Biologia
https://doi.org/10.1007/s11756-022-01244-y

ORIGINAL ARTICLE

DNA barcoding markers provide insight into species discrimination, genetic diversity and phylogenetic relationships of yam (*Dioscorea* spp.)

Nicholas Kipkiror[1] · Edward K. Muge[2] · Dennis M. W. Ochieno[3] · Evans N. Nyaboga[2]

Received: 7 March 2022 / Accepted: 13 October 2022

Abstract

Yams (*Dioscorea* species) are tuber crops that are grown in tropical regions of Africa, the Caribbean, South America, Asia and South Pacific islands. It is an important food security crop with economic, nutritional and medicinal values. However, many *Dioscorea* species have similar morphology leading to inaccurate identification, which hinders their conservation and adequate exploitation of the economically important species. The aim of this study was to test the ability of DNA barcoding [ribulose 1, 5-bisphosphate carboxylase/oxygenase (*rbcL*) and Maturase K (*matK*)] markers to distinguish species and as an alternative tool for correcting species misidentification. Phylogenetic analysis of *rbcL* and *matK* sequences revealed four strongly supported distinct species that included *Dioscorea bulbifera, Dioscorea alata, Dioscorea minutiflora* and *Dioscorea cayennensis*. The specific clade of each of the yam accession was informed by the species. The phylogenetic clustering was confirmed by Principal Component Analysis (PCA). DNA polymorphism in the yam species exhibited both synonymous and non-synonymous mutation. *RbcL* and *matK* gene sequences had nucleotide diversity of 0.00392 and 0.00632, respectively. There were seven haplotypes within the *rbcL* gene with a diversity index of 0.800 and variation of 0.00374. For *matK* gene, there were four haplotypes with a diversity index of 0.745 and variation of 0.00956. This study demonstrates that *rbcL* and *matK* are efficient DNA barcoding markers that can be used to identify and discriminate *Dioscorea* species. The identification and delimitation of *Dioscorea* species is crucial in utilization, conservation and in designing their breeding strategies.

Keywords *Dioscorea* · DNA barcodes · Genetic diversity · *matK* · *rbcL* · Species identification

Introduction

Yam (*Dioscorea* species) is cultivated in the subtropical and tropical regions of Africa, South America, Caribbean islands, Asia and South Pacific islands (Andres et al. 2017). In Africa, yam is the second most important tuber crop after cassava and therefore constitutes a major source of starch for the growing population where food security is continuously becoming a challenge (Fu et al. 2011). Globally, West Africa is the largest growing zone for yams with Nigeria, Ghana, Côte d'Ivoire, Benin, and Togo being the main growing belt (Fu et al. 2011). In 2018, approximately 72.6 million tons of yams were produced worldwide with 97.1% of the production in Africa (FAOSTAT 2020). Its production increased from 0.9 million hectares in 1961 to 8.7 million hectares in 2018 with the main species grown being *D. alata* L. and *D. rotundata* Poir. complex (FAOSTAT 2020). In Kenya, yams are grown mainly in three regions namely western, coastal and central highlands (Muiruri 2009). In 2019, its yield was approximated to be about 20,028 metric tons while the cultivated species included *Dioscorea bulbifera* L., *Dioscorea minutiflora* Engl. and *Dioscorea dumetorum* (Kunth) Pax (Ministry of Agriculture, Livestock and Fisheries Kenya 2019). Apart from being a major source of food, yams do

✉ Evans N. Nyaboga
nyaboga@uonbi.ac.ke

[1] Centre for Biotechnology and Bioinformatics (CEBIB), University of Nairobi, P.O. Box 30197, Nairobi 00100, Kenya

[2] Department of Biochemistry, University of Nairobi, P.O. Box 30197, Nairobi 00100, Kenya

[3] Department of Biological Sciences, School of Natural Sciences (SONAS), Masinde Muliro University of Science and Technology, Kakamega Webuye Highway, P.O. Box 190-50100, Kakamega, Kenya

Published online: 14 November 2022

have other functions which include medicinal/pharmacological and industrial applications (Mignouna et al. 2008; Andriamparany et al. 2014; Barlagne et al. 2017).

Of the more than six hundred species of yams with only twelve species cultivated for food, income and medicinal values (Lebot 2009; Sonibare et al. 2010; Verter and Bečvářová 2015). These species include *Dioscorea rotundata*, *D. alata*, *D. bulbifera*, *D. cayennensis* Lam., *D. dumetorum*, *D. esculenta* (Lour.) Burkill, *D. japonica* Thunb., *D. opposita* Thunb., *D. pentaphylla* L., *D. transversa* R.Br., *D. trifida* L.f., and *D. nummularia* Lam. (Lebot 2009; Sonibare et al. 2010; Verter and Bečvářová 2015). Each of the species has unique food quality traits and different bioactive compounds for their food and medicinal values (Padhan and Panda 2020) and therefore accurate identification of yam plant species is important for appropriate utilization and conservation of genetic resources.

Yam being an important source of food security requires that their genetic improvement effort is assured; however, this has been constrained by a long growth cycle (approximately 8 months), separate and non-uniform flowering of male and female and dioecy (Tamiru et al. 2017). The dioecy feature of yams limits their efficiency with regard to breeding (Sugihara et al. 2021). Similarly, the process of domesticating *Dioscorea* species has been made ambiguous by the dioecy nature which is responsible for frequent hybridization and polyploidization (Sugihara et al. 2021). Moreover, clonal propagation of the yam may reduce its genetic diversity, causes vulnerability to diseases and difficulty in removing deleterious mutation (Ramu et al. 2017) which hampers their accurate identification that is based on heritable variations. Therefore, efficient genetic improvement of yams requires an accurate and reliable method for correct identification of cultivated *Dioscorea* species.

One of the efficient methods for plant species identification and molecular phylogeny is DNA barcoding, a technology that uses short and standard DNA fragments of the genome (Hebert et al. 2003; CBOL Plant Working Group 2009). It has become a useful tool for biodiversity investigation and monitoring, molecular phylogeny and evolution. The use of DNA barcoding markers for identification and characterization of yams can be a useful tool because it is not subject to agro-climatic or environmental variations. It can also provide requisite information and more insight on yam plant evolution (Deschamps et al. 2012). DNA barcoding markers that are universal and have minimum evolution rates over time are considered suitable for elucidating the identity and characterization of yams or any other plant. Maturase K (*MatK*) and ribulose 1, 5-bisphosphate carboxylase/oxygenase (*rbcL*) markers are considered candidate primers for identification and characterization of yam species because they are universal, and can provide better resolution/discrimination among species (Patwardhan et al. 2014). It has also been reported that about 92% of the plants can be distinguished using these markers. Therefore, the use of DNA barcodes to ascertain plant species, mutations over evolutionary time and map their location is easy as it can be applied across species (Ngo Ngwe et al. 2015). In Kenya, there is limited information on the identity and characterization of cultivated *Dioscorea* species, useful information for the conservation of their biodiversity. The current study utilized *rbcL* and *matK* barcoding markers in the identification of *Dioscorea* species cultivated in Kenya. In addition, the markers were used to resolve phylogenetic relationships among *Dioscorea* species found in Kenya as well as elucidating their evolutionary and taxonomic relationships.

Materials and methods

Plant material

A total of 20 yam accessions were used in the present study. The accessions were collected from different yam growing regions in Kenya as well as from the National Genebank of Kenya (Table 1).

Extraction of genomic DNA, PCR amplification and sequencing

Genomic DNA for all the yam accessions was extracted from 200 mg of leaf samples using cetyltrimethylammonium bromide (CTAB) method with some modifications (Abdel-Latif and Osman 2017). The modifications included grinding the leaf sample in 600 µl of CTAB buffer with 150 µl of 10% sodium dodecyl sulfate (SDS) and centrifuging at 14,000 rpm for 10 min. The quality of isolated DNA was ascertained using agarose gel electrophoresis.

Polymerase chain reaction (PCR) amplifications were performed on Applied Biosystems 96-Well Veriti Thermal Cycler (ThermoFisher Scientific, USA). Two primers targeting *matK* (www.barcoding.si.edu) and *rbcL* genes (Cuénoud et al. 2002) were used. The sequences for the primers were: *matK*_F: 5'CCTATCCATCTGGAAATCTT3', *matK*_R: 5'GTTCTAGCACAAGAAAGTCG3', *rbcL*_1_F: 5'ATG TCACCACAAACAGAAAC3' and *rbcL*_74R: 5'TCG CATGTACCTGCAGTAGC3'. PCR was carried out in a total reaction volume of 20 µl using Taq DNA Polymerase 2 × Master Mix RED with 2 mM $MgCl_2$ final concentration PCR Mastemix (Ampliqon, Stenhuggervej 22, Denmark) containing: Tris–HCl pH 8.5, $(NH4)_2S0_4$, 4 mM $MgCl_2$, 0.2% Tween-20, 0.4 mM of each dNTP, Ampliqon Taq DNA polymerase and Inert red dye and stabilizer. Amplification conditions were: initial denaturation at 94 °C for 5 min, then 35 cycles at 94 °C denaturation for 30 s, 58 °C (for *rbcL*) and 48 °C (for *matK*) for 45 s and 72 °C for 30 s with final extension at 72 °C for 7 min. An aliquot (10 µl) of the PCR

Table 1 Samples of yam accessions used in the current study

No	Accession Code	Accession Name	Location of collection	GenBank accession No. (*rbcL*)	GenBank accession No. (*matK*)*	Name of species
1	A_N	Amola	National GeneBank of Kenya	MT522436	MT522414	*D. cayennensis*
2	B_N	Obiotungi	National GeneBank of Kenya	MT522437	MT522415	*D. cayennensis*
3	25_N	TDr2579	National GeneBank of Kenya	MT522434	-	*D. cayennensis***
4	19_N	TDr0097	National GeneBank of Kenya	MT522435	MT522413	*D. cayennensis*
5	6E_Murang'a	6E_Murang'a	Murang'a	MT522440	-	*D. cayennensis***
6	7E_Kirinyaga	7E_Kirinyaga	Kirinyaga	MT522426	MT522417	*D. cayennensis*
7	X1_N	TDr2436	National GeneBank of Kenya	MT522439	MT522416	*D. cayennensis*
8	M_N	Makakkwa	National GeneBank of Kenya	MT522438	-	*D. cayennensis***
9	06_N	TDr0060	National GeneBank of Kenya	MT522430	-	*D. alata***
10	2E_Meru	2E_Meru	Meru	MT522431	-	*D. alata***
11	00E_Meru	00E_Meru	Meru	MT522421	MT522441	*D. minutiflora*
12	1E_Meru	1E_Meru	Meru	MT522422	MT522442	*D. minutiflora*
13	3E_Meru	3E_Meru	Meru	MT522423	-	*D. minutiflora*
14	4E_Meru	4E_Meru	Meru	MT522424	-	*D. minutiflora*
15	5E_Murang'a	5E_Murang'a	Murang'a	MT522425	MT522443	*D. minutiflora*
16	8E_Nyeri	8E_Nyeri	Nyeri	MT522427	MT522444	*D. minutiflora*
17	10E_Kirinyaga	10E_Kirinyaga	Kirinyaga	MT522428	-	*D. minutiflora***
18	11E_Nyeri	11E_Nyeri	Nyeri	MT522429	-	*D. minutiflora***
19	M2_Molo	M2_Molo	Molo	MT522432	MT522418	*D. bulbifera*
20	M3_Molo	M3_Molo	Molo	MT522433	MT522419	*D. bulbifera*

*Column for *MatK*,—represent blanks because sequencing was not successful in those samples using *matK* marker

**Indicates that the yam species name in the column is only based on rbcL sequences

product was subjected to 1% (w/v) agarose gel electrophoresis stained with ethidium bromide (0.5 µg/ml).

