Original Research Article

ASSESSMENT OF THE GENETIC DIVERSITY OF THE TRADITIONAL TREE SPECIES *Kigelia africana (sausage tree)* USING MOLECULAR MARKERS FOR CONSERVATION GENOMICS IN KENYA

ABSTRACT:

This study is the first to explore the genetic composition of ancient *Kigelia africana* across a distribution range in Kenya. *Kigelia africana* a native forest species of Kenya as far as we know it; it is widely planted in the central regions of the country by the Kikuyu tribe who inhabit this region for making their traditional brew Muratina. Unlike other tree species like Acacia, *Kigelia africana* has seldom been studied, although there is ample evidence of its great ecological and economic value. Because of cultural reasons, natural populations are rare in the wild. In this study, four ancient tree populations were investigated to explore the genetic composition of *Kigelia africana* through DArTseq technology. Thirty-two (32) Plant seed samples were obtained from various locations, their DNA extracted, libraries prepared and sequenced using the Illumina 2500 High Throughput sequencer.

A total of 8,556 SilicoDArT and 3,703 SNP markers were selected and used. The average PIC was 0.45 and 0.41 for the SilicoDArT and SNPs respectively. The population structure and average linkage hierarchical clustering based on the SNPs revealed two distinct subpopulations and a few smaller admixture groups. Both marker types identified all 32 landraces as potential duplicates with very low heterozygosity based on the Gower's genetic dissimilarity. The heterozygosity defining the genetic variation within each subpopulation was around 0.25. A mantel test showed good harmony between DArTseq and SNP marker data sets. It also showed no significant correlation between genetic diversity and the geographical coordinates of the tree samples. The results of this study provide important information and insights for decision makers, farmers, and breeders to make the necessary actions to conserve this culturally important tree.

Keywords: genetic composition, Kenya, Kigelia Africana, polymorphism information content, SNP markers, genetic diversity.

Commented [HG1]: please state clearly the objective of your study before describing the methodology

INTRODUCTION

1.1 Background Information

Commented [HG2]: the general information section is long. you can summarise or delete certain sentences to make it easier to read.

Species that have small isolated populations are at higher risk of losing adaptive variation due to genetic drift and the genetic costs of inbreeding. Genetic diversity largely influences the ability of plant species to persist in the face of threats. Loss of genetic diversity has been considered as a crucial factor that results in inbreeding depression, reduced adaptation and fitness, and a decrease in long-term species survival. One can argue that the conservation of biodiversity is ultimately the conservation of genetic diversity and/or variation (Liu et al., 2020).

Kigelia africana, commonly known as the sausage tree has long, open sprays of large, wrinkled, maroon or dark red trumpet-shaped flowers that are velvety on the inside and usually overflowing with nectar. The short, squat trunk has light brown, sometimes flaky bark and supports a dense rounded to spreading crown (18 m high, 20 m wide) of leathery, slightly glossy foliage (deciduous). The huge, grey-brown fruits, 60 x 12 cm. hang from long stalks, from December (summer) to June (winter) and can weigh anything between 2 to 9 kg(Pitta, 2003).

The genus *Kigelia* has one species and occurs only in Africa. Many living species like the Sunbirds, Black-headed Oriole, Sombre, Black-eyed Bulbuls, Masked Weaver, Brown-headed Parrot, impala, Grey Lourie, Elephant, Kudu occasionally feed on *K. africana* leaves. Baboons, monkeys, bushpigs and porcupines eat the fruit. Epauletted fruit bats are thought to pollinate the flowers and Charaxes butterflies also visit the tree (Joffe Pitta, 2003).

The roots, wood, and leaves have been found to contain naphthoquinones, di-hydroisocoumarines, flavonoids, and aldehydic iridoids. From the root and its bark, the usual plant substances stigmasterol, β-sitosterol, ferulic acid, the naphthoquinones lapachol, 6-methoxymellein, and two new phenolic compounds have been isolated. Kigelin is the main component of the plant(Agyare et al., 2013). *K. africana* is used for the treatment of dysentery, venereal diseases, and as a topical application on wounds and abscesses. In the area around Nsukka, Nigeria, the bark is used for the treatment of venereal diseases. Verminoside has also been isolated from the fruit. Aqueous extract of *K. africana* has been shown to exhibit significant analgesic and anti-inflammatory effects (Tamokou and Kuete, 2014).

In Malawi, roasted fruits are used to flavor beer and aid fermentation. The tough wood is used for shelving and fruit boxes, and dugout canoes are made from the tree in Botswana and Zimbabwe. Roots are said to yield a bright yellow dye. In African folk medicine, traditional remedies are prepared from crushed, dried or fresh fruits and used to deal with ulcers, sores and syphilis - the fruit has antibacterial activity(Agyare et al., 2013). Today, beauty products and skin ointments are prepared from fruit extracts, mostly for protecting the skin from acne. Fresh fruit cannot be eaten as it is poisonous and has a strong purgative, and causes blisters in the mouth and on the skin. Green fruits are said to be poisonous. In time of food scarcity, seeds are roasted and eaten (Pitta, 2003).

In Central Kenya, especially among the Agikuyu and the Akamba tribes, the dried fruits are used to make an alcoholic beverage (muratina in Kikuyu, kaluvu in Kamba), which is a core component in cultural events in Central Kenya (Bussmann et al., 2021). The fruit is harvested then split into two along the grain, then dried in the sun. The dried fruit is then treated with bee pollen and honey. The treated fruit (miatine) is then used in fermentation process in making of sweet beer (Wikipedia, 2021).

Knowledge of population genetic diversity and structure is of fundamental importance for tree species conservation and breeding programs. Single-nucleotide polymorphisms (SNP) have proved to be the most abundant form of variation within a species at the genome level and can provide detailed insight into the genetic basis of a population. Combined with Next-Generation Sequencing (NGS) technology, SNP markers are having substantial impacts on population genetics as well as plant breeding. (Cai et al., 2020)

The assessment of genome-wide diversity by genotyping-by-sequencing (GBS) provides robust estimates of diversity and has been increasingly adopted as a fast, high-throughput, and cost-effective tool for whole-genome genetic diversity analysis in many germplasm sets. DArTseq technology (Diversity Arrays Technology), which is based on the GBS principle generates both SNP and DArTSeq markers. These have been shown to have higher consistency and reproducibility in diversity studies experiments (Liu *et al*., 2020). The DArTseq markers, based on GBS, efficiently target low-copy-number sequences via a complexity reduction method and have been successfully applied for genetic diversity studies in different species (Pascual et al., 2020).

