PLOS ONE

RESEARCH ARTICLE

SilicoDArT and SNP markers for genetic diversity and population structure analysis of *Trema orientalis*; a fodder species

Judith Ssali Nantongo*, Juventine Boaz Odoi, Hillary Agaba, Samson Gwali

National Forestry Resources Research Institute, Kifu, Mukono

* jsnantongo@yahoo.com

Citation: Nantongo JS, Odoi JB, Agaba H, Gwali S (2022) SilicoDArT and SNP markers for genetic diversity and population structure analysis of *Trema orientalis*; a fodder species. PLoS ONE 17(8): e0267464. https://doi.org/10.1371/journal.pone.0267464

Editor: Tzen-Yuh Chiang, National Cheng Kung University, TAIWAN

Received: June 14, 2021

Accepted: April 10, 2022

Published: August 22, 2022

Data Availability Statement: Relevant data are within the paper and its Supporting Information files. A full set of SNPs used to generate the values that have been presented in the paper were deposited at DOI: 10.6084/m9.figshare.19181729 (L172-173).

Funding: Funds were received through the National Agricultural Research Organisation/National Forestry Resources research institute Agricultural Technology and Agribusiness Advisory Services (ATAAS) Project. The funders had no role in study

Abstract

Establishing the genetic diversity and population structure of a species can guide the selection of appropriate conservation and sustainable utilization strategies. Next-generation sequencing (NGS) approaches are increasingly being used to generate multi-locus data for genetic structure determination. This study presents the genetic structure of a fodder species -*Trema orientalis* based on two genome-wide high-throughput diversity array technology (DArT) markers; silicoDArT and single nucleotide polymorphisms (SNPs). Genotyping of 119 individuals generated 40,650 silicoDArT and 4767 SNP markers. Both marker types had a high average scoring reproducibility (>99%). Genetic relationships explored by principal coordinates analysis (PCoA) showed that the first principal coordinate axis explained most of the variation in both the SilicoDArT (34.2%) and SNP (89.6%) marker data. The average polymorphic information content did not highly differ between silicoDArT (0.22) and SNPs (0.17) suggesting minimal differences in informativeness in the two groups of markers. The, mean observed (H_o) and expected (H_e) heterozygosity were low and differed between the silicoDArT and SNPs respectively, estimated at $H_o = 0.08$ and $H_e = 0.05$ for silicoDArT and $H_o = 0.23$ and $H_e = 0.19$ for SNPs. The population of *T. orientalis* was moderately differentiated (F_{ST} = 0.20–0.53) and formed 2 distinct clusters based on maximum likelihood and principal coordinates analysis. Analysis of molecular variance revealed that clusters contributed more to the variation (46.3–60.8%) than individuals (32.9–31.2%). Overall, the results suggest a high relatedness of the individuals sampled and a threatened genetic potential of *T. orientalis* in the wild. Therefore, genetic management activities such as ex-situ germplasm management are required for the sustainability of the species. Ex-situ conservation efforts should involve core collection of individuals from different populations to capture efficient diversity. This study demonstrates the importance of silicoDArT and SNP makers in population structure and genetic diversity analysis of *Trema orientalis*, useful for future genome wide studies in the species.

Introduction

Indigenous or naturalised fodder trees and shrubs are important feed sources for livestock in a wide range of farming systems in East Africa, where over 200 000 smallholder farmers plant

design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

fodder trees [1]. In Uganda, a wide range of species are used for fodder, and have been selected based on their palatability, medicinal values and coppicing ability [2]. *Calliandra calothyrsus*, *Leucaena trichandra or Gliricidia sepium* have been the most promoted fodder species [3]. While these species have provided a basis for increased tree fodder use, promoting alternative fodder trees to supplement the current livestock feeding strategies of smallholders in mixed farming systems is key to resilience. *Trema orientalis* is a potential multi-purpose fodder. However, lack of suitable seed is still a major challenge in most fodder promotion efforts [1]. For *T. orientalis*, seeds are collected from the wild, where populations have dwindled, in part due to degradation of natural habitats. Herbivory may also be important in determining distribution of pioneer species such as *T. orientalis* [4]. This likely affects the effective sizes with consequent stochastic changes in the genetic integrity of the seeds of this promising fodder species in the wild [5,6]. Establishing the genetic structure of *T. orientalis* can help to establish appropriate conservation, management, and sustainable utilization strategies [7–10].

Molecular markers have become valuable tools for quantifying genetic diversity, spatial genetic structure, mating systems, gene flow and breeding patterns of tree species and many wild and cultivated plants [11]. From restriction fragment length polymorphisms (RFLPs) to simple sequence repeats (SSRs) and then to next generation sequencing of single nucleotide polymorphisms (SNPs), the types of molecular markers used to characterise genetic diversity have evolved over the past several decades [12]. However, SNPs are becoming the choice marker for genetic analysis and breeding because of the large number of markers that can be generated at a reduced cost. SNPs are also the most frequent source of variation in eukaryotic genomes and their bi-allelic nature offers accuracy in variant calling [13]. In contrast to whole genome sequencing techniques, the recent genotyping-by-sequencing (GBS) techniques such as Diversity Array Technology (DArT) (http://www.diversityarrays.com/) enables simultaneous SNP discovery and sequencing from a targeted subset of the whole-genome. The more recent DArT sequencing (DArTseq) further reduces genome representation by sequencing only the most informative representations of genomic DNA, which improves the rate of genotype calling and the ability to sequence more samples for less cost [14]. DArTseq produces dominant (SilicoDArT) and co-dominant (SNP) markers that have been successfully applied for genetic structure analysis in several crops [15,16]. The markers especially allow the characterisation of population structure without prior knowledge of the genome or diversity [17,18].

Trema orientalis has very few genomic resources that can contribute to its improvement and domestication. Notably, the genome has been sequenced [19], providing valuable genetic information for accurately identifying the species, clarifying taxonomy and reconstructing the intergeneric phylogeny of Cannabaceae [19]. However, knowledge of the intraspecific genetic structure of *T. orientalis*.is required for its management. Therefore, we used high-throughput genotyping-by-sequencing (GBS) genotyping using the DArTseq platform to assess intraspecific genome-wide diversity and population structure of *T. orientalis*. The objectives of this study were: 1) to assess genetic diversity in *T. orientalis* using SilicoDArT and SNP markers; 2) to investigate fine-scale population structure of *T. orientalis*. This study lays a foundation for future genome-wide association studies or genomic selection in *T. orientalis*.