The amplification products with expected band sizes were purified using the GeneScript QuickClean DNA Gel extraction kit (Piscataway NJ, USA) as per the manufacturer's instructions. The samples were sent to Macrogen Genomic platform (Netherlands) for bidirectional Sanger sequencing.

Sequence assembly and analysis

The raw sequences generated were edited and assembled in Geneious Prime 2019.2 (https://www.geneious.com/) and viewed on Jalview version 2.11.0. Both de novo assembly and assembling by mapping to reference sequences were done. For de novo assembly, the border of each marker was selected based on the quality of the conservation sites set at 50% and above. Consensus sequences were generated from the sequenced fragments. Mapping to the respective reference genome sequences for both markers was done as the sequences were required for phylogeography, DNA polymorphism and divergence and haplotype analysis. *MatK* sequences were mapped to NC_039708.1:c3223-1664 (*D. bulbifera*), NC_039835.1:c3205-1646 (*D. minutiflora*) and NC_039836.1:c3214-1655 (*D. cayennensis*) reference sequences while *rbcL* generated sequences were mapped to NC_039707.1:54,667–56,100 (*D. alata*), NC_039708.1:54,453–55,886 (*D. bulbifera*), NC_039835.1:54,325–55,758 (*D. minutiflora*) and NC_039836.1:54,484–55,917 (*D. cayennensis*) sequences. These reference sequences were derived for each of the *Dioscorea* species as outlined herein.

Sequence alignment and phylogenetic analysis

Sequence similarity of de novo assembled sequences was searched by BLASTn and sequences aligned using MUSCLE Version 3.8 (Edgar 2004) implemented in Geneious Prime 2019.2. Phylogenetic construction was done using Maximum Likelihood (ML) technique for tree drawing with TN93 model (Tamura et al. 2013) for both markers. First, the sequences aligned on MUSCLE were exported and viewed on Jalview version 2.11.0.; then trimmed to a requisite conservatory quality of > 50%. The resultant aligned sequences were then exported to Geneious Prime (Trial Licensed- TRIAL-179A-DB79-399B-27F1) with which PhyML package had been installed. Phylogenetic tree reconstruction was undertaken at 1000 bootstrap value in TN93

model. The tree was visualized in an inbuilt Geneious Prime visualization system.

Phylogeographic divergence of yam lineages

Chloroplast DNA (cpDNA) *rbcL* and *matK* markers were used to examine the differentiation and phylogeographic history of the *Dioscorea* species grown in Kenya. Phylogeography patterns of *Dioscorea* species was done using Bayesian model implemented in Bayesian Evolutionary Analysis Sampling Tree (BEAST), Qin et al. (2013). Analysis was performed through ten million steps of Markov Chain Monte Carlo (MCMC) to build the tree posteriorly. Maximum clade credibility tree was annotated using Tree Annotator available in BEAST with 10% burn-in of all the trees built to remove the non significant trees. Subsequently, the resultant annotated tree was visualized on FigTree version 1.4.4.

DNA polymorphisms

Sequence polymorphism of cpDNA markers *rbcL* and *matK* were screened in all the amplified samples as described by Besnard et al. (2007). DNA Sequence Polymorphism (DnaSP) version 6.12.03 was used to analyze polymorphism in the *rbcL* and *matK* genes of the sequenced samples. The sequences were first mapped to reference sequences, aligned on MUSCLE version 3.8 (Edgar 2004), viewed and then trimmed on Jalview version 2.11.0 to a uniform sequence length. A permutation approach was used to estimate the significance of sequence differences between and within the yam species. This involved the estimation of statistical pairwise nucleotide differences.

DNA divergence between populations

DNA divergence between yam populations was measured by computing their variance Pi, Dxy and Da as outlined by Jukes and Cantor algorithm implemented in DnaSP version 6.12.03 described by Kartavtsev (2011). The DNA divergence between yam populations was measured using AMOVA and significance tested at 1000 permutations (Excoffier et al. 2009).

Haplotype analysis

Haplotype analysis of yam genotypes was carried out to ascertain the network of the cpDNA haplotypes (Guan 2014). This was done by constructing a cpDNA network using DnaSP version 6.12.03 which generated the haplotype data files. Arlequin 3.5.2.2 was used for haplotype inference by estimating allele frequencies at each locus. AMOVA was also undertaken at 1000 permutations (Beaumont and Nichols 1996; Excoffier et al. 2009). Significance for all the analysis was evaluated at 5% confidence level.

Principal component analysis (PCA)

Principal component analysis (PCA) was conducted to ascertain the relationship among the *Dioscorea* species. XLSTAT (License 7E1503-3C5B2C-43B5AA-1BFC14-F0C37F-5AB5EE) that employs multivariate analysis was used in the study. Biplot score was used to reveal the groupings (Bro and Smilde 2014). The dataset involved 24 and 14 accessions for *rbcL* and *matK* markers, respectively.

Results

Amplification, sequencing and multiple sequence alignment

Following PCR amplifications, single amplicons of ~ 710 and 830 base pairs (bp) were obtained with the markers *rbcL* and *matK*, respectively, for all the samples. The PCR amplification efficiency was 100% for both *rbcL* and *matK* regions. The success recovery rate of sequences was 90.91% and 50.0% for *rbcL* and *matK*, respectively. A total of 31 sequences were submitted to the NCBI with 20 and 11 being for *rbcL* and *matK* regions, respectively (Table 1). Samples of accessions (TDr2579, TDr0060, Makakkwa and 10_Kirinyaga) amplified using *matK* were not considered in the analysis because of their quality which was below the set sensitivity on sequence similarity. The edited sequences after homology searches were deposited in the Genbank and accession numbers for each *rbcL* and *matK* regions is given in Table 1.

Multiple sequence alignment and DNA polymorphism

Alignments of cleaned sequences produced 692 and 795 base pairs for *rbcL* and *matK*, respectively (Supplementary Fig. 1 and 2). Upon alignment and subsequent viewing on ESPript 3, single nucleotide polymorphic sites were identified. The polymorphic sites were recognized to be both synonymous and non-synonymous mutations on annotation. These mutations were found in all the four identified *Dioscorea* species. The codon position of the respective SNPs was also depicted.

Phylogenetic relationships based on rbcL and matK markers

To ascertain the identity of the resultant consensus sequences, a blast search against the NCBI database

sequences was done and yam species identified based on percentage similarity and respective E-values. All the blast sequences had similarity of 99.0% and above while their corresponding E-values was 0.0.

The tree showed a well–resolved phylogeny supported by bootstrap values and revealed four strongly supported species clades (Clade A, B, C and D) (Fig. 1). The species identified included *D. bulbifera*, *D. alata*, *D. minutiflora* and *D. cayennensis*. The tree was classified in an ascending order based on bootstrap values in each of the species' ancestral node. Clade A which represented *D. minutiflora* had a support value of 60.0 and the accessions in this clade were 11E_Nyeri, 10E_Kirinyaga, 8E_Nyeri, 5E_Murang'a, 4E_Meru, 3E_Meru, 1E_Meru and 00E_Meru. The relationship in this section was generally monophyletic. *D. cayennensis* (Clade B) was strongly supported by a bootstrap value of 78.0 and its accessions included 6E_Murang'a, 7E_Kirinyaga, Obiotungi, TDr2579, Amola, TDr2436, TDr0097 and Makakkwa. The relationship in this section was largely paraphyletic. *D. alata* (Clade C) had a support value of 90.0 with accessions being TDr0060 and 2E_Meru while its relationship was monophyletic. Clade D which was *D. bulbifera* had the highest support value of 99.0, a monophyletic relationship and its accessions were M3_Molo and M2_Molo (Fig. 1).

Three clades were obtained by Maturase K (*matK*) marker on phylogenetic reconstruction and each clade represented different *Dioscorea* species (Fig. 2). One subclade with more than two accessions was observed to segregate from clade A. The species identification included *D. minutiflora*, *D. cayennensis* and *D. bulbifera*. Clade A is *D. minutiflora* with a support value of 74.0, and polyphyletic relation while its cultivars included 1E_Meru, 8E_Nyeri, 5E_Murang'a and 0EMatK_Meru. Clade B consisted of *D. cayennensis* which was strongly supported by a value of 98.0 with a polyphyletic relationship and its accessions included 7E_Kirinyaga, Obiotungi, Amola, TDr2436 and TDr0097. *D. bulbifera* of Clade C was strongly supported by a value of 100 and its accessions were M3_Molo and M2_Molo and it had a monophyletic relationship.

For a better understanding and informative distinction of yam species, the use of combined markers to build a phylogenetic tree was explored (Fig. 3). The discriminatory power by individual markers upon combination was clearly notable. Only TDr2436 had ambiguous identity while 8E_Nyeri was correctly classified to belong to *D. minutiflora* species.

