Ma *et al. BMC Plant Biology* (2022) 22:467
https://doi.org/10.1186/s12870-022-03827-y

BMC Plant Biology

RESEARCH Open Access

De novo transcriptome assembly, gene annotation, and EST-SSR marker development of an important medicinal and edible crop, *Amomum tsaoko* (Zingiberaceae)

Mengli Ma[1], Hengling Meng[1], En Lei[2], Tiantao Wang[2], Wei Zhang[1,2] and Bingyue Lu[1,2*]

Abstract

Background: *Amomum tsaoko* is a medicinal and food dual-use crop that belongs to the Zingiberaceae family. However, the lack of transcriptomic and genomic information has limited the understanding of the genetic basis of this species. Here, we performed transcriptome sequencing of samples from different *A. tsaoko* tissues, and identified and characterized the expressed sequence tag-simple sequence repeat (EST-SSR) markers.

Results: A total of 58,278,226 high-quality clean reads were obtained and de novo assembled to generate 146,911 unigenes with an N50 length of 2002 bp. A total of 128,174 unigenes were successfully annotated by searching seven protein databases, and 496 unigenes were identified as annotated as putative terpenoid biosynthesis-related genes. Furthermore, a total of 55,590 EST-SSR loci were detected, and 42,333 primer pairs were successfully designed. We randomly selected 80 primer pairs to validate their polymorphism in *A. tsaoko*; 18 of these primer pairs produced distinct, clear, and reproducible polymorphisms. A total of 98 bands and 96 polymorphic bands were amplified by 18 pairs of EST-SSR primers for the 72 *A. tsaoko* accessions. The Shannon's information index (I) ranged from 0.477 (AM208) to 1.701 (AM242) with an average of 1.183, and the polymorphism information content (PIC) ranged from 0.223 (AM208) to 0.779 (AM247) with an average of 0.580, indicating that these markers had a high level of polymorphism. Analysis of molecular variance (AMOVA) indicated relatively low genetic differentiation among the six *A. tsaoko* populations. Cross-species amplification showed that 14 of the 18 EST-SSR primer pairs have transferability between 11 Zingiberaceae species.

Conclusions: Our study is the first to provide transcriptome data of this important medicinal and edible crop, and these newly developed EST-SSR markers are a very efficient tool for germplasm evaluation, genetic diversity, and molecular marker-assisted selection in *A. tsaoko*.

Keywords: *Amomum tsaoko*, Transcriptome sequencing, EST-SSR markers, Genetic diversity

Background

Amomum tsaoko is a perennial evergreen tufted herb of Zingiberaceae; the whole plant has a spicy taste. The dried fruit (also called Cao-Guo in China) of *A. tsaoko* is an important crude drug in traditional Chinese medicine (TCM), such as 'Cao-Guo-Zhi-Mu-Tang,' 'Da-Yuan-Yin,' 'Cao-Guo-Si-Wei-Tang,' and 'Li-Gan-Shi-Liu-Ba-Wei-San,' which clear dampness, resolve phlegm, warm the

*Correspondence: lby202@126.com

[2] College of Biological and Agricultural Sciences, Honghe University, Mengzi 661199, China
Full list of author information is available at the end of the article

spleen, and dispel colds. In addition, *A. tsaoko* is also a top-grade spice known as one of the "five spices" in food seasoning [1]. In recent years, *A. tsaoko* has been considered to have a broader utilization value and proven to have biological activities, such as anti-oxidation [2, 3], antibacterial [4], anti-inflammation [5], antidiabetic [6], anti-tumor [7], and anticonvulsant properties [8]. In addition, *A. tsaoko* is a TCM that has been prescribed for the treatment of COVID-19 [9–12]. As an aromatic Chinese herbal medicine used for both medicinal and edible purposes, essential oil is the most important active ingredient of *A. tsaoko*, and its content determines the quality of *A. tsaoko* (Pharmacopoeia of the People's Republic of China, 2020). Recent studies have shown that *A. tsaoko* essential oil has a significant inhibitory effect on COVID-19 [13]. The monoterpene 1,8-cineole was found to be the major constituent (34.6–45.24%) of the essential oil in *A. tsaoko*, and has well-known antiviral, anti-inflammatory, antimicrobial, and pain-relieving effects [14]. In recent years, the cloning and expression analysis of functional genes involved in terpenoid biosynthesis has become a popular topic of research. He et al. [15] used transcriptome sequencing to identify a number of terpenoid synthesis-related genes in *Amomum villosum*, including five monoterpene synthase genes *AvTPS1–AvTPS5*. However, until now, there have been no reports about regulated genes that are involved in the terpenoid biosynthesis of *A. tsaoko*.

A. tsaoko is a high-altitude medicinal crop that grows in humid forests, narrowly distributed in the southern Yunnan Province of China and northern Laos and Vietnam at high altitudes between 1100 and 1800 m in mountainous regions [16, 17]. Due to excessive harvesting and the destruction of the original habitat of *A. tsaoko*, wild resources are almost extinct, and it was listed as a "Nearly Endangered Species" on the IUCN Red List in 2012. Genetic diversity evaluation can provide important reference information for the identification and evaluation of germplasm resources and the selection of elite germplasm. Molecular marker technology is an effective tool for analyzing plant genetic diversity, which has the advantages of abundant quantity, high polymorphism, direct expression in the form of DNA, and is not affected by the environment [18, 19]. SSRs, also known as microsatellite DNA, are short, repeated DNA sequences present throughout the genome. It has the advantages of good reproducibility, codominance, abundant polymorphisms, and easy detection [20]. In recent years, SSR markers have been widely used in genetic diversity analysis [21, 22], linkage genetic map construction and QTL identification [23, 24], and marker-assisted breeding [25, 26].

As a TCM and spice, previous studies of *A. tsaoko* mainly focused on active component extraction, identification of chemical components, and pharmacological effects [27–32]; however, molecular studies of *A. tsaoko* are lacking. This limits the excellent germplasm selection and breeding utilization of *A. tsaoko*. Some researchers have begun to study the genetic diversity of *A. tsaoko* at the phenotypic and molecular levels. Zhang et al. [33] analyzed the phenotypic characteristics of *A. tsaoko* in nine producing areas of China. The results showed that the number of fruit ridges, the number of seeds per fruit, and the vertical diameter of fruit had the greatest variation. Genetic diversity and population genetic differentiation of *A. tsaoko* from eight different producing areas were analyzed by 12 RAPD markers [34]. More recently, the genetic diversity of 91 *A. tsaoko* accessions from southwest China was studied by SRAP and ISSR markers [17]. A few reports on SSR development and the genetic diversity of *A. tsaoko* have recently been published [1, 35], but the number of markers is insufficient to conduct comprehensive genetic studies.

With the rapid development of next-generation sequencing technology and lower sequencing costs, large-scale RNA-seq provides an important information for functional gene mining and molecular marker development in non-model species [36, 37]. Here, we present the first transcriptome of *A. tsaoko* using the Illumina HiSeqTM 4000 sequencing platform. The obtained transcriptome data advance our understanding of the function categories from the annotated genes on this species, and the development of EST-SSR markers will provide an important basis for germplasm evaluation, genetic diversity analysis, and molecular breeding of *A. tsaoko*.

Results

Sequencing and de novo assembly

In this study, 59,876,622 raw reads were obtained using the Illumina HiSeq 4000 platform. After strict quality control, 97.33% clean reads (58,278,226) were obtained with 96.28% Q20 and 94.25% Q30 bases. A total of 199,191 transcripts were obtained by de novo assembly of clean reads with Trinity software. After clustering transcripts and removing redundancy, 146,911 unigenes were obtained, with an average length of 1527 bp and an N50 value of 2002 bp (Table 1).

The assembled unigenes of *A. tsaoko* were functionally annotated against seven public databases (Table 2, Fig. 1A). A total of 123,420 unigenes (84.01%) were annotated successfully in the NR database, while 95,786 (65.20%) were annotated in the NT database. At the same time, 21,543 unigenes were annotated in all databases, and 128,174 unigenes (87.24%) were successfully annotated with at least one database. The species distribution showed that 80.4% matched *Musa acuminata*, and 5.6% matched *Elaeis guineensis*. The matching degrees of

Table 1 Characterization of *Amomum tsaoko* transcripts

Category	Item	Number
Sequenced reads	raw reads	59,876,622
	clean reads	58,278,226
	clean bases	8.74G
	error (%)	0.03
	Q20 (%)	96.28
	Q30 (%)	94.25
	GC content (%)	50.22
Transcripts	total number	199,191
	minimum length (bp)	201
	mean length (bp)	1219
	maximum length (bp)	10,949
	N50	1888
	N90	571
	total nucleotides	242,805,690
Unigenes	total number	146,911
	minimum length (bp)	201
	mean length (bp)	1527
	maximum length (bp)	10,949
	N50	2002
	N90	801
	total nucleotides	224,308,460

Table 2 Functional annotation of *A. tsaoko* in seven databases

Annotation database	Number of unigenes	Percentage (%)
Annotated in NR	123,420	84.01
Annotated in NT	95,786	65.20
Annotated in KEGG	53,059	36.11
Annotated in SwissProt	96,740	65.84
Annotated in PFAM	89,573	60.97
Annotated in GO	89,750	61.09
Annotated in KOG	36,621	24.92
Annotated in all databases	21,543	14.66
Annotated in at least one database	128,174	87.24
Total unigenes	146,911	100

Phoenix dactylifera, *Nelumbo nucifera*, and *Vitis vinifera* were 4.5, 0.6, and 0.5%, respectively, and 8.3% for other species (Fig. 1B).

Furthermore, for the KOG classification, 36,621 putative unigenes of *A. tsaoko* were classified into 25 clusters (Fig. 2). Among these categories, 'general function prediction only' (5039; 13.76%); 'posttranslational modification, protein turnover, chaperones' (4993; 13.63%); and 'translation, ribosomal structure and biogenesis' (3332; 9.10%) were the dominant groups, while only a few unigenes were annotated as 'cell motility' (35; 0.09%) and 'extracellular structures' (26; 0.07%). Furthermore, there were 89,750 unigenes categorized into three main GO categories: biological processes (233,859; 47.53%); cellular components (145,239; 29.52%); and molecular functions (112,917; 22.95%) (Fig. 3). Within the three categories, 'binding' (53,980), 'cellular process' (52,634), 'metabolic process' (49,492), 'catalytic activity' (41,468), and 'single-organism process' (37,122) were the most prevalent. Following searches against the KEGG database, 53,059 unigenes were classified into five categories, which were distributed in 130 metabolic pathways. The largest category was comprised of 'metabolism' (21,001; 53.69%), followed by 'genetic information processing' (10,413; 26.62%), 'cellular processes' (2887, 7.38%), 'organismal systems' (2559, 6.54%), and 'environmental information processing' (2255, 5.77%) (Fig. 4). In addition, 96,740 (65.84%) and 89,573 (60.97%) unigenes matched the SwissProt and PFAM databases, respectiv ely.

Importantly, KEGG analysis showed that 496 unigenes were annotated as being directly involved in the metabolism of terpenoids, including the terpenoid backbone (338 unigenes), monoterpenoid (22 unigenes), diterpenoid (81 unigenes), and sesquiterpenoid and triterpenoid (55 unigenes) biosynthesis pathways (Table 3). In the monoterpenoid biosynthesis pathway, two unigenes encoding alpha-terpineol synthase, 15 unigenes encoding neomenthol dehydrogenase, one unigene encoding 1,8-cineole synthase, and four unigenes encoding linalool synthase were significantly enriched. The FPKM values of Cluster-26,586.35631, Cluster-26,586.48848, Cluster-26,586.69915, Cluster-26,586.65004, Cluster-18,694.1, and Cluster-26,586.96889 were greater than 10 (Fig. 5A). Notably, 1,8-cineole synthase (Cluster-24,134.3) was initially identified in this species. Multiple sequence alignment revealed that Cluster-24,134.3 contained the DDXXD motif, which is conserved in angiosperm monoterpene synthases (Fig. 5B). Cluster analysis showed that Cluster-24,134.3 had the highest homology with the monoterpene synthase gene *AvTPS2* in *Amomum villosum* (Fig. 5C).

