Next-generation genomics-assisted characterization of genetic and phenotypic diversity in yams (*Dioscorea* spp) to support conservation and breeding for food security

Gezahegn Girma Tessema

Academic Supervisor: Prof. Charles Spillane, Botany and Plant Science, School of Natural Sciences, National University of Ireland Galway

IITA Research Supervisor: Dr. Melaku Gedil, IITA Biosciences Center, International Institute for Tropical Agriculture (IITA)

A thesis submitted to the

National University of Ireland, Galway

In fulfillment of the requirements for the degree of Doctor of Philosophy

April 2015

Table of Contents

List of Tables

List of Figures

Declaration

I certify that this thesis is my own work, and that I have not used this work in the course of another degree, either at National University of Ireland Galway, or elsewhere.

Signed: _____

Gezahegn Girma Tessema

Acknowledgements

I would like to express my special thanks to my supervisor Prof Charles Spillane, for accepting me in to his Genetics and Biotechnology research group and giving me the opportunity to pursue doctoral research. I am grateful for his encouragement and exemplary guidance throughout the study period. He has also given me the freedom to engage in various projects without objection. I would like to also thank all the members of graduate research committee, Dr. Manash Chatterjee, Dr. Zoe Popper and Dr. Danny Hunter for their comments and suggestions. I am grateful to Dr. Danny Hunter in particular for his support at the start of the project.

I am profoundly indebted to Dr. Melaku Gedil, my research supervisor from IITA, for his advice, support and friendship that has been invaluable on both academic and personal level.

This study would not have been possible without the financial support of the Netherlands Ministry of Foreign Affairs to register for my PhD program. I also recognize the APO fellowship grant from the ministry prior to this study. I acknowledge IITA for extended support to complete my research work. I appreciate the IITA management for the support and the HR unit in particular for facilitating the registration process. I am grateful to the National University of Ireland, Galway for waiving the PhD fee. The CGIAR Research Program on Roots, Tubers and Bananas funded the majority of these research studies.

I acknowledge financial support from Japan International Research Center for Agricultural Sciences (JIRCAS) for the research on Chapter 5. I am grateful to Dr.

Hiroko Takagi (EDITS-Yam Project Leader) in particular for giving me the initial motivation and encouragement that lead to the realization of this work in general and for all the supports during my stay in Japan. My heartfelt thanks are due to Dr. Ryohei Terauchi and his colleagues for helping us to generate the SuperSAGE data in his lab at the Iwate Biotechnology Research Center, Japan. I would like to also thank Dr. Muluneh Tamiru in particular for offering me the hospitability that makes my stay in Iwate more productive and enjoyable.

Several people have been involved in the execution of this research project. Dr. Robert Asiedu has been involved in various stages including the initial planning and editing manuscripts. Prof. Michael Abberton has kindly reviewed most of the manuscripts and he has been always instrumental in giving support for the successful completion of this project. Dr. Sharon Mitchell and Dr. Marc Deletre has generously reviewed and contributed on chapter 3.

Dr. Katie Hyma from institute for Genomic Diversity, Cornell University and Mr. Satoshi Natsume from Iwate Biotechnology Research Center kindly assisted me with Bioinformatic analysis.

My special thanks and appreciation to all the helpful people at IITA for their support. The following people deserve particular mention; Dr. Dominique Dumet, Dr. Abebe Menkir, Dr. Ismail Rabbi, Ms. Temitope Owoeye, Mr. Sunday Ebere, Dr. Antonio Lopez-Montes, Mr. Ayodele Alonge, Ms. Yemi Fasanmade, Mrs Lilian Mendoza and Dr. Yukiko Kashihara. I also express my appreciation for the feedback

and support I get from JIRCAS scientists; Dr. Satoru Muranaka, Dr. Ryo Matsumoto, Dr. Shinsuke Yamanaka and Dr. Pachakkil Babil.

All the research technicians, postgraduate students and staff members of IITA bioscience and genetic resources centers deserve thanks. They have always been willing and happy to support, which made my stay in Nigeria more successful and fruitful.

I am grateful to Professor Alexander Dansi, Dr. Vincent Lebot, Dr. Suchirat sirikul and Dr. William Solano for kindly providing plant materials from their respective institutes. I would like to also thank Dr. Adeniyi Jayeola for his support during yam species identification and sample collection.

I am deeply indebted to my beautiful and caring wife Tsega, my son Nathan and my parents. Their love, prayer and encouragement provided me the energy and motivation to accomplish.

Above all, I would like to thank God whose many blessings have made me who I am today.

Publications

Girma, G., K. Hyma, R. Asiedu, S. Mitchell, M. Gedil and C. Spillane. 2014. Next-Generation sequencing based genotyping, cytometry and phenotyping for understanding diversity and evolution of guinea yams. Theoretical and Applied Genetics 127: 1783-1794. doi:10.1007/s00122-014-2339-2.

Girma, G., Tamiru, M., Natsume, S., Uemura, A., Takagi, H., Gedil, M., Spillane, C. and Terauchi, R. (In Preparation). SuperSAGE-based transcriptome analysis of flowering and sex-related genes in white guinea yam (*D. rotundata* Poir.)

Girma, G., Gedil, M. and Spillane, C. (In Preparation). DNA barcoding of major yam species in the genus *Dioscorea*.

Girma, G., Gedil, M. and Spillane, C. (In Preparation). Morphological, SSR and Ploidy analysis of aerial tuber producing accessions of *D. alata* L. for its potential utilization as planting material.

Acronyms

AFLP- Amplified Fragment Length Polymorphism

AQPs- Aquaporin genes

ASCII- American Standard Code for Information Interchange

BLAST- Basic Local Alignment Search Tool

CATIE- Centro Agronómico Tropical de Investigación y Enseñanza

cDNA- Complementary DNA

DGE- Differential Gene Expression

E-value- Expect value in sequence BLAST

FDR- False Discovery Rate

GBS- Genotyping-By-Sequencing

HEPES- 4-(2-HydroxyEthyl)-1-PiperazineEthaneSulfonic Acid

IGD- Institute for Genomic Diversity

IPGRI-International Plant Genetic Resources Institute/Bioversity International

IITA-International Institute of Tropical Agriculture

ITS- Internal Transcribed Spacer

LOX-Lipoxygenase gene

LogCPM- Logarithmic Count Per Million/gene abundance

LogFC- Fold Change

MP-Maximum Parsimony

MCA- Multiple Correspondence Analysis

MDS-Multi-Dimensional Scaling analysis

matK- Megakaryocyte-Associated Tyrosine Kinase

NAC- from first letters of three different genes (NAM, ATAF and CUC)

NB- Negative Binomial model

NCBI- National Center for Biotechnology Information

NGS- Next Generation Sequencing

nr- Non Redundant

PCoA- Principal Coordinate Analysis

PIC-Polymorphic Information Content

PIF3- Phytochrome-Interacting Transcription

PPL-Percent Polymorphic Loci

QC- Quality Control

RAPD- Random Amplified Polymorphic DNA

rbcL- Ribulose-1, 5-Bisphosphate Carboxylase Large-subunit

RFLPs- Restriction Fragment Length Polymorphisms

RPS4- 40S ribosomal S4

SAGE- Serial Analysis of Gene Expression

SNP- Single Nucleotide Polymorphism

SSR- Simple Sequence Repeat

TK-Transketolase

TMM- Trimmed Mean of M values

UNEAK - Universal Network Enabled Analysis Kit

VPE- Vacular Processing Enzyme

Summary of contents

Yams (*Dioscorea* spp) are staple edible tuber food crop, a favored source of medicinal plants with socio-cultural value in West Africa, Asia, Far East, Oceania and the Caribbean regions. However, the knowledge on the extent of genetic diversity and relationship between the main cultivated *Dioscorea* species and wild relatives, molecular tools to support conventional taxonomic identification and genes responsible for key traits are limited.

In the current project first we have evaluated the performance of candidate DNA barcode regions of flowering plants for distinguishing *Dioscorea* species. Important markers were identified that fulfill better the criteria for good DNA barcode regions in terms of PCR amplification, sequence quality and discriminatory power.

The second research question demonstrated genetic relationship, variation in ploidy level, pattern of heterozygosity and allele sharing, contribution of wild relatives to cultivated species and confirmation on the previous and new suggestions in guinea yams taxonomy utilizing next generation sequencing based genotyping, GBS (genotyping by sequncing), combined with ploidy level and morphological data.

The research project also explored morphological, ploidy and genetic variation across aerial tuber and non-aerial tuber producing *D. alata* accessions. The different patterns in terms of ploidy level, morphological traits and genetic variation revealed across *D. alata* accessions generally higlighted the importance of developing cultivars with aerial tubers in yam improvement programs as a contribution to solve the challenge with yam planting materials.

The fourth research question aimed to analyze transcriptomes in relation to flowering and sex differentiation using high throughput Super-SAGE (Serial Anaysis of Gene Expression) technique and discovered several candidate genes. In addition the study explored variation in flower sex type and other morphological traits across *D. rotundata* genebank accessions.

In general this study presents novel approaches for the improvement of yams (*Dioscorea* spp) genetic resource conservation and breeding.