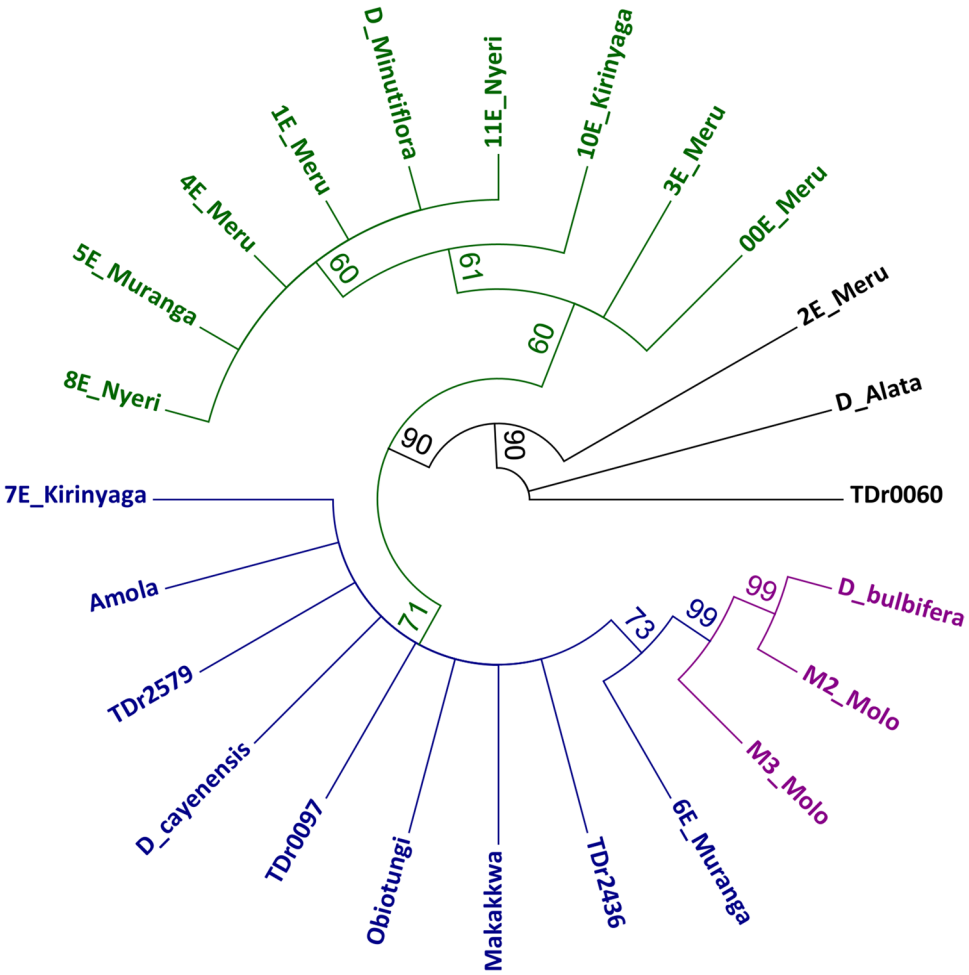


Fig. 1 Phylogenetic tree reconstruction based on *rbcL* marker. Each color indicates a different clade whereas a clade represents a distinct species. Clade A, green color represents *D. minutiflora*; Clade B, blue color is *D. cayennensis*; Clade C, black color represents *D. alata* and Clade D, purple color represents *D. bulbifera species*

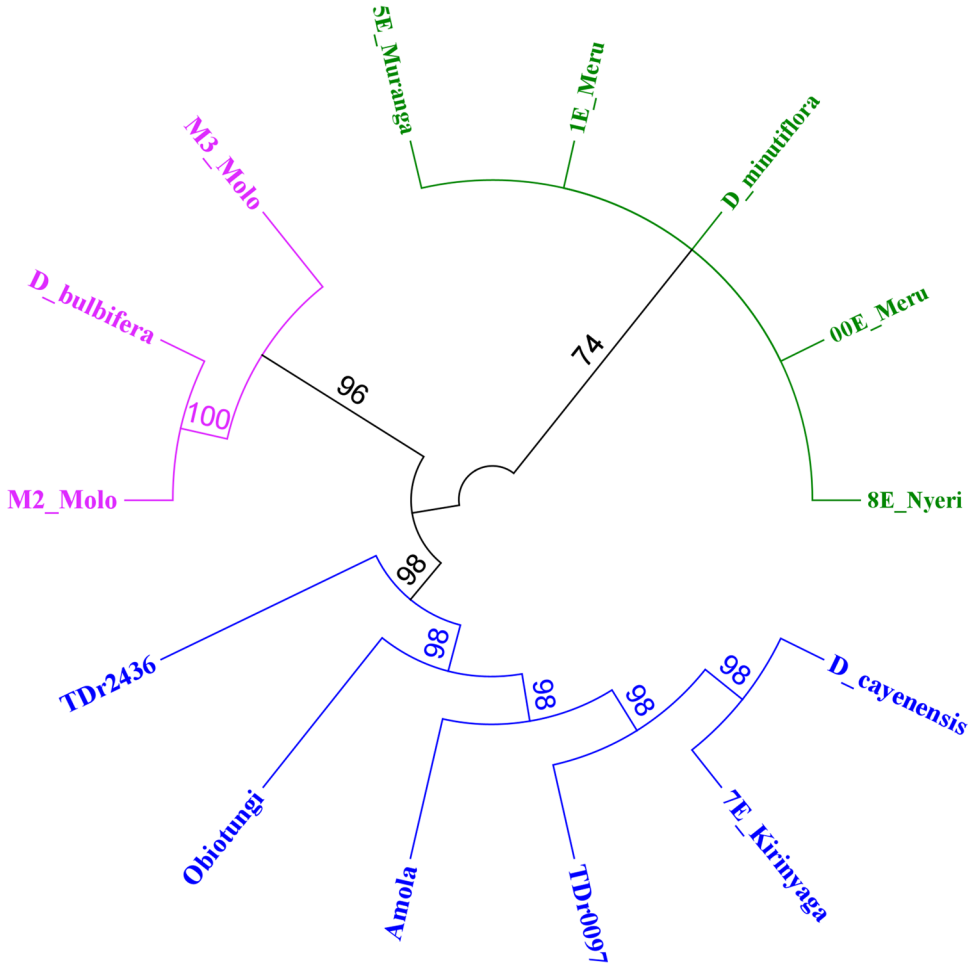


Fig. 2 Phylogenetic tree reconstruction based on *matK* marker. Different colors are used to represent the respective clades which represent different species. Clade A, green color represents *D. minutiflora*; Clade B, blue color is *D. cayennensis*; Clade C, purple color represents *D. bulbifera species*

The cultivars which belong to the species *bulbifera* were correctly classified by the two markers and their relationship resolved to be monophyletic. The accessions of the species *alata* were correctly classified by *rbcL* marker and its relationship was monophyletic. All the accessions for *D. minutiflora* were correctly resolved by the two markers.

Phylogeography analysis

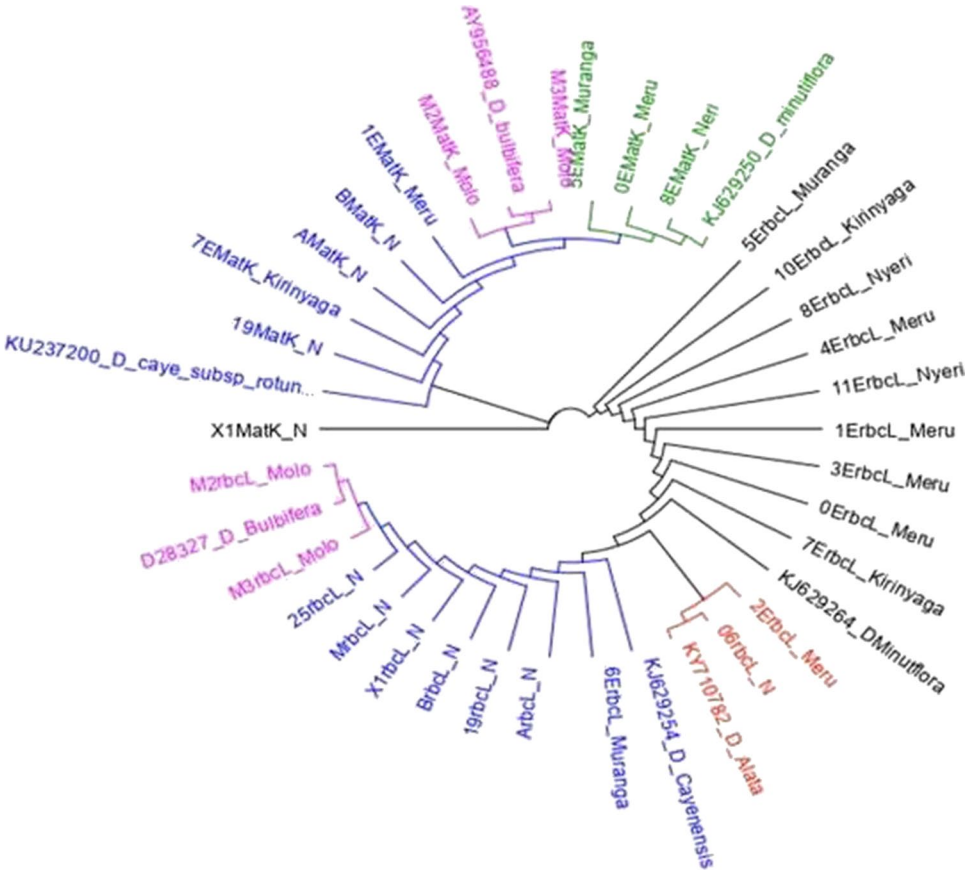
The yam accessions maximally differentiated into two distinct clades (Clade I and Clade II) with reference to *rbcL* sequences (Fig. 4). The differentiation used lognormal relaxed distribution under the uncorrelated relaxed clock. Clade I contained accessions classified based on the node age that indicated their divergence relationship. The accessions that belonged to Clade I include 0ErbcLMeru, 2E_Meru, 3E_Meru, 4E_Meru, TDr0060, 8E_Nyeri, 5E_Murang'a, 6E_Murang'a, 7E_Kirinyaga, 11E_Nyeri, 10E_Kirinyaga, TDr2579 and TDr0097. These accessions were from upper eastern (Meru) and central (Kirinyaga, Murang'a and Nyeri) regions of Kenya. Accessions from the Gene bank in this clade are TDr0060, TDr2579 and TDr0097 which were originally from Nigeria then cultivated in Kenya and preserved at the National Gene Bank of Kenya.