Development of EST-SSR markers

SSR loci were identified within the *A. tsaoko* transcriptome using the MISA software. Among the 146,911 non-redundant unigenes, we identified 55,590 potential SSRs, of which 3794 had a compound formation. There were 10,465 unigenes containing more than one SSR. An average of one SSR site was found every 1.53 kb. The largest fraction of SSRs were mononucleotide SSRs (26, 742, 48.11%), followed by trinucleotide repeats (13,849; 24.91%), dinucleotide repeats (12,716; 22.87%), tetranucleotide repeats (1169; 2.10%), hexanucleotide repeats (566; 1.02%), and

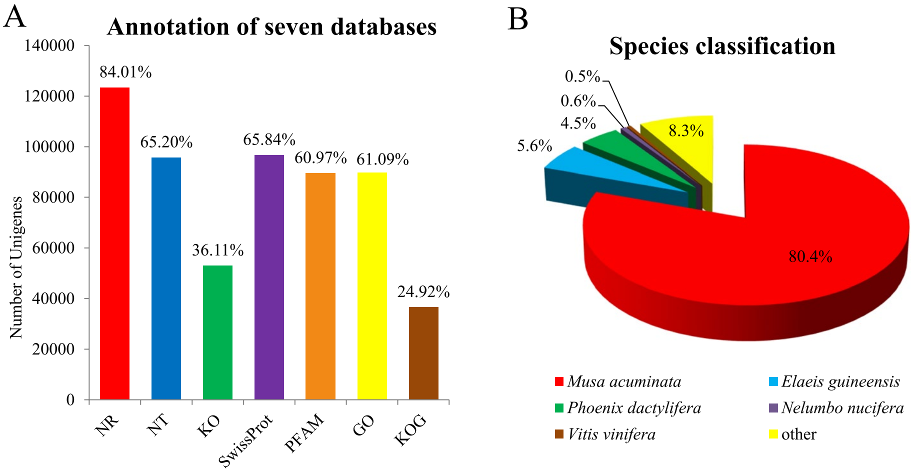


Fig. 1 Functional annotation (**A**) and species distribution (**B**) of unigenes in *Amomum tsaoko*

pentanucleotide repeats (548; 0.99%) (Fig. 6, Table 4). Among the different SSR repeat-type classes, the most dominant repeat motifs were A/T (25,819; 46.44%), AG/CT (7554; 13.59%), AT/AT (3971; 7.14%), AGG/CCT (3390; 6.10%), AAG/CTT (2709; 4.87%), and CCG/CGG (2330; 4.19%). The remaining motif types accounted for 17.67% of these repeats (Table 4). The tandem repeat numbers of these SSRs ranged from 5 to 86, and 10 tandem repeats (12,910; 23.22%) was the most common number, followed by 5 (7789; 14.01%), 6 (6749; 12.14%), 11 (5655; 10.17%), 12 (3975; 7.15%), and 7 (3792; 6.82%) (Table 5).

Validation of EST-SSR markers

A total of 42,333 primer pairs were successfully designed based on SSR flanking sequences. Four accessions of *A. tsaoko* were used for primer amplification specificity and efficiency testing. Eighty pairs of primers were

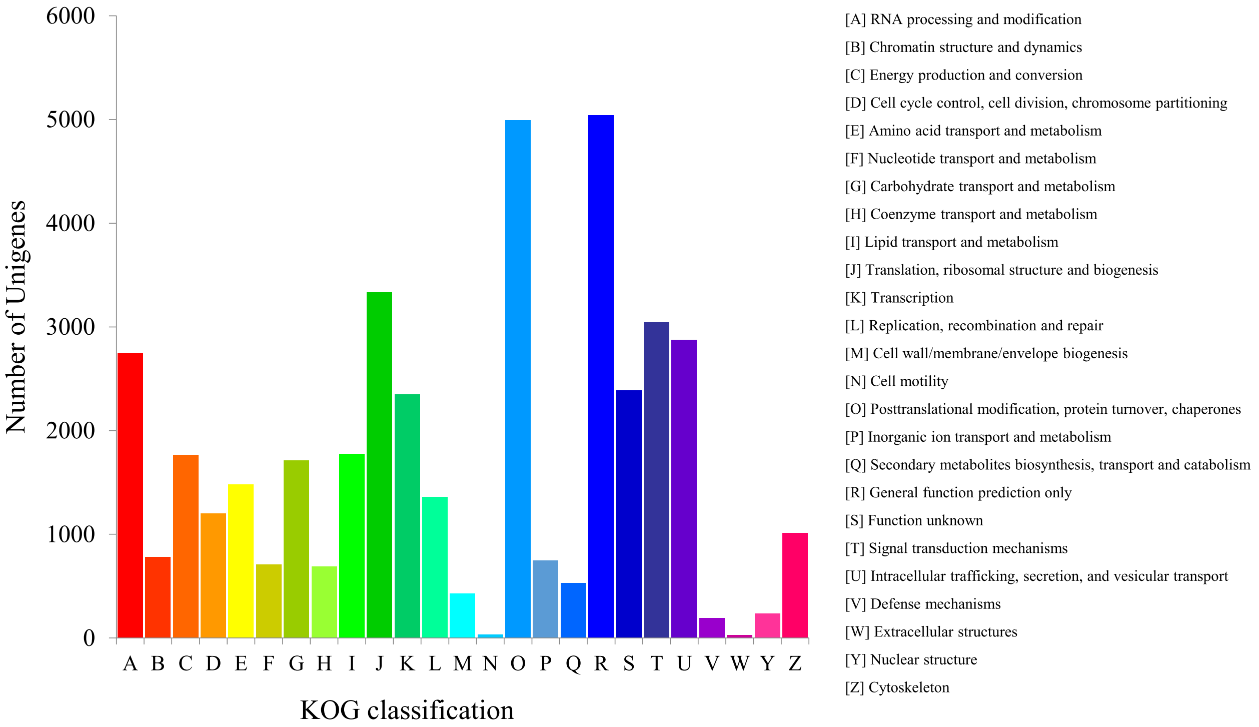


Fig. 2 Eukaryotic clusters of orthologous groups (KOG) classification of unigenes. A total of 36,621 unigenes with NR hits were grouped into 25 functional groups. The *y*-axis indicates the number of unigenes in a specific functional cluster. The *x*-axis indicates the function class

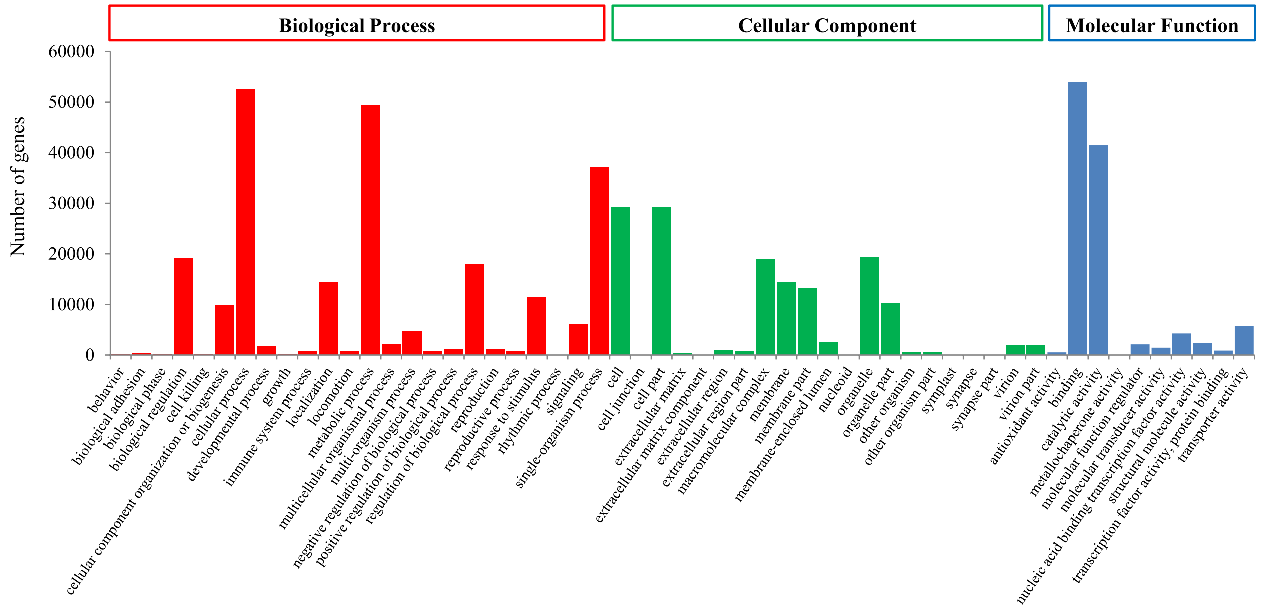


Fig. 3 Gene ontology (GO) classification of unigenes. A total of 89,750 unigenes were categorized into three main categories: biological process, cellular component, and molecular function. The *x*-axis indicates the subgroups in GO annotation, while the *y*-axis indicates the percentage of specific categories of genes in each main category

randomly selected for synthesis; 64 of the EST-SSRs successfully amplified DNA fragments with the expected band sizes. Eventually, 18 primer pairs yielding the best amplification were chosen to analyze the genetic diversity of 72 *A. tsaoko* accessions from six cultivated populations (Table 6). SSR amplification profiles of primer pairs AM242, AM272, AM273, and AM278 are shown in Fig. 7. A total of 98 bands were scored, with an average of 5.4 bands per primer and a range of 3 (AM213) to 12 (AM255). The average effective number (Ne) of alleles per locus was 2.961, ranging from 1.311 (AM208) to 4.852 (AM242). The Shannon' information index (I) ranged from 0.477 (AM208) to 1.701 (AM242) with a mean of 1.183. The observed heterozygosity (Ho) ranged from 0.264 (AM208) to 1.000 (AM225) with a mean of 0.594, and the expected heterozygosity (He) ranged from 0.237 (AM208) to 0.794 (AM242) with a mean of 0.613. The value of PIC ranged from 0.223 (AM208) to 0.779 (AM247) with a mean value of 0.580. The HWE analysis showed that six loci (AM224, AM225, AM237, AM247, AM248, and AM278) significantly deviated from HWE ($P<0.05$), and the remaining 12 loci (AM203, AM206, AM207, AM208, AM213, AM218, AM223, AM242, AM255, AM272, AM273, and AM279) were in accordance with HWE (Table 7). The frequencies of null alleles were low (<0.20) for each locus, except AM247.

Population genetic diversity

At the population level, the YY population scored higher values in Ne, I, and He, while Na and Ho were higher in the LVC and LC populations, respectively. Notably, the number of private alleles (PAr) in the JP and LVC populations (0.167) was higher than that in the PB, YY, LC, and YX populations. Mean fixation index (F) values ranged from −0.145 in the YX population to 0.015 in the YY population, with an average of −0.066 (Table 8). The samples did not cluster according to the six different sampling sites based on the UPGMA cluster and PCoA, and all samples from different populations were not separated from each other (Fig. 8). The studied populations showed relatively low between-population genetic differentiation (Fst average = 0.052; min = 0.029, max = 0.091) and small genetic distance (GD average = 0.155, min = 0.090, max = 0.263) (Table 9). Consistent with these results, AMOVA also showed that 90% molecular variance was found within populations (Table 10).