Materials and methods

Study species and sample collection

Trema orientalis also known as *Celtis orientalis* Linn., *Celtis guineensis* Schum. and Thonn., *Trema bracteolate* Hochst Blume, *Sponia orientalis* Linn. Decne, and *Trema guineensis* (Schum. and Thonn.) Ficalho is a species of flowering tree in the hemp family, Cannabaceae [20]. It is a shrub or small to medium size tree that can grow up to 18 m high in forest regions,

and up to 1.5 m tall in the savannah. The flowers are small, inconspicuous, and greenish, carried in short dense bunches. They are usually unisexual, i.e. male and female are separate, and occasionally bisexual. Flowers appear irregularly from late February to April, being pollinated by bees or wind [21]. Besides its use for fodder in Uganda as well as other African and Asian countries, the tree is useful for various wood and non-wood products [22]. *T. orientalis* was selected in Uganda through the National Forestry Resources Research Institute (NaFORRI), as a potential forage and therefore of interest for conservation and management.

In Uganda, *T. orientalis* occurs in forest fallows especially in the Central, Eastern and Western part of the country [23]. However, most forest reserves where it occurs have been degraded, which threatens the species [24], and designing conservation strategies for priority species has been identified as a key intervention. Therefore, we characterised the genetic structure of this species, to guide in-situ conservation as well as germplasm collection for ex-situ conservation. From West Bugwe Forest reserve and the surrounding woodlands (Fig 1), 119 leaf samples were randomly collected from mature trees for DNA extraction. Upon collection, the leaves were immediately preserved with silica gel.

DNA extraction

The leaf samples with silica gel were sent to Biosciences Eastern and Central Africa (BecA-ILRI) hub in Nairobi for DNA extraction. DNA extraction was done using Nucleomag plant genomic DNA extraction kit (Macherey-Nagel). The genomic DNA extracted was in the range of 50–100 ng/ul. DNA quality was checked on 0.8% agarose gel.

DArTseq genotyping

DNA was shipped to Diversity Arrays Technology Pty Ltd laboratories in Canberra, Australia for processing using the DArTseq™ platform using protocol optimised for *T. orientalis*. DNA samples were processed in digestion/ligation reactions using a combination of PstI and HpaII Restriction Enzymes (RE) [14] with modifications, where a single PstI-compatible adaptor was replaced with two different adaptors corresponding to two different RE overhangs. The PstI-

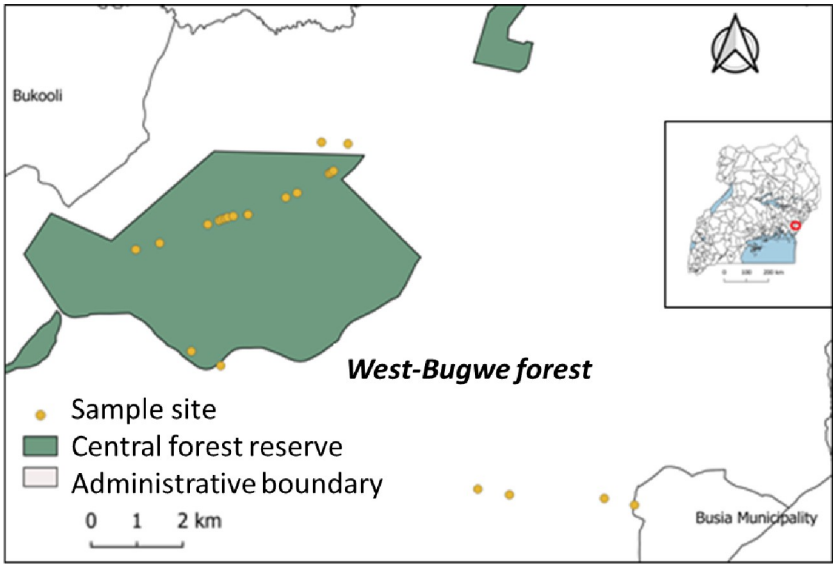

Fig 1. Map of Uganda (inset) and the enhanced site showing the location of West Bugwe forest and the surrounding woodlands (red mark on map) where leaf samples of *T. orientalis* were collected.

https://doi.org/10.1371/journal.pone.0267464.g001

compatible adapter was designed to include Illumina flowcell attachment sequence, sequencing primer sequence and "staggered", varying length barcode region, similar to the sequence that has been previously reported [17]. Reverse adapter contained flowcell attachment region and HpaII-compatible overhang sequence.

Only "mixed fragments" (PstI-HpaII) were effectively amplified in 30 rounds of polymerase chain reaction (PCR) using the following reaction conditions: 94˚C for 1 min, 30 cycles of; 94˚C for 20 sec, 58˚C for 30 sec, 72˚C for 45 sec, followed by a final hold of 72˚C for 7 min.

After PCR, equimolar amounts of amplification products from each sample of the 96-well microtiter plate were bulked and applied to c-Bot (Illumina) bridge PCR followed by sequencing on Illumina Hiseq2500. The sequencing (single read) was run for 77 cycles. Sequences generated from each lane were processed using proprietary DArT analytical pipelines. In the primary pipeline the poor-quality sequences were filtered away. The pipeline applied more stringent selection criteria to the barcode region compared to the rest of the sequence. In that way the assignments of the sequences to specific samples carried in the "barcode split" step were very reliable. Filtering was performed on the raw sequences using the following parameters:

Filter	Filter Parameters
Barcode region	Min Phred pass score 30, Min pass percentage 75
Whole read	Min Phred pass score 10, Min pass percentage 50

https://doi.org/10.1371/journal.pone.0267464.t001

Approximately 2,500,000 sequences per sample were used in marker calling. Single nucleotide polymorphisms were identified by aligning reads to create clusters across all individuals sequenced. As *T*. *orientalis* is a nonmodel species, reference alleles and SNP alleles for each locus were assigned arbitrarily—in most cases, reference alleles were indicated as the allele that was most frequent across all samples for that locus. SNP markers were aligned to the reference genomes in the accessions of Chickpea_ICC_v2 and Grape_v8 of the National Centre for Biotechnology Information (NCBI) in order to identify chromosome positions. The SNPs were also aligned to several bacteria genomes to identify bacterial contamination. The BLASTN algorithm with an e-value $\leq$ 5e-7 and percentage identity of 90% was used. SilicoDArTs and SNPs were scored as "dominant" markers, with "1" = Presence and "0" = Absence of a restriction fragment with the marker sequence in genomic representation of the sample. SNPs were scored as codominant markers with 0 for the homozygous allele aa, 1 for the heterozygous allele Aa and 2 for the homozygous allele AA.Finally, identical sequences were collapsed into "fastqcoll files". The fastqcoll files were "groomed" using DArT PL's proprietary algorithm which corrects low quality base from singleton tags into a correct base using collapsed tags with multiple members as a template. The "groomed" fastqcoll files were used in the DArTs proprietary SNP and presence/absence variation (SilicoDArT) calling pipeline, DArTsoft14. For SNP calling all tags from all libraries included in the DArTsoft14 analysis are clustered using DArT PL's C++ algorithm at the threshold sequence distance of 3 base pairs, followed by parsing of the clusters into separate SNP loci using a range of technical parameters, especially the balance of read counts for the allelic pairs. In addition, multiple samples were processed as technical replicates (from DNA to allelic calls) and scoring consistency was used as the main selection criteria for high quality/low error rate markers.