Clade II had accessions whose divergence was close as indicated by their node age relationship. The accessions included 1E_Meru, Obiotungi, Amola, TDr2436, Makakkwa M3_Molo and M2_Molo. These accessions were from Rift Valley and those originally obtained from Nigeria and cultivated in Kenya. Only one cultivar from Meru was included in clade II.

Yam accessions were distinctively categorized into two clades (Clade I and Clade II) based on the node age relations on *matK* region (Fig. 5). The differentiation denotes the phylogeographuc relationship between the cultivars under study. Clade I contained accessions 1E_Meru and 00E_Meru that were from upper eastern (Meru) region of Kenya. Clade II contained accessions such as M2_Molo, TDr2436, M3_Molo, 5E_Murang'a, 8E_Nyeri, Amola, 7E_Kirinyaga, Obiotungi and TDr0097. The accessions in clade II were from Rift Valley, Central (Murang'a, Kirinyaga and Nyeri), upper Eastern (Meru) regions of Kenya and National GeneBank of Kenya.

Genetic diversity of *rbcL* and *MatK* polymorphism

Twelve (12) segregation sites (S) or variable sites within the *rbcL* gene of the accessions were identified (Table 2). The nucleotide diversity (π) was calculated to be 0.00392 while the



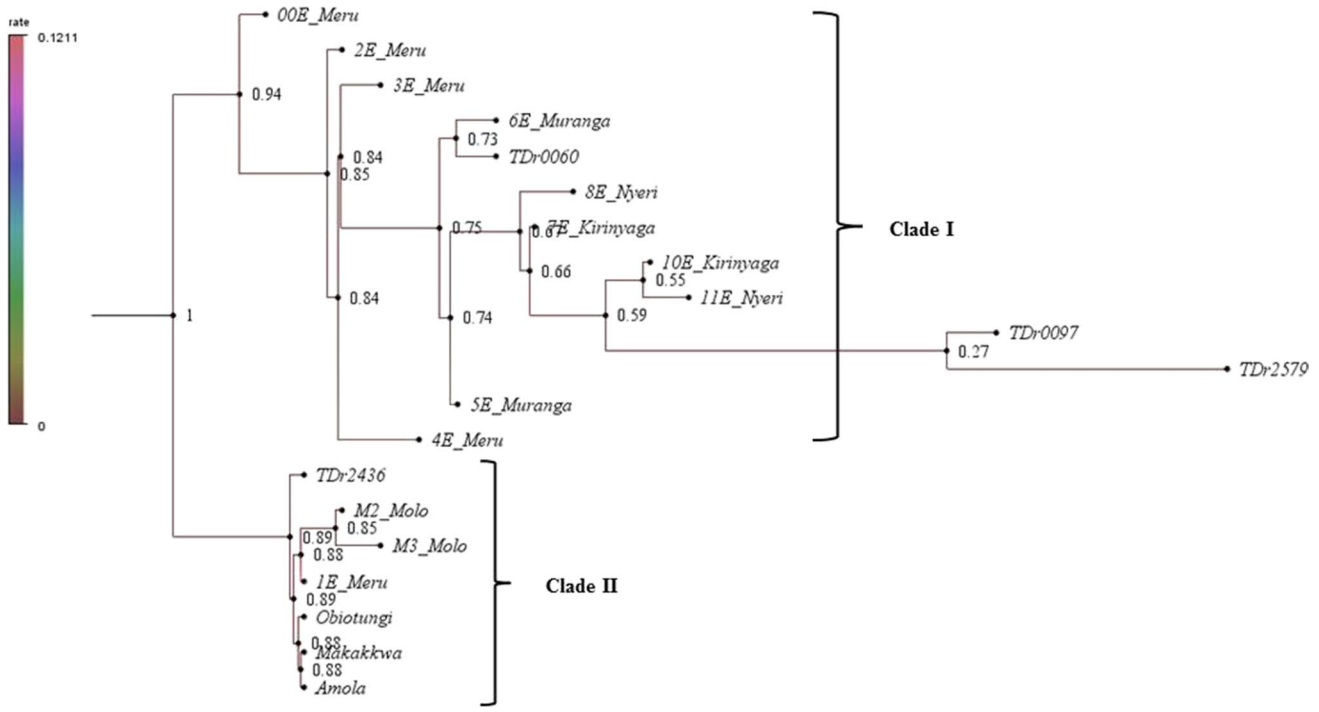


Fig. 3 Phylogenetic tree reconstruction based on both *rbcL* and *matK* markers. Different colors are used to represent the respective clades which represent different species. Clade A, blue color represents *cayennensis*; Clade B, purple color is *D. bulbifera*; Clade C, black color represents *D. alata species*. Color green is Clade D which represents *D. minutiflora species*

Fig. 4 Phylogeography representation based on *rbcL* marker. Two distinct clades resulting from the species differentiation were obtained. Clade I and Clade II generally represented species from different geographical region. The differentiation time indicated informed the relatedness of the cultivars in each clade. Clade I and II are represented by black and blue text, respectively

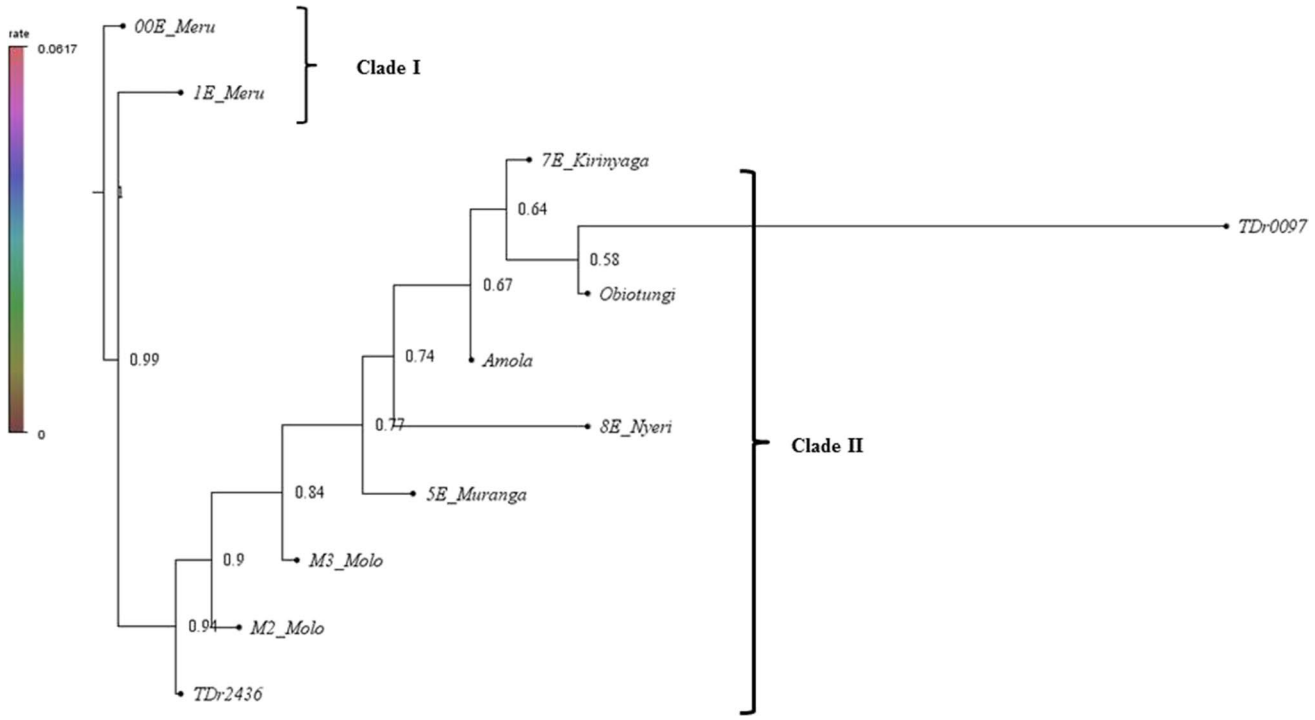


Fig. 5 Phylogeography representation based on *matK* marker. Two distinct clades resulting from the species differentiation were obtained. Clade I and Clade II generally represented species from respective geographical regions. The differentiation time indicated informed the relatedness of the cultivars in each clade. Clade I and II are represented by black and blue text, respectively

Table 2 DNA polymorphism based on *rbcL* marker

Polymorphic sites/ Segregation sites (S)	12	Position in the Gene	Reference	Amino Acid	Variant Codon	
Singleton	3	5	TGT	Cys	TTT	Phe
		7	TGG	Trp	GGG	Gly
		682	GGG	Gly	AGG	Arg
Parsimony informative sites	9	12	ATT	Ile	ATC	Ile
		58	CTA	Leu	ATA	Ile
		163	TGG	Trp	CGG	Arg
		280	CTA	Leu	TTA	Leu
		430	ACA	Thr	GCA	Ala
		433	CGG	Arg	TGG	Trp
		469A	TGG	Trp	CGG	Arg
		551A	TCT	Ser	TTT	Phe
		670	GGG	Gly	AGG	Arg

average number of nucleotide differences was 2.668. The *rbcL* gene had a total of 669 monomorphic sites and a sequence length of 692 base pairs while twelve polymorphic sites identified comprised of 3 singleton mutations and 9 parsimony informative bases. The corresponding codons in the *rbcL* gene were mutants and belonged to both synonymous and non-synonymous mutations. These mutations were considered to be both singleton and parsimony informative sites. Synonymous mutation occurred in positions 12, 58, 280 and 430. The synonymous mutations were attributed to selection that occurred within the respective species populations. Mutations considered non synonymous occurred in positions 5, 7, 163, 433 469, 551 670 and 682 (Table 2). These non-synonymous mutations were attributed to genetic diversity of the accessions.