Cross-species transferability

To test the transferability of the novel EST-SSR markers in different Zingiberaceae species, 18 polymorphic EST-SSR markers were tested for amplification in 11 other species. Fourteen primer pairs were successfully amplified, and the transferability rates (TR) ranged from

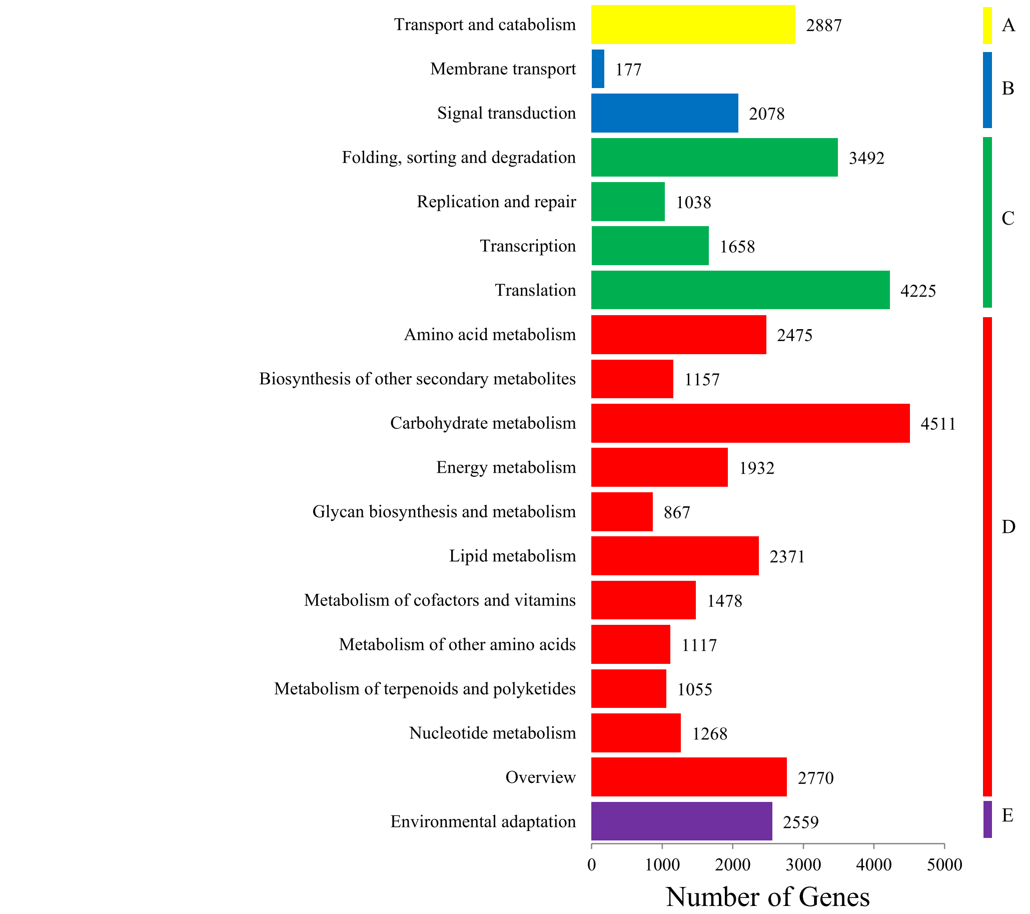


Fig. 4 Clusters of orthologous groups based on KEGG classification. **A** Cellular process; **B** Environmental information processing; **C** Genetic information processing; **D** Metabolism; **E** Organismal systems

27.78% (*Alpinia coriandriodora*) to 77.78% (*Alpinia zerumbet* and *Curcuma phaeocaulis*), with an average of 68.18% (Fig. 9A). Furthermore, UPGMA phylogenetic analysis based on transferable SSR markers showed that 12 Zingiberaceae species clustered into two groups (Fig. 9B): the first class (the green shaded area indicating the genera *Amomum* and *Alpinia*) and the second class (the red shaded area indicating the genera *Kaempferia*, *Hedychium*, and *Curcuma*).

Table 3 Overview of the terpene biosynthetic pathways

Pathway	No. of unigenes	Pathway ID
Terpenoid backbone biosynthesis	338	ko00900
Monoterpenoid biosynthesis	22	ko00902
Diterpenoid biosynthesis	81	ko00904
Sesquiterpenoid and triterpenoid biosynthesis	55	ko00909

Discussion

Characterization of the *Amomum tsaoko* transcriptome

For species without reference genomes, transcriptome sequencing is considered the most effective way of mining functional genes and developing novel molecular markers [38, 39]. The results of this study represent the first report of transcriptome analysis and EST-SSR detection in *A. tsaoko*. To maximize the transcriptome information gained, mixed samples of five organs, namely roots, stem, leaves, flowers, and fruit, were collected from *A. tsaoko* for paired-end transcriptome sequencing. A total of 58,278,226 clean and high-quality reads with a 94.25% Q30 level, which ensures the quality of sequencing. In the present study, the N50 sizes of generated unigenes (2002 bp) were obviously longer than those in the Zingiberaceae family, such as *A. villosum* (N50 = 1381 bp) [15], *C. wenyujin* (N50 = 1566 bp) [40], *Curcuma longa* (N50 = 1515 bp) [41], *Curcuma alismatifolia* (N50 = 1501 bp) [42], *Zingiber officinale* (N50 = 1077 bp) [43], and *Elettaria cardamomum* (N50 = 616–664 bp) [44].

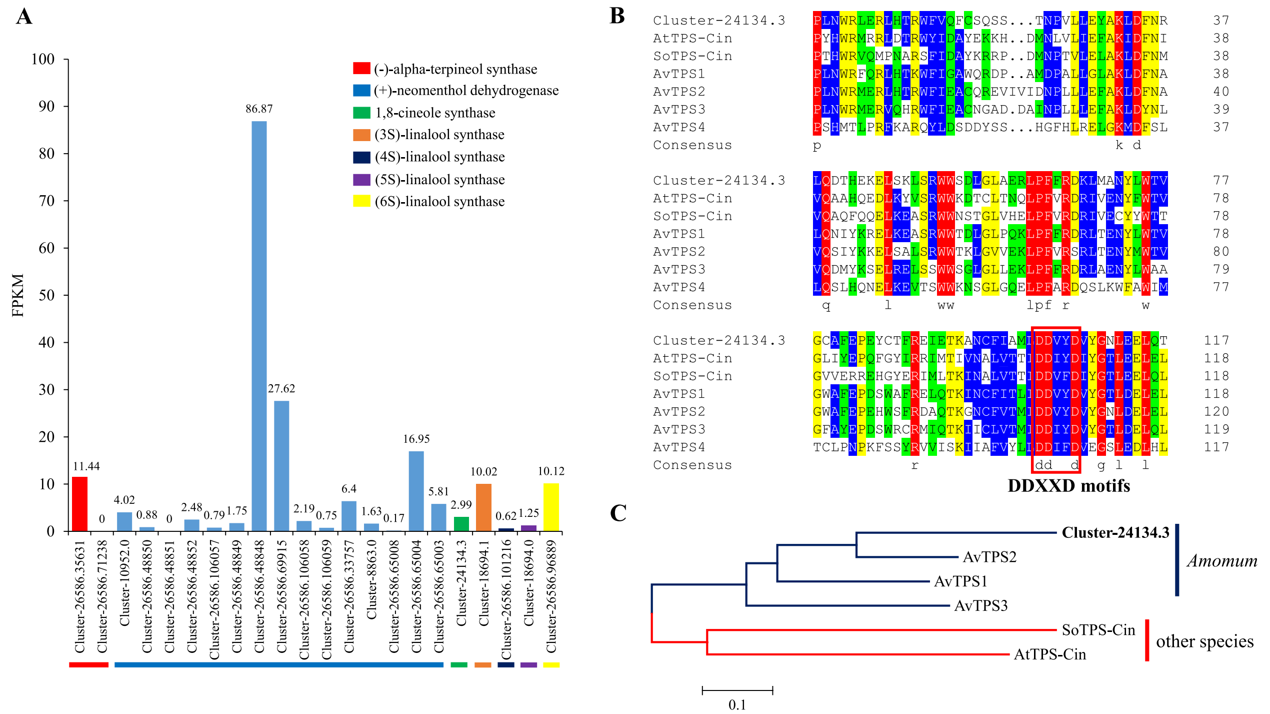


Fig. 5 Unigene analysis of potential monoterpene metabolic pathway in *A. tsaoko*. Expression of candidate unigenes related to monoterpene synthesis of *A. tsaoko* (**A**). Alignment of amino acid sequences of Cluster-24,134.3 (putative 1,8-cineole synthase unigene) with other TPS proteins from *Arabidopsis thaliala* (AtTPS-Cin), *Salvia officinalis* (SoTPS-Cin), and *Amomum villosum* (AvTPS1-AvTPS4) using DNAMAN 8.0 (**B**). Phylogenetic tree constructed based on amino acid multiple sequence alignment using MEGA 5.2 with the neighbor-joining method (**C**), dark blue branches represent genus *Amomum* species (*A. tsaoko* and *A. villosum*), and red branches represent other species (*A. thaliala* and *S. officinalis*)

Table 4 Distribution of EST-SSR loci in the transcriptome of *A. tsaoko*

Repeat unit type	Motif	SSR number	Ratio (%)	Dominating motif		
				Motif type	Number	Ratio (%)
Mononucleotide	2	26,742	48.11	A/T	25,819	96.55
Dinucleotide	4	12,716	22.87	AG/CT	7554	59.41
Trinucleotide	10	13,849	24.91	AGG/CCT	3390	24.48
Tetranucleotide	30	1169	2.10	AAAG/CTTT	197	16.85
Pentanucleotide	58	548	0.99	AAAAT/ATTTT	66	12.04
Hexanucleotide	72	566	1.02	AGGCGG/CCGCCT	42	7.42
Total	176	55,590				

Among them, 128,174 unigenes were annotated to 7 major databases, such as GO, KEGG, NR, and KOG, accounting for 87.24% of the total unigenes. These annotated unigenes provide a reference basis for the study of metabolic pathways, gene function classification, plant hormone signal transduction, and quality character analysis of *A. tsaoko*. In addition, there are 18,737 unigenes unannotated. It is speculated that these unigenes may be specific new genes of *A. tsaoko* or that the non-coding RNA sequence and public gene database are imperfect [45–47]. Based on NR alignment results, most of the unigenes were annotated to *Musa acuminata* (80.4% similarity), similar to the report in *Curcuma alismatifolia* (Zingiberaceae family) (80.4% similarity) [42]. Zingiberaceae (*A. tsaoko*) and Musaceae (*Musa acuminata*) belong to Zingiberacea, and both are

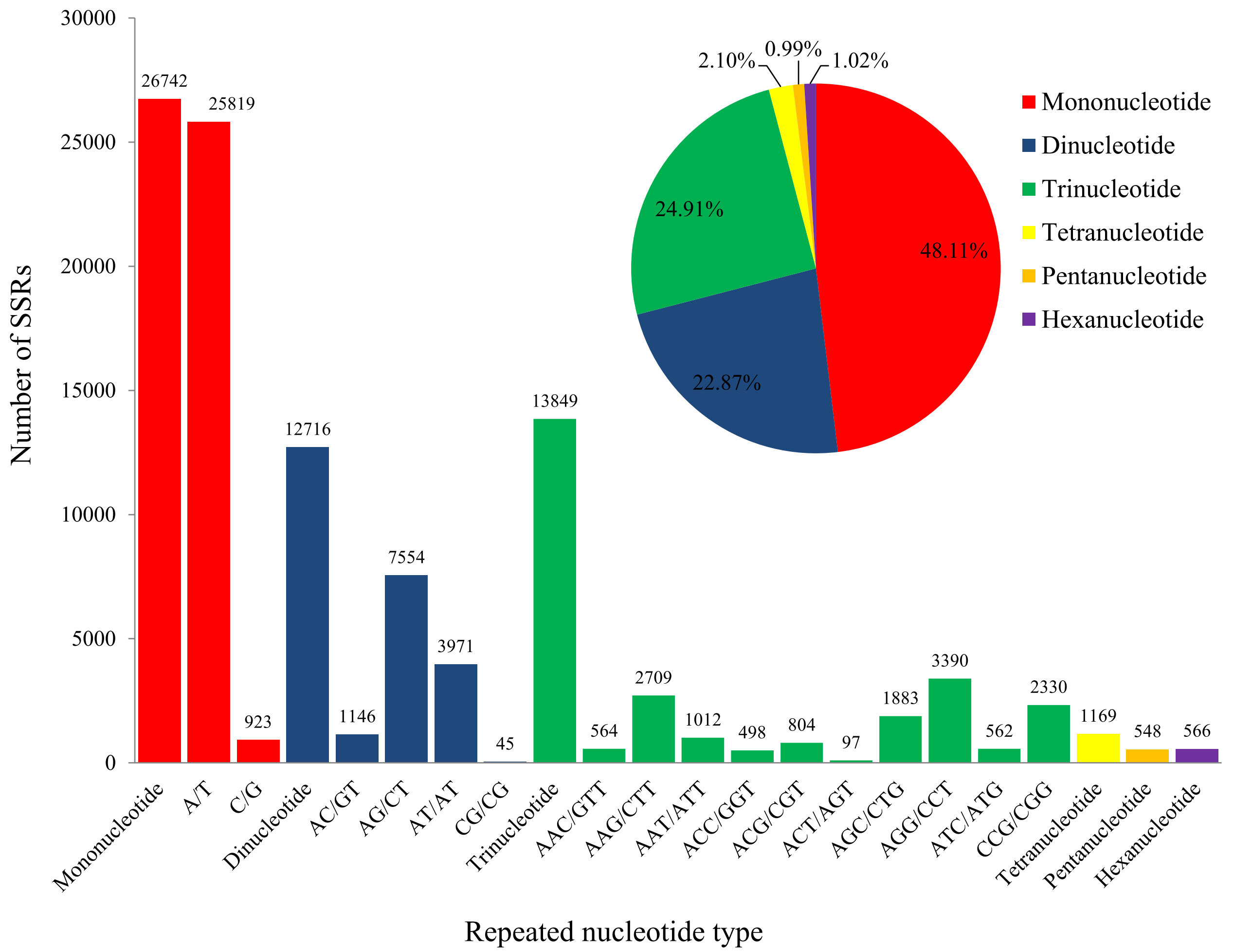


Fig. 6 Frequency and distribution of EST-SSRs from the transcriptome of *A. tsaoko*. The most abundant dinucleotide and trinucleotide motifs were AG/CT and AGG/CCT