Quality analysis of marker data

The markers were tested for reproducibility (%)–the proportion of technical replicate assay pairs for which the marker score exhibited consistency; call rate (%)–the success of reading the

marker sequence across the sample; polymorphism information content (PIC)—the degree of diversity of the marker in the population and the usefulness of the marker for linkage analysis; and one ratio–the proportion of the samples for which genotype scores equalled '1'. The Spearman correlation between the Euclidean distances of the matrices of DArTseq and SNP markers was determined using the Mantel test in R. The raw SNP data were deposited at doi: 10.6084/m9.figshare.19181729.

Data filtering process

The data was filtered using the dartR v 1.9.9.1 package [25] in R to remove all SNPs and silicoDArT markers that had > 5% missing data and individuals with > 10% missing data. Markers with a reproducibility score (RepAvg) < 100% were also removed as well as those that originated from the same fragment. Non-informative monomorphic markers were also removed. SNPs with a minor allele frequency (MAF) of < 1% were also discarded MAF filtration was not done for presence/absence silicoDArT. The markers were further filtered based on the one ratio value, where markers with extremely low one ratio (<0.05) were not included in the analysis.

To elaborate the genetic structure of the populations, a model-based Bayesian clustering was conducted using STRUCTURE 2.3.4 software. STRUCTURE uses a hierarchical Bayesian model to identify subpopulations and estimate global ancestry for each sampled individual based on allele frequency data [26]. The analysis was run separately for silicoDArT and SNPs. Numbers in the range from 1 to 10 were assumed for K. The initial burn-in period, for each run, was set to 100,000 with 100,000 MCMC (Markov chain Monte Carlo) iterations [27]. The admixture model was applied without using any prior population information. To find the suitable value of K, the number of clusters (K) was tested in the range from 1 to 10, and were then plotted against ΔK in STRUCTURE HARVESTER [28] to identify the most likely value of K.

Using dartR, principal coordinate analyses (PCoA) was used to investigate genetic relationships among individuals. PCoA was performed separately on the SilicoDArT and SNP datasets. To further explore the genetic relationships of *T*. *orientalis* individuals evaluated in this study, a maximum likelihood dendrogram was constructed in MEGA X using SNP markers with no prior population assumptions [29]. Using MEGA X, maximum likelihood fits of 24 different nucleotide substitution models to estimate substitution rates were developed.

Genetic diversity analyses

Using selected markers, all genetic diversity indices were estimated using the R package "ADEGENET" [30]. The R package ADEGENET uses discriminant analysis of principal components to allow for data dimensionality reduction in large genomic datasets. The following diversity indices were therefore computed to illustrate the overall genetic divergence among the subpopulations: observed (H_o) and expected heterozygosity (H_e), total gene diversity (H_t), genetic differentiation (F_{st}) and population inbreeding coefficient (F_{is}), fixation index (F_{st}). Marker allele frequency–the frequency at which the second most common allele occurs in a given population [31], was also computed as the number of minor alleles in the population/total number of alleles in the population. Analysis of molecular variance was done using hierfstat package in R [32].

Sequence similarity search

To put the study sequences in the context of other published sequences, 100 sequences of SNPs were randomly selected at different nodes and their similarity with published sequences

searched in the NCBI database using BLASTN algorithm. A minimum e-value of $1e^{-5}$ and >80% identity, query coverage as well as total score were considered. Another dendrogram of *T. orientalis* and selected sequences from other species was generated using MEGA X [29].

Results

T. orientalis silicoDArT and SNP detection

A total of 4767 SNPs and 40,650 and silicoDArT markers were generated from 119 individuals of *T. orientalis*. The call rate of the silicoDArT markers varied between 72–100%, with an average of 98%. Missing values ranged from 5 to 10% for individual trees, and 0 to 33% for the markers. Reproducibility of the silicoDArT markers averaged to 99% (range 91% - 100%). For SNPs, missing values ranged from 0 to 50% for individual trees, and 0 to 42% for the markers. The call rate ranged from 35 to 100% with an average of 90%. The reproducibility of markers ranged from 90% to 100% with an average of 99%. The quality of marker calling was further verified by the ratio of transitions (Ts; i.e. A/G or T/C substitutions) versus transversions (Tv; i.e. A/T, A/C, T/G or C/G substitutions) which approximated to 0.5 (for both SNPs and silicoDArTs) in most of the 24 different nucleotide substitution models (S1 Table).

Genetic diversity and Polymorphism Information Content (PIC)

Overall, silicoDArT markers retained, the PIC value ranged from 0.02–0.5 (average = 0.22). However, there was 29% of the PIC values between 0.1–0.5 (Fig 2). The polymorphic information content (PIC) of SNPs ranged from 0 to 0.49 (average = 0.17), with 84% ranging between 0.1–0.5.

The mean minor allele frequency (MAF) based on SNPs ranged between 0.004–0.5 with an average of 0.16. Only 5% of the SNP markers had minor allele frequency less than 0.05 indicating that most markers were common genetic variants. MAF was not estimated for the dominant silicoDArT markers. After the filtration criteria above, 117 individuals were retained and 2061 SNP markers, while all individuals and 18, 163 silicoDArT markers were retained. These were used for the proceeding analyses.

The genetic diversity values calculated as expected heterozygosity (He) in the population varied from 0.05 for silicoDArTs and 0.27 for SNPs (Table 1). The low mean observed (H_o) and expected (H_e) heterozygosity (Table 1) corroborates with the low PIC values above.

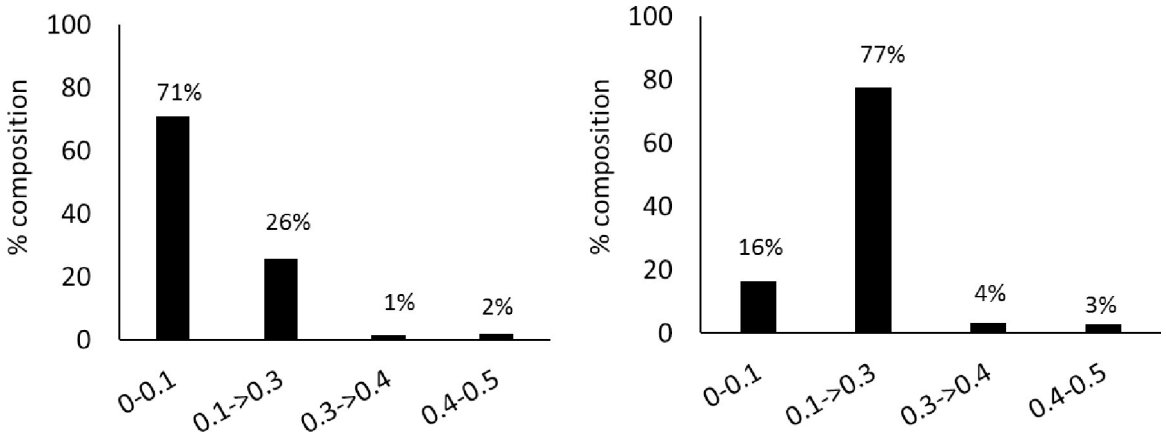

Fig 2. The polymorphic information content of the a) silicoDArT and b) SNP markers.