Table 3 DNA polymorphism based on *matK* marker

Polymorphic sites/ Segregation Sites (S)	14	Position in the Gene	Reference		Variant Codon	
Singleton	2	758	CAC	Hist	CTC	Leu
		784	TTG	Leu	CTG	Leu
Parsimony informative sites	12					
		83	CAG*	Gln	CCG	Pro
		257	GGA*	Gly	GAA	Glu
		403	GGA*	Gly	AGA	Arg
		412	CTT*	Leu	TTT	Phe
		468	CAT	Hist	CAC	Hist
		495	TTC	Phe	TTT	Phe
		532	CCT	Pro	TTT	Phe
		533	CCT*	Pro	TTT	Phe
		545	AGG	Arg	AAG	Lys
		561	GAA	Glu	GAG	Glu
		658	ATA	Ile	GTA	Val
		672	CCC	Pro	CCT	Pro

*Positions of mutations considered non-synonymous

Fourteen (14) segregation sites (S) and 757 monomorphic sites /invariant base pairs were identified within the *matK* sequences of the yam accessions (Table 3). The nucleotide diversity (π) in the eleven (11) sequences was 0.00632 while the average number of nucleotide differences was 4.87273. The polymorphic sites comprised of both singleton mutations (2) and twelve parsimony informative bases (Table 3). The corresponding codons in the *MatK* sequence were mutants and belonged to both synonymous and non-synonymous mutations. Synonymous mutation occurred in positions 468, 495, 545, 561, 658, 672 and 784. The synonymous mutations were attributed to selection that occurred within the respective species population. Mutations considered non synonymous occurred in positions 83, 257, 403, 412, 532, 533 and 758. These non-synonymous mutations were attributable to genetic diversity of the accessions.

DNA divergegence between populations

DNA divergence between yam populations based on *rbcL* marker revealed that there were no shared mutations between the yam species populations (Table 4). The average number of nucleotide substitutions per site between populations ranged from 0.00375 to 0.01019 which corresponded to its number of net nucleotide substitution per site between populations that ranged from 0.00298 to 0.01019 (Table 4). *D. minutiflora* species had polymorphic mutations while the other yam species populations had no mutations. The total number of differences between populations was: *D. bulbifera* and *D. minutiflora* = 6, *D. bulbifera* and *D. alata* = 7, *D. bulbifera* and *D. cayennensis* = 4, *D. alata* and *D. minutiflora* = 4, *D. cayennensis* and *D. minutiflora* = 3, *D. alata* and *D. cayennensis* = 3. The number of fixed differences between populations was inferred from the total polymorphic sites between populations. The fixed differences number alluded to the diversity and relatedness of the accessions used in this study. Species *D. minutiflora* had the highest number of polymorphic sites between populations, which may relate to the number of accessions used and their diversity.

There were no shared mutations between the yam species based on *matK* marker (Table 5). The average number of nucleotide substitutions per site between populations ranged from 0.00454 to 0.01492 which corresponded to its number of net nucleotide substitutions per site between populations that ranged from 0.00389 to 0.01427 (Table 5). Only mutations in *D. minutiflora* species were polymorphic but monomorphic in the other yam species. The total number of fixed differences between species were *D. bulbifera* and *D. minutiflora* = 11, *D. bulbifera* and *D. cayennensis* = 10, *D. cayennensis* and *D. minutiflora* = 3. The number of fixed differences was inferred from the total polymorphic sites between the different species. Relatedness among the yam species can be deduced from their polymorphic site differences/number of fixed differences. The relatedness of *D. minutiflora* and *D. cayennensis* was due to differences in 3 bases while that of *D. cayennensis* and D. *bulbifera* was 10 bases difference. *D. bulbifera* and *D. minutiflora* was 11 bases genetically distant.

Haplotype analysis

Haplotype analysis of yam species was conducted to ascertain the network of the cpDNA haplotypes. Haplotypes were

Table 4 DNA divergegence between *Dioscorea* species populations based on *rbcL* marker

Population	*D. bulbifera* (*P1*)	*D. minutiflora* (*P2*)	*D. bulbifera* (*P1*)	*D. alata* (*P2*)	*D. bulbifera* (*P1*)	*D. cayennensis* (*P2*)	*D. minutiflora* (*P1*)	*D. alata* (*P2*)	*D. minutiflora* (*P1*)	*D. cayennensis* (*P2*)	*D. alata*(*P1*)	*D. cayennensis* (*P2*)
Polymorphic sites in each population	0	4	0	0	0	0	4	0	4	0	0	0
Total number of polymorphic sites	10		7		4		7		6		3	
Average number of nucleotide difference (kt)	2.836		4.667		1.556		1.855		1.658		1.167	
Nucleotide diversity(pt)	0.00416		0.00679		0.00227		0.00272		0.00244		0.0017	
Number of fixed differences	6		7		4		4		2		3	
Mutations polymorphic in P1 but Monomorphic in P2	0		0		0		4		4		0	
Mutations polymorphic in P2, but monomorphic in P1	4		0		0		0		0		0	
Shared Mutations	0		0		0		0		0		0	
Average number of nucleotide differences between populations	6.556		7.000		4.000		3.556		2.556		3.000	

Table 4 (continued)

Population	*D. bulbifera (P1)*	*D. minutiflora (P2)*	*D. bulbifera (P1)*	*D. alata (P2)*	*D. bulbifera (P1)*	*D. cayennensis (P2)*	*D. minutiflora (P1)*	*D. alata (P2)*	*D. minutiflora (P1)*	*D. cayennensis (P2)*	*D. alata(P1)*	*D. cayennensis (P2)*
Average nucleotide substitution per site between population	0.00961		0.01019		0.00584		0.00521		0.00375		0.00438	
Number of net nucleotide substitution per site between populations	0.00884		0.01019		0.00584		0.00444		0.00298		0.00438	

generated by analyzing the polymorphic sites within the *rbcL* and *MatK* genes. Twelve polymorhic sites within the *rbcL* gene and 14 polymorphic sites on the *MatK* gene were involved in detemination of haplotypes of the sequenced accesions.

There were 7 haplotypes within the *rbcL* gene with a diversity index of 0.800 and variation of 0.00374 (Table 6). Haplotypes 4 and 7 had the highest number of accessions with seven and six, respectively. Haplotypes 1 and 3 had two accessions each while each of the haplotypes 2, 5 and 6 had a single accession each (Table 6). The distribution of haplotypes was in accordance with their geographic location and species. Haplotypes 1 contained accessions from Rift valley (Molo) which belonged to *D. bulbifera* species. Haplotype 2 and 6 had accessions from Meru that belonged to *D. minutiflora* species. Haplotype 3 had accessions from both Meru and GeneBank of Kenya that belonged to species *D. alata.* Haplotype 4 comprised of accessions from GeneBank of Kenya and one from Murang'a which were *D. cayennensis* species. Haplotype 5 had one accessions from Nyeri of *D. minutiflora* species. Haplotype 7 comprised of accessions *D. minutiflora* from Nyeri, Murang'a, Kirinyaga and Meru. Haplotype 3 and 4 comprised of accessions from different geographic locations. No accession had multiple haplotypes.

There were 4 haplotypes within the *matK* gene with a diversity index of 0.745 and variation of 0.00956 (Table 6). Haplotype 3 and 4 had the highest number of accessions with five and three respectively (Table 6). Haplotypes 1 had two cultivars while haplotype 1 had only one cultivar. The haplotype distribution elucidated the species and their geographic locations. Haplotype 1 had accessions from Rift valley which belonged to species *D. bulbifera*. Haplotype 2 had accession from Nyeri which belonged to *D. minutiflora* species. Haplotype 3 comprised of accessions from both Kirinyaga and GeneBank of Kenya that belonged to *D. cayenenis.* Haplotype 4 comprised of accessions from Murang'a and Meru which belong to the species *D. minutiflora*. Haplotype 3 comprised of accessions from different geographic locations. No accession had multiple haplotypes.

Principal component analysis (PCA)

Principal component analysis was also done to determine the geneti relationships among the accessions. Four distinct clusters were generated for *rbcL* marker with each cluster respresnting *D. cayennensis, D. minutiflora* and *D. bulbifera* and *D. alata*. The three distinct clusters for the *matK* marker were represented by *D. cayennensis*, *D. minutiflora* and *D. bulbifera*. For both *rbcL* and *matK* markers, each of the cluster represented different yam species (Figs. 6 and 7). The results from the PCA were similar to the phylogenetic clustering.

Table 5 DNA divergence between populations based on *matK* marker

Population	*D. bulbifera* (*P1*)	*D minutiflora* (*P2*)	*D. bulbifera* (*P1*)	*D. caye_rotundata (P2)*	*D. minutiflora* (*P1*)	*D. caye_rotundata* (*P2*)
Polymorphic sites in Each population	0	4	0	0	2	0
Total Number of Polymorphic sites	13		10		5	
Average Number of nucleotide difference (kt)	6.533		5.333		2.214	
Nucleotide diversity(pt)	0.00847		0.00682		0.00287	
Number of fixed differences	11		10		3	
Mutations polymorphic in P1 but monomorphic in P2	0		0		2	
Mutations polymorphic in population 2, but monomorphic in population 1	2		0		0	
Shared mutations	0		0		0	
Average number of nucleotide differences between populations	11.500		10.000		3.50	
Average nucleotide substitution per site between population	0.01492		0.01279		0.00454	
Number of net nucleotide substitution per site between populations	0.01427		0.01279		0.00389	

Table 6 Haplotype analysis based on *rbc*L and *matK* marker

Marker	Haplotype Number	Dioscorea species	Haplotype	Yam accessions
rbcL	Hap_1: 2	*D. bulbifera*	GTTCTCACTCGG	[M2_Molo and M3_Molo]
	Hap_2: 1	*D. minutiflora*	TGCACTGTTCAG	[4E_Meru]
	Hap_3: 2	*D. alata*	GTTCCTGTCTAG	[2E_Meru and TDr0060]
	Hap_4: 7	*D. cayennensis*	GTTCCTGCTCAG	[6E_Murang'a, Amola, TDr2579, TDr0097, Obiotungi, Makakkwa and TDr2436]
	Hap_5: 1	*D. minutiflora*	GTCACTGTTCAG	[11E_Nyeri]
	Hap_6: 1	*D. minutiflora*	GTTACTGTTCAA	[1E_Meru]
	Hap_7: 6	*D. minutiflora*	GTTACTGTTCAG	[5E_Murang'a, 3E_Meru3E_Meru, 00E_Meru, 7E_Kirinyaga, 8E_Nyeri and 10E_Kirinyaga]
matK	Hap_1: 2	*D. bulbifera*	AGGCTCCCGAACAT	[M2_Molo and M3_Molo]
	Hap_2: 1	*D. minutiflora*	CAACCTTTAGGTTC	[8E_Nyeri]
	Hap_3: 5	*D. cayennenis rotundata*	CGATTTTTAGGTAT	[7E_Kirinyaga, TDr0097, Amola, Obiotungi and TDr2436]
	Hap_4: 3	*D. minutiflora*	CAACCTTTAGGTAT	[00E_Meru, 5E_Murang'a and 1E_Meru]