1.2 Statement of the Problem and Justification of the study

Studies employing NGS to address conservation genomics and subsequent conservation strategies for threatened plants are still rare(Liu et al., 2020). The *Kigelia africana*, has been a part of the traditional practices of the Agikuyu and Akamba people of Kenya for decades. As such, there is need to give it a genetic identity to be able to conserve its precious economic and cultural value. Since this species is also grown in other parts of Africa such as Zimbabwe and Malawi, the local landraces should be genotyped to allow further classification and insight on their genetic constituents and distribution. Most of the tree species are sparsely distributed in central and eastern regions of Kenya. Genetic data generated from this allows us to determine whether there is any genetic variation within the species from different locations within the Kenyan borders.

Advances in molecular biology and high-throughput genotyping technologies have significantly impacted the field of plant conservation, shifting from a phenotype-based to a genotype-based characterization. Molecular markers have shown to be invaluable tools for assessing plants' genetic resources by improving our understanding with regards to the distribution and the extent of genetic variation within and among species(Porth and El-Kassaby, 2014). Therefore, there being no study on the genetic diversity of *Kigelia africana*, this study sets to determine the genetic diversity, given the tree's economic and ecological importance. This plant has

Commented [HG3]: please delete this sentence and reword it to state your objective clearly

great potential to be developed as a source of drugs by pharmaceutical industries according to (Saini et al., 2008).

1.3 Research Questions

i. There is no difference between genetic variation and genetic differentiation in the various collected samples of *Kigelia africana*.

ii. There is no difference in genetic diversity among the collected samples of *Kigelia africana* across the specified geographical spectrum.

1.4 Objectives

1.4.1 General Objective

To explore the genetic composition of ancient *Kigelia africana* trees across the Kenyan demographic for genomic conservation purpose.

Commented [HG4]: replace with evaluate

1.4.2 Specific Objective

i. Determine the genetic diversity within *Kigelia africana* using DArTseq technology.

ii. Determine the genetic differentiation in allelic frequencies among *Kigelia africana* populations in Kenya.

LITERATURE REVIEW

Commented [HG5]: this part is also too long. please summarize in one paragraph. you are repeating information that was given in the introduction.

2.1 Kigelia Africana

Commented [HG6]: please write the full name of the species Kigelia africana including the authors.

2.1.1 Classification

Kigelia africana, a species belonging to the taxonomic tree, is classified under the domain Eukaryota, which consists of organisms with complex cells. It falls within the kingdom Plantae, encompassing plants, and the phylum Spermatophyta, which includes seed-producing plants. As a member of the subphylum Angiospermae, Kigelia africana is a flowering plant. It is further classified under the class Dicotyledonae, order Scrophulariales, and family Bignoniaceae. Lastly, it is part of the genus Kigelia, which contains the species *Kigelia Africana* (Areces-Berazain, 2022).

2.1.2 Morphology

The Kigelia Africana tree can grow up to 25 meters in height and typically has a wide, rounded crown. Its leaves are arranged opposite or in groups of three, with imparipinnate leaflets clustered near the branch tips. The leaflets, numbering between 5 and 11, are sessile or subsessile, except for the terminal leaflet, which has a 1-4(-6.5) cm long petiolule. The leaflet lamina measures 3.5-20 x 2.5-11 cm and is ovate-elliptic in shape. The apex is obtuse, broadly tapering, rounded, or retuse, while the base is rounded to cuneate and may be asymmetric, except for the terminal leaflet, which is symmetrical. Leaflet surfaces range from glabrous to somewhat hairy, with entire, serrate, or toothed margins that can be noticeably wavy. Lateral nerves are impressed above and prominent below. The petiole is 3.5-15 cm long, with the rhachis sulcate above and terete below.

The inflorescence takes the form of a pendulous, lax, terminal panicle measuring 30-100(-150) cm in length. The flowers are zygomorphic, large, and longly pedunculate, with pedicels 1-11(-13.5) cm long that curve upward at the tip. The calyx is tubular to campanulate, 2-4.3 cm long, irregularly 4-5-lobed with ribbed lobes up to 1 cm long(Areces-Berazain, 2022).

2.1.3 Distribution of Kigelia africana

Kigelia africana is a plant that is indigenous to continental sub-Saharan Africa and is widely dispersed there. Madeira, the Canary Islands, Cape Verde, Réunion, Mauritius, several African islands, California, Florida, Hawaii, Central and South America (Mexico, Honduras, El Salvador, Nicaragua, Costa Rica, Panama, Colombia, Venezuela, French Guiana, Ecuador, Peru, and Brazil), Western Asia (Israel, Iraq), South and Southeast Asia (Pakistan, India, Maldives, Sri Lanka, Myanmar, Thailand, Laos, Vietnam, China, Taiwa), and the Caribbean islands.

2.1.4 Implications of genetic diversity on conservation efforts

Genetic diversity is important for *K. africana* as it affects the livelihoods of indigenous and local communities that use it for traditional and entrepreneurial purposes. Rich genetic diversity within and among forest tree species provides an important basis for maintaining culture, food security and enabling sustainable development. A small number of genetically different, locally well-adapted *Kigelia*

africana landraces have been generated through natural and/or human-mediated activities during previous years.
Clinical trials and tests have been done on *K. africana* to accurately authenticate its chemical constituents and pharmacological properties. The plant is used in many traditional medicine systems to control various diseases including cancer. In western Kenya, stem bark of *K. africana* is boiled and one glass (300 ml) taken orally twice a day for three months to suppress breast, lung and skin cancers (Mukavi et al., 2020).

Kigelia africana has biologically active phytochemicals, many of which have been isolated. Whilst the fruits are most often cited in pharmacological studies, other plant parts are also used in herbal preparations. Commercially available products have been formulated from *K. africana*, though many have not been fully standardized. Despite many efforts by researchers to scientifically validate traditional uses of *K. africana*, many remain merely claims. With this information, there is need for genetic characterization of these ethnobotany, phytochemistry and pharmacology traits, to enhance further understanding of the *K. africana* plant, scientifically validate other traditional uses, isolate new bioactive phytochemicals and standardize *K. africana* products(Nabatanzi et al., 2020).

In Benin Republic, *K. africana* is used in treating wounds with microbial infections, as well as treating of diabetes, soothing toothache pain, and resolving skin diseases(Dossou-Yovo et al., 2022).

The conservation genomics component aims to provide genomic information to support conservation of the Kenyan flora. Information of the genetic diversity of most tree species in any region of the world helps to contribute to the creation and adoption of effective strategies for their preservation and future use.

2.2 Molecular Markers used for diversity studies in Trees

In recent times, molecular markers have proven to be invaluable tools for assessing genetic resources of tree plants by improving understanding of the users with regards to the distribution and the extent of genetic variation within and among the species. Knowledge of the genetic diversity of the threatened tree species in any region of the world may contribute to the creation of effective strategies for their preservation, improvement, and future use(Bedassa, 2018).

