polymorphisms (REMAP) for distinguishing the ginger chemotype of Thua Thien Hue (*Zingiber officinale* Roscoe) from other vietnamese ginger types. *Jordan Journal of Biological Sciences*, 15(02), 295–302. <https://doi.org/10.54319/jjbs/150218>
- Ardiyani, M., Senjaya, S. K., Maruzy, A., Widiyastuti, Y., Sulistyarningsih, L. D., & Susila, S. (2021). Genetic diversity of 'Bangle' (*Zingiber montanum* (J.Koenig) Link ex A.Dietr.) inferred from sequence-related amplified polymorphism markers. *Agriculture and Natural Resources*, 55(1), 105–112. <https://doi.org/10.34044/j.anres.2021.55.1.14>
- Ashraf, K., Ahmad, A., Chaudhary, A., Mujeeb, M., Ahmad, S., Amir, M., & Mallick, N. (2014). Genetic diversity analysis of *Zingiber officinale* Roscoe by RAPD collected from subcontinent of India. *Saudi Journal of Biological Sciences*, 21(2), 159–165. <https://doi.org/10.1016/j.sjbs.2013.09.005>
- Babu, K. N., Sheeja, T. E., Minoo, D., Rajesh, M. K., Samsudeen, K., Suraby, E. J., & Kumar, I. P. V. (2021). Random amplified polymorphic DNA (RAPD) and derived techniques. In P. Besse (Ed.), *Molecular plant taxonomy. Methods in molecular biology* (Vol. 2222, pp. 219–247). Springer. [https://doi.org/10.1007/978-1-0716-0997-2\\_13](https://doi.org/10.1007/978-1-0716-0997-2_13)
- Bag, B. (2018). Ginger processing in India (*Zingiber officinale*): A review. *International Journal of Current Microbiology and Applied Sciences*, 7(04), 1639–1651. <https://doi.org/10.20546/ijcmas.2018.704.185>
- Baliga, M. S., Haniadka, R., Pereira, M. M., D'Souza, J. J., Pallaty, P. L., Bhat, H. P., & Popuri, S. (2011). Update on the chemopreventive effects of ginger and its phytochemicals. *Critical Reviews in Food Science and Nutrition*, 51(6), 499–523. <https://doi.org/10.1080/10408391003698669>
- Baruah, J., Pandey, S. K., Begum, T., Sarma, N., Paw, M., & Lal, M. (2019). Molecular diversity assessed amongst high dry rhizome recovery ginger germplasm (*Zingiber officinale* Roscoe) from NE-India using RAPD and ISSR markers. *Industrial Crops and Products*, 129, 463–471. <https://doi.org/10.1016/j.indcrop.2018.12.037>
- Beiki, A. H., Saboor, S., & Ebrahimi, M. (2012). A new avenue for classification and prediction of olive cultivars using supervised and unsupervised algorithms. *Plos One*, 7(9), Article e44164. <https://doi.org/10.1371/journal.pone.0044164>
- Blanco, E. Z., Bajay, M. M., Siqueira, M. V. B. M., Zucchi, M. I., & Pinheiro, J. B. (2016). Genetic diversity and structure of Brazilian ginger germplasm (*Zingiber officinale*) revealed by AFLP markers. *Genetica*, 144(6), 627–638. <https://doi.org/10.1007/s10709-016-9930-1>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Chen, M., Liu, Q., Chen, S., Liu, Y., Zhang, C. H., & Liu, R. (2019). XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. *IEEE Access*, 7, 13149–13158. <https://doi.org/10.1109/ACCESS.2019.2893448>
- Costa, M. O., Capel, L. S., Maldonado, C., Mora, F., Mangolin, C. A., & Machado, M. D. F. P. D. S. (2019). High genetic differentiation of grapevine rootstock varieties determined by molecular markers and artificial neural networks. *Acta Scientiarum. Agronomy*, 42, Article e43475. <https://doi.org/10.4025/actasciagron.v42i1.43475>
- Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, 19(1), 11–15.
- El baroty, G., Abd El Baky, H., Farag, R., & Saleh, M. (2010). Characterization of antioxidant and antimicrobial compounds of cinnamon and ginger essential oils. *African Journal of Biochemistry Research*, 4, 167–174.
- Engdal, S., Klepp, O., & Nilsen, O. G. (2009). Identification and exploration of herb-drug combinations used by cancer patients. *Integrative Cancer Therapies*, 8(1), 29–36. <https://doi.org/10.1177/1534735408330202>

- Friel, J., Bombarely, A., Fornell, C. D., Luque, F., & Fernández-Ocaña, A. M. (2021). Comparative analysis of genotyping by sequencing and whole-genome sequencing methods in diversity studies of *Olea europaea* L. *Plants*, 10(11), Article 2514. <https://doi.org/10.3390/plants10112514>
- Geiger, J. L. (2005). The essential oil of ginger, *Zingiber officinale*, and anaesthesia. *International Journal of Aromatherapy*, 15(1), 7–14. <https://doi.org/10.1016/j.ijat.2004.12.002>
- Govindarajan, V. S., & Connell, D. W. (1982). Ginger-chemistry, technology, and quality evaluation: Part 1. *Critical Reviews in Food Science and Nutrition*, 17(1), 1–96. <https://doi.org/10.1080/10408398209527343>
- Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41, 95–98.