Table 5 The length distribution of EST-SSRs based on the number of nucleotide repeat units

Repeats	Repetition times									
	5	6	7	8	9	10	11	12	13	14	≥15
Mononucleotide						11,545	4455	2905	1896	1370	4571
Dinucleotide		3002	1961	1433	1132	982	921	948	521	426	1390
Trinucleotide	6330	3207	1694	1185	442	371	275	116	67	43	119
Tetranucleotide	754	310	52	31	6	4	1	6	3		2
Pentanucleotide	385	101	33	20	5	4					
Hexanucleotide	320	129	52	36	20	4	3		1		1
Total	7789	6749	3792	2705	1605	12,910	5655	3975	2488	1839	6083
Distribution frequency (%)	14.01	12.14	6.82	4.87	2.89	23.22	10.17	7.15	4.48	3.31	10.94

tropical plants, which may explain the large number of homologous genes between them [48]. In addition, 8.3% of the homologous sequences have not been matched, which may because some unigene fragments are too small to match a single data sequence. *A. tsaoko* can be divided into 3 major categories and 55 subcategories in

Table 6 Characterization of 18 EST-SSRs among 72 *A. tsaoko* accessions

Locus	Primer sequences (5′–3′)	Repeat motif	Ta (°C)	Size (bp)	NB	NPB	PPB
AM203	F: TTCAACCCCACGACACACAA R: AGGATGAAGGCTGAGGAGGT	(AAGGCG)5	55	156	5	5	100.00%
AM206	F: ACAGAGAACCACAGGCGAAG R: GACGAGTTGGGAAGGGATGG	(AGA)13	55	164	7	7	100.00%
AM207	F: GAGCCGATCGAGCTAACGAA R: TGGAAAGGTAGGTGCAACGT	(AGAA)5	55	168	5	4	80.00%
AM208	F: CGCACCCCAATTCTCCTTCT R: GCGGATCAGTATCATCCCCG	(AGC)7	55	160	4	3	75.00%
AM213	F: GTTTCCGCCGTGAACTAACG R: TATGGCCTAGTCACGTTGGC	(ATC)5	55	167	3	3	100.00%
AM218	F: GCTTCCTCTCCTATGACGCC R: TGCTTTCGTCGCTTGTCTCT	(CCGCCT)6	55	159	5	5	100.00%
AM223	F: TGTTCTTCTACTGCTGCGCT R: TGGTGTGAAGGAGAGGAGGA	(CCTCT)6	55	166	5	5	100.00%
AM224	F: TTGCTAGAGCTCAAGCCACC R: CAAAGCTCGAGGATAGGCGT	(CGA)9	55	170	6	6	100.00%
AM225	F: TCGTGATCCCTTCGCTTTGT R: GCTCCATCGCCTCCAACATA	(CGAAGG)5	55	166	6	6	100.00%
AM237	F: GTGTGATGGGGGTAAGGGTG R: AGCGCCGTAAAAGGTCCTTT	(GAA)8	55	158	4	4	100.00%
AM242	F: GAGAGCGAGTTGAGCCATGA R: GCCGGTGACAAAGATGGAGA	(GCA)12	55	166	7	7	100.00%
AM247	F: AAGAGAGCATCAAGAGGCCG R: CGACCGAACCTTGTAACCCA	(GGA)10	55	163	6	6	100.00%
AM248	F: CGTCTAGTGCTCCGGAATCC R: GCCATAGCTCCCTCTCCCTT	(GGA)6	55	157	4	4	100.00%
AM255	F: CGGGAACAACGACGGTAAGA R: CGGAATCACAATCGCCATCG	(GGCGGA)7	55	166	12	12	100.00%
AM272	F: GCGATCTCCAGGGCGAAATA R: TTTAGCCCTCCTCCTCCTCC	(TGC)7	55	156	4	4	100.00%
AM273	F: AAAGAAGGAATCTGGCCCGG R: ACGGCGAAAGGTCAGCAATA	(TGCCTC)6	55	164	5	5	100.00%
AM278	F: TGCCTGCGCTTAGTCAATCA R: CCCTCCAATGTTCCCAACGA	(TTC)7	55	157	6	6	100.00%
AM279	F: ATTGCAATTGAAGCAGCGCA R: GGTGGTTTGCATCCATGGTG	(TTCT)6	55	156	4	4	100.00%

Ta annealing temperature, *NB* Number of amplified bands, *NPB* Number of polymorphic bands; *PPB* Percentage of polymorphic bands

GO functional classification, mainly focusing on cellular part, catalytic activity, and metabolic process. The same outcome was found in the transcriptomic study of *A. villosum* from congeneric species [15]. Volatile terpenoids are the active metabolites in the essential oil of *A. tsaoko*, but the terpene synthases (TPS) responsible for their biosynthesis remain unknown. In our study, the KEGG enrichment analysis found that 496 unigenes were enriched in the metabolic pathways of terpenoids, which will provide a better understanding of terpenoid biosynthesis in *A. tsaoko*. The DDXXD conserved motifs contained in the amino acid sequence of Cluster-24,134.3 binds catalytic substrates (such as GPP) through complexation with metal ions (such as Mg^{2+}), which is shared by ionization-dependent terpenoid synthases belonging to the monoterpene synthase functional domain [49, 50]. The phylogenetic tree based on amino acid sequences showed that Cluster-24,134.3 had the highest homology with the AvTPS2 identified in *A. villosum*, which is in the same genus, and was quite

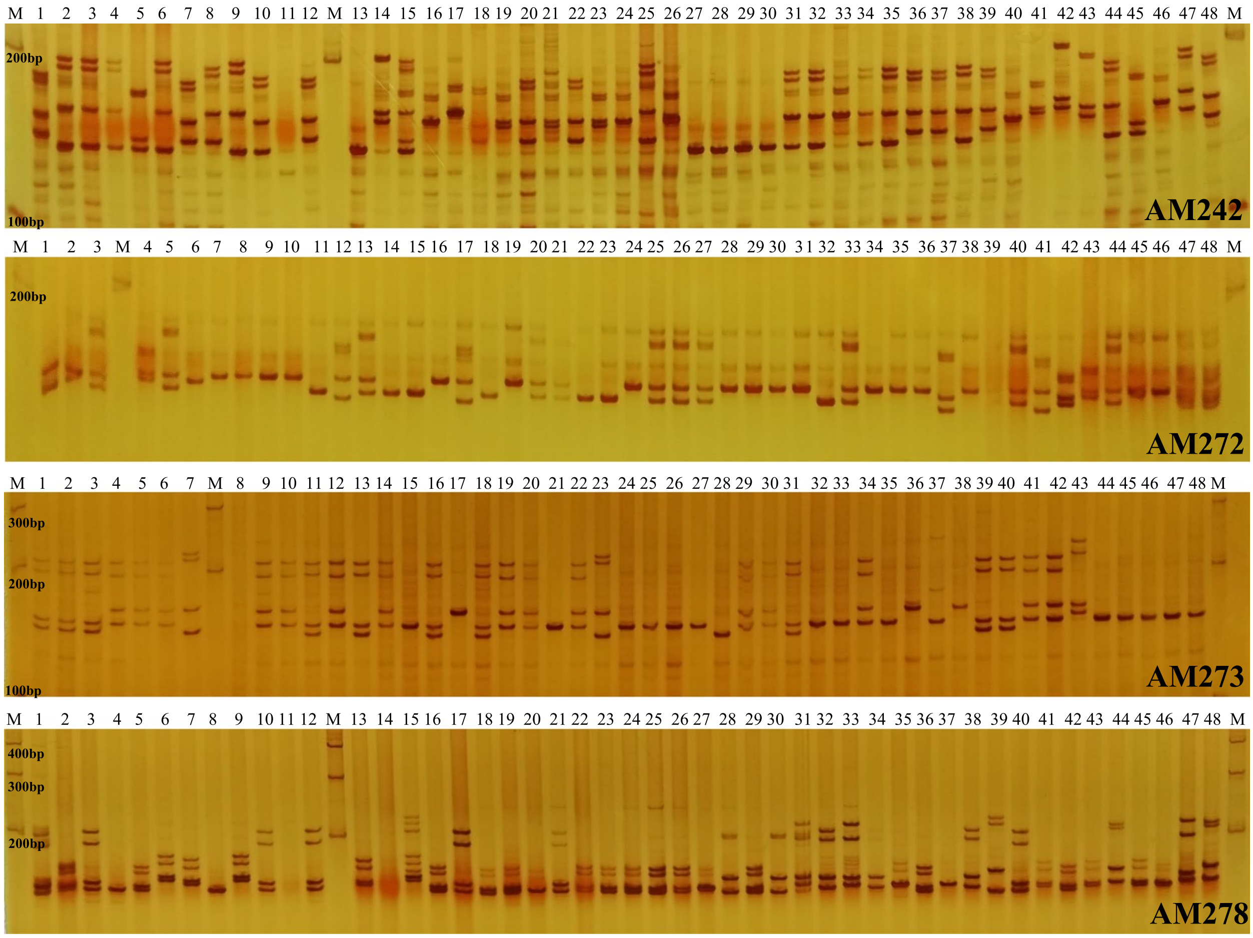

M 1 2 3 4 5 6 7 8 9 10 11 12 M 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 M
200bp
100bp
AM242
M 1 2 3 M 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 M
200bp
AM272
M 1 2 3 4 5 6 7 M 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 M
300bp
200bp
100bp
AM273
M 1 2 3 4 5 6 7 8 9 10 11 12 M 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 M
400bp
300bp
200bp
AM278

Fig. 7 Electrophoretic profiles of 48 *A. tsaoko* accessions using four pairs of EST-SSR primers. Denatured PCR products were separated using an 8% non-denaturing polyacrylamide gel and then stained with 1% silver nitrate

different from the 1,8-cineole synthase gene of *Arabidopsis* and *Salvia officinalis*. This may be because the phylogeny and lineage differentiation of angiosperm TPS are closely related to the natural plant classification system, and the similarity of TPS from similar species is much higher than that of TPS from different sources with the same function [51]. Follow-up cloning and functional verification of the corresponding genes of Cluster-24,134.3 are necessary.