Chapter 1

1 General Introduction

1.1. Taxonomy and Botany

Yams are members of the genus *Dioscorea* under Dioscoreaceae, a family of monocotyledonous flowering plants. The genus is further grouped into 70 sections, a taxonomic rank below the genus mainly based on vine twining (Sun, et al., 2012) and comprises some 450 species (Govaerts, et al., 2007) to over 600 species (Coursey, 1967) of which 10 are agriculturally important species (Lebot, 2009). These agriculturally important species include; *D. alata* L.*, D. rotundata* Poir.*, D. cayenensis* Lam.*, D. trifida* L.f. *, D. esculenta* (Lour.) Burkill, *D. bulbifera* L.*, D. opposita* Thunb.*, D. pentaphylla* L.*, D. transversa* R.Br. *and D. nummularia* Lam. Some of the morphological diversity characteristics of yams include twining habit, bulbil formation, spininess (Alexander and Coursey, 1969) and size and number of tubers (Lebot, 2009). These were used to group the major food yams into five different sections (Table 1.1).

The *Dioscorea* species are herbaceous climbing monocots. They have perennial vines that produce starchy tubers. However, the yam tuber lacks the anatomical characteristics of a modified stem structure unlike the Irish potato (*Solanum*

tuberosum L.); it has no buds or eyes, no scale leaves and no terminal bud at the distal end of the tuber (Lebot, 2009). Yam stems typically are twining clockwise or anticlockwise. However, there are no specialized organs such as tendrils. The yam stem can be either winged or without wing, spiny or spineless, glabrous or hairy, circular, rectangular or polygonal (Coursey, 1967). The leaves are simple and cordate for some species (e.g. *D. rotundata*) or compound for others, consisting of three-leaf lets (e.g. *D. dumetorum*) or five leaf lets (e.g. *D. pentaphylla*). Each leaf let has three conspicuous veins or ribs joining at the tip. The flowers, arranged in spikes or racemes, are small and generally unisexual although some species are observed to have lines with monoecious flowers. The pollen grains are sticky and cannot be dispersed by wind; small insects are probably the main agents of pollination (Terauchi and Kahl, 1999). Pollen grain viability in yams appears to be low. Fruit and seed set are also very low and usually not more than five viable seeds are produced on one plant (Sadik and Okereke, 1975). The fruits are dry dehiscent, 3-angled or winged capsules, containing six flat, light and winged seeds.

1.2. Origin, Geographic distribution and Domestication of yams

The following three centers of origin are generally agreed upon for the economically important cultivated edible yams (Figure 1.1): Southeast Asia for *D. alata*, *D. esculenta*, *D. opposita* and *D. bulbifera* (Alexander and Coursey, 1969, Burkill, 1951), West Africa for *D. rotundata*, *D. cayenensis*, and *D. dumetorum* (Kunth) Pax (Coursey, 1967) and tropical South America for *D. trifida* (Alexander and Coursey, 1969). Onwueme (1978) considered *D. bulbifera* as a species appearing at the same time in Asia and in Africa, and it is worth noting that *D. bulbifera* is the only yam species with wild populations known from both Africa and Asia (Ramser, et al., 1996, Ramser, et al., 1997).

Yams grow throughout the tropical and subtropical regions of the world. The crop is particularly well adapted to warm, sunny climates with temperatures between 25^0C and 30^0C and require ample moisture. They require deep, loose, textured loamy soil that is rich in organic matter, but they do not tolerate waterlogged conditions (Mignouna, et al., 2009).

Yam domestication has occurred independently within each of the three centers. D. *alata*, considered the most diverse species (Mignouna, et al., 2002), is believed to have originated from spontaneous hybrids between *D. hamiltonii* Hook. f. and *D. persimilis* Prain & Burkill in South-East Asia (Coursey, 1967). During the domestication process, there was an East to West movement of *D. alata* and *D. esculenta*, another Asiatic species now growing widely in Africa and the Americas (Ng, et al., 2007).

Table 1.1: Sections of major cultivated yams within the genus *Dioscorea* based on morphological characteristics.

Section	Species	Major morphological characteristics
Enantiophyllum	*D. alata*	- one to three large tubers
	D. cayenensis	
	D. rotundata	- twin to the right
	D. opposita	- winged stems
	D. japonica	
	D. nummularia	- occasional bulbils
	D. transversa	
Combilium	*D. esculenta*	- large number of individually small tubers,
		- twin to the left
Opsophyton	*D. bulbifera*	- aerial bulbils
		- twin to the left
Macrogynodium	*D. trifida*	- small tubers
		- twin to the left
		- spineless stem
Lasiophyton	*D. hispida*	- cluster of medium-sized tubers
	D. dumetorum	
	D. pentaphylla	- twin to the left
		- large thorns on stems

D. rotundata and *D. cayenensis* are the two most important species in West Africa, although there has been controversy whether they are same or different species. They have both been described as resulting from a process of domestication of wild yams of the section Enantiophyllum (Mignouna and Dansi, 2003). Indeed, the process of yam domestication by farmers is still ongoing in Benin (Mignouna and Dansi, 2003, Zannou, et al., 2004). In West Africa farmers collect wild yam tubers and perform different practices which lead to changes in shape and taste and consider the yams as edible after 2-3 consecutive cycles of planting and harvest (Zannou, et al., 2006). These West African species were taken to the tropical and subtropical Americas and became an important crop in that region, particularly in the Caribbean (Ng, et al., 2007).

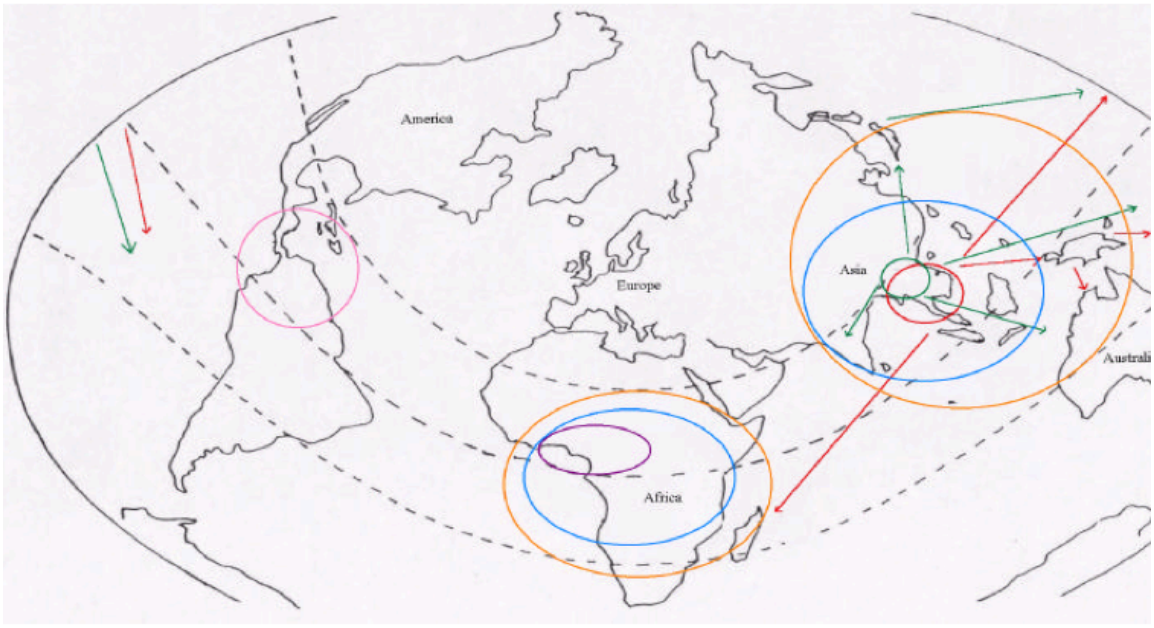


Figure 1.1. Origin and geographic distribution of *Dioscorea* species. Adapted from Degras (1993). Circles represented the origin of eight cultivated *Dioscorea* species with different colors including *D. alata* (red), *D. esculenta* (green), *D. dumetorum* and *D. hispida* (yellow), *D. cayenensis* and *D. rotunda* (purple), *D. trifida* (pink) and *D. bulbifera* (blue). In addition the red and green arrows indicate the direction of distribution of *D. alata* and *D. esculenta* species respectively.

1.3. Importance and Utilization

The cultivated *Dioscorea* species are an economically important staple source of starch in the diet while many of the wild yams are also important plants in times of food scarcity (Bahuchet, et al., 1991, Sato, 2001). Several authors have reported the direct use of wild yams as a food source in West Africa (Bahuchet, et al., 1991, Sato, 2001, Zannou, et al., 2006). Likewise, Lebot (2009) reported the use of wild *Dioscorea*, *D. hispida* Dennst. in times of famine because its large tubers are easily dug and detoxicated by prolonged soaking. The genus is also a favored source of medicinal plants, used to extract precursors of cortisone and other steroid hormones (Kaimal and Kemper, 1999, Martin, 1969, Omoruyi, 2008, Vendl, et al., 2006). In West Africa, yam is cultivated not only for consumption and as a source of income but it has also sociocultural values (Zannou, et al., 2004), being used for traditional religious observances and as social gifts. Akinboro, et al. (2008) has suggested that consumption of yam (*Dioscorea* spp), as a factor that could possibly increase twinning rate in humans as it is believed to contain a natural hormone, phytoestrogen, which may stimulate multiple ovulation.

Yam utilization is mostly as boiled or pounded. Usually fresh yam is peeled, boiled and pounded until sticky elastic dough is produced. This is called pounded yam or yam "fufu". Drying of tubers soon after harvest and converting into slices or milling into flour for pounded yam ensures availability of yam in various forms. Yam is also consumed by roasting of tuber but rare. Yams have not been processed to any

significant extent commercially. The main processed yam product traditionally made at village level is yam flour.