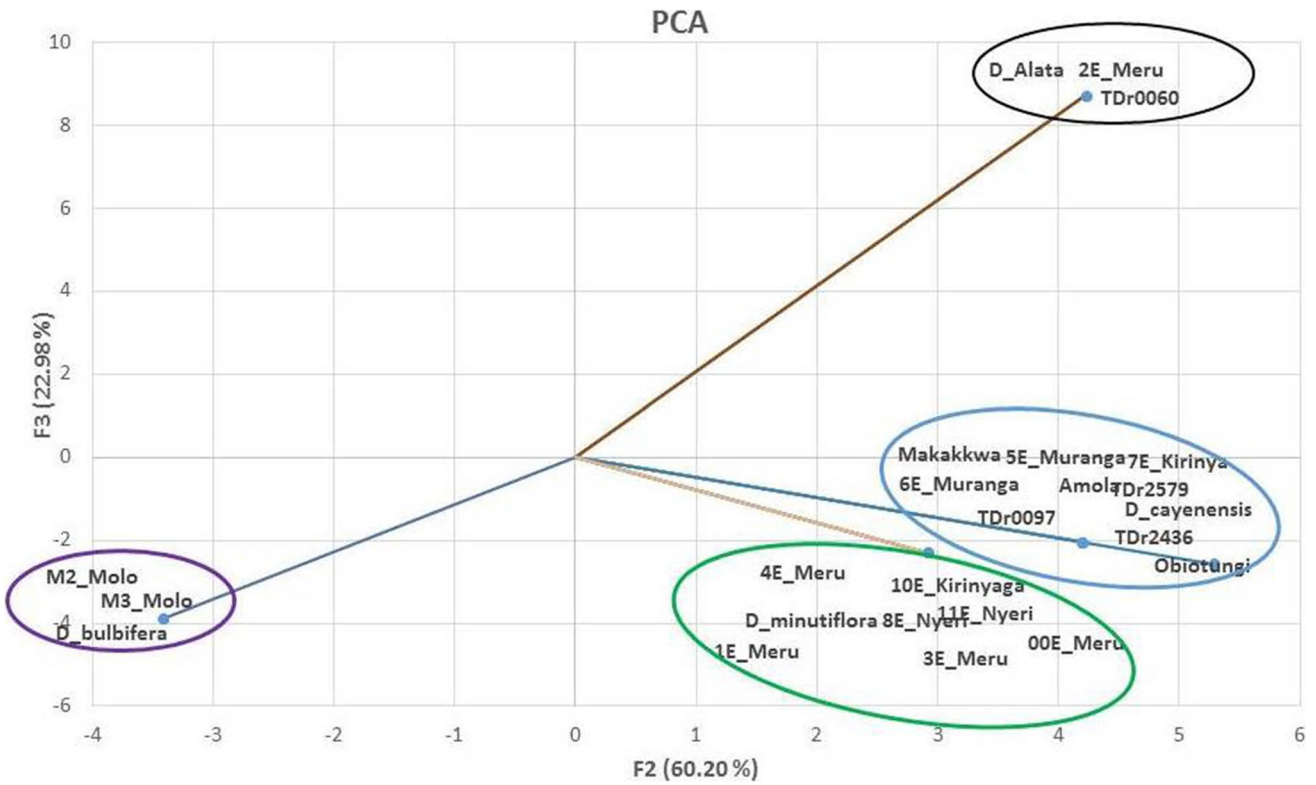


Fig. 6 Principal component analysis based on *rbcL* marker sequence analysis. Four separate groups represent the different yam species identified using *rbcL* marker. The grouping accounted for 83.20% of total variance in the data set. Colored ellipses represent the groups identified in the phylogenetic analysis

Discussion

Species identification, delimitation and molecular phylogeny of yam accessions cultivated in Kenya, was carried out using *rbcL* and *matK* barcoding markers. Both markers were able to discriminate the yam accessions into their respective species. *MatK* was able to discriminate yam species in the study with much less ambiguity making it a good barcoding marker candidate for inter-specific divergence resolution. It was able to resolve the identity of accession 7E_Kirinyaga identified as *D. minutiflora* by *rbcL* to *D. cayennensis.* This agrees with Carneiro et al. (2019) that *matK* proved to be the most efficient marker in plant authentication at species level.

Similarly, *rbcL* is also a good candidate for yam plant barcoding as it can universally amplify and produce quality sequences across taxa (CBOL Plant Working Group 2009). However, its discriminatory ability is fairly efficient because of its highly conserved sequences and relatively low inter-specific divergence. Hollingsworth et al. (2011) observed that *rbcL* marker is an efficient barcoding marker but lowly efficient in inter-specific divergence discrimination. This was contrary to Li et al. (2014) that *rbcL* and *matK* failed to discriminate *Calligonum* species. By combining the two (*rbcL*+*matK*) markers, barcoding and discrimination of yam species can be achieved. Girma et al. (2016) determined that combining *rbcL*+*matK* markers was 76.2% optimum for yam barcoding and that *matK* marker detected high interspecific variation and identified 63.2% of yam species. Sun et al. (2012) demonstrated that a combination of the two markers achieved a higher success rate of species discrimination than a single marker. Overall, the two markers were efficient with regard to discrimination of *Dioscorea* species.

The combination of *rbcL* and *matK* has been promoted as one of the promising universal two-region plant barcodes because it greatly improves species identification (CBOL Plant Working Group 2009). Both *rbcL* and *matK* markers were able to discriminate four *Dioscorea* species from farmers' fields in Kenya and two species from the National Genebank of Kenya (GBK). The identified yam species from farmers' fields included *D. minutiflora*, *D. cayennensis*, *D. alata* and *D. bulbifera.* The *Dioscorea* species from the GBK included *D. cayennensis*, *D. alata* and *D. minutiflora* was the most cultivated yam species by farmers in Kenya. The results agree with the available information which indicates that yam species grown in Kenya are *D. rotundata*, *D. minutiflora*, *D. bulbifera* and *D. dumetorum* (FAOSTAT 2009).

For *rbcL* marker the clades were supported by bootstrap values of 60, 71, 90 and 100 representing the yam species *D. minutiflora*, *D. cayennensis*, *D. alata* and *D. bulbifera*, respectively. The values support the reliability of the species

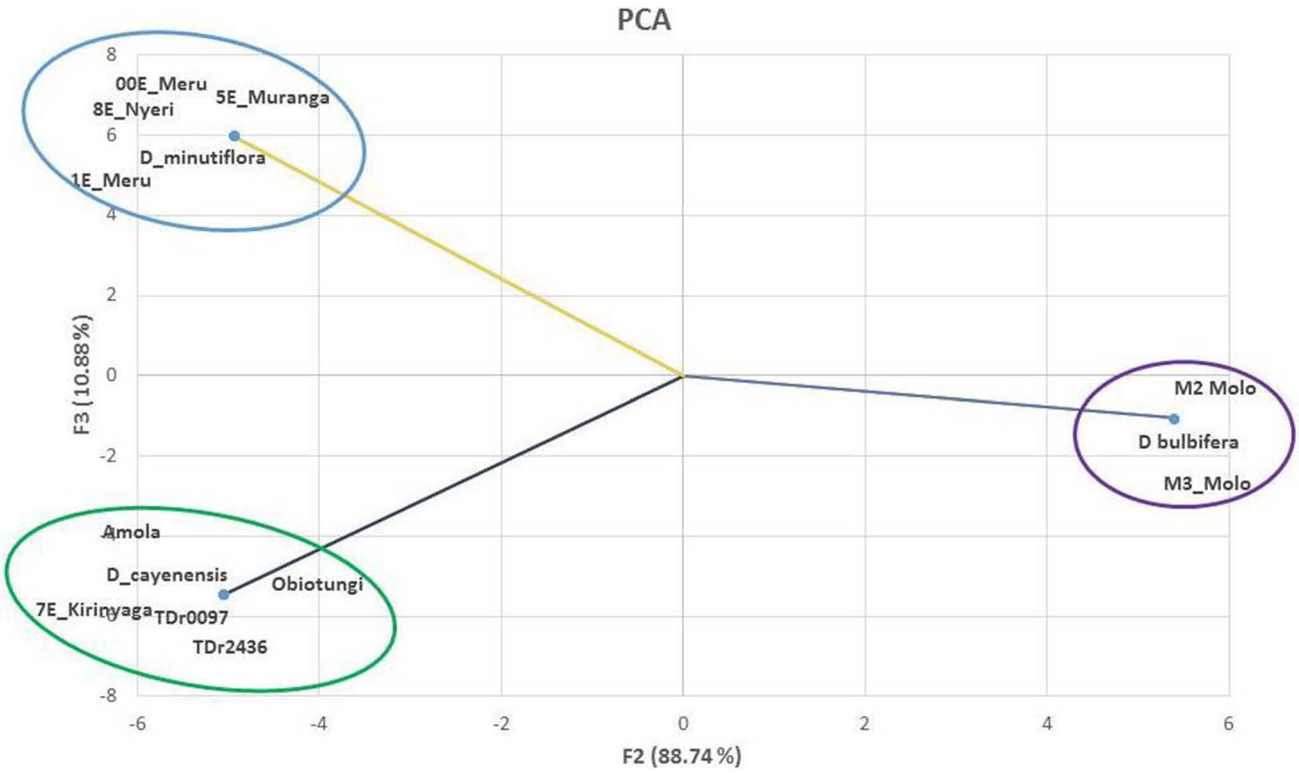


Fig. 7 Principal component analysis based on *matK* marker sequence analysis. Three separate groups represent the different yam species identified using *matK* marker. The grouping accounted for 99.62% of total variance in the data set. Colored ellipses represent the groups identified in the phylogenetic analysis

identification by *rbcL* marker. The accessions of the yam species in the respective clades can have a relationship which may be paraphyletic, polyphyletic and monophyletic. The *matK* marker had three clades that were strongly supported by bootstrap values of greater than 74. Specifically, *D. minutiflora*, *D. cayennensis* and *D. bulbifera* had bootstrap values of 74, 98 and 100, respectively. The bootstrap values strongly supported that clustering of the accessions to the respective clades which were representatives of the different yam species. The clades inform taxonomy which is important in the classification of plant species. For the biodiversity conservation of yam species, adequate information on their taxonomy is fundamental. In addition, breeding of crop plants relies on the genetic diversity of the species. Therefore, information on species taxonomy is useful for development of biodiversity resources and this requires distinctive identification of the plant species for their conservation (Ngo Ngwe et al. 2015).