https://doi.org/10.1371/journal.pone.0267464.g002

Table 1. Genetic diversity of *T. orientalis* based on silicoDArT and SNP markers. Estimates with p indicate that these are corrected e.g. corrected Fst = Fstp.

	silicoDArT	SNPs
H_o	0.08	0.23
H_e	0.05	0.19
H_t	0.06	0.40
H_{tp}	0.08	0.61
D_{st}	0.01	0.21
D_{stp}	0.03	0.43
F_{st}	0.20	0.53
F_{stp}	0.33	0.70
F_{is}	-0.51	-0.23
D_{est}	0.03	0.52

https://doi.org/10.1371/journal.pone.0267464.t002

Population structure analysis

Genetic relationships among the *T. orientalis* individuals were assessed using a model-based clustering method that infers population structure using genotype data consisting of unlinked markers. Results from silicoDArT markers revealed 2 clusters (K = 2) (Figs 3 and S1), where cluster I consisted of more individuals than cluster II (Table 2). Therefore, the STRUCTURE results at K = 2 were subject to population genetics analyses. Similarly, SNPs clustering revealed that there were more individuals in cluster 1 than in cluster 2. Similar clustering was also visible in the dendrogram that identified two major clusters based on SNP markers (S2 Fig).

Genetic relationships among individuals were further explored by principal coordinates analysis (PCoA) (Fig 4). Using silicoDArT and SNP markers, PCoA identified two subpopulations, revealing the influence of tree location on the genetic diversity within *T. orientalis*. The first principal coordinate axis explained a higher proportion of variation (34.2% and 89.6%) than the second principal coordinate axis (18.3% and 2.9%) for both silicoDArT and SNPs (Fig 4a & 4b). For the SNP data, the clustering was tighter, and clusters had less overlap than the silicoDArT markers.

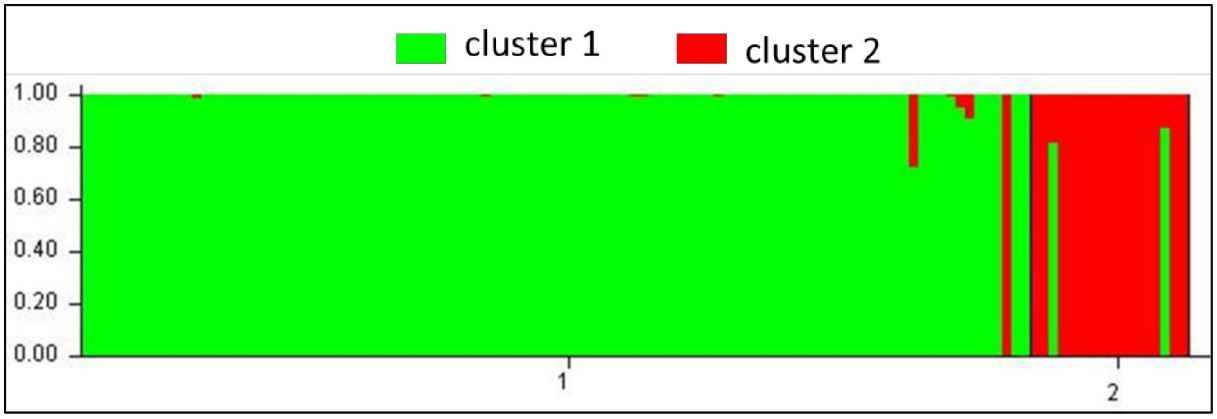

Fig 3. Number of clusters of the *T. orientalis* population using silicoDArT marker data estimated using the model-based Bayesian algorithm implemented in the STRUCTURE program. A similar graph was obtained for the SNP markers (graph not shown).

https://doi.org/10.1371/journal.pone.0267464.g003

Table 2. Genetic divergence among (net nucleotide distance) and within (expected heterozygosity) populations, and the proportion of membership of the population samples based on silicoDArT and SNP markers.

	silicoDArT		SNPs	
	Group 1	**Group 2**	**Group 1**	**Group 2**
Proportion of membership	83.7	16.3	86.2	13.8
Nucleotide distance	0.44		0.53	
Expected heterozygosity	0.09	0.20	0.06	0.26
Genetic differentiation (F_{ST})	0.76	0.55	0.96	0.55

https://doi.org/10.1371/journal.pone.0267464.t003

Genetic differentiation of *T. orientalis*

Based on the two clusters identified in STRUCTURE, the silicoDArT markers also showed lower estimates of total genetic diversity (Ht) and genetic diversity (Dst) among groups/populations (H_t = 0.06, D_{st} = 0.01) compared with SNP markers (H_t = 0.40, D_{st} = 0.21) (Table 1). The estimates for genetic differentiation (F_{st}) were also lower with silicoDArT markers (Fst = 0.20) compared to SNPs (Fst = 0.53) (Table 1). The low PIC values observed above and differences between H_o and H_e was consistent with the moderate inbreeding coefficient (F_{is}), where F_{is} = -0.51[silicoDArT] and -023 [SNPs].

Overall, results indicated the presence of higher variation (AMOVA results) contained between clusters inferred using silicoDArTs (46.3%) and SNPs (60.8%) than individuals. Variation among individuals was 32.9% and 31.2% based on silicoDArTs and SNPs respectively. The consistency of these results is also reflected in the Mantel test that revealed strong association ($r = 0.61$; $P < 0.0001$) between both markers.

Sequence similarity

To put the resulting SNPs in the context of other sequences produced using other sequencing methods, the length of the short sequence reads corresponding with SilicoDArT markers ranged from 20 to 69 nucleotides (nt), with an average of 55.2nt and for SNPs the range was 22–69 (average 64.6 nt).