1.4. Cytogenetics of yams

Yam chromosomes are dot like, often clumped together and make chromosome counting difficult. Yams are generally polyploid, where different species and different clones within a species can display different ploidy levels (Table 1.2). It has been proposed that the guinea yam, *D. rotundata* is a tetraploid with a basic chromosome number of 10 (x = 10) (Dansi, et al., 2001, Gamiette, et al., 1999, Obidiegwu, et al., 2009). Hexaploid and octaploid individuals have been reported for *D. cayenensis* based on DNA flow cytometry, using *Solanum lycopersium* L(Obidiegwu, et al., 2009) and tetraploid *D. rotundata* (Dansi, et al., 2001, Gamiette, et al., 1999) as internal standards. However, a study based on segregation patterns of isozyme and microsatellite loci has indicated that *D. rotundata* is diploid (x=20, 2n=40) (Scarcelli, et al., 2005). The flow cytometry histograms for *D. cayenensis-rotundata* were not separated from those of its related wild species (*D. abyssinica*, *D. mangenotiana*, *D. burkilliana* and *D. praehensilis*) (Gamiette, et al., 1999). Likewise, the basic chromosome numbers of the wild guinea yams are expected to be 20. A report based on microsatellite segregation analysis in four different progenies of *D. alata* accessions were demonstrated as diploid, triploid and tetraploid (2n = 2x, 3x, 4x), respectively, and not tetraploid, hexaploid and octoploid, as previously assumed

(Arnau, et al., 2009). A study by Nemorin, et al. (2012) further confirmed the autotetraploid nature of the 2n = 80 clones of *D. alata*.

Similar ploidy studies have been performed for a tropical American species, *D. trifida* which is confirmed to have a basic chromosome number of 20 not 10 (Bousalem, et al., 2006). *D. trifida*, once thought to be octoploid, is now considered to be an auto-tetraploid (Bousalem, et al., 2006). The situation is not yet confirmed for other *Dioscorea* species.

There is considerable variation in genome size among *Dioscorea* species so far reported, ranging from 342 Mbp in *D. dumetorum* (1C=0.35 pg) (Obidiegwu, et al., 2009b) to as high as 6602 Mbp (1C=6.75) for *D. elephantipes* (Zonneveld, et al., 2005). The database of plant genome size developed and maintained by Bennett and Leitch (2012) contains report on DNA C-values for 19 yam species.

Table 1.2. Summary on estimated genome size, chromosome numbers and reported ploidy levels of the most important *Dioscorea* species.

Geographic origin	Species	1C (Mbp)*	1C (pg)*	Chromosome numbers (2n)	Ploidy levels reported
Africa	*D. rotundata*	697	0.71	40,60	2x,4x,6x
	D. cayenensis	748	0.77	18,36,54,60,	2x,3x,6x,8x,14x
		1257	1.29	80,140	
	D. dumetorum	342	0.35	36, 40, 45, 54	2x,3x
	D. bulbifera	1174	1.20	40,60,70,80,100	4x,6x,7x,8x,10x
South East Asia	*D. alata*	562	0.58	40,60,80	2x,3x,4x,6x,8x
	D. esculenta	1027	1.05	40,60,90,100	4x,6x,9x,10x
	D. hispida	NA**	NA	40, 60	4x,6x
	D. pentaphylla	NA	NA	40,70,80,140	4x,7x,8x,14x
Japan	*D. opposita*	NA	NA	40,140	4x,14x
	D. japonica	NA	NA	40	4x
Melanesia	*D. nummularia*	NA	NA	60,80,100,	6x,8x,10x,
				120	12x
	D. transversa	NA	NA	80	8x
Latin America	*D. trifida*	NA	NA	54, 72, 80, 81	4x,8x

*database release 8.0, Dec. 2012, http://www.kew.org/cvalues/(Bennett and Leitch, 2012)

** data not available

Source for ploidy level reports*: (Arnau, et al., 2009, Bousalem, et al., 2006, Gamiette, et al., 1999, Obidiegwu, et al., 2009, Obidiegwu, et al., 2009b, Scarcelli, et al., 2005)*

1.5. Yam germplasm conservation

Several national institutes (representing 20 countries) were reported to have a collection of yam germplasm ranging from 15 accessions in Costa Rica to 1012 accessions in Benin (unpublished survey data). The International Institute of Tropical Agriculture (IITA) comprising more than 3000 accessions (Girma, et al., 2012) maintains the largest yam germplasm collection. However, the IITA genebank germplasm is restricted to collections from West Africa, and includes only 5 of 10 agriculturally important species (Figure 1.2). The conservation method is mainly field genebank. *In vitro* regeneration as tissue culture is also being used to lesser extent (depending on protocols for different genotypes) and the use of cryopreservation has not yet applied for yam long-term conservation.

The major needs in yam germplasm conservation includes minimizing and/or avoiding duplication and mismatches, implementing good conservation practices to reduce pathogen loads in field genebanks, further targeted collections by giving priority to regions/countries not yet addressed, developing molecular tools to explore and better understand yam genetic diversity, and developing protocols for the production of virus-free planting materials.

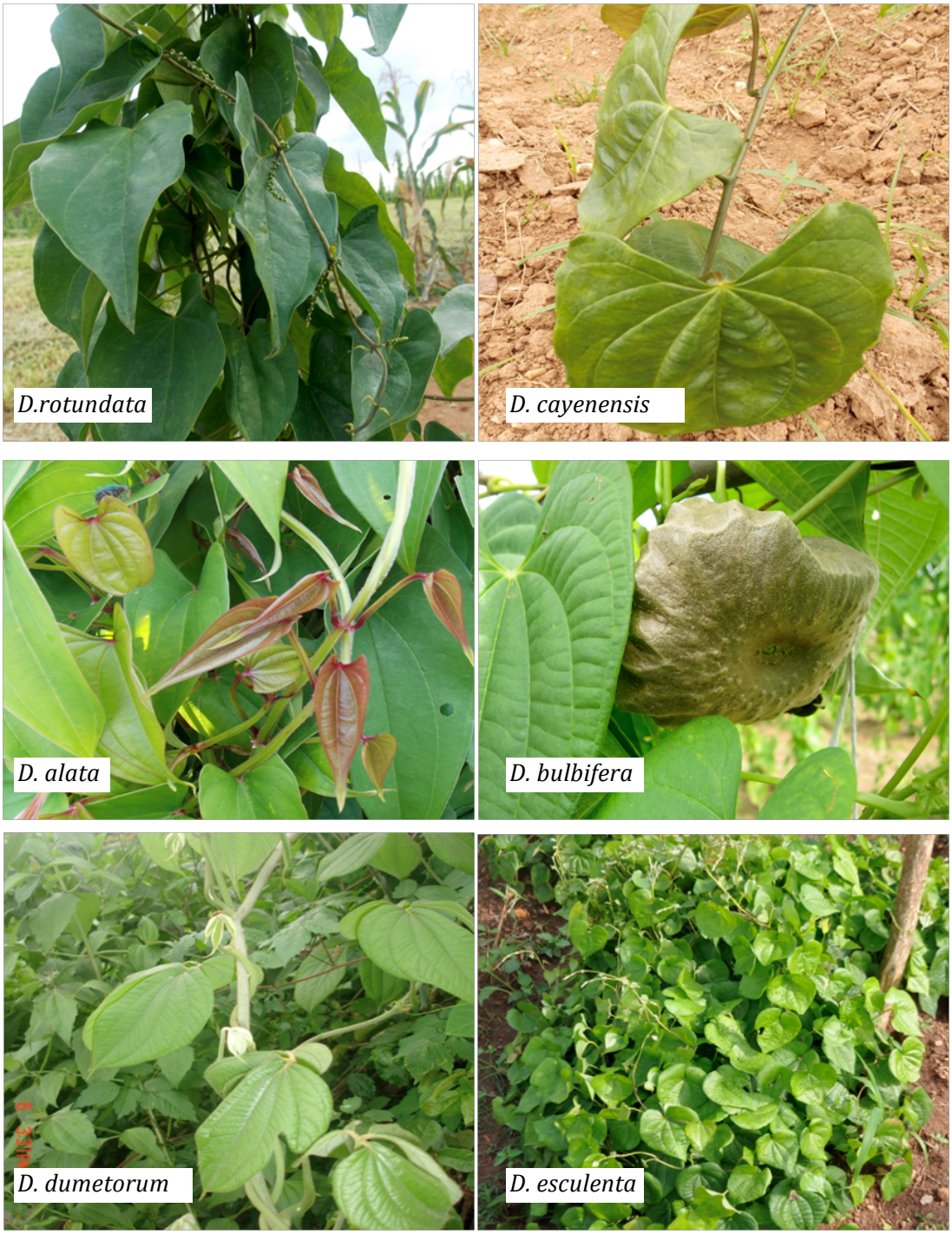


Figure 1.2. Some of the most important *Dioscorea* species, conserved at IITA field genebank. Photo by Gezahegn Girma.

1.6. Yam improvement

The genetic improvement of yams is constrained by several factors including the vegetative propagation, long growing cycle, polyploidy, dioecy, poor to non-flowering and high heterozygosity (Egesi, et al., 2002, Mignouna, et al., 2007).