Phylogeography establishes the congruence of geographical and phylogenetic relationships of species in order to explain the process that defines their populations' genetic diversity in space and time. The yam accessions were maximally differentiated into two distinct clades based on both *rbcL* and *matK* sequences. Clade I accessions based on *rbcL* gene were from Meru, Kirinyaga, Murang'a, Nyeri and GBK. Yam accessions in this clade belonged to *D. minutiflora*, *D. alata* and *D. cayannensis* and were obtained from farmers' fields while accessions from GBK included in this clade, belonged to *D. cayannensis* and *D. alata*. Yam accessions in Clade II based on *rbcL* sequence were from farmers' fields and GBK. The accessions in this clade belonged to *D. bulbifera* and *D. cayennensis* species.

For *matK* sequences, clade I had cultivars from Meru which belonged to *D. minutiflora*, while Clade II comprised of cultivars from all the sampled regions and belonged to *D. bulbifera*, *D. minutiflora* and *D. cayennensis*. Each of the species in this clade was a derivative of a sub-clade that fulfilled the phylogeography affirmation of congruence that exists between species geographic location and its phylogeny. It is evident from the results that the clade an accession belonged to, was informed by its geographic location and its species. The yam species clades originated from their time of divergence. It is important to note that phylogeography information can be used to predict the gene flow within populations in a given region. The results agree with Arnau et al. (2017) that there are two genepools for *D. alata* based on SSR markers that are divergent in India and Vanuatu which provides a clear genetic differentiation between yam species in Asia and South Pacific which both have secondary diversification.

Yams were independently domesticated in Asia, America and Africa in three different times with the species being

Dioscorea alata, *Dioscorea trifida* and *Dioscorea rotundata*, respectively (Scarcelli et al. 2019). However, the ennoblement process and domestication of *D. rotundata* has been characterized by hybridization and polyploidization. The presence of *D. cayennensis* cultivated in Kenya in a different clade but together with *D. alata* from GBK and originally obtained from Nigeria, confirmed that the species origin is Nigeria but through divergence, is domesticated in Kenya.

The domestication of *D. bulbifera* was mainly in the rift valley Kenya. The species is native to Asia but introduced to tropical Africa and mainly grows in areas with high humidity, temperatures of 25–35 °C and annual rainfall of above 1000 mm (Wunderlin et al. 2008). These environmental conditions are found in the Rift valley (Molo) region where the species is domesticated in Kenya. The variability of this accession is small as its monophyletic relationship was supported by a bootstrap value of > 98% similarity in both markers. Moreover, their genetic variation is close as indicated by their node age relation for both *rbcL* and *matK* markers. This is also affirmed by the presence of the accessions in the same clade for the two genes.

The presence of both *D. minutiflora* and *D. alata* in the same clade for *rbcL* marker demonstrated their closeness with regard to genetic diversity. *D. alata* origin is not well documented although it is believed to have originated from South East Asia in the sixteenth century and dispersed to other regions of the world including West Africa (Arnau et al. 2017). The differentiation of *D. alata* in both farmers' fields yam accessions and GBK (originally from Nigeria) indicated that they were related as affirmed by their existence in the same clade. This also implied that their genetic variation could be low.

Yam accessions for *D. minutiflora* were distributed in both clades in the two markers with their divergence time being close as indicated by node age relation for *rbcL* and *matK*, respectively. These accessions closely clustered with *D. alata* and *D. cayennensis* species implying that their differentiation time was near each other and therefore their genetic variation could be small for the accession involved. It also confirms that the *D. cayennensis* is a triploid product of hybridization of male *D. burkilliana* and female *D. rotundata* (Terauchi et al. 1992; Girma et al. 2014; Scarcelli et al. 2019; Sugihara et al. 2021). The origin of the cultivar could also be traced to tropical regions of Eastern Africa which is why it is a dominant yam accession in Kenya.

Polymorphism in the *matK* gene was slightly higher than that in *rbcL* gene, therefore, *rbcL* has more conserved sites thus serving as a good candidate for barcoding while *matK* was good for interspecific discrimination. This is related to the number of parsimony sites and the singletons and nucleotide diversity index which were all higher for *MatK* compared to *rbcL*. Haplotype diversity was slightly higher in the *rbcL* gene sequence compared to *matK* gene of the yam accessions. The slightly higher haplotype diversity in the *rbcL* sequence can be linked to the total number of haplotypes identified (seven) compared to four haplotypes determined by *matK* sequence. *RbcL* gene sequence of yam accessions had a lower nucleotide diversity and higher haplotype diversity while *matK* sequence had a higher nucleotide diversity and lower haplotype diversity. Mulualem et al. (2018) found that the genetic diversity of yam landraces in Ethiopia ranged from 0.00 to 0.80 with a polymorphism of 0.30 information content. Arnau et al. (2017) using SSRs demonstrated that the diversity among *D. alata* ranged from 0.20 to 0.86 which is close to the haplotype diversity of the two markers used in the current study.

The polymorphism among the yam accessions resulted in both synonymous and non-synonymous mutations. The synonymous mutations can be attributed to selection that occurred within the respective species population. Ude et al. (2019) noted that transversional mutation of G/T occurred at a consensus position of 335 then transitions at 362 (A/G), 368 (A/G), 371 (C/T) and 391 (C/T) within the yam accessions in Nigeria. The nature of the substitutions that occurred were both trans-versions and transition. Transition mutations can be attributed to selection while transversion mutation can be linked to genetic diversity (Lyons and Lauring 2017).

Tajima D result by *rbcL* sequences indicated that yam accessions were under sweep selection while *matK* gene sequences demonstrated that yam populations were under balanced selection. The Tajima D for *rbcL* was -0.75777 while that for *matK* was 0.08564. Positive Tajima value indicates a balanced selection where alleles with intermediate frequency thrive. It also denotes a population that is formed recently from two different populations which *Dioscorea* populations can exhibit. Negative Tajima value denotes purifying selection where excessive polymorphisms of low frequency occur. It also implies that population growth is being exhibited. Both markers have pointed out that the yam plant is under selection pressure and growth. There were 3 singleton sites in *rbcL* gene sequences and 2 singleton sites in *matK* sequence which corresponded to the nature of selection among the yam species. These findings concur with Akakpo et al. (2017) that, cultivated yam showed a skewed distribution to positive Tajima D value. Similarly, Wicke et al. (2011) observed that *rbcL* has a strong purifying selection in autotrophic plants which reduces its evolution rate and ability to distinguish related species. Conversely, *matK* is under relaxed purifying selection and has a high rate of nucleotide substitution therefore better ability in species discrimination (Duffy et al. 2009).

Principal component analysis demonstrated the clustering of yam accessions into their respective species. Yam accessions were distinctively identified into *D. alata*, *D. bulbifera*, *D. minutiflora* and *D. cayennensis* species. The grouping for both *rbcL* and *matK* markers accounted for more than 80.0% (99.62% for *matK* and 83.20% for *rbcL)* of the total variance which is useful in the species identification of the yam accessions. *D. alata* and *D. bulbifera* were distantly related to the

other yam species identified. *D. minutiflora* and *D. cayennensis* were closely related. The relationship among the species demonstrates their genetic variation which informs their origin and domestication in the respective geographic locations. Cao et al. (2021) observed that PCA generated similar results to that of phenotypic trait classification and thus it can be used to evaluate the genetic variation among species. Therefore, PCA analysis further provides information on the classification of yam accessions into different species.

Overall, the genetic diversity of *Dioscorea* species in Kenya is low as demonstrated by their genetic variations based on *rbcL* and *matK* markers. This can be attributed to low cross pollination and sexual recombination among the different yam species. The propagation method of yams in the field makes it susceptible to pathogens and thus prevents the formation and adaptation of new varieties. Pests attack and propagation methods contribute to genetic loss of yam species. Similarly, domestication of *Dioscorea* species is ambiguous because of its dioecism nature which is responsible for frequent hybridization and polyploidization (Ramu et al. 2017). Moreover, clonal propagation of the yam may reduce its genetic diversity, causes vulnerability to diseases and difficulty in removing deleterious mutations (Hebert et al. 2003). Other factors contributing to the low diversity include genetic loss/erosion caused by drought, gross destruction, changes in the natural habitat and neglect by farmers to continue growing the species (Bressan et al. 2014).

Conclusions

DNA barcoding markers, *matK* and *rbcL* have proven to be efficient for *Dioscorea* species discrimination and identification. The two markers were able to identify four species (*D. bulbifera*, *D. alata*, *D. minutiflora* and *D. cayennensis*) that are domesticated in Kenya. The cultivation of *D. minutiflora* is seemingly dominant in both central and upper eastern Kenya. The other three species are cultivated in the rift valley, eastern and central Kenya. The distribution of the species in Kenya is clustered into two main groups with regard to their phylogeography. Polymorphism in the yam species exhibits both synonymous and non-synonymous mutations. *RbcL* gene sequence of yam accessions had a lower nucleotide diversity and higher haplotype diversity while *matK* gene sequence recorded a higher nucleotide diversity and lower haplotype diversity. The DNA divergence between the yam species populations was similar based on the two markers and with no shared mutations. Therefore, DNA barcoding using combined *rbcL* and *MatK* can identify species but there is need to identify alternative DNA loci that help elucidate the identity and phylogeography profiles of *Dioscorea* species.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11756-022-01244-y.