Blasting the 100 sequences selected over the branches of the dendrogram, 52 SNPs could not match any other sequence, while 15 SNPs matched *Cannabis sativum* (Cannabaceae)

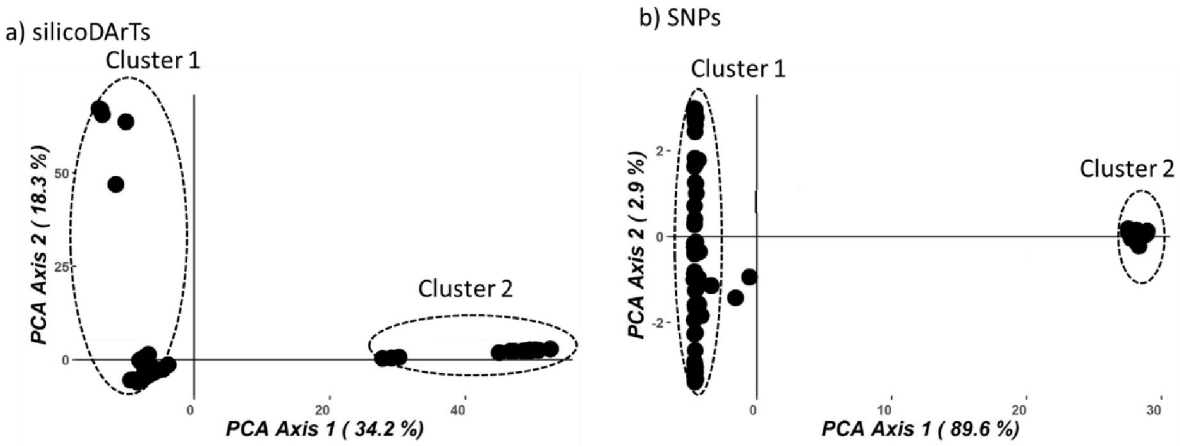

Fig 4. Principal coordinates analysis plot to infer group structure of *T. orientalis* based on a) silicoDArT b) SNP markers. Axis explained respectively 34.2% and 89.6% of the total variation in the samples based on respectively silicoDArT and SNP markers.

https://doi.org/10.1371/journal.pone.0267464.g004

sequences, 9 sequences matched *Morus notabilis* (Moraceae) while the rest were more similar to sequences *T. orientale*, *Prunus dulcis*, *Juglans regia*, *Ziziphus jujuba*, *Fragaria vesca*, *Corylus avellana*, *Vigna radiata*, *Quercus lobata*, *Populus eupratica*, *Pistacia vera*, *Chenopodium quinoa and Nymphaea colorata*. The genetic relationship among the sequences of *T. orientalis* and the above species is illustrated in Fig 5. The close relationship of the SNPs in this study with close members in the same lineages suggests that the identified silicoDArT and SNP markers were of high quality.

Discussion

The importance of understanding the genetic diversity of fodder species is critical for conservation and utilization of their germplasm in breeding programs. While most studies that have used the DArT platform have mainly worked with cultivated species [16,33,34], our study highlights the suitability of DArT platform for the genomic dissection of a variety of wild plant species. Given that the average cost per data point of silicoDArT is less than SNP markers [35], the DArT platform provides opportunities for genetic-based management of diverse species in less developed countries. The DArT system enabled the detection of two types of markers, the SNPs and silicoDArT markers which; (i) exhibited high call rates and reproducibility, (ii) showed reduced genetic diversity (iii) exhibited strong genetic differentiation; and (iv) were consistent with other published sequences of taxa related to *T. orientalis*. Such high call rate and reproducibility has been recorded for DArT technologies in different plant species [27,36] indicating the reliability of the DArT methods for genotyping several plant species.

The results from the silicoDArT and SNP markers indicated low genetic variation in *T. orientalis* with potential consequences on the species ability to recover from demographic, environmental and genetic stochasticity [10]. Genetic variation in populations is measured in several ways, the most common of which has traditionally been the proportion of polymorphic loci and patterns of observed and expected heterozygosity. The polymorphism information content (PIC) values range from 0 to 0.5, where the following classification on the informativeness based on PIC values has been derived: low (0 to 0.10), medium (0.10 to 0.25), high (0.30 to 0.40) and very high (0.40 to 0.50) [37,38]. The results from the study showed that both silicoDArTs and SNPs exhibited medium to high informativeness (average PIC = 0.17–0.22) suggesting that they can detect the polymorphism among the individuals of *T. orientalis*. The PIC values were in the range of those established for other trees like Macadamia, where PIC for silicoDArT and SNP markers were 0.29 and 0.21 respectively, although the distribution was different [27]. The PIC values were however mostly lower than what has been detected in food crops such as beans, chickpeas, cassava and wheat [33,39–41] possibly signifying inherently low PIC values associated these markers in trees.

The average observed heterozygosity H_o for the markers was low but was in range of what has been reported in other tropical forest trees the same region [42,43] which could be due to anthropogenic disturbances in most natural vegetation that potentially erode the genetic diversity. However, contrary to these studies [42,43] that indicated $H_o < H_e$, which is normally indicative of inbreeding, our study showed $H_o > H_e$, for both SNP and silicoDArT markers. This suggests presence of an isolate-breaking effect (the mixing of two previously isolated populations or presence of hybrids) [44], consistent with the negative inbreeding coefficient that was observed for both markers, which points to presence of excessive heterozygotes. However, other hypotheses for presence of negative breeding coefficients have been highlighted [45]; including a lack of selfed progeny in small populations of outcrossing species, negative assortative mating when reproduction occurs between individuals bearing phenotypes more dissimilar than by chance and selection during the life cycle of the most heterozygous individuals.