- Harisaranraj, R., Kumarasamy, S., & Saravanababu, S. (2009, January 1). *DNA finger printing analysis among eight varieties of Zingiber officinale Rosc. By using RAPD markers*. [https://www.researchgate.net/publication/237810194\\_DNA\\_Finger\\_Printing\\_Analysis\\_among\\_Eight\\_Varieties\\_of\\_Zingiber\\_officinale\\_Rosc\\_By\\_Using\\_RAPD\\_Markers](https://www.researchgate.net/publication/237810194_DNA_Finger_Printing_Analysis_among_Eight_Varieties_of_Zingiber_officinale_Rosc_By_Using_RAPD_Markers)
- Hien, L. T. B., Quy, L. T. M., Thuy, N. L. L., & Hoai, N. T. (2018). Study on extraction process, chemical composition and antibacterial activity of ginger oil in Thua Thien Hue. *Journal of Medicine and Pharmacy*, 8(3), 24–30. <https://doi.org/10.34071/jmp.2018.3.4>
- Ismail, N. A., Rafii, M. Y., Mahmud, T. M. M., Hanafi, M. M., & Miah, G. (2016). Molecular markers: A potential resource for ginger genetic diversity studies. *Molecular Biology Reports*, 43(12), 1347–1358. <https://doi.org/10.1007/s11033-016-4070-3>
- Jatoi, S. A., Kikuchi, A., Yi, S. S., Naing, K. W., Yamanaka, S., Watanabe, J. A., & Watanabe, K. N. (2006). Use of rice SSR markers as RAPD markers for genetic diversity analysis in *Zingiberaceae*. *Breeding Science*, 56(2), 107–111. <https://doi.org/10.1270/jsbbs.56.107>
- Joblib: Computing with Python functions*. (n.d.). (2023, September 8) <https://github.com/joblib/joblib>
- Kang, Y., Choi, C., Kim, J. Y., Min, K. D., & Kim, C. (2023). Optimizing genomic selection of agricultural traits using K-wheat core collection. *Frontiers in Plant Science*, 14, Article 1112297. <https://www.frontiersin.org/articles/10.3389/fpls.2023.1112297>
- Kubra, I. R., & Jaganmohanrao, L. (2012). An overview on inventions related to ginger processing and products for food and pharmaceutical applications. *Recent Patents on Food, Nutrition & Agriculture*, 4(1), 31–49.
- Mao, Q. Q., Xu, X. Y., Cao, S. Y., Gan, R. Y., Corke, H., Beta, T., & Li, H. B. (2019). Bioactive compounds and bioactivities of ginger (*Zingiber officinale* Roscoe). *Foods*, 8(6), Article 185. <https://doi.org/10.3390/foods8060185>
- Mia, M. S., Patwary, A. K., Hassan, L., Hasan, M. M., Alam, M. A., Latif, M. A., Monjurul Alam Mondal, M., & Puteh, A. B. (2014). Genetic diversity analysis of ginger (*Zingiber officinale* Roscoe.) Genotypes using RAPD markers. *Life Science Journal*, 11(8), 90–94.
- Mohanty, S., Panda, M. K., Acharya, L., & Nayak, S. (2014). Genetic diversity and gene differentiation among ten species of *Zingiberaceae* from Eastern India. *3 Biotech*, 4(4), 383–390. <https://doi.org/10.1007/s13205-013-0166-9>
- Motlagh, M. R. S., Kulus, D., Kaviani, B., & Habibollahi, H. (2023). Exploring fungal endophytes as biocontrol agents against rice blast disease. *Acta Agrobotanica*, 76, Article 182943. <https://doi.org/10.5586/aa/182943>
- Nair, K. P. P. (2013). *The agronomy and economy of turmeric and ginger*. Elsevier. <https://doi.org/10.1016/C2011-0-07514-2>
- Nayak, S., Naik, P. K., Acharya, L., Mukherjee, A. K., Panda, P. C., & Das, P. (2005). Assessment of genetic diversity among 16 promising cultivars of ginger using cytological and molecular markers. *Zeitschrift Für Naturforschung C*, 60(5–6), 485–492. <https://doi.org/10.1515/znc-2005-5-618>
- Nguyen-Hoang, A., Sandell, F. L., Himmelbauer, H., & Dohm, J. C. (2024). Spinach genomes reveal migration history and candidate genes for important crop traits. *NAR Genomics and Bioinformatics*, 6(2), Article lqae034. <https://doi.org/10.1093/nargab/lqae034>
- Nicoll, R., & Henein, M. Y. (2009). Ginger (*Zingiber officinale* Roscoe): A hot remedy for cardiovascular disease? *International Journal of Cardiology*, 131(3), 408–409. <https://doi.org/10.1016/j.ijcard.2007.07.107>
- Peakall, R., & Smouse, P. E. (2006). Genalex 6: Genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, 6(1), 288–295. <https://doi.org/10.1111/j.1471-8286.2005.01155.x>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85), 2825–2830.