A molecular or DNA marker is the difference in DNA nucleotide sequences between individual organisms or species, that is in proximity or closely linked to a target gene that expresses a trait. Usually the target gene, expressed trait or biological function and the associated closely linked molecular marker are inherited together. The specific genomic location of the molecular marker within chromosomes is referred to as a locus or loci, and it may be known or unknown. The tight association of molecular markers to a trait or gene of particular biological function, makes the markers serve as practical signs or flags that signal a particular gene locus and aid the detection or identification of the associated traits whether the genes involved are known or unknown and whether the gene(s) can be detected or not. Molecular or DNA markers do not influence traits associated with the expression or function of the linked gene or genes. DNA markers are useful for telling the individual genotypic differences (polymorphisms) in similar or different species. These differences are due to varied types of mutations of the DNA creating nucleotide sequence variations (Amiteye, 2017).

The mutations causing these differences could be single nucleotide substitutions, rearrangements involving insertions or deletions, DNA section duplication, translocations and inversions as well as mistakes in replication of DNA that are tandemly repeated. Molecular marker signals that are used to reveal genotypic differences between individuals due to marker sequence differences are called polymorphic markers. On the other hand, DNA markers that cannot be used to differentiate between or among genotypes are referred to as monomorphic markers. The characteristics of a good and very useful DNA marker are that the marker is ubiquitous and evenly distributed throughout the genome, easy to assay, replicable, cost effective, multiplexed and can be automated. An ideal molecular marker must also be highly polymorphic, co-dominant in expression to enable effective discrimination between homozygotes and heterozygotes, should be highly reproducible and possible to share data generated among laboratories. Also, a very good molecular DNA marker creates no detrimental effect on phenotype, is genome-specific in nature, and multi-functional. DNA markers are categorized into various classes depending on the detection method: hybridization, polymerase chain reaction (PCR) and DNA sequence dependent molecular markers (Amiteye, 2017).

2.3 DArTSeq Technology

A good example of sequence dependent molecular markers are the DArT (Diversity Array Technology Pty Ltd) markers. DArT markers were developed as one of the ultra-high-throughput, no prior sequence data-independent, cost effective, whole-genome genotyping technique with large number of markers that cover the entire genome. DArT markers have been applied successfully in genomic studies in many species including those with large and complex genomes such as barley, sugarcane, wheat, oat and strawberry. The DArTseq method has been used in discriminating different species for population studies, diversity studies, characterization of germplasm and studies involving genome-wide association (Badu-Apraku et al., n.d.).

DArT markers are developed through the use of combinations of restriction enzyme digestions to reduce genome complexity, followed by next-generation sequencing of complexity reduced representations or fragments to identify DNA polymorphisms and SNPs leading to the production of thousands of polymorphic loci in a single assay. The DArT platform generates two variants of markers, the SilicoDArT and DArTSeq SNP markers. SilicoDArT markers are dominant and are mostly scored for the absence (0) or presence (1) of a single allele while as DArTSeq SNPs are co-dominant markers (Adu et al., 2021).

A good quality genomic DNA of 50–100 ng amount is enough for purposes of DArT analysis. DArT overcomes many of the limitations of currently available marker technologies (Amiteye, 2017).

MATERIAL AND METHODS

3.1 Plant Materials

Most forests have been exposed to severe disturbance as a result of human activities, and the *K. africana* species is now found in patches in villages and national forest parks. To avoid materials from unknown sources, only ancient trees with a DBH (diameter at breast height) greater than 100 cm were selected for this study. Since this

is a qualitative study, a total of 32 individuals were randomly collected from various regions in Kenya based on human interactions from the local people, especially those who brew the traditional *Muratina* beer. The formular used is: $Sample\ Size = \frac{Z^2 * p(1-p)}{C^2}$. Where Z is the confidence level, p is the expected proportion in population based on previous studies and expressed as decimal, and c is the confidence interval, expressed as decimal (Charan and Biswas, 2013).

Commented [HG7]: what are these previous studies? please describe them briefly to make it easier to understand the method.

The name, geographic location, altitude, for each sample is recorded and described in table 1 below.

Table 1: Origin, collection sites and geographical coordinates of Kigelia africana landraces from Kenya used in this study.

Area name	County	Co-ordinates	Genotype	Quantity
Ruaka	Kiambu	-1.200527, 36.776289	Mur1,Mur2	2
Ruiru	Kiambu	-1.143403, 37.027777	Mur3	1
Juja	Kiambu	-1.2734316,36.7280686	Mur4-6	3
Witeithie	Kiambu	-1.062939, 36.995229	Mur7	1
Gatundu	Kiambu	-1.2734316,36.7280688	Mur8-9	2
Kangundo	Machakos	-1.2734316,36.7280689	Mur10-12	3
Matuu	Machakos	-1.2734316,36.7280690	Mur13-15	3
Katumani	Machakos	-1.612352, 37.203988	Mur16-18	3
Kieni	Nyeri	-0.318396, 36.753943	Mur19	1
Kanyariri	Embu	-1.2734316,36.7280693	Mur20	1
Siakago	Embu	-0.581557, 37.635987	Mur21-22	2
Maua	Meru	0.252559, 37.929558	Mur23	1
Kahuho	Kiambu	-1.195837, 36.674395	Mur24	1
Kandara	Muranga	-0.896207, 36.999131	Mur25-26	2
Kianjiruini, Maragua	Muranga	-0.795372, 37.117579	Mur27-29	3
Mida	Kilifi	-3.352570, 39.915182	Mur30	1
Dumbule	Kwale	-4.151469, 39.402422	Mur31-32	2

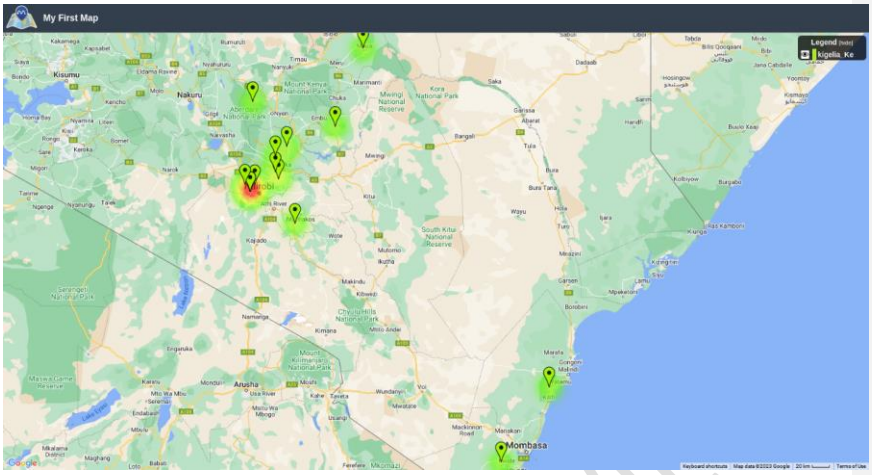

Figure 1: A geographical map of *Kigelia africana* sample collection locations in Kenya used in this study.