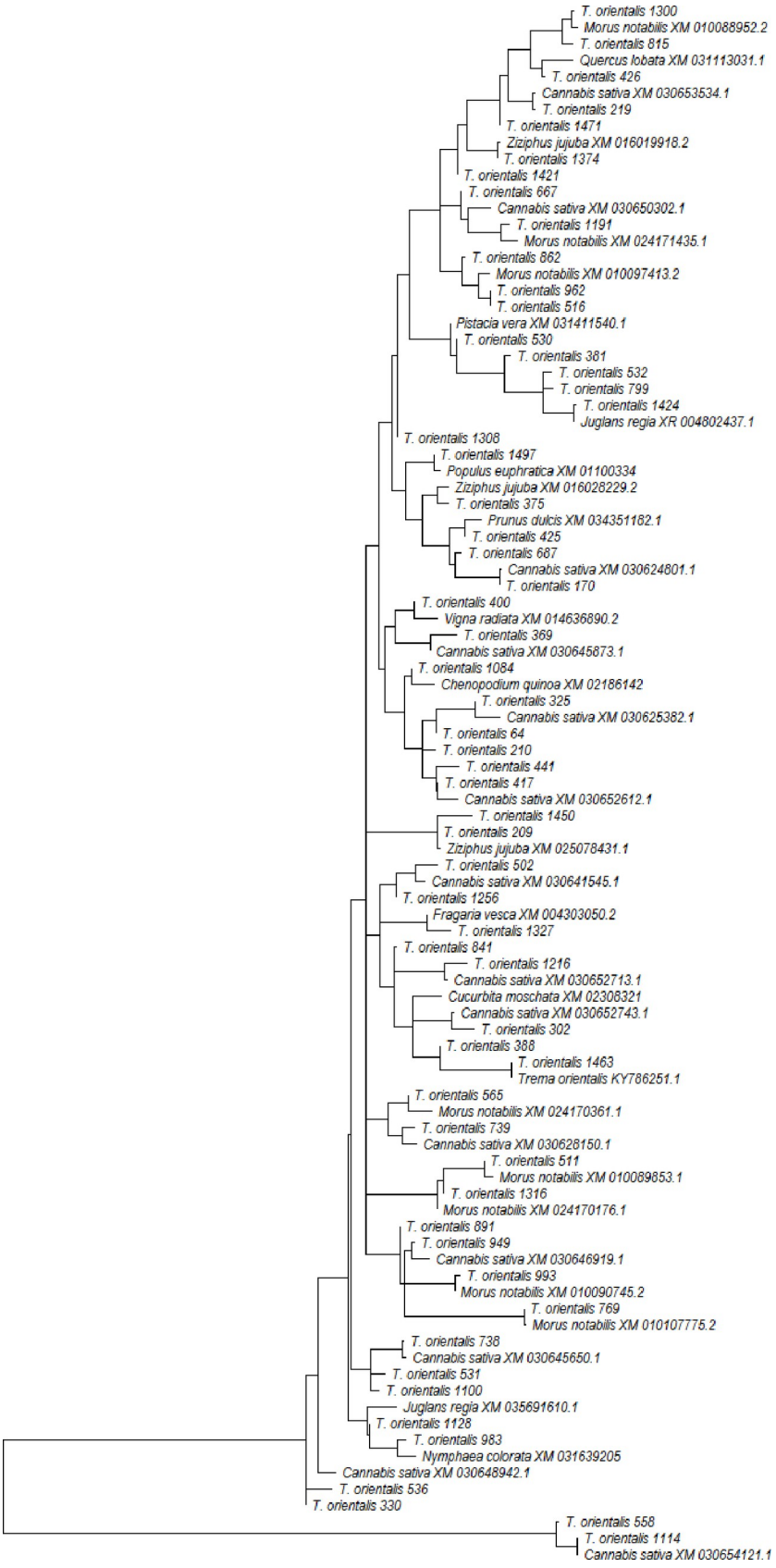
T. orientalis 1300
Morus notabilis XM 010088952.2
T. orientalis 815
Quercus lobata XM 031113031.1
T. orientalis 426
Cannabis sativa XM 030653534.1
T. orientalis 219
T. orientalis 1471
Ziziphus jujuba XM 016019918.2
T. orientalis 1374
T. orientalis 1421
T. orientalis 667
Cannabis sativa XM 030650302.1
T. orientalis 1191
Morus notabilis XM 024171435.1
T. orientalis 862
Morus notabilis XM 010097413.2
T. orientalis 962
T. orientalis 516
Pistacia vera XM 031411540.1
T. orientalis 530
T. orientalis 381
T. orientalis 532
T. orientalis 799
T. orientalis 1424
Juglans regia XR 004802437.1
T. orientalis 1308
T. orientalis 1497
Populus euphratica XM 01100334
Ziziphus jujuba XM 016028229.2
T. orientalis 375
Prunus dulcis XM 034351182.1
T. orientalis 425
T. orientalis 667
Cannabis sativa XM 030624801.1
T. orientalis 170
T. orientalis 400
Vigna radiata XM 014636890.2
T. orientalis 369
Cannabis sativa XM 030645873.1
T. orientalis 1084
Chenopodium quinoa XM 02186142
T. orientalis 325
Cannabis sativa XM 030625382.1
T. orientalis 64
T. orientalis 210
T. orientalis 441
T. orientalis 417
Cannabis sativa XM 030652612.1
T. orientalis 1450
T. orientalis 209
Ziziphus jujuba XM 025078431.1
T. orientalis 502
Cannabis sativa XM 030641545.1
T. orientalis 1256
Fragaria vesca XM 004303050.2
T. orientalis 1327
T. orientalis 841
T. orientalis 1216
Cannabis sativa XM 030652713.1
Cucurbita moschata XM 02308321
Cannabis sativa XM 030652743.1
T. orientalis 302
T. orientalis 388
T. orientalis 1463
Trema orientalis KY786251.1
T. orientalis 565
Morus notabilis XM 024170361.1
T. orientalis 739
Cannabis sativa XM 030628150.1
T. orientalis 511
Morus notabilis XM 010089853.1
T. orientalis 1316
Morus notabilis XM 024170176.1
T. orientalis 891
T. orientalis 949
Cannabis sativa XM 030646919.1
T. orientalis 993
Morus notabilis XM 010090745.2
T. orientalis 769
Morus notabilis XM 010107775.2
T. orientalis 738
Cannabis sativa XM 030645650.1
T. orientalis 531
T. orientalis 1100
Juglans regia XM 035691610.1
T. orientalis 1128
T. orientalis 983
Nymphaea colorata XM 031639205
Cannabis sativa XM 030648942.1
T. orientalis 536
T. orientalis 330
T. orientalis 558
T. orientalis 1114
Cannabis sativa XM 030654121.1

Fig 5. Dendrogram based on maximum likelihood showing genetic relationships *Trema orientalis* SNPs in this study and published sequences of related taxa. Sequences starting with SNP are derived from this current study while the rest are from related taxa selected from the NCBI BLAST (see methods).

https://doi.org/10.1371/journal.pone.0267464.g005

These observations are also in line with the clustering observed with both silicoDArT and SNP markers, where *T*. *orientalis* is moderately differentiated and formed 2 distinct clusters. The SNP data clustered the groups more tightly, with less overlap and explained more variation in the samples possibly because SNPs are abundant in plant genomes. This clustering was supported by results of the genetic differentiation metric (F_{st} = 0.20–0.53) between pairs of clusters. Ideally, F_{st} values below 0.05 indicate low genetic differentiation, while values between 0.05–0.15, 0.15–0.25, and above 0.25 indicate moderate, high, and very high genetic differentiation respectively [46]. The total gene diversity (H_t = 0.06–0.40) across markers was lower than what has been established for forest trees in the wild [43,47]. Although the mating system (unisexual flowers) of *T*. *orientalis* [22] should reduce self-fertilization, the excessive heterozygosity may be associated with restricted pollen and seed dispersal possibly resulting from fragmented landscapes [24]. The degradation may also reduce population sizes, especially the actively reproducing trees such that few trees contribute to the seedling recruitment, hence most of the trees that were sampled seemed related. Studies on the population structure and recruitment of this species in the wild are encouraged. The constraints on gene flow were also unexpected since *T*. *orientalis* disperses its seed by birds [22] and pollinated by bees which are expected to span over a large geographical area aiding the gene flow.