Yam is primarily a clonally propagated crop. Hence, its production is restricted to underground tubers, the reason being poor botanic seed production and germination. Regardless of inconsistent flowering and poor crossing success, intra-specific hybridization in cultivated species is relatively less challenging. Hence, several varieties have been developed and released (at least for *D. rotundata, D. cayenensis* and *D.alata).*

Inter-specific hybridizations have also been made, with the aim of transferring some of the important traits from wild relatives. *D. rotundata* has been crossed with wild *D. praehensilis* and cultivated *D. cayenensis* (Akoroda, 1985), and there have also been successful interspecific hybridizations between *D. rotundata* and wild relatives (*D. abyssinica* Hochst. & Kunth, *D. togoensis* R. Knuth, and *D. praehensilis)* made at International Institute of Tropical Agriculture (Robert Asiedu, personal communication). However, none of the wild relatives are so far extensively used in yam variety development. No ploidy manipulation (crossing between different ploidy groups) has been reported yet in any of the *Dioscorea* species.

Traits including tolerance and adaptability to moisture stress and low soil fertility, resistance to pathogens (e.g. yam mosaic virus, tuber rots, anthracnose, fungi), pest (e.g. nematodes), insects (e.g. scale insects), tolerance to abiotic stresses (e.g. moisture stress; low soil fertility), and suitability to cropping systems (e.g. plant architecture, vigour, and maturity period) are the most important in yam improvement programs.

1.7. Problem description and research objectives

Yams offer a huge benefit to humankind as an economically and socio-culturally important edible tuber food crop in tropical regions worldwide. However, studies on yam genetic diversity and relationship, ploidy level and the effect of polyploidy and information regarding genes responsible for key traits are limited. Some of the main research problems addressed in the current study followed by research methodologies used are described below and includes 1) lack of molecular systematics efforts to determine the evolutionary relationship of yam species, and there are few molecular tools to support conventional taxonomic identification to understand the taxonomy of yam. DNA barcoding is a taxonomic method that uses a short genetic marker in an organism's DNA to identify it as belonging to a particular species. DNA sequences are increasingly being used in systematics and surveys of biological diversity; both to find clusters that can be called species and to assign new specimens to previously identified species (Hebert, et al., 2003). Organisms divided up into clusters of individuals that are similar to each other and different from

individuals in other clusters. Moreover, members of a cluster of sexual organisms interbreed mainly or exclusively with other members of the same cluster. The clustering is based on not only visible phenotypes but also at the level of DNA sequences that also fall into discrete clusters (Birky, et al., 2010). . These clusters of very similar individual organisms are considered as species (Coyne, 2004) and each species might have a unique DNA sequence, which allows identifying it from all other species. Generally species might have individuals carrying slight variants/polymorphisms of the sequence such as SNPs as a result of mutations in DNA region. When the ancestral species becomes two daughter species, each of them shares genetic variation with one another and with the ancestor and the two daughter species are noticeably different at particular DNA regions. As time goes this will result in fixed genetic differences between species, which provide the way for species identification. However, defining species has been a challenge in systematics (Birky, 2013). Unlike asexuals, the out-crossing nature in sexual organisms creates an additional problem: once speciation has begun to split one species into two, different recombining genes complete their segregation into two populations at different rates. In addition to hybridization the complex evolutionary processes such as polyploidy are common in plants, making species boundaries difficult to define (Fazekas, et al., 2009, Rieseberg, et al., 2006). This problem can be largely addressed by using organelle genes to detect speciation, because the organelle genome is usually not affected by recombination and achieves reciprocal monophyly in about 1/4 the time of the average nuclear gene in diploid sexual organisms (Birky, 1991). In the current study we have followed the most common

DNA barcoding studies that attempts to use DNA sequences to identify species already defined by traditional systematics. Barcode identification of a species is based on empirically determined limits of sequence differences, and is usually not justified by any theory (Birky, 2013).

2) The polyploidy nature of yam, a heritable condition of possessing more than two complete sets of chromosomes, can provide advantages as it could enable polyploid plants to grow in a wide range of environments and can be used as sources of variability for yam improvement. However, not much is known regarding the ploidy level and type of polyploidy, autopolyploidy or allopolyploidy. The former refers to polyploids that arise within a species and the later to those that arise due to the hybridization of two distinct species. Moreover, the effect of increased ploidy level on phenotypic performance across yam species has not been investigated. Increased ploidy levels are known to have link with stress in plants. A study on environmental aridity and polyploidy occurrence in *Brachypodium distachyon* L. showed variation in water use efficiency across ploidy levels, with tetraploids being more efficient in the use of water than diploids under water-restricted growing conditions (Manzaneda, et al., 2012). Similarly, an increase in ploidy level from diploid to triploids was reported to cause alteration of leaf morphology and decrease in stomatal density that leads to considerable reduction in water loss from the leaves of triploids as compared to diploids of *Citrus clementina* Hort.ex. Tan. (Padoan, et al., 2013). Polyploid plants have shown resistance to biotic (pests and pathogens) and abiotic (drought and cold etc.) stress factors in some cases and this resistance enables them to have greater adaptability to wider ecological regions. The higher

chromosome number and gene expression was suggested as a possible cause to increase in the concentration of particular secondary metabolites and chemicals that are responsible for defense mechanism (Yildiz, 2013). A reduction in fertility among increased ploidy, triploids, as compared to diploids individuals of *Miscanthus sinensis* Andersson was reported (Rounsaville, et al., 2011).

Flow cytometry is a high-throughput analytical tool that simultaneously detects and quantifies multiple optical properties (fluorescence, light scatter) of single particles, usually cells or nuclei labeled with fluorescent probes, as they move in a narrow liquid stream through a powerful beam of light. Ease of sample preparation, reliability and high sample throughput make flow cytometry using DNA-selective fluorochromes as the method of choice and better suited than other methods such as Feulgen densitometry to estimate genome size, level of generative polyploidy, nuclear replication state and endopolyploidy (polysomaty) (Dolezel, et al., 2007).

3) The extent of genetic diversity and relationship between the main cultivated *Dioscorea* species and wild relatives has not been well investigated. Recent progress in high-throughput sequencing technologies has revolutionized the field of genomics, creating possibility to generate large amounts of sequence data very rapidly, accurately and at a substantially lower cost. Next-Generation Sequencing (NGS) based genotyping procedures such as Genotyping-By-Sequencing (GBS) represent high-marker density approaches, which can help reveal the extent of genetic relatedness and genetic variation within and between cultivated species and their wild relatives (Spindel, et al., 2013). The GBS approach is based on reducing genome complexity with restriction enzymes, coupled with multiplex NGS for high-

density single nucleotide polymorphism (SNP) markers discovery (Elshire, et al., 2011). The genome-wide molecular marker discovery, highly multiplexed genotyping, flexibility and low cost of GBS makes it an excellent tool in plant genetics and breeding (Deschamps, et al., 2012, Poland and Rife, 2012).

The development of a robust SNP calling pipeline, Universal Network Enabled Analysis Kit (UNEAK) (Lu, et al., 2013) facilitates the use of GBS for genomic diversity and genetic relationship studies in polyploid and species that lack a reference genome sequence, such as guinea yams, but its reliability in SNPs calling remains to be determined. The non-reference UNEAK pipeline developed for SNP markers discovery and genotyping was used in the current study as described by Lu et al (2013). Illumina Qseq or Fastq files were used as the inputs of UNEAK. All of the reads were computationally trimmed to 64 bp. Identical reads were classified as a tag. Pairwise alignment was performed to find tag pairs differing by only a single bp mismatch. Tag networks were built and reciprocal, real tag pairs were retained as SNPs.

Maximum parsimony analysis and calculation of nucleotide distances (substitution rates per site) between and within groups are the most important methodologies for understanding the genetic relationship between populations and individuals within populations. Phylogenetics is the study of evolutionary relationships among organisms or genes. The purposes of phylogenetic studies are mainly to reconstruct evolutionary ties between organisms and to estimate the time of divergence between organisms since they last shared a common ancestor. Phylogenetic tree

construction methods are mainly grouped into two categories: distance based and character based. The most common distance based methods are the Unwieghted Pair Group Method using Arithmetic Mean (UPGMA) (Sneath P.H.A. and Sokal, 1973) and Neighbor Joining (Saitou and Nei, 1987) algorithms that are based on the initial creation of a distance matrix. The second category is the character-based method also used in the current study is maximum parsimony (Fitch, 1971), which take a probabilistic approach to tree construction, and searches all possible tree topologies for the optimal tree.

Multi Dimensional Scaling (MDS) and Principal Components Analysis (PCA) are both grouping techniques that are classified as dimensioning technique. These techniques produce low dimensional plots in which the individuals are spread according to their relatedness. MDS is one of the most commonly used multivariate techniques in evolutionary relationship studies (Pelé, et al., 2011). MDS is a means of visualizing or exploring individual and/or group differences or the level of similarities/dissimilarities of individual cases of a dataset. MDS take high-dimensional vectors and map them down to two- or three-dimensional vectors, trying to preserve all the relevant distances. In summary the idea is that we start with vectors $v_1, v_2 \ldots v_n$ in a p-dimensional space, where p is large, and we want to find new vectors $x_1, x_2 \ldots x_n$ in R^2 or R^3 such that

$$\sum_{i=1}^{n} \sum_{j \neq i} (\delta(v_1, v_2) - d(x_1, x_2))^2$$

is as small as possible, where δ is distance in the original space and d is Euclidean distance in the new space.