- Prasad, S., & Tyagi, A. K. (2015). Ginger and its constituents: Role in prevention and treatment of gastrointestinal cancer. *Gastroenterology Research and Practice*, 2015, Article 142979. <https://doi.org/10.1155/2015/142979>

- Rajathi, A. A., Sundarraj, A. A., Leslie, S., & Shree, M. P. (2017). Processing and medicinal uses of cardamom and ginger – A review. *Journal of Pharmaceutical Sciences and Research*, 9(11), 2117–2122.
- Sajeev, S., Roy, A. R., Ingirai, B., Pattanayak, A., & Deka, B. C. (2011). Genetic diversity analysis in the traditional and improved ginger (*Zingiber officinale* Rosc.) clones cultivated in North-East India. *Scientia Horticulturae*, 128(3), 182–188. <https://doi.org/10.1016/j.scienta.2011.01.024>
- Sasidharan, I., & Menon, A. N. (2010). Comparative chemical composition and antimicrobial activity fresh & dry ginger oils (*Zingiber officinale* Roscoe). *International Journal of Current Pharmaceutical Research*, 2(4), 40–43.
- Sharma, P., Singh, V., Ali, M., & Ali, M. (2016). Chemical composition and antimicrobial activity of fresh rhizome essential oil of *Zingiber officinale* Roscoe. *Pharmacognosy Journal*, 8(3), 185–190. <https://doi.org/10.5530/pj.2016.3.3>
- Shukla, Y., & Singh, M. (2007). Cancer preventive properties of ginger: A brief review. *Food and Chemical Toxicology*, 45(5), 683–690. <https://doi.org/10.1016/j.fct.2006.11.002>
- Sirsat, M. S., Oblessuc, P. R., & Ramiro, R. S. (2022). Genomic prediction of wheat grain yield using machine learning. *Agriculture*, 12(9), Article 9. <https://doi.org/10.3390/agriculture12091406>
- Vasala, P. (2012). Ginger. In K. Peter (Ed.), *Handbook of herbs and spices* (2nd ed., pp. 319–335). Woodhead Publishing.
- Vásquez, J. L., Vásquez, J., Briceño, J. C., Castillo, E., & Travieso, C. M. (2010). Feature selection of RAPD haplotypes for identifying peach palm (*Bactris gasipaes*) landraces using SVM. *WSEAS Transactions on Computers*, 9(3), 205–214.
- Yeh, H., Chuang, C., Chen, H., Wan, C., Chen, T., & Lin, L. (2014). Bioactive components analysis of two various gingers (*Zingiber officinale* Roscoe) and antioxidant effect of ginger extracts. *LWT - Food Science and Technology*, 55(1), 329–334. <https://doi.org/10.1016/j.lwt.2013.08.003>
- You, Q., Yang, X., Peng, Z., Xu, L., & Wang, J. (2018). Development and applications of a high throughput genotyping tool for polyploid crops: Single nucleotide polymorphism (SNP) array. *Frontiers in Plant Science*, 9, Article 104. <https://doi.org/10.3389/fpls.2018.00104>
- Zahid, N. A., Jaafar, H. Z. E., & Hakimian, M. (2021). Micropropagation of ginger (*Zingiber officinale* Roscoe) 'Bentong' and evaluation of its secondary metabolites and antioxidant activities compared with the conventionally propagated plant. *Plants*, 10(4), Article 630. <https://doi.org/10.3390/plants10040630>
- Zhang, F., Kang, J., Long, R., Li, M., Sun, Y., He, F., Jiang, X., Yang, C., Yang, X., Kong, J., Wang, Y., Wang, Z., Zhang, Z., & Yang, Q. (2023). Application of machine learning to explore the genomic prediction accuracy of fall dormancy in autotetraploid alfalfa. *Horticulture Research*, 10(1), Article uhac225. <https://doi.org/10.1093/hr/uhac225>
- Zheng, W. H., Zhuo, Y., Liang, L., Ding, W. Y., Liang, L. Y., & Wang, X. F. (2015). Conservation and population genetic diversity of *Curcuma wenyujin* (Zingiberaceae), a multifunctional medicinal herb. *Genetics and Molecular Research: GMR*, 14(3), 10422–10432. <https://doi.org/10.4238/2015.September.8.3>
- Zhu, S., Liu, Q., Qiu, S., Dai, J., & Gao, X. (2022). DNA barcoding: An efficient technology to authenticate plant species of traditional Chinese medicine and recent advances. *Chinese Medicine*, 17(1), Article 112. <https://doi.org/10.1186/s13020-022-00655-y>