Conclusion

Trema orientalis exhibits low genetic diversity and a potentially threatened genetic integrity. The strong population structure suggests that collection of germplasm should be done in different populations to maximise genetic variation in the collections. Characterisation of other populations is also recommended as well as studies on the population structure and recruitment of this species. The statistical analysis of DArT data sets showed high consistency with the results based on SNPs highlighting the suitability of DArT platforms for genomic dissection of *T*. *orientalis*.

Supporting information

S1 Fig. Estimation of number of groups of the *T*. *orientalis* population using silicoDArT marker data, as estimated using the model-based Bayesian algorithm implemented in the STRUCTURE program. A similar graph was obtained for the SNP markers (graph not shown).
(DOCX)

S2 Fig. Dendrogram based on maximum likelihood showing genetic relationships *Trema orientalis* sequences used in this study.
(DOCX)

S1 Table. Maximum Likelihood fits of 24 different nucleotide substitution models.
(DOCX)

Acknowledgments

The authors thank Sarah Nalumansi and Sulaiman Kato for helping with sample collection, and Samuel Ongerep for generating Fig 1. Appreciation also goes to Biosciences Eastern and

Central Africa (BECA) at the International Livestock Research Centre (ILRI), Nairobi for the technical support.

Author Contributions

Conceptualization: Judith Ssali Nantongo, Juventine Boaz Odoi, Hillary Agaba, Samson Gwali.

Data curation: Judith Ssali Nantongo.

Formal analysis: Judith Ssali Nantongo, Samson Gwali.

Funding acquisition: Hillary Agaba, Samson Gwali.

Investigation: Judith Ssali Nantongo, Juventine Boaz Odoi.

Methodology: Judith Ssali Nantongo, Juventine Boaz Odoi, Samson Gwali.

Project administration: Judith Ssali Nantongo, Hillary Agaba, Samson Gwali.

Resources: Hillary Agaba.

Software: Samson Gwali.

Supervision: Hillary Agaba, Samson Gwali.

Validation: Samson Gwali.

Visualization: Judith Ssali Nantongo.

Writing – original draft: Judith Ssali Nantongo.

Writing – review & editing: Judith Ssali Nantongo, Juventine Boaz Odoi, Hillary Agaba, Samson Gwali.

References

1. Franzel S, Carsan S, Lukuyu B, Sinja J, Wambugu C. Fodder trees for improving livestock productivity and smallholder livelihoods in Africa. Current Opinion in Environmental Sustainability. 2014; 6:98–103.
2. Sekaatuba J, Kugonza J, Wafula D, Musukwe W, Okorio J. Identification of indigenous tree and shrub fodder species in the lake Victoria shore region of Uganda. Uganda Journal of Agricultural Sciences. 2004; 9(1):372–8.
3. Kabirizi J, Ejobi F. Indigenous fodder trees and shrubs as feed resources for intensive goat production in Uganda. Farmers Handbook.; 2006.
4. Goodale UM, Berlyn GP, Gregoire TG, Tennakoon KU, Ashton MS. Differences in survival and growth among tropical rain forest pioneer tree seedlings in relation to canopy openness and herbivory. Biotropica. 2014; 46(2):183–93.
5. Nantongo JS, Gwali S. Long-term viability of populations of Prunus africana ((hook. f.) kalm.) in Mabira forest: implications for in situ conservation. African Journal of Ecology. 2018; 56(1):136–9.
6. Schippmann U, Leaman DJ, Cunningham A. Impact of cultivation and gathering of medicinal plants on biodiversity: global trends and issues. Biodiveristy and the ecosystem approach in agriculture, forestry and fishries Satellite event on the occasion of the Ninth regular session of the commission on genetic resources for food and agriculture Rome 12–13 October 2002 Inter departmental working group on biological diversity for food and agriculture, Rome. FAO2002.
7. Nantongo JS, Eilu G, Geburek T, Schueler S, Konrad H. Detection of self incompatibility genotypes in Prunus africana: Characterization, evolution and spatial analysis. Plos one. 2016; 11(6):e0155638. https://doi.org/10.1371/journal.pone.0155638 PMID: 27348423
8. Coates DJ, Byrne M, Moritz C. Genetic Diversity and Conservation Units: Dealing With the Species-Population Continuum in the Age of Genomics. Frontiers in Ecology and Evolution. 2018; 6(165).
9. Nantongo JS, Potts BM, Hugh F, Jessica N, Stephen E, Don A, et al. Quantitative Genetic Variation in Bark Stripping of Pinus radiata. Forests. 2020; 11(12):1356.