3.2 DNA Isolation

DNA was isolated and purified using the NucleoMag 96 Plant genomic DNA extraction kit (Macherey–Nagel, Du¨ren, Germany), following the manufacturer's instructions. Concentration of the extracted DNA were normalised within the range of 50–100 ng/ul. The quality and quantity of the DNA samples was then checked on 0.8% agarose gel.

Commented [HG8]: which organs did you extract the DNA from? you need to describe your extraction method in more detail.

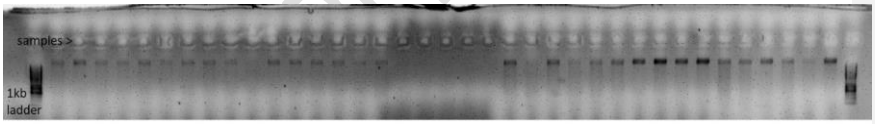

Figure 2: DNA bands on 0.8% Agarose Gel for the 32 *K. africana* samples

3.3 Library Construction and Sequencing

Libraries were constructed following the protocol described in (Kilian et al., 2012). Two DArTseq complexity reduction methods had to tested since this was the first time these three species was being sequenced. A rare cutting restriction endonuclease enzyme PstI (50 -CTGCA|G-30) in combination with two different frequently cutting restriction enzymes HpaII (50-C|CGG-30) and MseI (50 -T|TAA-30) were tested. The PstI/HpaII combination was selected as the best performing method. For each sample, 2 ul of DNA was digested with the PstI/HpaII restriction enzyme combination. Digestion products were ligated to barcoded adapters pair annealed to the two restriction enzyme overhangs. The PstI-compatible adapters include the partial attachment sequence for the 'Read 1 End' of the Illumina flow cell, a barcode of variable length (4–8 bp) and the PstI-compatible overhang sequence. The reverse adapters include the partial sequence for the 'Read 2 End' of the Illumina flow cell and MawI compatible overhang sequence. The adapter-ligated fragments were amplified in a Polymerase Chain Reaction (PCR) using optimized settings for a total

Commented [HG9]: have you extracted DNA from three different species?

of 35 cycles. After PCR, equimolar amounts of the amplified products from each sample were pooled together, purified, and loaded on the cBot (Illumina, Inc., San Diego, CA, USA) for clustering on an Illumina Single Read flow cell. Libraries were then sequenced in the Illumina Hiseq 2500 using the single read sequencing protocol. A proprietary automatic genotypic data analytical pipeline, DArTsoft14, developed by DArT Pty Ltd, Canberra, Australia, was used to generate allele calls for SNP and DArT markers from the sequence data generated (Kafoutchoni et al., 2021). Ameker were scored a '0', '1', 'and '- 'representing presence, absence, and no-zero count for the silico dart markers. The SNP markers were scored as '1' for the SNP allele homozygote, '0' for reference allele homozygote, and '2' for heterozygotes (Adu et al., 2021). For this study, SNP markers were used as the preferred marker of choice.

3.4 Marker Quality Parameters

SNP markers were selected for best performance based on their polymorphic information content (PIC), percentage call rate, and marker percentage reproducibility from the duplicated sample replicates. The PIC shows the diversity of the marker within the populations, while showing its ability to detect polymorphism among the individuals in a population. Since DArTseq and SNP markers are based on dominance (presence/ absence), PIC ranges from zero for monomorphic markers, to 0.5 for markers present in 50% of individuals and are absent in the remaining 50%. Markers quality parameters were trimmed automatically using the DArTsoft v14

The DArT software automatically has computed several quality parameters for each DArTseq and SNP marker, such as call rate, polymorphic information content (PIC), and reproducibility of both markers(Baloch et al., 2017) .

3.5 Genetic diversity and population relationship analysis

Population structure and genetic diversity was calculated from each of the 32 samples' DArTSeq and SNP data. The newly developed and released dartR version 2 for conservation genetic analysis was used for the statistical analysis and visualization of the data. Diversity indices were estimated to show the clear diversity, if any, between populations. These indices include observed and expected heterozygosity (Ho, He), population inbreeding coefficient (Fis), total gene diversity (Ht), and the gene diversity among collected samples (Dst)(Mijangos et al., 2022).

To get a clear picture of the genetic structure of Kigelia africana in Kenya, STRUCTURE software was used using the Bayesian clustering algorithm. This was flexibly estimated inside the dartR package. A neighbor-joining tree was constructed using the SNP and DArTSeq, principal components analysis (PCA) based on a pairwise genetic distance matrix of the accessions, and Hierarchical analysis of molecular variance (AMOVA) was used to support the hierarchical structure analysis. The genetic differentiation between populations was analyzed by estimating the pairwise fixation index (Fst) (Wadl et al., 2018). Similarities between trees will be estimated using Dice coefficients of similarity. The genetic similarity among genotypes will be estimated from the dissimilarity (distance) matrix generated from simple matching coefficient. The resulting dissimilarity matrix will be further analyzed using the probability that the alleles at a random locus are identical in state (IBS). Principal component analysis (PCA) was used to assess the diversity among the *Kigelia africana* accessions (Padmaja, 2009).

RESULTS

Commented [HG10]: you discuss some results in the "results" section which decreases the consistency of your "discussion" section. please reorganize this part and just make a description of your results.

4.1 DArTseq and SNP detection

A total of 11,793 SNP markers were generated after sequencing. A final selection of 3,703 markers were selected with an >90% reproducibility, and >80% call rate. DArTseq markers were reduced to 8,556 from a total of 26,352. This was due to a lot of low call rate markers below 80%. The average call rate was observed at 0.99% while reproducibility for the markers was observed at 1 meaning a 100% consistency in the marker scoring.

4.2 Genetic diversity and population structure

All markers had a PIC ranging between 0.39 to 0.45 and an average of 0.41 which is very informative. Overall polymorphism information content (PIC) of the DArTseq markers was 0.45 and 0.41for the SNP markers. The average expected heterozygosity (He) in the population varied from 0.30 for DArTseq and 0.41for SNPs (Table 1). The mean observed (Ho) and expected (He) heterozygosity (Table 1) corroborates with the high PIC values above.

Table 2. Basic statistics and genetic diversity of *K. africana* based on SNP and SilicoDArT markers.