Frequency and distribution of EST-SSRs

Molecular marker technology has been widely used in various fields of plant sciences, including germplasm resource identification, genetic diversity, new variety breeding, and map construction [18]. Among them, SSRs are known as a marker type widely used at present. Furthermore, SSRs can be divided into two categories: genomic and expressed sequence tags (EST-SSRs). The construction of a transcriptome platform has promoted the development of DNA molecular markers to a great extent. To date, there are only a few reports about molecular markers in *A. tsaoko* [1, 17, 35]. Recently, 123 SSRs were identified in the *A. tsaoko* chloroplast genome; the mean density of SSR loci is up to 1/1.33 kb; however, they have not been validated for *A. tsaoko* [48]. In recent years, the use of transcriptome data to obtain sequences containing microsatellites, and the study of their genetic diversity has been successfully reported [52–54]. Excluding mononucleotide repeats, trinucleotides were found to be the most abundant repeats in the present study. This result was consistent with previous studies in other plants, such as *Curcuma alismatifolia* [42], *Vicia amoena* [22], and *Pseudotaxus chienii* [55]. Importantly, the abundance of trinucleotides in coding regions does not change the coding frame and therefore may not affect the functions of the genes [56, 57]. The most abundant dinucleotide SSR motifs were AG/CT (59.41%), a similar finding has been reported in other plant genomes [42, 53, 58, 59]. This may

Table 7 Allelic diversity of 18 EST-SSR markers used in 72 accessions of *A. tsaoko*

Locus	Ne	I	Ho	He	PIC	HWE	FNA
AM203	3.014	1.171	0.653	0.668	0.600	0.962	0.018
AM206	3.977	1.529	0.743	0.749	0.723	0.520	−0.067
AM207	1.371	0.532	0.310	0.271	0.270	0.881	−0.026
AM208	1.311	0.477	0.264	0.237	0.223	0.645	−0.049
AM213	2.127	0.830	0.543	0.530	0.463	0.583	0.038
AM218	2.293	0.963	0.592	0.564	0.490	0.973	−0.054
AM223	3.187	1.298	0.704	0.686	0.643	0.776	0.178
AM224	2.668	1.274	0.606	0.625	0.599	0.024*	−0.034
AM225	3.452	1.423	1.000	0.710	0.684	0.000***	−0.046
AM237	2.741	1.121	0.571	0.635	0.591	0.006**	0.066
AM242	4.852	1.701	0.775	0.794	0.771	0.177	−0.062
AM247	4.393	1.638	0.348	0.772	0.779	0.000***	0.335
AM248	2.814	1.169	0.614	0.645	0.610	0.003**	−0.046
AM255	4.001	1.686	0.689	0.750	0.770	0.957	−0.111
AM272	2.331	1.065	0.507	0.571	0.538	0.286	0.051
AM273	2.726	1.114	0.681	0.633	0.563	0.692	0.032
AM278	4.263	1.576	0.704	0.765	0.738	0.001**	0.034
AM279	1.771	0.733	0.394	0.435	0.386	0.761	−0.058
Mean	2.961	1.183	0.594	0.613	0.580	0.458	0.011

Ne effective number of alleles, *I* Shannon's information index, *Ho* observed heterozygosity, *He* expected heterozygosity, *PIC* Polymorphic information content, *HWE* Hardy–Weinberg equilibrium, *FNA* Frequencies of null alleles. Significant deviation from HWE at * $P<0.05$, ** $P<0.01$, and *** $P<0.001$

Table 8 Genetic diversity of *A. tsaoko* populations based on EST-SSR markers

Population	PPB/%	Na	Ne	I	Ho	He	PAr	F
JP	100.000	4.000	2.662	1.078	0.573	0.582	0.167	− 0.009
PB	88.890	3.167	2.481	0.927	0.596	0.535	0.056	−0.110
LVC	100.000	4.111	2.714	1.073	0.584	0.578	0.167	−0.009
YY	100.000	4.000	2.747	1.086	0.576	0.591	0.056	0.015
LC	100.000	3.722	2.450	0.981	0.620	0.540	0.111	−0.140
YX	100.000	3.722	2.415	0.967	0.617	0.537	0.111	−0.145
Mean	98.150	3.787	2.578	1.019	0.594	0.560	0.111	−0.066

PPB Percentage of polymorphic bands, *Na* observed number of alleles, *Ne* effective number of alleles, *I* Shannon's information index, *Ho* observed heterozygosity, *He* expected heterozygosity, *PAr* number of private alleles, *F* Fixation index

be because alanine (Ala) and leucine (Leu) are the amino acids with the highest frequency in proteins, while the AG/CT dinucleotide motif is present in the codons of Ala and Leu [60]. Moreover, some studies have shown that the CT motif frequently occurs in the 5′ UTRs; this motif plays an important role in regulating nucleic acid metabolism and gene expression in plants [61, 62]. Among trinucleotide repeat motifs, AGG/CCT (24.48%) occurred most commonly, agreeing with the results reported in *Zantedeschia rehmannii* [63], *Phragmites karka* [64], *Amorphophallus* [65], and *Aspidistra saxicola* [66]. It has been suggested that AGG/CCT interruptions play a protective role in genomic packaging [67, 68].

Validation of EST-SSR markers

In our study, 64 primer pairs (80%) were successfully amplified among the 80 randomly selected EST-SSRs (designated AM201-AM280) across 4 *A. tsaoko* accessions. Some studies have pointed out that, compared with genomic SSRs, EST-SSRs have a better amplification effect, but the polymorphism rate is relatively low [69, 70]. However, these newly developed EST-SSRs (Na = 5.444; Ho = 0.594; He = 0.613) have higher resolutions than microsatellite markers (Na = 2.000; Ho = 0.250–0.372; He = 0.432–0.456) [35] and genomic SSR markers (Na = 3.913; Ho = 0.468; He = 0.500) [1] in *A. tsaoko*. This phenomenon has also been reported in poplar [71].

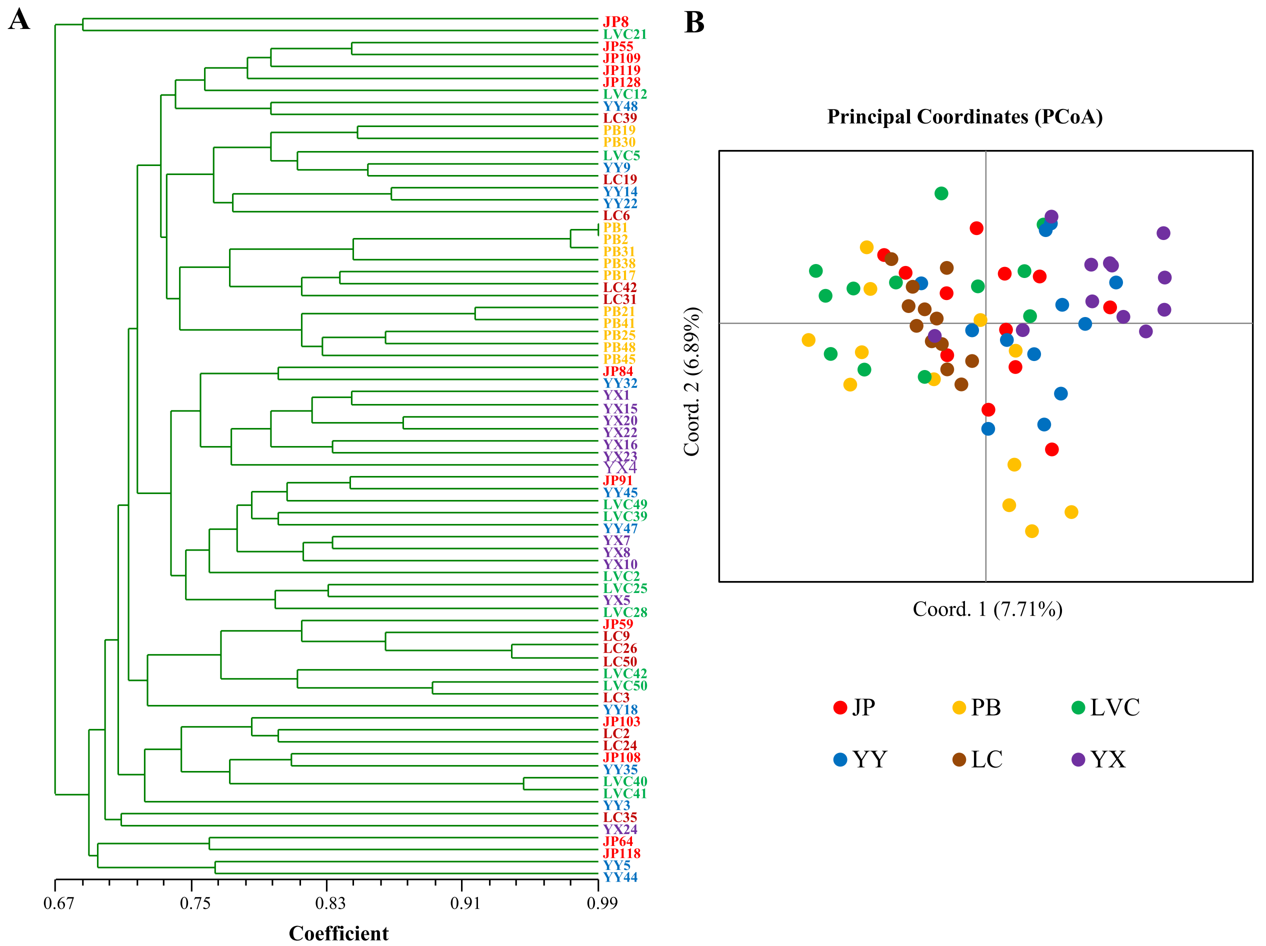


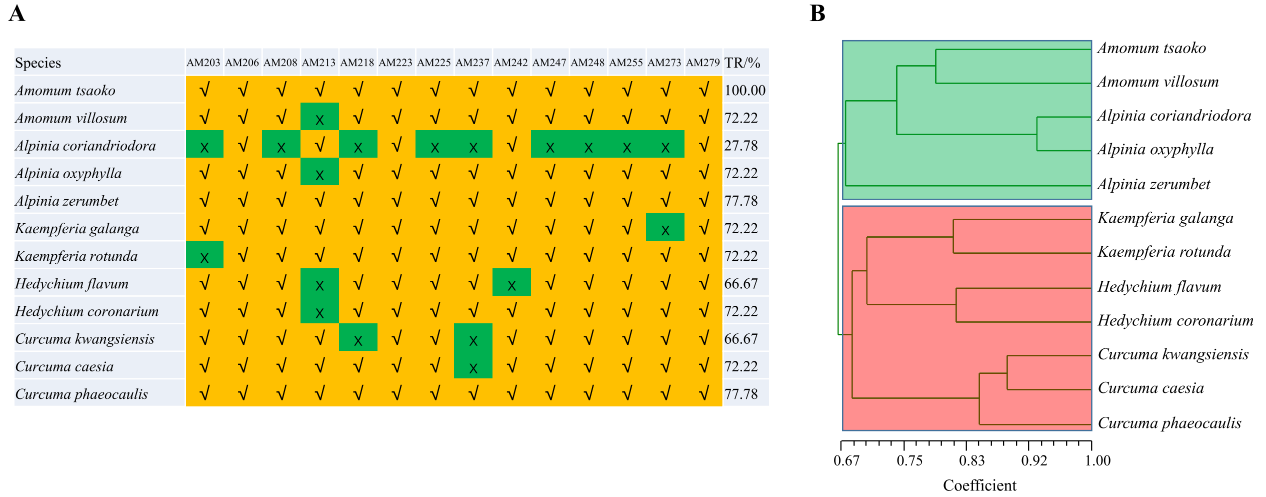
Fig. 8 Cluster analysis and principal coordinate analysis (PCoA) based on the genotyping of 72 accessions of *A. tsaoko*

Table 9 Pairwise comparison of F_{ST} (above diagonal) and genetic distance (below diagonal) for six *A. tsaoko* populations

Population	JP	PB	LVC	YY	LC	YX
JP	–	0.058	0.029	0.030	0.051	0.057
PB	0.166	–	0.050	0.058	0.059	0.091
LVC	0.090	0.144	–	0.040	0.041	0.069
YY	0.103	0.175	0.132	–	0.041	0.044
LC	0.147	0.156	0.115	0.127	–	0.068
YX	0.170	0.263	0.204	0.130	0.207	–

Table 10 Analysis of molecular variance (AMOVA) in *A. tsaoko* from six populations

Source of variation	df	Sum of squares	Variance components	Total variation (%)	*P* value
Among populations	5	65.535	0.306	5%	
Among individuals within populations	66	380.708	0.283	5%	
Within individuals	72	374.500	5.201	90%	< 0.001
Total	143	820.743	5.791		