10. Frankham R, Ballou SEJD, Briscoe DA, Ballou D. Introduction to conservation genetics: Cambridge university press; 2002.
11. Schulman AH. Molecular markers to assess genetic diversity. Euphytica. 2007; 158(3):313–21.
12. Vinson CC, Mangaravite E, Sebbenn AM, Lander TA. Using molecular markers to investigate genetic diversity, mating system and gene flow of Neotropical trees. Brazilian Journal of Botany. 2018; 41 (2):481–96.
13. Vignal A, Milan D, SanCristobal M, Eggen A. A review on SNP and other types of molecular markers and their use in animal genetics. Genetics selection evolution. 2002; 34(3):275–305. https://doi.org/10.1186/1297-9686-34-3-275 PMID: 12081799
14. Kilian A, Wenzl P, Huttner E, Carling J, Xia L, Blois H, et al. Diversity arrays technology: a generic genome profiling technology on open platforms. Data production and analysis in population genomics: Springer; 2012. p. 67–89.
15. Macko-Podgórni A, Iorizzo M, Smółka K, Simon PW, Grzebelus D. Conversion of a diversity arrays technology marker differentiating wild and cultivated carrots to a co-dominant cleaved amplified polymorphic site marker. Acta Biochimica Polonica. 2014; 61(1). PMID: 24644550
16. Brinez B, Blair MW, Kilian A, Carbonell SAM, Chiorato AF, Rubiano LB. A whole genome DArT assay to assess germplasm collection diversity in common beans. Molecular breeding. 2012; 30(1):181–93.
17. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one. 2011; 6(5):e19379. https://doi.org/10.1371/journal.pone.0019379 PMID: 21573248
18. Muktar MS, Teshome A, Hanson J, Negawo AT, Habte E, Entfellner J-BD, et al. Genotyping by sequencing provides new insights into the diversity of Napier grass (Cenchrus purpureus) and reveals variation in genome-wide LD patterns between collections. Scientific reports. 2019; 9(1):1–15.
19. Zhang H, Jin J, Moore MJ, Yi T, Li D. Plastome characteristics of Cannabaceae. Plant diversity. 2018; 40(3):127–37. https://doi.org/10.1016/j.pld.2018.04.003 PMID: 30175293
20. Adinortey MB, Galyuon IK, Asamoah NO. Trema orientalis Linn. Blume: A potential for prospecting for drugs for various uses. Pharmacogn Rev. 2013; 7(13):67–72. https://doi.org/10.4103/0973-7847.112852 PMID: 23922459
21. Abe T. Threatened Pollination Systems in Native Flora of the Ogasawara (Bonin) Islands. Annals of Botany. 2006; 98(2):317–34. https://doi.org/10.1093/aob/mcl117 PMID: 16790463
22. Orwa CM, A; Kindt R; Jamnadass R; Simons A. Agroforestry tree Database:a tree reference and selection guide version 4.0 (http://www.worldagroforestry.org/af/treedb/). 2009.
23. Mosango M, Mwanjalolo Majaliwa J. Phytosociological study of Trema orientalis and Vernonia auriculifera highland community in Southwestern Uganda [East Africa]. Polish Botanical Journal. 2008; 53 (2):125–38.
24. Otieno A, Buyinza M, Kapiyo R, Oindo B. Local communities and collaborative forest management in West Bugwe Forest Reserve, Eastern Uganda. 2013.
25. Gruber B, Unmack PJ, Berry OF, Georges A. dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. Molecular Ecology Resources. 2018; 18 (3):691–9. https://doi.org/10.1111/1755-0998.12745 PMID: 29266847
26. Pritchard JK, Wen W, Falush D. Documentation for STRUCTURE software: Version 2. University of Chicago, Chicago, IL. 2010.
27. Alam M, Neal J, O'Connor K, Kilian A, Topp B. Ultra-high-throughput DArTseq-based silicoDArT and SNP markers for genomic studies in macadamia. PloS one. 2018; 13(8):e0203465. https://doi.org/10.1371/journal.pone.0203465 PMID: 30169500
28. Earl DA. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation genetics resources. 2012; 4(2):359–61.
29. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular biology and evolution. 2018; 35(6):1547–9. https://doi.org/10.1093/molbev/msy096 PMID: 29722887
30. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008; 24(11):1403–5. https://doi.org/10.1093/bioinformatics/btn129 PMID: 18397895
31. Tabangin ME, Woo JG, Martin LJ, editors. The effect of minor allele frequency on the likelihood of obtaining false positives. BMC proceedings; 2009: Springer.
32. Hierfstat Goudet J., a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes. 2005; 5(1):184–6.

33. Akbari M, Wenzl P, Caig V, Carling J, Xia L, Yang S, et al. Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. Theoretical and applied genetics. 2006; 113 (8):1409–20. https://doi.org/10.1007/s00122-006-0365-4 PMID: 17033786

34. Huang Y-F, Poland JA, Wight CP, Jackson EW, Tinker NA. Using genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat. PloS one. 2014; 9(7):e102448. https://doi.org/10.1371/journal.pone.0102448 PMID: 25047601

35. Kilian A, Huttner E, Wenzl P, Jaccoud D, Carling J, Caig V, et al., editors. The fast and the cheap: SNP and DArT-based whole genome profiling for crop improvement. Proceedings of the international congress in the wake of the double helix: from the green revolution to the gene revolution; 2003.

36. Hassani SMR, Talebi R, Pourdad SS, Naji AM, Fayaz F. In-depth genome diversity, population structure and linkage disequilibrium analysis of worldwide diverse safflower (Carthamus tinctorius L.) accessions using NGS data generated by DArTseq technology. Molecular Biology Reports. 2020; 47 (3):2123–35. https://doi.org/10.1007/s11033-020-05312-x PMID: 32062796

37. Serrote CML, Reiniger LRS, Silva KB, Rabaiolli SMdS, Stefanel CM. Determining the Polymorphism Information Content of a molecular marker. Gene. 2020; 726:144175. https://doi.org/10.1016/j.gene.2019.144175 PMID: 31726084

38. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American journal of human genetics. 1980; 32(3):314. PMID: 6247908

39. Xia L, Peng K, Yang S, Wenzl P, De Vicente MC, Fregene M, et al. DArT for high-throughput genotyping of cassava (Manihot esculenta) and its wild relatives. Theoretical and Applied Genetics. 2005; 110 (6):1092–8. https://doi.org/10.1007/s00122-005-1937-4 PMID: 15742202

40. Farahani S, Maleki M, Mehrabi R, Kanouni H, Scheben A, Batley J, et al. Whole genome diversity, population structure, and linkage disequilibrium analysis of chickpea (Cicer arietinum L.) genotypes using genome-wide DArTseq-based SNP markers. Genes. 2019; 10(9):676. https://doi.org/10.3390/genes10090676 PMID: 31487948

41. Valdisser PAMR Pereira WJ, Almeida Filho JE Müller BSF, Coelho GRC, de Menezes IPP, et al. In-depth genome characterization of a Brazilian common bean core collection using DArTseq high-density SNP genotyping. BMC Genomics. 2017; 18(1):423. https://doi.org/10.1186/s12864-017-3805-4 PMID: 28558696

42. Gopaulchan D, Motilal LA, Bekele FL, Clause S, Ariko JO, Ejang HP, et al. Morphological and genetic diversity of cacao (Theobroma cacao L.) in Uganda. Physiol Mol Biol Plants. 2019; 25(2):361–75. https://doi.org/10.1007/s12298-018-0632-2 PMID: 30956420

43. Gwali S, Vaillant A, Nakabonge G, Okullo JBL, Eilu G, Muchugi A, et al. Genetic diversity in shea tree (Vitellaria paradoxa subspecies nilotica) ethno-varieties in Uganda assessed with microsatellite markers. Forests, Trees and Livelihoods. 2015; 24(3):163–75.

44. Zalapa JE, Brunet J, Guries RP. The extent of hybridization and its impact on the genetic diversity and population structure of an invasive tree, Ulmus pumila (Ulmaceae). Evol Appl. 2010; 3(2):157–68. https://doi.org/10.1111/j.1752-4571.2009.00106.x PMID: 25567916

45. Stoeckel S, Grange J, FERNÁNDEZ-MANJARRES JF, Bilger I, FRASCARIA-LACOSTE N, Mariette S. Heterozygote excess in a self-incompatible and partially clonal forest tree species—Prunus avium L. Molecular Ecology. 2006; 15(8):2109–18. https://doi.org/10.1111/j.1365-294X.2006.02926.x PMID: 16780428

46. Evolution Wright S. and the genetics of populations: a treatise in four volumes: Vol. 4: variability within and among natural populations: University of Chicago Press; 1978.

47. Nantongo JS, Lamoris Okullo JB, Eilu G, Ratsimiala Ramonta I, Odee D, Cavers S. Structuring of genetic diversity in Albizia gummifera C.A.Sm. among some East African and Madagascan populations. African Journal of Ecology. 2010; 48(3):841–3.