	Ho	He	Hs	Ht	Dst	Htp	Dstp	Fst	Fstp	Fis	Dest
SNP	0.33	0.41	0.50	0.50	0.00	0.50	0.00	-0.01	-0.01	0.35	-0.01
silicoDArT	0.39	0.31	0.38	0.38	0.00	0.38	0.00	0.00	0.00	-0.33	0.00

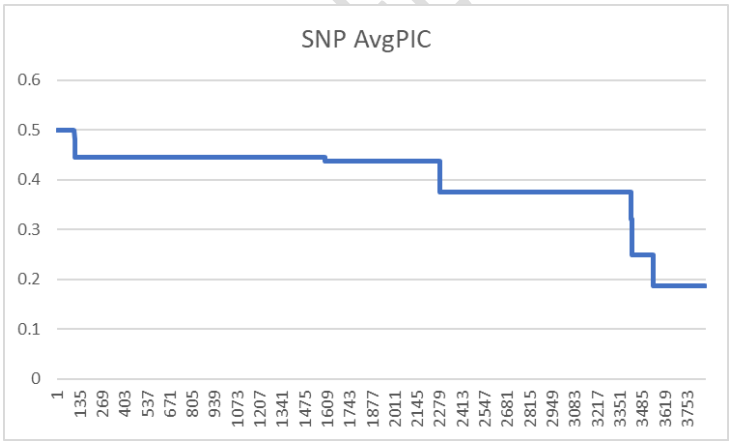

Figure 3. The polymorphic information content of the SNP markers.

The minor allele frequencies by locus for SNP data scored a minimum of 0.23 and a mean of 0.44. MAF for DArTseq dominant markers was not calculated.

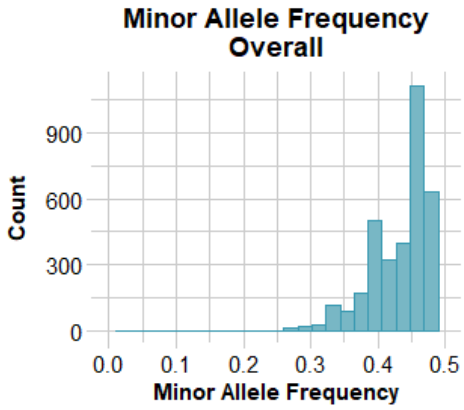

Figure 4. The mean minor allele frequency (MAF) based on SNPs

4.3 Population structure analysis

Genetic similarities among the *K. africana* individuals were assessed using the SNP markers and the results revealed 3 clusters, which was also supported by the Delta-K plot. With more individuals in one cluster than the other two clusters of Kilifi and Nyeri populations, which had one sample each. A neighbor joining tree was constructed and showed similar clustering based on the SNP and silicoDArT data (figure).

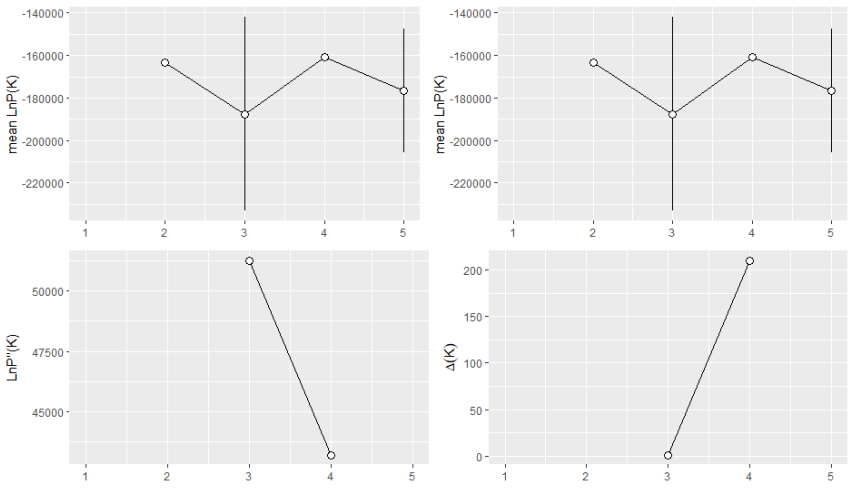

Figure 5: Mean LnP(K), LnP(K), and Delta-K(ΔK) observed in Structure analyses for K values of 1-5 in the K. africana populations.

A Neighbor joining tree was constructed from the Euclidean distances calculated from the DArTSeq and SNP data. The samples were grouped into 3 clusters based on

location as seen in the figure below.

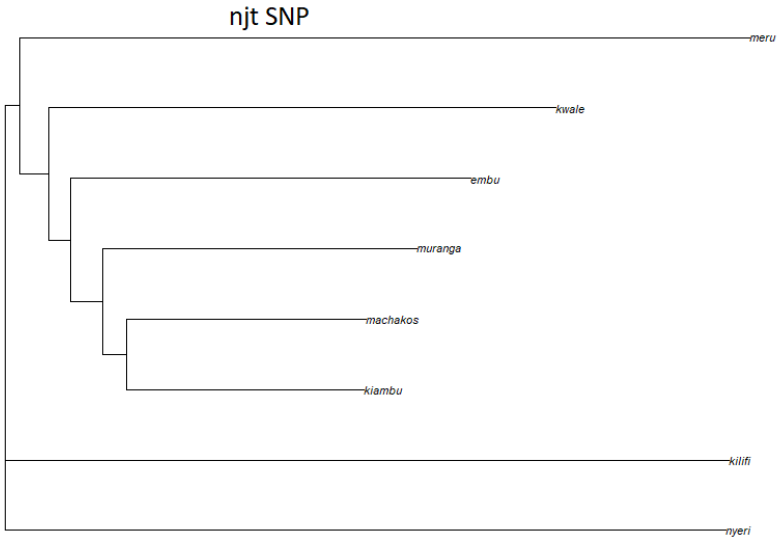

Figure 6: A neighbor joining tree of K. africana SNP data

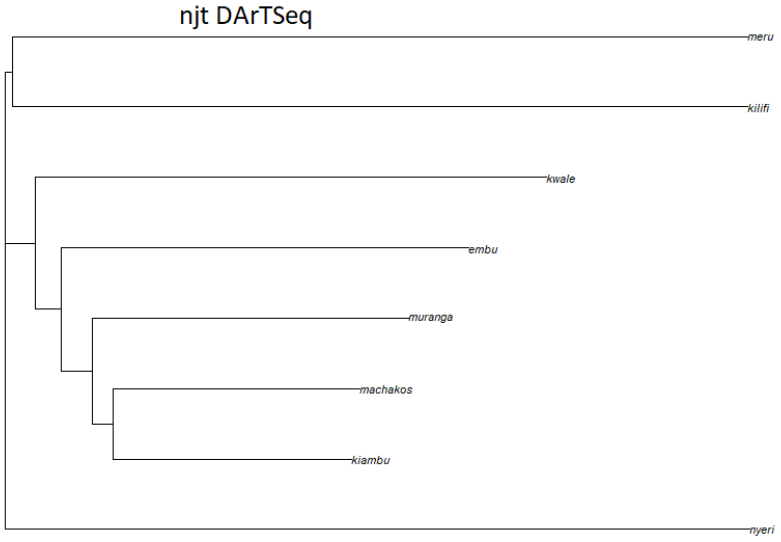

Figure 7: A neighbor joining tree of K. africana silicoDArT data

Based on (Sherwin et al., 2021), the diversity summary of the provided *K. africana* samples was calculated including the allelic richness (q = 0), Shannon information (q = 1), and heterozygosity (q = 2).