PCA is another way to visualize relationships among individuals. In PCA the principal components are found by calculating the eigenvectors and eigen values of the data covariance matrix. This process is equivalent to finding the axis system in which the co-variance matrix is diagonal. The eigenvector with the largest eigen value is the direction of greatest variation; the one with the second largest eigen value is the (orthogonal) direction with the next highest variation and so on. The PCA computation steps involves transforming an $N \times d$ matrix X into an $N \times m$ matrix Y by centralizing the data (subtract the mean), followed by calculating the $d \times d$ covariance matrix:

$$C_{i,j} = \frac{1}{N-1}\sum_{q=1}^{N} X_{q,i}.X_{q,j}$$

$C_{i,i}$ (diagonal) is the variance of variable i.

$C_{i,j}$ (off-diagonal) is the covariance between variables i and j.

and calculating the eigenvectors of the covariance matrix and selecting m eigenvectors that correspond to the largest m eigen values to be the new basis.

In a given covariance matrix A, anon-zero vector v is an eigenvector of A if there is a scalar λ (eigen value) such that

$$Av=\lambda v$$

The equation is also called characteristic equation and has *n* roots. Roots are eigen values and corresponding eigen vectors are principal components. First principal component is the eigen vector associated with the largest eigen value of *A*.

4) The poor flowering, seed production and germination of cultivated yams (Lebot, 2009, Mignouna, et al., 2007) restricts farm-level production to clonal propagation (Scarcelli, et al., 2013). The guinea yam, *D. rotundata* also commonly called white yam, is the most preferred and predominantly cultivated yam species in West Africa (Scarcelli, et al., 2011). However, the dioecy and "poor to non-flowering" nature of the crop are among the main constraints limiting its genetic improvement (Mignouna, et al., 2007). Very little is known regarding what genes are responsible for key traits in yams. For instance, the molecular mechanism underlying flowering patterns in yam is not understood.

A number of sequencing based transcriptome comparisons have been used as important tools for gene expression profiling, novel gene discovery, and genome annotation studies. Commonly used techniques include the Serial Analysis of Gene Expression (SAGE) method, which is based on the isolation of unique short sequence tags (14–15 bp) (Velculescu, et al., 1995), Long SAGE that uses a different type IIS enzyme, MmeI, which releases 21-bp fragments from each transcripts (Saha, et al., 2002), Robust-LongSAGE (RL-SAGE) (Gowda, et al., 2004), Expressed Sequence Tag Analysis (EST) (Nielsen, et al., 2006) Digital Gene Expression TAG (DGE-TAG), DeepSAGE (Nielsen, et al., 2006) and RNA-Seq (Marioni, et al., 2008).

The high throughput SuperSAGE (Serial Analysis of Gene Expression) technique that involves sequencing of longer fragments (26-bp) and simultaneous analysis of multiple samples by using indexing (barcoding) has been indicated to have an advantage over other techniques based on next generation sequencing (such as DGE-TAG that provides a relatively short tag reads (21-bp) which sometimes create tag-to-gene annotation more difficult) and RNA-Seq that requires a large amount of sequence reads to fully cover the dynamic range and to provide a truly quantitative gene expression profiling (Matsumura, et al., 2010). The present study represents a first attempt to identify sex-related genes in white Guinea yam (*D. rotundata*) based on SuperSAGE (serial analysis of gene expression) analysis of male, female and monoecious accessions.

5) Unavailability of planting material is another challenges in yam production despite the increasing demand for local consumption. In addition to its wide adaptation and cultivation, *D. alata* is known to produce aerial tubers in some accessions, which can serve as an alternative source to planting material. The Study was conducted to investigate the molecular, morphological and ploidy variation across *D. alata* accessions producing aerial tubers, potential and alternative planting material, including accessions without aerial tubers.

Overall, the aim of this PhD research was to: (1) identify chloroplast or nuclear regions that can help to identify yam species, (2) investigate the genetic diversity and population structure of cultivated guinea yams and their relationship with its

wild relatives utilizing Genotyping By Sequencing (GBS), cytometry and morphology; (3) investigate the molecular genetics of flowering in yams and (4) conduct a morphological and molecular diversity study of *D. alata* with desirable traits (producing aerial tuber in addition to underground tuber).

1.8. References

Akinboro, A., M. Azeez and A. Bakare. 2008 Frequency of twinning in southwest Nigeria. Indian J Hum Genet. 14: 41-47.

Akoroda, M.O. 1985. Pollination management for controlled hybridization of white yam. Scientia Horticulturae. 25: 201-209.

Alexander, J. and D. Coursey. 1969. The origins of yam cultivation. In: Ucko PJ and Dimbleby GW, editors, The domestication and exploitation of plants. London: Duckworth. p. 405–425.

Alexander, J. and D.G. Coursey. 1969. The origins of yam cultivation. In: The domestication and exploitation of plants and animals. In: G. W. D. Edited by: P.J. Ucko, editor In: The domestication and exploitation of plants and animals. Aldine Publishing Company., Chicago. p. 405-425.

Arnau, G., A. Nemorin, E. Maledon and K. Abraham. 2009. Revision of ploidy status of *Dioscorea alata* L. (Dioscoreaceae) by cytogenetic and microsatellite segregation analysis. Theoretical and applied genetics. 118: 1239-1249. doi:10.1007/s00122-009-0977-6.

Bahuchet, S., D. McKey and I. Garine. 1991. Wild yams revisited: Is independence from agriculture possible for rain forest hunter-gatherers? Human Ecology 19: 213-243. doi:10.1007/BF00888746.

Bennett, M. and I. Leitch. 2012. Angiosperm DNA C-values database(release 8.0, Dec. 2012) http://www.kew.org/cvalues/.

Birky, C.W., Jr. 2013. Species Detection and Identification in Sexual Organisms Using Population Genetic Theory and DNA Sequences. PLoS ONE 8: e52544. doi:10.1371/journal.pone.0052544.

Birky, C.W., Jr., J. Adams, M. Gemmel and J. Perry. 2010. Using Population Genetic Theory and DNA Sequences for Species Detection and Identification in Asexual Organisms. PLoS ONE 5: e10609. doi:10.1371/journal.pone.0010609.

Birky CWJ (1991) Evolution and population genetics of organelle genes: Mechanisms and models. In: Selander RK, Clark AG, Whittam TS, editors. Evolution at the Molecular Level. Sunderland, MA: Sinauer Associates, Inc. pp. 112–134.

Bousalem, M., G. Arnau, I. Hochu, R. Arnolin, V. Viader, S. Santoni, et al. 2006. Microsatellite segregation analysis and cytogenetic evidence for tetrasomic inheritance in the American yam *Dioscorea trifida* and a new basic chromosome number in the Dioscoreaceae. Theoretical and applied genetics 113: 439-451. doi:10.1007/s00122-006-0309-z.

Burkill, I.H. 1951. Dioscoreaceae. Flora Malesiana Ser. 4: 293-335.

Coursey, D.G. 1967. Yams, An account for the Nature, Origins, Cultivation and Utilization of the Useful Members of the Dioscoreaceae.Longmans, Green and Co. Ltd. Londers, UK., Londers, UK.

Coyne JA, Orr HA (2004) Speciation. Sunderland, Massachusetts: Sinauer Associates, Inc.

Dansi, A., H.D. Mignouna, M. Pillay and S. Zok. 2001. Ploidy variation in the cultivated yams (*Dioscorea cayenensis-Dioscorea rotundata* complex) from Cameroon as determined by flow cytometry. Euphytica 119: 301-307. doi:10.1023/A:1017510402681.

Degras, L. 1993. The yam: A Tropical Root Crop. The Macmillan Press Ltd., London and Basingstoke, UK. pp.408.

Deschamps, S., V. Llaca and G.D. May. 2012. Genotyping-by-Sequencing in Plants. Biology 1: 460-483.

Dolezel, J., J. Greilhuber and J. Suda. 2007. Estimation of nuclear DNA content in plants using flow cytometry. Nat. Protocols 2: 2233-2244.

Egesi, C.N., M. Pillay, R. Asiedu and J.K. Egunjobi. 2002. Ploidy analysis in water yam, *Dioscorea alata* L. germplasm. Euphytica 128: 225-230. doi:10.1023/A:1020868218902.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, et al. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6: e19379. doi:10.1371/journal.pone.0019379.

Fazekas, A.J., P.R. Kesanakurti, K.S. Burgess, D.M. Percy, S.W. Graham, S.C. Barrett, et al. 2009. Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? Molecular ecology resources 9 Suppl s1: 130-139. doi:10.1111/j.1755-0998.2009.02652.x.

Fitch, W. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Syst. Zool 20: 406-416.

Gamiette, F., F. Bakry and G. Ano. 1999. Ploidy determination of some yam species (*Dioscorea* spp.) by flow cytometry and conventional chromosomes counting. Genetic Resources and Crop Evolution 46: 19-27. doi:10.1023/A:1008649130567.

Girma, G., S. Korie, D. Dumet and J. Franco. 2012. Improvement of accession distinctiveness as an added value to the global worth of the yam (*Dioscorea* spp) genebank. International Journal of Conservation Science 3: 199-206.

Govaerts, R., P. Wilkin and R.M.K. Saunders. 2007. World Checklist of Dioscoreales. Yams and their allies. The Board of Trustees of the Royal Botanic Gardens, Kew. . p. 1-65.

Gowda, M., C. Jantasuriyarat, R.A. Dean and G.L. Wang. 2004. Robust-LongSAGE (RL-SAGE): a substantially improved LongSAGE method for gene discovery and transcriptome analysis. Plant Physiol 134: 890-897.