Species	AM203	AM206	AM208	AM213	AM218	AM223	AM225	AM237	AM242	AM247	AM248	AM255	AM273	AM279	TR/%
Amomum tsaoko	√	√	√	√	√	√	√	√	√	√	√	√	√	√	100.00
Amomum villosum	√	√	√	x	√	√	√	√	√	√	√	√	√	√	72.22
Alpinia coriandriodora	x	√	x	√	x	√	x	x	√	x	x	x	x	√	27.78
Alpinia oxyphylla	√	√	√	x	√	√	√	√	√	√	√	√	√	√	72.22
Alpinia zerumbet	√	√	√	√	√	√	√	√	√	√	√	√	√	√	77.78
Kaempferia galanga	√	√	√	√	√	√	√	√	√	√	√	√	x	√	72.22
Kaempferia rotunda	x	√	√	√	√	√	√	√	√	√	√	√	√	√	72.22
Hedychium flavum	√	√	√	x	√	√	√	√	x	√	√	√	√	√	66.67
Hedychium coronarium	√	√	√	x	√	√	√	√	√	√	√	√	√	√	72.22
Curcuma kwangsiensis	√	√	√	√	x	√	√	x	√	√	√	√	√	√	66.67
Curcuma caesia	√	√	√	√	√	√	√	x	√	√	√	√	√	√	72.22
Curcuma phaeocaulis	√	√	√	√	√	√	√	√	√	√	√	√	√	√	77.78

Fig. 9 Cross-species transferability of 14 polymorphic EST-SSR markers in different Zingiberaceae species. The transferability rates of EST-SSR markers in 11 Zingiberaceae species (**A**). UPGMA phylogenetic analysis of 12 Zingiberaceae species based on transferable SSR markers (**B**)

PIC is one of the main indices used to evaluate locus polymorphisms, including high (PIC >0.5), moderate (0.5 > PIC >0.25), and low (PIC <0.25) polymorphism [72]. Among the 72 accessions, the PIC values indicated a good informative level of these EST-SSRs, including 13 highly polymorphic (AM203, AM206, AM223, AM224, AM225, AM237, AM242, AM247, AM248, AM255, AM272, AM273, and AM278), four moderately polymorphic (AM207, AM213, AM218, and AM279), and only one low polymorphic (AM208) markers. EST-SSR markers exhibit much higher polymorphism than RAPD, ISSR, SRAP, and genomic SSR markers for *A. tsaoko* [1, 17, 34]. HWE analysis showed that 12 primer pairs (AM203, AM206, AM207, AM208, AM213, AM218, AM223, AM242, AM255, AM272, AM273, and AM279) were in accordance with HWE, while the remaining primer sets (AM224, AM225, AM237, AM247, AM248, and AM278) significantly deviated from HWE ($P < 0.05$). Taken together, the newly developed EST-SSRs are suitable for exploring the genetic diversity and relationships of *A. tsaoko*. These molecular markers will be effective tools in fingerprinting germplasm collections from different sources to guide germplasm evaluation and conservation in *A. tsaoko*. We also analyzed the characteristics of SSR loci within the terpenoid metabolism-related unigenes (Table S1). Next, we will develop molecular markers around these SSR loci, which is very important for gene mining and marker-assisted selection breeding for controlling important traits in *A. tsaoko*.

Genetic diversity of the *A. tsaoko* population

Genetic diversity is an important indicator of the effective management and utilization of plant germplasm resources. The level of genetic diversity is often correlated with its range of distribution; the wide distribution range of plants usually shows higher levels of genetic diversity compared with narrowly distributed plants [73, 74]. However, while the known distribution of *A. tsaoko* is narrow, it retains relatively high levels of genetic diversity at the population level (PPB = 98.150%, I = 1.019, Ho = 0.594, and He = 0.560). Similar results have been reported for many other plant species, such as *Petunia secreta* [75], *Nouelia insignis* [76], and *Calanthe tsoongiana* [77]. In our study, Jingping and Lvchun counties, as possible places of origin of *A. tsaoko*, harbored the highest number of private alleles (PAr = 0.167), which indicates the presence of specific sequences or genes in JP and LVC populations [78]; this was also consistent with our previous findings [17]. The overall mean value of the fixation index was −0.066, showing an excess of heterozygosity present in the *A. tsaoko* population; this result is in agreement with a previous study carried out with genomic SSR markers [1]. The pairwise F_{ST} analysis (average pairwise $F_{ST} = 0.052$) detected no significant genetic differentiation among six *A. tsaoko* populations. Correspondingly, the AMOVA results also showed that much more genetic variance exists among individuals within populations. High genetic diversity, excess heterozygotes, and low genetic differentiation are characteristics of outbreeding populations [79]. *A. tsaoko* is a perennial cross-pollinating plant pollinated by insects (native bumblebee), allowing allele exchange among individuals in the population [17]. In this study, the accessions did not cluster according to different sampling sites by UPGMA cluster, and the PCoA analysis further confirmed low genetic differentiation among populations.

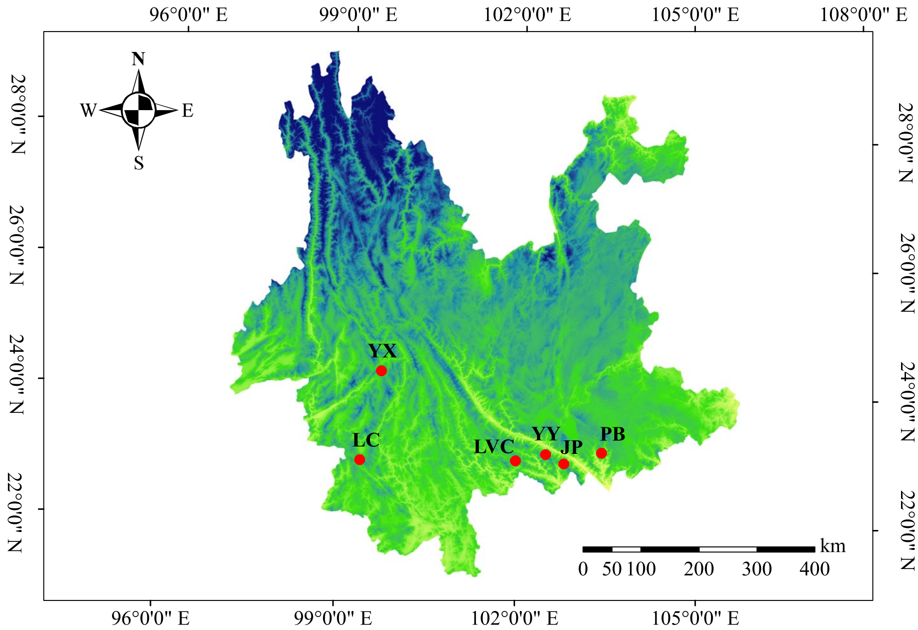


Fig. 10 Geographic locations of the 6 populations of *A. tsaoko* included in this study

Compared with bulk crops, the molecular markers available for Zingiberaceae are very limited. In our study, the ESR-SSR markers developed in *A. tsaoko* have high cross-species transferability (with an average transferability of 68.18%) with other Zingiberaceae species, and these markers can accurately classify different species, making them useful for increasing the number of available molecular markers of Zingiberaceae species.

Conclusions

In this study, we reported the first transcriptome sequencing analysis of *A. tsaoko* using Illumina sequencing technology. The de novo assembly generated a total of 146,911 unigenes with an average length of 1527 bp and an N50 of 2002 bp. In total, 128,174 unigenes were successfully annotated to the NR (84.01%), NT (65.20%), KEGG (36.11%), Swiss-prot (65.84%), PFAM (60.97%), GO (61.09%), and KOG (24.92%) databases. 496 unigenes involved in the terpenoid biosynthesis pathway were identified. Based on the transcriptome assembly, 55,590 potential EST-SSRs were identified and characterized. Seventy-two *A. tsaoko* accessions using 18 novel polymorphic EST-SSR primers showed rich genetic diversity and low genetic differentiation among populations. The transcriptome data from this study will provide valuable resources for investigating gene functions, and the EST-SSR markers developed will provide a foundation for germplasm identification, genetic diversity assessment, and marker-assisted breeding in *A. tsaoko*.

Methods

Plant material

The transcriptome sequencing materials were collected from the young roots, stems, leaves, flowers, and fruits of 5-year-old *A. tsaoko* plants in Caoguoshan Village, Adebo Township, Jinping County, Honghe Prefecture, Yunnan Province, China (22°54′30.34″N, 103°13′16.39″E). To verify EST-SSR markers, young leaves from a total *of* 72 *A. tsaoko* accessions were collected from different sites covering the plant's major distribution areas in China (Fig. 10; Table S2). For cross-species transferability analysis, 12 Zingiberaceae species were used, which consisted of two species in the genus *Amomum* (*A. tsaoko* and *A. villosum*), two species in *Kaempferia* (*K. galangal* and *K. rotunda*), two species in *Hedychium* (*H. flavum* and *H. coronarium*), and three species in *Curcuma* (*C. kwangsiensis*, *C. caesia*, and *C. phaeocaulis*) (Table S3). No approval or permission was required to collect these samples. All the species were identified by Professor Bingyue Lu (Honghe University). The genomic DNA was extracted from leaf tissues using a modified cetyltrimethylammonium bromide (CTAB) method [80].

RNA isolation, library preparation, and transcriptome sequencing

Total RNA was extracted from three biological replicates (with two technical replicates per each biological replicate) from mixed samples of the five different organs using the RNeasy extraction kit (QIAGEN, Beijing, China). The integrity and purity of extracted RNA were measured by a NanoDrop spectrophotometer and agarose gel electrophoresis. After total RNA extraction, eukaryotic mRNA was enriched with Oligo (dT) beads. Next, fragmentation buffer was added to break mRNA into short fragments. Then, these mRNA fragments were used to synthesize first-strand cDNA with hexameric random primers. Subsequently, second-strand cDNA was synthesized by adding DNA polymerase I, RNase H, dNTPs, and buffer. The purified double-stranded cDNA was repaired at the end, dA-tailed fragments were added, and it was connected to the sequencing connector. After library inspection, the quality-qualified libraries were sequenced using the Illumina HiSeq platform (Novogene Biotech Co., Ltd., Beijing, China).

De novo assembly and annotation

After sequencing, raw data were filtered, and high-quality clean data were obtained by removing the joint sequence and low-quality reads. High-quality reads were used for de novo assembly using Trinity software [81]. Finally, we obtained sequences that could not be extended on either end, which were considered unigenes. The resulting clean data were deposited in the Sequence Read Archive of National Center for Biotechnology Information (NCBI) (Bioproject no. PRJNA735890; Biosample no. SAMN19601270). The assembled *A. tsaoko* transcripts were compared with the NCBI non-redundant protein sequences (NR), NCBI non-redundant nucleotide sequences (NT), protein family (PFAM), eukaryotic clusters of orthologous groups (KOG), a manually annotated and reviewed protein sequence (SwissProt), and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases using the BLASTX algorithm (significant thresholds of *E*-value $<10^{-5}$), and Gene Ontology (GO) annotations were obtained using Blast2GO software (http://www.geneontology.org) based on NR annotations.

Development and characterization of EST-SSRs

Based on the transcriptome sequencing data of *A. tsaoko*, MISA software (https://webblast.ipk-gatersleben.de/misa/) was used to search the SSR loci of *A. tsaoko*. The search criteria were as follows: mono- to hexa-nucleotide repeats ≥10, 6, 5, 5, 5, and 5; mono- to hexa-nucleotide repeat types are labeled P1, P2, P3, P4, P5, and P6. Primer3 (http://primer3.sourceforge.net/releases.php) was used to design SSR primers based on the detected sites.

SSR analysis

Eighty pairs of EST-SSR primers were randomly synthesized, and four *A. tsaoko* germplasm resources with different fruit shapes (round type, oval type, long type, and shuttle type) were selected to verify the effectiveness and polymorphisms of the primers. The primers producing clear bands and obvious polymorphisms were subsequently used for genetic diversity assessments. The PCR reaction system, amplification program, and gel electrophoresis were as previously described [1].