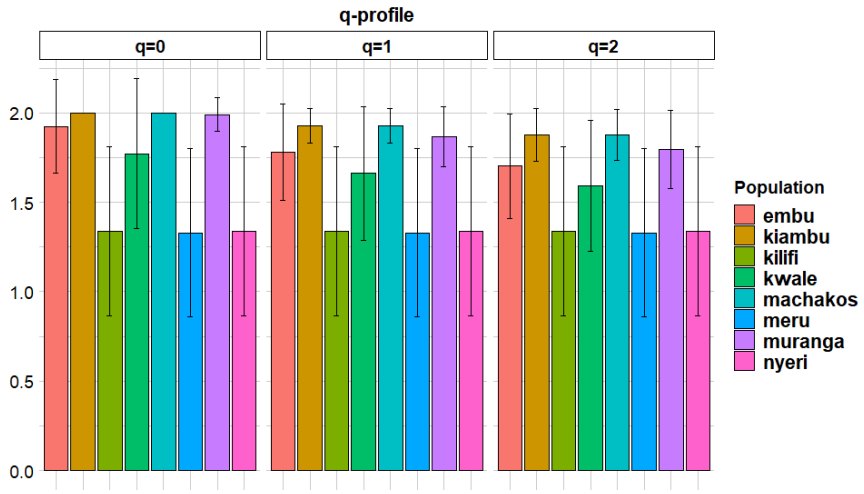

Figure 8: Population Diversity Summary based on SNP data

Individual genetic diversity was analyzed by principal coordinate analysis (PCA) as shown in figure 9 and 10 below. The PCA analysis showed very low average variance of 3.5% for silicoDArT, and 4.7% for SNP data.

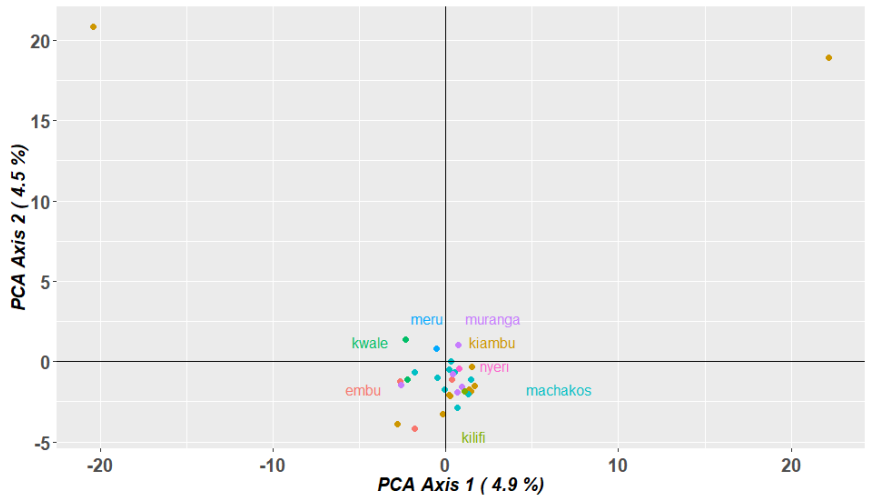

Figure 9: Principal coordinates analysis plot to infer group structure of *K. africana* based on SNP markers.

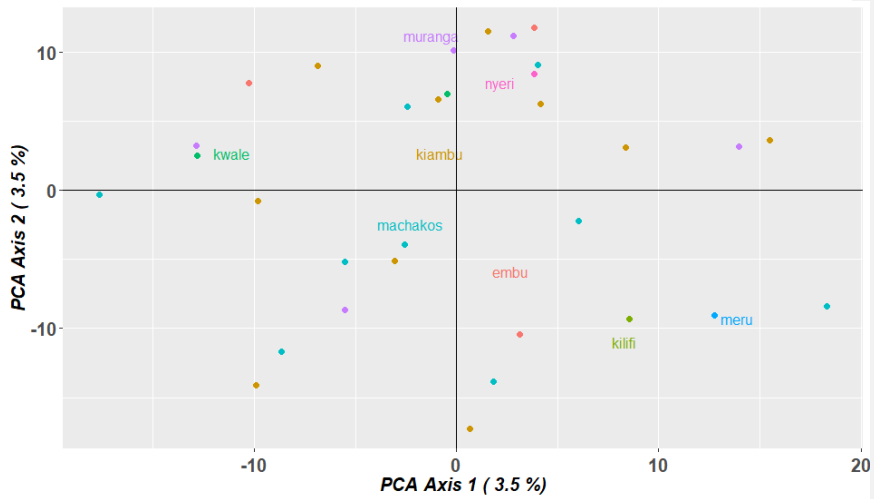

Figure 10: Principal coordinates analysis plot to infer group structure of *K. africana* based on silicoDArT markers

4.4 Genetic differentiation of *K. africana*

Based on the cluster identified by the STRUCTURE analysis, low estimates of total genetic diversity (Ht), and genetic diversity (Dst) were observed more on the silicoDArT than in the SNP data (Table 3). Genetic differentiation (Fst) was lower in SNP data than in the silicoDArTs. There was also low inbreeding coefficient (Fis) from both data sets. The summary of the results shows low variation among individuals and between populations using AMOVA analysis of the silicoDArTs (7.9 %), and SNPs (8.3%). SNP and silicoDArT data showed consistency as their association rated at 0.54 significance based on the Mantel test.

4.5 Sequence Similarity

Blasting all 3703 SNP and 8556 silicoDArT sequences revealed a much interesting result. Closely related matchs with e-value greater than 1.0E-11 were matching to *Sesamum indicum, Durio zibethinus, Carica papaya, Arachis duranensis, Erythranthe guttatus, Kolkwitzia amabilis, Solanum pennellii, Hevea brasiliensis, Utricularia reniformis, Hesperelaea palmer, Gossypium trilobum, Mimulus guttatus, Betula pendula, Capsicum annuum, Castilleja paramensis, Boea hygrometrica, Utricularia reniformis, Butomus umbellatus, Vitis vinifera, Primulina liboensis, Crescentia cujete, Lophophytum mirabile, and Tectona grandis.* Most of which are tree, shrub, and herb species. This close similarity with these species suggests that the silicoDArT and SNP markers were of high quality. These blast results were from only 383 SNP markers, and 77 SilicoDArT markers from the total.

DISCUSSION, CONCLUSION AND RECOMMENDATIONS

5.1. Discussion

Commented [HG11]: you're less precise in your argumentation. you need to make comparisons between different populations and individuals. i think that's your objective. you can also show the implications of your study. please show how your company contributes to the conservation of *Kigelia africana*.