Kaimal A. and Kemper KJ. 1999. Wild yam (Dioscoreaceae). http://www.mcp.edu/herbal/default.htm

Lebot, V. 2009. Tropical root and tuber crops: cassava, sweet potato, yams and aroids. CABI Publishers, Wallingford,UK: CABI pp. 413.

Lu, F., A.E. Lipka, J. Glaubitz, R. Elshire, J.H. Cherney, M.D. Casler, et al. 2013. Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. PLoS Genet 9: e1003215. doi:10.1371/journal.pgen.1003215.

Manzaneda, A.J., P.J. Rey, J.M. Bastida, C. Weiss-Lehman, E. Raskin and T. Mitchell-Olds. 2012. Environmental aridity is associated with cytotype segregation and polyploidy occurrence in Brachypodium distachyon (Poaceae). The New phytologist 193: 797-805. doi:10.1111/j.1469-8137.2011.03988.x.

Martin, F.W. 1969. The Species of *Dioscorea* Containing Sapogenin. Economic Botany 23: 373-379.

Marioni, J.C., C.E. Mason, S.M. Mane, M. Stephens and Y. Gilad. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18: 1509-1517. doi:10.1101/gr.079558.108.

Matsumura, H., K. Yoshida, S. Luo, E. Kimura, T. Fujibe, Z. Albertyn, et al. 2010. High-Throughput SuperSAGE for Digital Gene Expression Analysis of Multiple Samples Using Next Generation Sequencing. PLoS ONE 5: e12010. doi:10.1371/journal.pone.0012010.

Mignouna, H., M. Abang and R. Asiedu. 2007. Yams. In: C. Kole, editor Genome mapping and molecular breeding Pulses, Sugar and Tuber Crops. Springer, Heidelberg, Berlin, New York, Tokyo. p. 271–296.

Mignouna, H.D., M.M. Abang, R. Asiedu and R. Geeta. 2009. Yam (*Dioscorea*) husbandry: cultivating yams in the field or greenhouse. Cold Spring Harbor protocols 2009: pdb prot5324. doi:10.1101/pdb.prot5324.

Mignouna, H.D., M.M. Abang, A. Onasanya and R. Asiedu. 2002. Identification and application of RAPD markers for anthracnose resistance in water yam (*Dioscorea alata*). Annals of Applied Biology 141: 61-66. doi:10.1111/j.1744-7348.2002.tb00195.x.

Mignouna, H.D. and A. Dansi. 2003. Yam (*Dioscorea* ssp.) domestication by the Nago and Fon ethnic groups in Benin. Genetic Resources and Crop Evolution 50: 519-528. doi:10.1023/A:1023990618128.

Nemorin, A., K. Abraham, J. David and G. Arnau. 2012. Inheritance pattern of tetraploid *Dioscorea alata* and evidence of double reduction using microsatellite marker segregation analysis. Mol Breeding 30: 1657-1667. doi:10.1007/s11032-012-9749-0.

Ng, S.Y.C., S.H. Mantell and N.Q. Ng. 2007. Biotechnology in Germplasm Management of Cassava and Yams. In: E. E. Benson, editor Plant Conservation Biotechnology. Taylor and Francis CRC ebook account. p. 179-187.

Nielsen, K.L., A.L. Hogh and J. Emmersen. 2006. DeepSAGE--digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. Nucleic Acids Res 34: e133. doi:10.1093/nar/gkl714.

Obidiegwu, J., J. Loureiro, E. Ene-Obong, E. Rodriguez, M. Kolesnikova-Allen, C. Santos, et al. 2009. Ploidy level studies on the *Dioscorea cayenensis/Dioscorea rotundata* complex core set. Euphytica 169: 319-326.

Obidiegwu, J., E. Rodriguez, E. Ene-obong, J. Loureiro, C. Muoneke, C. Santos, et al. 2009b. Estimation of the nuclear DNA content in some representative of genus *Dioscorea*. Scientific Research and Essay 4 448-452.

Omoruyi, F.O. 2008. Jamaican bitter yam sapogenin: potential mechanisms of action in diabetes. Plant foods for human nutrition 63: 135-140. doi:10.1007/s11130-008-0082-z.

Onwueme, I.C. 1978. The tropical tuber crops : yams, cassava, sweet potato, and cocoyamsWiley, Chichester;Toronto;New York, pp.234.

Padoan, D., A. Mossad, B. Chiancone, M.A. Germana and P.S.S.V. Khan. 2013. Ploidy levels in Citrus clementine affects leaf morphology, stomatal density and water content. Theoretical and Experimental Plant Physiology 25: 283-290.

Pelé, J., H. Abdi, M. Moreau, D. Thybert and M. Chabbert. 2011. Multidimensional Scaling Reveals the Main Evolutionary Pathways of Class A G-Protein-Coupled Receptors. PLoS ONE 6: e19094. doi:10.1371/journal.pone.0019094.

Poland, J.A. and T.W. Rife. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Gen. 5: 92-102.

Ramser, J., K. Weising, G. Kahl, C. Lopez-Peralta and R. Wetzel. 1996. Genomic variation and relationships in aerial yam (*Dioscorea bulbifera* L.) detected by random amplified polymorphic DNA. Genome 39: 17-25.

Ramser, J., K. Weising, R. Terauchi, G. Kahl, C. Lopez-Peralta and W. Terhalle. 1997. Molecular marker based taxonomy and phylogeny of Guinea yam (*Dioscorea rotundata - D. cayenensis*). Genome 40: 903-915.

Rieseberg, L.H., T.E. Wood and E.J. Baack. 2006. The nature of plant species. Nature 440:524-527.

Rounsaville, T., D. Touchell and T. Ranney. 2011. Fertility and Reproductive Pathways in Diploid and Triploid *Miscanthus sinensis*. HortScience 46: 1353-1357.

Sadik, S. and O. Okereke. 1975. Flowering, Pollen Grain Gerrnina-tion, Fruiting, Seed Germination and Seedling Development of White Yam, *Dioscorea rotundata* Poir. Ann Bot 39: 597-604.

Saha, S., A.B. Sparks, C. Rago, V. Akmaev, C.J. Wang, B. Vogelstein, et al. 2002. Using the transcriptome to annotate the genome. Nat Biotech 20: 508-512. doi:http://www.nature.com/nbt/journal/v20/n5/suppinfo/nbt0502-508_S1.html.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.

Sato, H. 2001. The potential of edible yams and yam like plants as a staple food resource in the African tropical rain forest. African Study Monographs, Suppl. 26: 123-134.

Scarcelli, N., O. Dainou, C. Agbangla, S. Tostain and J.L. Pham. 2005. Segregation patterns of isozyme loci and microsatellite markers show the diploidy of African yam *Dioscorea rotundata* (2n = 40). Theoretical and applied genetics 111: 226-232. doi:10.1007/s00122-005-2003-y.

Scarcelli, N., A. Barnaud, W. Eiserhardt, U.A. Treier, M. Seveno, A. d'Anfray, et al. 2011. A Set of 100 Chloroplast DNA Primer Pairs to Study Population Genetics and Phylogeny in Monocotyledons. PLoS ONE 6: e19954. doi:10.1371/journal.pone.0019954.

Scarcelli, N., M. Couderc, M. Baco, J. Egah and Y. Vigouroux. 2013. Clonal diversity and estimation of relative clone age: application to agrobiodiversity of yam (*Dioscorea rotundata*). BMC Plant Biology 13: 178.

Sneath P.H.A. and Sokal, R.R. 1973. Numerical taxonomy. The principles and practice of numerical taxonomy, W. H. Freeman & Co, San Francisco, 513 pp.

Spindel, J., M. Wright, C. Chen, J. Cobb, J. Gage, S. Harrington, et al. 2013. Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density

SNP markers and new value to traditional bi-parental mapping and breeding populations. Theoretical and Applied Genetics 126: 2699-2716.

Srivastava, A.K., T. Gaiser, H. Paeth and F. Ewert. 2012. The impact of climate change on Yam (*Dioscorea alata)* yield in the savanna zone of West Africa. Agriculture, Ecosystems & Environment 153: 57-64.

Sun, X.Q., Y.J. Zhu, J.L. Guo, B. Peng, M.M. Bai and Y.Y. Hang. 2012. DNA barcoding the *Dioscorea* in China, a vital group in the evolution of monocotyledon: use of matK gene for species discrimination. PLoS One 7: e32057. doi:10.1371/journal.pone.0032057.

Terauchi, R. and G. Kahl. 1999. Sex determination in *Dioscorea tokoro*, a wild yam species. In: C. Ainsworth, editor Sex Determination in Plants. BIOS, Oxford OX4 1RE, UK.

Velculescu, V., L. Zhang, B. Vogelstein and K. Kinzler. 1995. Serial analysis of gene expression. Science 270: 484 - 487.

Vendl, O., C. Wawrosch, C. Noe, C. Molina, G. Kahl and B. Kopp. 2006. Diosgenin contents and DNA fingerprint screening of various yam (*Dioscorea* sp.) genotypes. Journal of biosciences 61: 847-855.

Mustafa Yildiz (2013). Plant Responses at Different Ploidy Levels, Current Progress in Biological Research, Dr. Marina Silva-Opps (Ed.), ISBN: 978-953-51-1097-2, InTech, DOI: 10.5772/55785.