Data analysis

According to the polyacrylamide electrophoresis results, at the same migration location, a band is marked as "1," no band is marked as "0," and a missing band is indicated by "-". Subsequently, 0/1 data were preprocessed using DataFormater software [82], which transformed SSR data into readable input files for NTSYS-pc and PowerMarker. The GenAlEx 6.501 program [83] was employed to calculate genetic diversity parameters, namely the observed number of alleles (Na), the effective number of alleles (Ne), Shannon's information index (I), expected heterozygosity (He), observed heterozygosity (Ho), number of private alleles (PAr), and fixation index (F), which was used to assess Hardy–Weinberg genetic equilibrium. PowerMarker v3.25 [84] was used to estimate the polymorphism information content (PIC) of each EST-SSR primer pair. We estimated the frequencies of null alleles (FNA) using Micro-checker 2.2.3 [85]. The genetic differentiation between populations was carried out using analysis of molecular variance (AMOVA), pairwise $F_{ST,}$ and pairwise Nei's genetic distance in GenAlEx software. Cluster analysis (unweighted pair group method with arithmetic mean algorithm) and principal coordinate analysis (PCoA) were performed using NTSYS-pc software [86] and GenAlEx v6.5, respectively. Multiple amino acid sequence alignment and phylogenetic analysis was performed using DNAMAN 8.0 and MEGA version 5.2 software. The amino acid sequences of the 1,8-cineole synthase gene of *Arabidopsis thaliala* (GenBank accession no. NM_113483.5) and *Salvia officinalis* (GenBank accession no. AF051899.1) were downloaded from the NCBI (http://www.ncbi.nlm.nih.gov/). The amino acid sequences of four monoterpene synthase genes (*AvTPS1*, *AvTPS2*, *AvTPS3*, and *AvTPS4*) of *A. villosum* were obtained from the published literature [87].

Abbreviations

EST-SSR: Expressed sequence tag-simple sequence repeat; AMOVA: Analysis of molecular variance; TCM: Traditional Chinese medicine; COVID: Coronavirus disease; IUCN: International union for conservation of nature; RFLP: Restriction fragment length polymorphism; RAPD: Random amplified polymorphic DNA; ISSR: Inter-simple sequence repeat; SSR: Simple sequence repeat; SNP: Single nucleotide polymorphism; QTL: Quantitative trait locus; CTAB: Cetyltrimethylammonium bromide; NR: NCBI non-redundant protein sequences; NT: NCBI non-redundant nucleotide sequences; PFAM: Protein family; KOG: Eukaryotic clusters of orthologous groups; SwissProt: A manually annotated and reviewed protein sequence; KEGG: Kyoto encyclopedia of genes and genomes; GO: Gene ontology; Na: Observed number of alleles; Ne: Effective number of alleles; I: Shannon's information index; He: Expected heterozygosity; Ho: Observed heterozygosity; PAr: Number of private alleles; F: Fixation index; FNA: Frequencies of null alleles; UPGMA: Unweighted pair-group method with arithmetic means; PCoA: Principal coordinate analysis; HWE: Hardy–Weinberg genetic equilibrium.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-022-03827-y.

Additional file 1: Table S1. Analysis of SSR loci of terpenoid metabolic pathway-related unigenes in *A. tsaoko*. **Table S2.** Populations of *A. tsaoko* from different locations in the study. **Table S3.** Sampling location information of 12 Zingiberaceae species.

Additional file 2: Fig. S1. The original image of electrophoretic using AM242, AM272, AM273, and AM278 primer pairs.

Acknowledgements

We thank Accdon (www.accdon.com) for its linguistic assistance during the preparation of this manuscript.

Authors' contributions

BL conceived and designed the experiments; BL, EL, and TW collected *A. tsaoko* samples; MM, HM, and WZ performed the experiments; BL, and MM analyzed the data and wrote the manuscript. All authors reviewed and approved the manuscript.

Funding

This work was funded the Special Basic Cooperative Research Programs of Yunnan Provincial Undergraduate Universities' Association (No. 202001BA070001–181), the Youth Top-notch Talent Support Program of Yunnan Province (No. YNWR–QNBJ–2020–207) and the National Natural Science Foundation of China (No. 82260735, 31460380).

Availability of data and materials

The original sequencing data generated in the study have been deposited into the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA735890). The EST sequences in this study have been submitted to GenBank with accession numbers OM468473–OM468552 (https://www.ncbi.nlm.nih.gov/genbank/). Other datasets supporting the conclusions of this article are included within the article.

Declarations

Ethics approval and consent to participate

Amomum tsaoko is neither wild nor endangered in China, and no specific permission was required for the collection. The study complied with relevant institutional, national, and international guidelines and legislation.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflicts of interest.

Author details

[1]Key Laboratory for Research and Utilization of Characteristic Biological Resources in Southern Yunnan, Honghe University, Mengzi 661199, China. [2]College of Biological and Agricultural Sciences, Honghe University, Mengzi 661199, China.

Received: 7 February 2022 Accepted: 30 August 2022
Published online: 29 September 2022

References

1. Lu BY, Ma ML, Zhang W, Meng HL, Lei E, Wang TT, et al. Development of 23 novel microsatellite markers of *Amomum tsao-ko* (Zingiberaceae) based on restriction-site-associated DNA sequencing. Mol Biol Rep. 2021;48(2):1943–9.
2. Zhang TT, Lu CL, Jiang JG. Antioxidant and anti-tumour evaluation of compounds identified from fruit of *Amomum tsaoko* Crevost et Lemaire. J Funct Foods. 2015;18(423–4):31.
3. Zhang XF, Tang YJ, Guan XX, Lu X, Li J, Chen XL, et al. Flavonoid constituents of *Amomum tsao-ko* Crevost et Lemarie and their antioxidant and antidiabetic effects in diabetic rats - in vitro and in vivo studies. Food Funct. 2022;13(1):437–50.
4. Guo N, Zang YP, Cui Q, Gai QY, Jiao J, Wang W, et al. The preservative potential of *Amomum tsaoko* essential oil against E. coil, its antibacterial property and mode of action. Food Control. 2017;75:236–45.
5. Hong SS, Lee JE, Jung YW, Park JH, Lee JA, Jeong W, et al. Monoterpenoids from the fruits of *Amomum tsao-ko* have inhibitory effects on nitric oxide production. Plants (Basel). 2021;10(2):257.
6. He XF, Chen JJ, Huang XY, Hu J, Zhang XK, Guo YQ, et al. The antidiabetic potency of *Amomum tsao-ko* and its active flavanols, as PTP1B selective and alpha-glucosidase dual inhibitors. Ind Crop Prod. 2021;160:112908.
7. Chen C, You F, Wu FH, Luo YS, Zheng GH, Xu HL, et al. Antiangiogenesis efficacy of ethanol extract from *Amomum tsaoko* in ovarian cancer through inducing ER stress to suppress p-STAT3/NF-kB/IL-6 and VEGF loop. Evid-based Compl Alt. 2020;2020:2390125.
8. Wang KN, Liu YN, Shi Y, Yan MZ, Rengarajan T, Feng X. *Amomum tsaoko* fruit extract exerts anticonvulsant effects through suppression of oxidative stress and neuroinflammation in a pentylenetetrazol kindling model of epilepsy in mice. Saudi J Biol Sci. 2021;28(8):4247–54.
9. Zong Y, Ding ML, Jia KK, Ma ST, Ju WZ. Exploring active compounds of Da-Yuan-Yin in treatment of COVID-19 based on network pharmacology and molecular docking method. Chin Tradit Herb Drugs. 2020;51:836–44.
10. Lee DYW, Li QY, Liu J, Efferth T. Traditional Chinese herbal medicine at the forefront battle against COVID-19: clinical experience and scientific basis. Phytomedicine. 2021;80:153337.
11. Li XR, Wen ZS, Si MD, Jia YX, Liu HX, Zheng YG, et al. Exploration of Hanshi Zufei prescription for treatment of COVID-19 based on network pharmacology. Chin Herb Med. 2022;14(2):294–302.
12. Zhu YW, Yan XF, Ye TJ, Hu J, Wang XL, Qiu FJ, et al. Analyzing the potential therapeutic mechanism of Huashi Baidu decoction on severe COVID-19 through integrating network pharmacological methods. J Tradit Complement Med. 2021;11(2):180–7.
13. Liu JZ, Lyu HC, Fu YJ, Cui Q. *Amomum tsao-ko* essential oil, a novel anti-COVID-19 omicron spike protein natural products: a computational study. Arab J Chem. 2022;15(7):103916.
14. Yang SY, Xue YF, Chen DJ, Wang ZT. *Amomum tsao-ko* Crevost & Lemarié: a comprehensive review on traditional uses, botany, phytochemistry, and pharmacology. Phytochem Rev. 2022;10:1–35.
15. He XY, Wang H, Yang JF, Deng K, Wang T. RNA sequencing on *Amomum villosum* Lour. Induced by MeJA identifies the genes of WRKY and terpene synthases involved in terpene biosynthesis. Genome. 2018;61(2):91–102.
16. de Boer H, Newman M, Poulsen AD, Droop AJ, Fer T, Hien LTT, et al. Convergent morphology in *Alpinieae* (Zingiberaceae): Recircumscribing *Amomum* as a monophyletic genus. Taxon. 2018;67(1):6–36.
17. Ma ML, Wang TT, Lu BY. Assessment of genetic diversity in *Amomum tsao-ko* Crevost & Lemarie, an important medicine food homologous crop from Southwest China using SRAP and ISSR markers. Genet Resour Crop Ev. 2021;68(6):2655–67.