Acknowledgements The authors thank the Centre for Biotechnology and Bioinformatics (CEBIB) and Department of Biochemistry, University of Nairobi for providing research facilities.

Authors' contributions EKM, DMWO and ENN designed the research. NK conducted experiments. NK analyzed data with the guidance of EKM, DMWO and ENN. NK drafted the manuscript and all authors reviewed and approved the manuscript.

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declarations

Research involving human and/or animals rights This research did not involve Human Participants and/or Animals.

Informed consent Not applicable as this research did not involve human participants or clinical trials.

Conflict of Interest The authors declare no competing interests that could have influenced the work reported in this paper.

References

Abdel-Latif A, Osman G (2017) Comparison of three genomic DNA extraction methods to obtain high DNA quality from maize. Plant Methods 13(1):1

Akakpo R, Scarcelli N, Dansi A, Djedatin G, Thuillet AC, Rhoné B, François O, Alix K, Vigouroux Y (2017) Molecular basis of African yam domestication: analyses of selection point to root development, starch biosynthesis, and photosynthesis related genes. BMC Genomics 18(1):1–9

Andres C, AdeOluwa OO, Bhullar GS (2017) Yam (Dioscorea spp.) - A rich staple crop neglected by research. Encyclopedia of Applied Plant Sciences, 2, 435-441). Academic Press, San Diego, USA, pp 435–441

Andriamparany JN, Brinkmann K, Jeannoda V, Buerkert A (2014) Effects of socio-economic household characteristics on traditional knowledge and usage of wild yams and medicinal plants in the Mahafaly region of south-western Madagascar. J Ethnobiol Ethnomedicine 10(1):1–21

Arnau G, Bhattacharjee R, Mn S, Chair H, Malapa R, Lebot V, Perrier X, Petro D, Penet L, Pavis C (2017) Understanding the genetic diversity and population structure of yam (Dioscorea alata L.) using microsatellite markers. PLoS One 12(3):e0174150

Barlagne C, Cornet D, Blazy JM, Diman JL, Ozier-Lafontaine H (2017) Consumers' preferences for fresh yam: a focus group study. Food Sci Nutr 5(1):54–66

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Proc Royal Soc B: Biol Sci 263(1377):1619–1626

Besnard G, Rubio De Casa R, Vargas P (2007) Plastid and nuclear DNA polymorphism reveals historical processes of isolation and reticulation in the olive tree complex (*Olea europaea*). J Biogeogr 34(4):736–752

Bressan EA, Briner Neto T, Zucchi MI, Rabello RJ, Veasey EA (2014) Genetic structure and diversity in the Dioscorea cayennensis/D. rotundata complex revealed by morphological and isozyme markers. Genet Mol Res 13(1):425–437

Bro R, Smilde AK (2014) Principal component analysis. Anal Methods 6(9):2812–31

Cao T, Sun J, Shan N, Chen X, Wang P, Zhu Q, Xiao Y, Zhang H, Zhou Q, Huang Y (2021) Uncovering the genetic

diversity of yams (Dioscorea spp.) in China by combining phenotypic trait and molecular marker analyses. Ecol and Evol 11(15):9970–9986
Carneiro de Melo Moura C, Moura C, Brambach F, Jair Hernandez Bado K, Krutovsky KV, Kreft H, Tjitrosoedirdjo SS, Siregar IZ, Gailing O (2019) Integrating DNA barcoding and traditional taxonomy for the identification of dipterocarps in remnant lowland forests of Sumatra. Plants 8:461
CBOL Plant Working Group (2009) A DNA barcode for land plants. PNAS USA 106:12794–12797
Cuénoud P, Savolainen V, Chatrou LW, Powell M, Grayer RJ, Chase MW (2002) Molecular phylogenetics of *Caryophyllales* based on nuclear 18S rDNA and plastid *rbc*L, *atp*B, and *mat*K DNA sequences. Am J Bot 89:132–144
Deschamps S, Llaca V, May GD (2012) Genotyping-by-Sequencing in Plants. Biol 1(3):460–483
Duffy AM, Kelchner SA, Wolf PG (2009) Conservation of selection on matK following an ancient loss of its flanking intron. Gene 438(1–2):17–25
Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797
Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. Heredity 103(4):285–298
FAOSTAT (2009) Food and Agricultural Organization. FAOSTAT-DATA. FAO, Rome. https://www.faostat.fao.org/. Accessed December 2021
FAOSTAT (2020) Food and Agriculture Organization of the United Nations Statistics database, FAOSTAT. Retrieved from https://www.fao.org/. Accessed February 2022
Fu RH, Kikuno H, Maruyama M (2011) Research on yam production, marketing and consumption of Nupe farmers of Niger State, central Nigeria. Afri J Agric Res 6(23):5301–5313
Girma G, Hyma KE, Asiedu R, Mitchell SE, Gedil M, Spillane C (2014) Next-generation sequencing based genotyping, cytometry and phenotyping for understanding diversity and evolution of guinea yams. Theor Appl Genet 127(8):1783–1794
Girma G, Spillane C, Gedil M (2016) DNA barcoding of the main cultivated yams and selected wild species in the genus *Dioscorea*. J Syst Evol 54(3):228–237
Guan Y (2014) Detecting structure of haplotypes and local ancestry. Genetics 196(3):625–642
Hebert PD, Cywinska A, Ball SL, Dewaard JR (2003) Biological identifications through DNA barcodes. Proc Royal Soc B: Biol Sci 270(1512):313–321
Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. PLoS ONE 6(5):e19254
Kartavtsev YP (2011) Divergence at Cyt-b and Co-1 mtDNA genes on different taxonomic levels and genetics of speciation in animals. Mitochondrial DNA 22(3):55–65
Lebot V (2009) Tropical root and tuber crops: Cassava, sweet potato, yams and aroids. CABI, Wallingford, CT, p 17
Li Y, Feng Y, Wang XY, Liu B, Lv GH (2014) Failure of DNA barcoding in discriminating *Calligonum* species. Nordic J Bot 32(4):511–517
Lyons DM, Lauring AS (2017) Evidence for the selective basis of transition-to-transversion substitution bias in two RNA viruses. Mol Biol Evol 34(12):3205–3215
Mignouna HD, Abang MM, Asiedu R (2008) Genomics of yams, a common source of food and medicine in the tropics. Genomics of Tropical Crop Plants. New York, NY: Springer. pp. 549–570.
Ministry of Agriculture, Livestock and Fisheries Kenya (2019) National Root and Tuber Crops Development Strategy 2019–2022.
Muiruri KS (2009) Molecular phylogeny of Kenyan Dioscorea species (Yams) and the quantification of their Dioscin Levels (Masters Dissertation, University of Nairobi)
Mulualem T, Mekbib F, Shimelis H, Gebre E, Amelework B (2018) Genetic diversity of yam (Dioscorea spp.) landrace collections from Ethiopia using simple sequence repeat markers. Aust J Crop Sci 12(8):1222–1230
Ngo Ngwe MFS, Omokolo DN, Joly S (2015) Evolution and phylogenetic diversity of yam species (Dioscorea spp.): Implication for conservation and agricultural practices. PLoS one 10(12):e0145364
Padhan B, Panda D (2020) Potential of neglected and underutilized Yams (Dioscorea spp.) for improving nutritional security and health benefits. Front Pharmacol 11:496
Patwardhan A, Ray S, Roy A (2014) Molecular markers in phylogenetic studies - A review. J Phylogenet Evol Biol 2:131
Qin A, Wang M, Cun Y, Yang F, Wang S, Ran J, Wang X (2013) Phylogeographic evidence for a link of species divergence of ephedra in the qinghai-tibetan plateau and adjacent regions to the miocene asian aridification. PLoS One8(2):e56243
Ramu P, Esuma W, Kawuki R, Rabbi IY, Egesi C, Bredeson JV, Bart RS, Verma J, Buckler ES, Lu F (2017) Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. Nature Genet 49(6):959–963
Scarcelli N, Cubry P, Akakpo R, Thuillet AC, Obidiegwu J, Baco MN, Otoo E, Sonké B, Dansi A, Djedatin G, Mariac C (2019) Yam genomics supports West Africa as a major cradle of crop domestication. Sci Adv. 5(5):eaaw1947
Sonibare MA, Asiedu R, Albach DC (2010) Genetic diversity of Dioscorea dumetorum (Kunth) Pax using amplified fragment length polymorphisms (AFLP) and cpDNA. Biochem Syst Ecol 38(3):320–334
Sugihara Y, Kudoh A, Oli MT, Takagi H, Natsume S, Shimizu M, Abe A, Asiedu R, Asfaw A, Adebola P, Terauchi R (2021) Population Genomics of Yams: Evolution and Domestication of Dioscorea Species. In: Population Genomics. Cham: Springer.
Sun XQ, Zhu YJ, Guo JL, Peng B, Bai MM, Hang YY (2012) DNA barcoding the *Dioscorea* in China, a vital group in the evolution of monocotyledon: use of mat K gene for species discrimination. PLoS ONE 7(2):e32057
Tamiru M, Natsume S, Takagi H, White B, Yaegashi H, Shimizu M, Urasaki N (2017) Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination. BMC Biol 15(1):86
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30(12):2725–2729
Terauchi R, Chikaleke VA, Thottappilly G, Hahn SK (1992) Origin and phylogeny of Guinea yams as revealed by RFLP analysis of chloroplast DNA and nuclear ribosomal DNA. Theor Appl Genet 83(6–7):743–751
Ude GN, Igwe DO, McCormick J, Ozokonkwo O, Harper J, Ballah D, Aninweze C, Obih C, Okoro M, Ene C, Chiezey VO (2019) Genetic Diversity and DNA Barcoding of Yam Accessions from Southern Nigeria. Am J Plant Sci 10:179–207
Verter N, Bečvářová V (2015) An analysis of yam production in Nigeria. Acta Univ Agric Et Silvic Mendelianae Brun 63(2):659–665
Wicke S, Schneeweiss GM, Depamphilis CW, Muller KF, Quandt D (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol 76:273–297
Wunderlin RP, Hansen BF, Hansen BF (2008) Atlas of Florida Vascular Plants (https://www.plantatlas.usf.edu/).[SM Landry and KN Campbell (application development), Florida Center for Community Design and Research.]. ISB. USF. Tampa. Accessed August 2021