Zannou, A., A. Ahanchede, P. Struik, P. Richard, J. Zoundjihekpon, R. Tossou, et al. 2004. Yam and cowpea diversity management by farmers in the Guinea-Sudan transition zone of Benin. NJAS 52: 393-420.

Zannou, A., P. Richards and P. Struik. 2006. Knowledge on yam variety development: insights from farmers' and researchers' practices. Knowledge Management for Development Journal 2: 30-39.

Zonneveld, B.J., I.J. Leitch and M.D. Bennett. 2005. First nuclear DNA amounts in more than 300 angiosperms. Ann Bot 96: 229-244. doi:10.1093/aob/mci170.

Chapter 2

2. DNA barcoding of major yam species in the genus *Dioscorea*

2.1 Background and Justification

Yams (*Dioscorea* spp.) belong to the monocotyledons within the family Dioscoreaceae of flowering plants. *Dioscorea* is the largest genus comprising some 450 species (Govaerts, et al., 2007) to over 600 species (Coursey, 1967) of which 10 are staple yams (Lebot, 2009). These species are mainly found in tropical or subtropical regions of the world. Species differentiation is currently exclusively based on morphological descriptors where ambiguities are inevitable. Moreover, the difficulty to find reliable and stable morphological traits limits species discrimination.

Morphological characters are still the most applied and used characteristics for identification and taxonomy of yams. However, for yams and other species morphology-based procedures have several disadvantages including considerable morphological plasticity between organisms of the same species and these methods are hampered by the existence of convergent evolution, in which the same

phenotypic feature can emerge independently in phylogenetically unrelated organisms (Pereira, et al., 2008). Moreover, morphology-based approaches require assessment of whole plants and its usefulness diminishes when specimens/tissues such as leaf sample, tuber, *in vitro* materials and etc are dealt with. It is even more challenging in crops like yam that require more than 6 months of growth cycle on the field.

Developing molecular tools supported by taxonomic identification is very important for unambiguous species naming or classification. A DNA barcode is an aid to taxonomic identification, which uses a standard short genomic region that is universally present in target lineages and has sufficient sequence variation to discriminate among species (Kress and Erickson, 2007). DNA bar-coding techniques are a useful tool for taxonomists as it allows objective specimens identification more quickly and cheaply and provides a central catalog of species diversity, which can be accessed by anyone. Hence, improves biodiversity databases (Miller, 2007, Schaefer and Strimmer, 2005).

A variety of loci have been suggested as DNA bar codes for plants. Kress, et al. (2005) described three criteria that must be satisfied when evaluating genetic loci appropriate for plant DNA barcoding: (i) significant species- level genetic variability and divergence, (ii) an appropriately short sequence length so as to facilitate DNA extraction and amplification, and (iii) the presence of conserved flanking sites for developing universal

primers. In addition the following key criteria was indicated for loci selection; a) sequence quality: the number of positions at, or above, a user defined quality threshold (Little, 2010). Sequencing quality metrics such as PHRED quality score (Q score) (Cock, et al., 2010) can provide important information about the accuracy of base calling and it is the quality-scoring standard for different sequencing technologies. It indicates the probability that the sequencer calls a given base incorrectly by assigning a Q score represented as American Standard Code for Information Interchange (ASCII) characters to a base, which is equivalent to the probability of the number of times an incorrect base is called. The quality value Q assigned to a base-call was defined as a property that is logarithmically related to the base calling error probabilities (p)

$$Q = -10 \log_{10} p$$

where p is the estimated error probability for that base-call (Ewing and Green, 1998). The higher the PHRED quality scores the higher base call accuracy and high quality values correspond to low error probabilities, and conversely. The process of generating a PHRED quality-scoring scheme is largely the same in next-generation sequencing and Sanger sequencing. A Q score of 30 (Q30) assigned by PHRED to a base, also considered a benchmark for quality in next-generation sequencing (Trivedi, et al., 2014) is equivalent to the probability of an incorrect base call 1 in 1000 times which means the probability of a correct base call is 99.9%. When sequencing quality reaches Q30, virtually all of the reads will be perfect, having zero errors and ambiguities. A lower base call accuracy of 99% (Q20) will have an

incorrect base call probability of 1 in 100, which means that every 100 bp sequencing read will likely contain an error. By comparison, Sanger sequencing systems generally produce base call accuracy of ~99.4%, or ~Q20 (http://www.xcelrisgenomics.com/PDF/XcelSeq/XcelSeqBrochure.pdf). Sequencing data with low Q scores can increase false-positive variant calls, which can result in inaccurate conclusions, wasted time and expense. b) universality: the suitability of loci for routine sequencing across different plant species and c) discrimination: the capacity of loci to distinguish species (Hollingsworth, et al., 2009). For instance, the chloroplast genome; rbcL (Ribulose-1, 5-Bisphosphate *Carboxylase Large-subunit*) (Blaxter, et al., 2005), *matK* (*Megakaryocyte-Associated Tyrosine Kinase*) (Janzen, et al., 2009, Selvaraj, et al., 2008), 2-locus combination of *rbcL* + *matK* (Hollingsworth, et al., 2009), non-coding plastid *trnH-psbA* intergenic spacer region (Pang, et al., 2012) and non- coding *trnH-psbA* paired with one of the coding loci, *rbcL* (Kress and Erickson, 2007, Kress, et al., 2005) are candidate barcode regions proposed by different authors. From the nuclear genome, *ITS /ITS1* and *ITS2* (Chen, et al., 2010, Gao, et al., 2010) were considered as other leading candidates as universal plant barcodes (Figure 2.1). A recent report by Sun, et al. (2012) on DNA barcoding of the *Dioscorea* species from China indicated *matK* as a potential barcode region for species identification based on the inter-specific divergence revealed. However, the study was not extensive enough in terms of addressing diverse *Dioscorea* species, as it is limited to only Asian origin. Moreover, only few agriculturally important species was included.

In the present PhD study, the candidate DNA barcode regions including *rbcL, matK*, *trnH-psbA*, *ITS* and the combination of *rbcL* and *matK* regions were evaluated for PCR amplification, sequence quality and discriminatory power among yam species. The main objective of the study was therefore to test and validate existing plant DNA barcodes in main *Dioscorea* species and its wild relatives.

2.2 Materials and Methods

2.2.1 Plant materials and DNA isolation

A total of 69 individuals of 21 different *Dioscorea* species identified and maintained by different genebank institutes representing all main cultivated species and close wild relatives were used in this study (Table 2.1). Genomic DNA was extracted using Qiagen DNeasy plant mini kit following manufacturers protocol with minor modification that involves pre-washing before starting the DNA extraction using HEPES buffer to remove secondary compounds/polysaccharides.

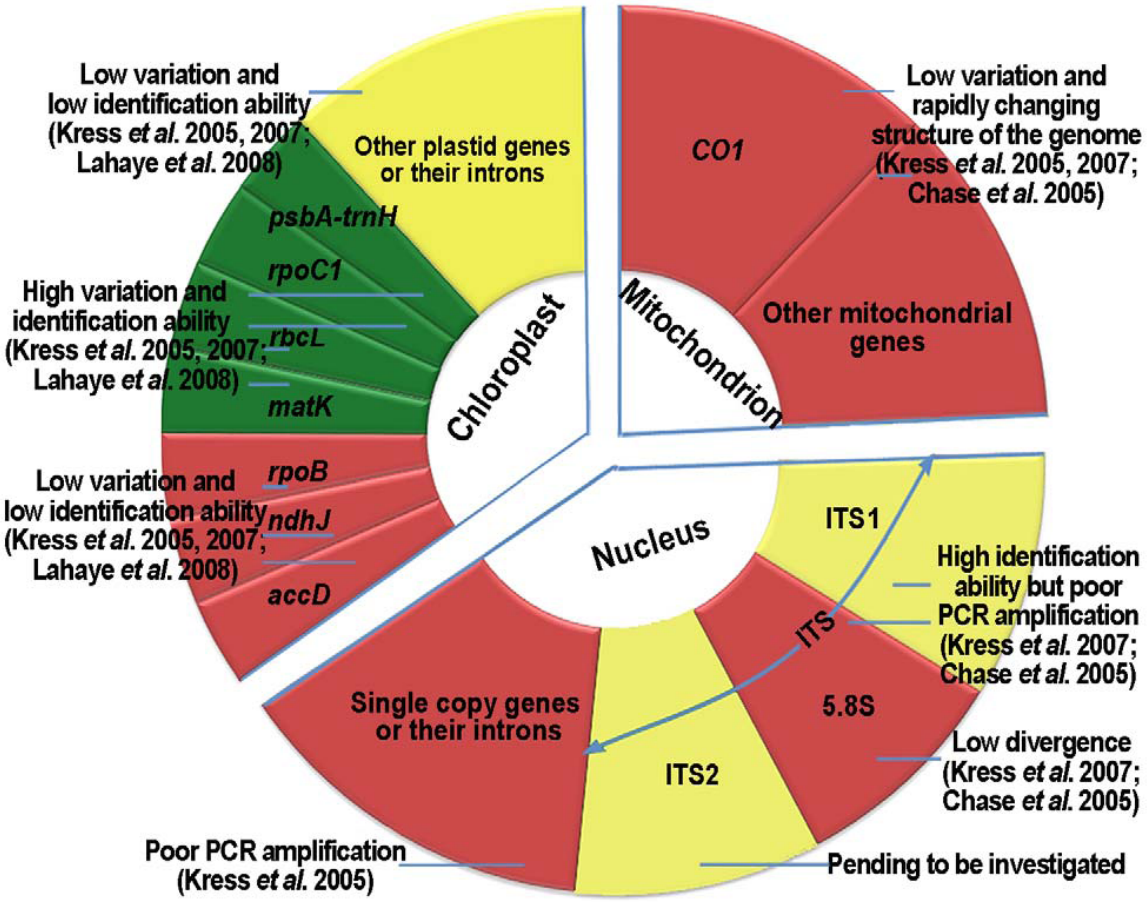


Figure 2.1. Summary on genes from three genomes in plants that are candidate barcodes. Green markers are potential barcodes, red markers are poor candidates and yellow markers are pending to be investigated. Source: (Chen, et al., 2010).