18. Agarwal M, Shrivastava N, Padh H. Advances in molecular marker techniques and their applications in plant sciences. Plant Cell Rep. 2008;27(4):617–31.
19. Ismail NA, Rafii MY, Mahmud TMM, Hanafi MM, Miah G. Molecular markers: a potential resource for ginger genetic diversity studies. Mol Biol Rep. 2016;43(12):1347–58.
20. Powell W, Machray GC, Provan J. Polymorphism revealed by simple sequence repeats. Trends Plant Sci. 1996;1(7):215–22.
21. Lu J, Zhang YC, Diao X, Yu K, Dai XW, Qu P, et al. Evaluation of genetic diversity and population structure of *Fragaria nilgerrensis* using EST-SSR markers. Gene. 2021;796:145791.
22. Wu FF, Zhang SX, Gao Q, Liu F, Wang JL, Wang XG. Genetic diversity and population structure analysis in a large collection of *Vicia amoena* in China with newly developed SSR markers. BMC Plant Biol. 2021;21(1):544.
23. Wu HB, He XL, Gong H, Luo SB, Li MZ, Chen JQ, et al. Genetic linkage map construction and QTL analysis of two interspecific reproductive isolation traits in sponge gourd. Front Plant Sci. 2016;7:980.
24. Lu BY, Yang CY, Xie K, Zhang L, Wu T, Li LF, et al. Quantitative trait loci for grain-quality traits across a rice F2 population and backcross inbred lines. Euphytica. 2013;192(1):25–35.
25. Bharadwaj C, Tripathi S, Soren KR, Thudi M, Singh RK, Sheoran S, et al. Introgression of "QTL-hotspot" region enhances drought tolerance and grain yield in three elite chickpea cultivars. Plant Genome. 2021;14(1):e20076.
26. Karunarathna KHT, Mewan KM, Weerasena O, Perera S, Edirisinghe ENU. A functional molecular marker for detecting blister blight disease resistance in tea (*Camellia sinensis* L.). Plant Cell Rep. 2021;40(2):351–9.
27. Cui Q, Wang LT, Liu JZ, Wang HM, Guo N, Gu CB, et al. Rapid extraction of *Amomum tsao-ko* essential oil and determination of its chemical composition, antioxidant and antimicrobial activities. J Chromatogr B. 2017;1061:364–71.
28. Liu LJ, Zhao YM, Ming J, Chen J, Zhao GH, Chen ZY, et al. Polyphenol extract and essential oil of *Amomum tsao-ko* equally alleviate hypercholesterolemia and modulate gut microbiota. Food Funct. 2021;12(23):12008–21.
29. Kim JG, Jang H, Le TPL, Hong HR, Lee MK, Hong JT, et al. Pyranoflavanones and pyranochalcones from the fruits of *Amomum tsao-ko*. J Nat Prod. 2019;82(7):1886–92.
30. Wang Y, You CX, Wang CF, Yang K, Chen R, Zhang WJ, et al. Chemical constituents and insecticidal activities of the essential oil from *Amomum tsaoko* against two stored-product insects. J Oleo Sci. 2014;63(10):1019–26.
31. He XF, Wang HM, Geng CA, Hu J, Zhang XM, Guo YQ, et al. Amomutsaokols A-K, diarylheptanoids from *Amomum tsao-ko* and their alpha-glucosidase inhibitory activity. Phytochemistry. 2020;177:112418.
32. Qin HW, Wang YZ, Yang WZ, Yang SB, Zhang JY. Comparison of metabolites and variety authentication of *Amomum tsao-ko* and *Amomum paratsao-ko* using GC-MS and NIR spectroscopy. Sci Rep. 2021;11(1):15200.
33. Zhang W, Lu BY, Meng HL, Wei X, Yang ZQ, Yang SC. Phenotypic diversity analysis of the fruit of *Amomum tsao-ko* Crevost et Lemarie, an important medicinal plant in Yunnan, China. Genet Resour Crop Ev. 2019;66(5):1145–54.
34. Lu BY, Ma ML, Wang TT, Meng HL, Lei E, Zhang W. Genetic diversity and genetic relationships of *Amomum tsao-ko* based on random amplified polymorphic DNA markers. Int J Agric Biol. 2018;20(9):2032–8.
35. Yang YW, Yang ZY, Yan MR, Qian ZG, Guan KY. Isolation and characterization of microsatellite markers for *Amomum tsaoko* (Zingiberaceae), an economically important plant in China. Genet Mol Res. 2014;13(4):8220–4.
36. Guo JD, Huang Z, Sun JL, Cui XM, Liu Y. Research progress and future development trends in medicinal plant transcriptomics. Front Plant Sci. 2021;12:691838.
37. Madritsch S, Burg A, Sehr EM. Comparing de novo transcriptome assembly tools in di- and autotetraploid non-model plant species. BMC Bioinformatics. 2021;22(1):146.
38. Unamba CIN, Nag A, Sharma RK. Next generation sequencing technologies: the doorway to the unexplored genomics of non-model plants. Front Plant Sci. 2015;6:1074.
39. Zalapa JE, Cuevas H, Zhu HY, Steffan S, Senalik D, Zeldin E, et al. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. Am J Bot. 2012;99(2):193–208.
40. Chen R, Wei QH, Liu YH, Wei X, Chen XB, Yin XP, et al. Transcriptome sequencing and functional characterization of new sesquiterpene synthases from *Curcuma wenyujin*. Arch Biochem Biophys. 2021;709:108986.
41. Annadurai RS, Neethiraj R, Jayakumar V, Damodaran AC, Rao SN, Katta M, et al. De novo transcriptome assembly (NGS) of *Curcuma longa* L. rhizome reveals novel transcripts related to anticancer and antimalarial terpenoids. PLoS One. 2013;8(2):e56217.
42. Taheri S, Abdullah TL, Rafii MY, Harikrishna JA, Werbrouck SPO, Teo CH, et al. De novo assembly of transcriptomes, mining, and development of novel EST-SSR markers in *Curcuma alismatifolia* (Zingiberaceae family) through Illumina sequencing. Sci Rep. 2019;9:3047.
43. Jiang YS, Huang MJ, Wisniewski M, Li HL, Zhang MX, Tao X, et al. Transcriptome analysis provides insights into gingerol biosynthesis in ginger (*Zingiber officinale*). Plant Genome. 2018;11(3):1–11.
44. Nadiya F, Anjali N, Thomas J, Gangaprasad A, Sabu KK. Transcriptome profiling of *Elettaria cardamomum* (L.) Maton (small cardamom). Genom Data. 2017;11:102–3.
45. Liu MY, Qiao GR, Jiang J, Yang HQ, Xie LH, Xie JZ, et al. Transcriptome sequencing and de novo analysis for Ma bamboo (*Dendrocalamus latiflorus* Munro) using the Illumina platform. PLoS One. 2012;7(10):e46766.
46. Shi X, Sun HJ, Chen YT, Pan HW, Wang SF. Transcriptome sequencing and expression analysis of cadmium (cd) transport and detoxification related genes in cd-accumulating *Salix integra*. Front Plant Sci. 2016;7:1577.
47. Shen SS, Chai XT, Zhou Q, Luo D, Wang YR, Liu ZP. Development of polymorphic EST-SSR markers and characterization of the autotetraploid genome of sainfoin (*Onobrychis viciifolia*). Peer J. 2019;7:e6542.
48. Ma ML, Lu BY. The complete chloroplast genome of *Amomum tsao-ko*. Mitochondrial DNA B. 2020;5(1):848–9.
49. Oldfield E, Lin FY. Terpene biosynthesis: modularity rules. Angew Chem Int Ed Engl. 2012;51(5):1124–37.
50. Christianson DW. Structural and chemical biology of Terpenoid Cyclases. Chem Rev. 2017;117(17):11570–648.
51. Bohlmann J, Meyer-Gauen G, Croteau R. Plant terpenoid synthases: molecular biology and phylogenetic analysis. Proc Natl Acad Sci U S A. 1998;95(8):4126–33.
52. Ruan XX, Wang Z, Wang T, Su YJ. Characterization and application of EST-SSR markers developed from the transcriptome of *Amentotaxus argotaenia* (Taxaceae), a relict vulnerable conifer. Front Genet. 2019;10:1014.
53. Li S, Ji FF, Hou FF, Cui HL, Shi QQ, Xing GM, et al. Characterization of Hemerocallis citrina transcriptome and development of EST-SSR markers for evaluation of genetic diversity and population structure of *Hemerocallis* collection. Front Plant Sci. 2020;11:686.
54. Vu DD, Shah SNM, Pham MP, Bui VT, Nguyen MT, Nguyen TPT. De novo assembly and transcriptome characterization of an endemic species of Vietnam, *Panax vietnamensis* ha et Grushv., including the development of EST-SSR markers for population genetics. BMC Plant Biol. 2020;20(1):358.
55. Xu RX, Wang Z, Su YJ, Wang T. Characterization and development of microsatellite markers in *Pseudotaxus chienii* (Taxaceae) based on transcriptome sequencing. Front Genet. 2020;11:574304.
56. Bazzo BR, de Carvalho LM, Carazzolle MF, Pereira GAG, Colombo CA. Development of novel EST-SSR markers in the macauba palm (*Acrocomia aculeata*) using transcriptome sequencing and cross-species transferability in Arecaceae species. BMC Plant Biol. 2018;18(1):276.
57. Qi WH, Jiang XM, Yan CC, Zhang WQ, Xiao GS, Yue BS, et al. Distribution patterns and variation analysis of simple sequence repeats in different genomic regions of bovid genomes. Sci Rep. 2018;8:14407.
58. Zhang Y, Zhang X, Wang YH, Shen SK. De novo assembly of transcriptome and development of novel EST-SSR markers in *Rhododendron rex* Levl. Through Illumina sequencing. Front. Plant Sci. 2017;8:1664.
59. Zhang ZY, Xie WG, Zhao YQ, Zhang JC, Wang N, Ntakirutimana F, et al. EST-SSR marker development based on RNA-sequencing of *E. sibiricus* and its application for phylogenetic relationships analysis of seventeen Elymus species. BMC Plant Biol. 2019;19:235.
60. Chen LY, Cao YN, Yuan N, Nakamura K, Wang GM, Qiu YX. Characterization of transcriptome and development of novel EST-SSR makers based on next-generation sequencing technology in *Neolitsea sericea* (Lauraceae) endemic to east Asian land-bridge islands. Mol Breed. 2015;35(9):1–15.

61. Martienssen RA, Colot V. DNA methylation and epigenetic inheritance in plants and filamentous fungi. Science. 2001;293(5532):1070–4.
62. Morgante M, Hanafey M, Powell W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet. 2002;30(2):194–200.
63. Wei Z, Sun ZZ, Cui BB, Zhang QX, Xiong M, Wang X, et al. Transcriptome analysis of colored calla lily (*Zantedeschia rehmannii* Engl.) by Illumina sequencing: de novo assembly, annotation and EST-SSR marker development. Peer J. 2016;4:e2378.
64. Nayak SS, Pradhan S, Sahoo D, Parida A. De novo transcriptome assembly and analysis of *Phragmites karka*, an invasive halophyte, to study the mechanism of salinity stress tolerance. Sci Rep. 2020;10(1):5192.
65. Zheng XF, Pan C, Diao Y, You YN, Yang CZ, Hu ZL. Development of microsatellite markers by transcriptome sequencing in two species of *Amorphophallus* (Araceae). BMC Genomics. 2013;14:490.
66. Huang DN, Zhang YQ, Jin MD, Li HK, Song ZP, Wang YG, et al. Characterization and high cross-species transferability of microsatellite markers from the floral transcriptome of *aspidistra saxicola* (Asparagaceae). Mol Ecol Resour. 2014;14(3):569–77.
67. Jarem DA, Huckaby LV, Delaney S. AGG interruptions in (CGG)(n) DNA repeat tracts modulate the structure and thermodynamics of non-B conformations in vitro. Biochemistry. 2010;49(32):6826–37.
68. Volle CB, Delaney S. AGG/CCT interruptions affect nucleosome formation and positioning of healthy-length CGG/CCG triplet repeats. BMC Biochem. 2013;14:33.
69. Biswas MK, Bagchi M, Nath UK, Biswas D, Natarajan S, Jesse DMI, et al. Transcriptome wide SSR discovery cross-taxa transferability and development of marker database for studying genetic diversity population structure of *Lilium* species. Sci Rep. 2020;10(1):18621.
70. Varshney RK, Graner A, Sorrells ME. Genic microsatellite markers in plants: features and applications. Trends Biotechnol. 2005;23(1):48–55.
71. Wu QC, Zang FQ, Xie XM, Ma Y, Zheng YQ, Zang DK. Full-length transcriptome sequencing analysis and development of EST-SSR markers for the endangered species *Populus wulianensis*. Sci Rep. 2020;10(1):16249.
72. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980;32(3):314–31.
73. Hamrick JL, Godt MJW. Effects of life history traits on genetic diversity in plant species. Phil Trans R Soc Lond B. 1996;351(1345):1291–8.
74. Nybom H. Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. Mol Ecol. 2004;13(5):1143–55.
75. Turchetto C, Segatto ALA, Mader G, Rodrigues DM, Bonatto SL, Freitas LB. High levels of genetic diversity and population structure in an endemic and rare species: implications for conservation. Aob Plants. 2016;8:plw002.
76. Luan SS, Chiang TY, Gong X. High genetic diversity vs. low genetic differentiation in *Nouelia insignis* (Asteraceae), a narrowly distributed and endemic species in China, revealed by ISSR fingerprinting. Ann Bot. 2006;98(3):583–9.
77. Qian X, Wang CX, Tian M. Genetic diversity and population differentiation of *Calanthe tsoongiana*, a rare and endemic orchid in China. Int J Mol Sci. 2013;14(10):20399–413.
78. Saboori S, Noormohammadi Z, Sheidai M, Marashi S. SCoT molecular markers and genetic fingerprinting of date palm (*Phoenix dactylifera* L.) cultivars. Genet Resour Crop Evol. 2020;67(1):73–82.
79. Garcia-Arias FL, Osorio-Guarin JA, Zarantes VMN. Association study reveals novel genes related to yield and quality of fruit in cape gooseberry (*Physalis peruviana* L.). front. Plant Sci. 2018;9:362.
80. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 1987;19(1):11–5.
81. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–U130.
82. Fan WQ, Gai HM, Sun X, Zhang ZF, Ren M. DataFormater, a software for SSR data formatting to develop population genetics analysis. Mol Plant Breed. 2016;14(1):265–70.
83. Peakall R, Smouse PE. GenAlEx 6.5: genetic analysis in excel. Population genetic software for teaching and research-an update. Bioinformatics. 2012;28(19):2537–9.
84. Liu KJ, Muse SV. PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics. 2005;21(9):2128–9.
85. Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P. MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes. 2004;4(3):535–8.
86. Rohlf F. NTSYSpc: numerical taxonomy and multivariate system. Version 2.1 Exeter software. New York: Applied Biostatics Inc; 2000.
87. Wang H, Ma DM, Yang JF, Deng K, Li M, Ji XY, et al. An integrative volatile Terpenoid profiling and transcriptomics analysis for Gene Mining and functional characterization of *AvBPPS* and *AvPS* involved in the Monoterpenoid biosynthesis in *Amomum villosum*. Front Plant Sci. 2018;9:846.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.