It's very important to understand the genetic diversity of indigenous tree species as this will shed some light on their relationships with other plant species, and important genetic and phytochemical potentials they might possess. The DArT platform proved to provide useful information on a never-before genotyped tree species, at an affordable price point. Two types of markers were used for detection, the silicoDArTs and the SNP markers. Both showed high call rate and reproducibility showed reduced genetic diversity, and strong genetic differentiation among other plant species. The high call rates and reproducibility is common among other tree species genotyped using the DArTseq technology, showing their reliability and consistency.

The results from the silicoDArT and SNP markers indicated low genetic variation in *K. africana* with potential consequences on the species ability to recover from human population dynamics, genetic recombination, and environmental effects. Genetic diversity is measured commonly using the proportion of polymorphic loci and patterns of the observed vs expected heterozygosity. This therefore makes the PIC value ranges be described as low ranging from 0.0 to 0.10, medium as 0.10 to 0.25, and high as 0.30 to 0.40, and very high as 0.40 to 0.50. The results showed both silicoDArTs and SNP had PIC ranging between 0.39 to 0.45 and an average of 0.41. This shows high to very high polymorphisms and high informativeness.

Some tree samples were older than others, with at least 30 years age difference, as this is the case with the Kilifi and Nyeri samples. A small insignificant genetic difference was observed between the tree species as seen by the allele frequencies below.

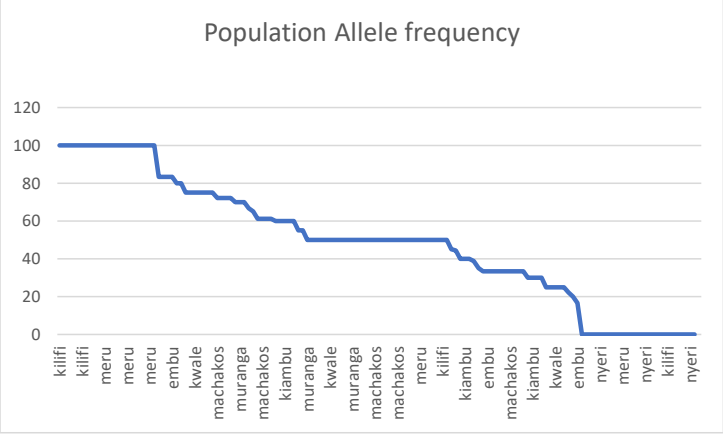

Figure 11: Allele Frequencies of various *K. africana* populations

The high PIC values observed and differences between Ho and He was consistent with the inbreeding coefficient (Fis), where Fis = 0.35 for silicoDArT and -0.33 for SNPs. Positive Fis values are an indication that individuals in a population are more related than expected. And for SNP data having a -0.33 score shows the difference in

detail data extracted, as SNP data is derived from SilicoDArT markers. However, these figures as compared to those from (Nantongo et al., 2022) which was above 0.5. This also shows that the species *K. africana* has not been adversely affected by anthropogenic factors during its existence. Which makes sense, as collection of these samples was often in remote locations with few human presence, hence low genetic diversity erosion. This was also backed up by the almost equal Ho and He values averaging 0.36 for both. When Ho is lower than He, this means there is presence of inbreeding, also supported by the negative inbreeding coefficient (Fis) -0.33 for the SNP data.

The observations from the neighbor joining tree showed that *K. africana* is moderately differentiated forming three distinct clusters. With the SNP clustering more tightly showing more variation as SNP markers are more abundant in plant genomes. This clustering was supported by the genetic differentiation values (Fst) which was below 0.01 showing low genetic differentiation according to (Nantongo et al., 2022). The genetic diversity index (Ht) was high for both SNP and silicoDArT at 0.50 and 0.38 respectively. This is an indication of high genetic diversity of the tree species due to high heterozygosity as shown by the high average MAF of 0.44%. This high heterozygosity maybe due to restricted seed dispersal due to *K. africana' s* reliance on animal vectors to transfer pollen from the flower to the stigma of another different individual as the trees as usually in isolated locations.

From the blast results, we observed closely related plant species that were mostly shrubs, trees, and herbs that at least have their sequence information available. This showed the potential pharmaceutical prospecting opportunities. The relatedness to these biodiverse plants shows potential high biodiversity that is beneficial to the world at large as biodiversity is essential for global food security(ENDEVR, 2023). The next step would be to isolate the genes these species have in common and evaluate their genetic value.

5.2. Conclusion

The study represents the first exploration of the genetic composition of ancient *Kigelia africana* populations in Kenya. The potential diminishing population size of the species despite its high genetic diversity is a threat to its genetic integrity. There is need for *Kigelia africana* germplasm to be collected, characterized, and preserved from different populations across the African continent to maximize genetic variation. The use of DArTseq technology enabled generation of high quality and reliable data for downstream analysis. The use of DartR package for statistical analysis proved to be a friendlier way for DArTseq and SNP data analysis using R software. To explore further the genetic potential of K. africana, and in-depth research needs to be performed relating the genetic constituents and the therapeutical and/or medicinal properties it's said to possess.

Commented [HG12]: you're talking about DNA extraction and analysis methods. don't lose sight of your objective.

ABBREVIATIONS AND ACRONYMS

DArT: **Diversity Arrays Technology**

DArTseq: **Diversity Array Technology sequence**

GBS: **Genotyping by Synthesis**

PCA: **Principal component analysis**

SNP: **Single Nucleotide Polymorphism**

UPGMA: **Unweighted pair group method with arithmetic mean.**

PIC: **Polymorphism Information Content**

AMOVA: **Analysis of Molecular Variance**

References

Adu, B. G., Akromah, R., Amoah, S., Nyadanu, D., Yeboah, A., Aboagye, L. M., Amoah, R. A., & Owusu, E. G. (2021). High-density DArT-based SilicoDArT and SNP markers for genetic diversity and population structure studies in cassava (Manihot esculenta Crantz). *PLOS ONE*, *16*(7), e0255290. https://doi.org/10.1371/journal.pone.0255290

Agyare, C., Obiri, D. D., Boakye, Y. D., & Osafo, N. (2013). Anti-Inflammatory and Analgesic Activities of African Medicinal Plants. *Medicinal Plant Research in Africa: Pharmacology and Chemistry*, 725–752. https://doi.org/10.1016/B978-0-12-405927-6.00019-9

Amiteye, S. (2017). *Basic concepts and methodologies of DNA marker systems in plant molecular breeding*. https://doi.org/10.1016/j.heliyon.2021.e08093

Areces-Berazain, F. (2022). Kigelia africana (sausage tree). *CABI Compendium*, *CABI Compendium*. https://doi.org/10.1079/cabicompendium.29403

Badu-Apraku, B., Luísa Garcia-Oliveira, A., Petroli, D., Hearne, S., Adewale, S. A., & Gedil, M. (n.d.). *Genetic diversity and population structure of early and extra-early maturing maize germplasm adapted to sub-Saharan Africa*. https://doi.org/10.1186/s12870-021-02829-6

Baloch, F. S., Alsaleh, A., Shahid, M. Q., Çiftçi, V., E. Sáenz de Miera, L., Aasim, M., Nadeem, M. A., Aktaş, H., Özkan, H., & Hatipoğlu, R. (2017). A Whole Genome DArTseq and SNP Analysis for Genetic Diversity Assessment in Durum Wheat from Central Fertile Crescent. *PLOS ONE*, *12*(1), e0167821-. https://doi.org/10.1371/journal.pone.0167821

Bedassa, T. (2018). Molecular marker based genetic diversity in forest tree populations. *Forestry Research and Engineering: International Journal*, *2*(4), 176–182. https://doi.org/10.15406/freij.2018.02.00044

Bussmann, R. W., Paniagua-Zambrana, N. Y., & Njoroge, G. N. (2021). *Kigelia africana (Lam.) Benth. Bignoniaceae* (pp. 641–648). https://doi.org/10.1007/978-3-030-38386-2_97

Cai, M., Wen, Y., Uchiyama, K., Onuma, Y., & Tsumura, Y. (2020). Population Genetic Diversity and Structure of Ancient Tree Populations of Cryptomeria japonica var. sinensis Based on RAD-seq Data. *Forests*, *11*(11), 1192. https://doi.org/10.3390/f11111192

Charan, J., & Biswas, T. (2013). How to calculate sample size for different study designs in medical research? *Indian Journal of Psychological Medicine*, *35*(2), 121–126. https://doi.org/10.4103/0253-7176.116232

Dossou-Yovo, H. O., Vodouhè, F. G., Kindomihou, V., & Sinsin, B. (2022). Investigating the Use Profile of Kigelia africana (Lam.) Benth. through Market Survey in Benin. *Conservation*, *2*(2), 275–285. https://doi.org/10.3390/conservation2020019

ENDEVR. (2023, September). *Why Fruits Have Lost Their Vitamins*. ENDEVR Documentary. https://www.youtube.com/watch?v=2H3VhsnyCdI

Joffe Pitta. (2003, August). *Kigelia africana | PlantZAfrica*. http://pza.sanbi.org/kigelia-africana

Kafoutchoni, K., Agoyi, E., Symphorien, A., Assogbadjo, A., & Agbangla, C. (2021). Genetic diversity and population structure in a regional collection of Kersting's groundnut (Macrotyloma geocarpum (Harms) Maréchal & Baudet). *Genetic Resources and Crop Evolution*, *68*, 3285–3300. https://doi.org/10.1007/s10722-021-01187-4

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., Caig, V., Heller-Uszynska, K., Jaccoud, D., Hopper, C., Aschenbrenner-Kilian, M., Evers, M., Peng, K., Cayla, C., Hok, P., & Uszynski, G. (2012). Diversity Arrays Technology: A Generic Genome Profiling Technology on Open Platforms. *Methods in Molecular Biology (Clifton, N.J.)*, *888*, 67–89. https://doi.org/10.1007/978-1-61779-870-2_5

Liu, D., Zhang, L., Wang, J., & Ma, Y. (2020). Conservation Genomics of a Threatened Rhododendron: Contrasting Patterns of Population Structure Revealed From Neutral and Selected SNPs. *Frontiers in Genetics*, *11*, 757. https://www.frontiersin.org/article/10.3389/fgene.2020.00757

Mijangos, J. L., Gruber, B., Berry, O., Pacioni, C., & Georges, A. (2022). dartR v2: An accessible genetic analysis platform for conservation, ecology and agriculture. *Methods in Ecology and Evolution*, *13*(10), 2150–2158. https://doi.org/10.1111/2041-210X.13918

Nabatanzi, A., Nkadimeng, S., Lall, N., Kabasa, J., & Mcgaw, L. (2020). Ethnobotany, Phytochemistry and Pharmacological Activity of Kigelia africana (Lam.) Benth. (Bignoniaceae). *Plants*, *9*, 753. https://doi.org/10.3390/plants9060753

Nantongo, J., Odoi, J., 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*, e0267464. https://doi.org/10.1371/journal.pone.0267464

Padmaja, G. (2009). Uses and Nutritional Data of Sweetpotato. In G. Loebenstein Gad and Thottappilly (Ed.), *The Sweetpotato* (pp. 189–234). Springer Netherlands. https://doi.org/10.1007/978-1-4020-9475-0_11

Pascual, L., Ruiz, M., López-Fernández, M., Pérez-Peña, H., Benavente, E., Vázquez, J. F., Sansaloni, C., & Giraldo, P. (2020). Genomic analysis of Spanish wheat landraces reveals their variability and potential for breeding. *BMC Genomics 2020 21:1*, *21*(1), 1–17. https://doi.org/10.1186/S12864-020-6536-X

Porth, I., & El-Kassaby, Y. (2014). Assessment of the Genetic Diversity in Forest Tree Populations Using Molecular Markers. *Diversity*, *6*, 283–295. https://doi.org/10.3390/d6020283

Saini, S., Kaur, H., Verma, B., & Singh, S. (2008). Kigelia africana (Lam.) Benth. — An overview. *Nat. Prod. Radiance*, *8*(2), 190–197.

Sherwin, W. B., Chao, A., Jost, L., & Smouse, P. E. (2021). Information theory broadens the spectrum of molecular ecology and evolution: (Trends in Ecology and Evolution 32:12, p:948–963, 2017). *Trends in Ecology & Evolution*, *36*(10), 955–956. https://doi.org/https://doi.org/10.1016/j.tree.2021.07.005

Tamokou, J.-D., & Kuete, V. (2014). Toxic Plants Used in African Traditional Medicine. In *Toxicological Survey of African Medicinal Plants* (pp. 135–180). Elsevier. https://doi.org/10.1016/B978-0-12-800018-2.00007-8

Wadl, P. A., Olukolu, B. A., Branham, S. E., Jarret, R. L., Yencho, G. C., & Jackson, D. M. (2018). Genetic Diversity and Population Structure of the USDA Sweetpotato (Ipomoea batatas) Germplasm Collections Using GBSpoly. *Frontiers in Plant Science*, *9*, 1166. https://www.frontiersin.org/article/10.3389/fpls.2018.01166

Wambua Mukavi, J., Wafula Mayeku, P., Muhoro Nyaga, J., & Naulikha Kituyi, S. (2020). In vitro anti-cancer efficacy and phyto-chemical screening of solvent extracts of Kigelia africana (Lam.) Benth. *Heliyon*, *6*(7), e04481–e04481. https://doi.org/10.1016/j.heliyon.2020.e04481

Wikipedia. (2021, September 18). *Kigelia*. Https://En.Wikipedia.Org/Wiki/Kigelia. https://en.wikipedia.org/wiki/Kigelia