Table 2.1. List of individuals used in this study with its respective species name, botanical section and source.

Sample ID	Species	Cultivation status	Section	Source
Dabys_11	*D.abyssinica* Hochst ex Kunth	wild	Enantiophyllum	Benin
Dabys_12	*D.abyssinica*	wild	Enantiophyllum	Benin
Dabys_13	*D.abyssinica*	wild	Enantiophyllum	Benin
Dburk_11	*D.burkilliana* J.Miege	wild	Enantiophyllum	Benin
Dburk_2	*D.burkilliana* J.Miege	wild	Enantiophyllum	Benin
Dburk_6	*D.burkilliana* J.Miege	wild	Enantiophyllum	Benin
Dburk_7	*D.burkilliana* J.Miege	wild	Enantiophyllum	Benin
Dhirt_1	*D.hirtiflora* Benth	wild	NA	IITA
Dhirt_2	*D.hirtiflora* Benth	wild	NA	IITA
Dhisp_1	*D.hispida* Dennst	wild	Lasiophyton	Thailand
Dhisp_2	*D.hispida* Dennst	wild	Lasiophyton	Thailand
Djapo_1	*D.japonica* Thunb	cultivated	Enantiophyllum	Thailand
Dj173835	*D.japonica* Thunb	cultivated	Enantiophyllum	Japan
Dj173836	*D.japonica* Thunb	cultivated	Enantiophyllum	Japan
Dn1581	*D.nummularia* Lamarck	cultivated	Enantiophyllum	Vanuatu

Dn1625	*D.nummularia* Lamarck	cultivated	Enantiophyllum	Vanuatu
Do169362	*D.opposita* Thunb.	cultivated	Enantiophyllum	Japan
Doppo_2	*D.opposita* Thunb.	cultivated	Enantiophyllum	Thailand
Dpenta_1	*D.pentaphylla* L.	cultivated	Lasiophyton	Thailand
Dpenta_2	*D.pentaphylla* L.	cultivated	Lasiophyton	Thailand
ACC6670	*D.pentaphylla* L.	cultivated	Lasiophyton	CATIE
Dpreu_1	*D.preussii* Pax	wild	NA	IITA
Dpreu_2	*D.preussii* Pax	wild	NA	IITA
Dpreu_3	*D.preussii* Pax	wild	NA	IITA
Dtogo_1	*D.togoensis* R. Knuth	wild	Enantiophyllum	Nigeria
Dtogo_2	*D.togoensis* R. Knuth	wild	Enantiophyllum	Nigeria
Dtogo_3	*D.togoensis* R. Knuth	wild	Enantiophyllum	Nigeria
Dtogo_4	*D.togoensis* R. Knuth	wild	Enantiophyllum	Nigeria
Dtogo_5	*D.togoensis* R. Knuth	wild	Enantiophyllum	Nigeria
Dtoko_1	*D.tokoro* Makino	wild	Stenophora	Japan
Dtoko_2	*D.tokoro* Makino	wild	Stenophora	Japan
Dt621	*D.transversa* R.Br.	cultivated	Enantiophyllum	Vanuatu
Dt711	*D.transversa* R.Br.	cultivated	Enantiophyllum	Vanuatu
Dt_10730	*D.trifida* L.	cultivated	Macrogynodium	CATIE
Dt_7231	*D.trifida* L.	cultivated	Macrogynodium	CATIE
Dt_662	*D.trifida* L.	cultivated	Macrogynodium	Guadeloupe

TDa1007	*D.alata* L.	cultivated	Enantiophyllum	IITA
TDa1237	*D.alata* L.	cultivated	Enantiophyllum	IITA
TDa1295	*D.alata* L.	cultivated	Enantiophyllum	IITA
TDa1352	*D.alata* L.	cultivated	Enantiophyllum	IITA
TDa4129	*D.alata* L.	cultivated	Enantiophyllum	IITA
TDb3075	*D.bulbifera* L.	cultivated	Opsophyton	IITA
TDb3077	*D.bulbifera* L.	cultivated	Opsophyton	IITA
TDb3085	*D.bulbifera* L.	cultivated	Opsophyton	IITA
TDc2793	*D.cayenensis* Lamarck	cultivated	Enantiophyllum	IITA
TDc2794	*D.cayenensis* Lamarck	cultivated	Enantiophyllum	IITA
TDc2800	*D.cayenensis* Lamarck	cultivated	Enantiophyllum	IITA
TDd3093	*D.dumetorum* (Kunth) Pax	cultivated	Lasiophyton	IITA
TDd3097	*D.dumetorum* (Kunth) Pax	cultivated	Lasiophyton	IITA
TDd3107	*D.dumetorum*	cultivated	Lasiophyton	IITA
TDd3110	*D.dumetorum*	cultivated	Lasiophyton	IITA
TDd3771	*D.dumetorum*	cultivated	Lasiophyton	IITA
TDe2463	*D.esculenta* (Lour.) Burkill	cultivated	Combilium	IITA
TDe3040	*D.esculenta*	cultivated	Combilium	IITA

TDe3028	*D.esculenta*	cultivated	Combilium	IITA
TDm2938	*D.mangenotiana* J.Miege	wild	Enantiophyllum	IITA
TDm3051	*D.mangenotiana* J.Miege	wild	Enantiophyllum	IITA
TDm3054	*D.mangenotiana* J.Miege	wild	Enantiophyllum	IITA
TDp3019	*D.praehensilis* Benth	wild	Enantiophyllum	IITA
TDp3020	*D.praehensilis* Benth	wild	Enantiophyllum	IITA
TDp3021	*D.praehensilis* Benth	wild	Enantiophyllum	IITA
TDp3022	*D.praehensilis* Benth	wild	Enantiophyllum	IITA
Dp_forest	*D.praehensilis* Benth	wild	Enantiophyllum	IITA
TDr1477	*D.rotundata* Poiret	cultivated	Enantiophyllum	IITA
TDr1490	*D.rotundata* Poiret	cultivated	Enantiophyllum	IITA
TDr1957	*D.rotundata* Poiret	cultivated	Enantiophyllum	IITA
TDr3508	*D.rotundata* Poiret	cultivated	Enantiophyllum	IITA
TDr3854	*D.rotundata* Poiret	cultivated	Enantiophyllum	IITA
TDr4040	*D.rotundata* Poiret	cultivated	Enantiophyllum	IITA

NA=information not available; IITA=International Institute of Tropical Agriculture; CATIE= Centro Agronómico Tropical de Investigación y Enseñanza

2.2.2 PCR amplification and sequencing reaction

Polymerase chain reaction (PCR) amplification of the *rbcL, matK, trnH-psbA* and *ITS* regions was conducted in Applied Biosystems, Veriti 96 well thermal cycler (Applied

Biosystems, USA) using 25ng DNA template in a 20μl reaction mixture (1x Taq buffer, 1u Taq Polymerase, 0.2mM dNTPs, 2.5mM Mgcl2, and each 0.3μM of forward and reverse primers (synthesized by Integrated DNA Technologies (IDT), Belgium). Most of the primers were obtained from previous studies on yam and other flowering plants. Two additional primer pairs (GT1 and GT2) were designed based on about 80 *Dioscorea* species sequence retrieved from NCBI under genebank accession numbers (AM889705.1 to AM889506.1, KJ922775.1 to KJ922822.1, KF372555.1 to KF372556.1, JQ259957.1 to JQ260103.1, AY973832.1 to AY973832.1, HQ637581.1 to HQ637725.1, JQ733670.1 to JQ733722.1, JX501470.1 to JX501494.1, DQ974175.1 to DQ974189.1, EU407548.1 to EU407549.1, EF028329.1 to EF028333.1, KJ922770.1 to KJ922838.1, HQ637579.1 to HQ637724.1 and JX501472.1 to JX501500.1). These two primers were only used for closely related guinea yam species that could not be identified by all the other primers. The amplification program was 3 min preheating and initial denaturation at 94°C, 35 cycles of 45 sec denaturation at 94°C, 30 sec primer annealing at different temperatures depending on primer used (Table 2.2), and 90 min extension at 72°C with a final extension of 7 min at 72°C. The amplified DNA fragments were separated by electrophoresis on 1.5% agarose in 1xTBE buffer stained with 5μl ethidium bromide. The PCR product was cleaned following standard ethanol precipitation procedure and sequenced in both directions with the primers used for PCR amplification. The sequencing reaction was done using 1.5μl of purified PCR product, 1.0 μl big dye terminator ready reaction mix, 1.0μl 5x sequencing buffer, 1.0μl (5pmol/μl) primer and 5.0μl water.

Table 2.2. List of primers and reaction conditions used in the study.

Marker	Name of primers	Sequences (5'-3')	Annealing T^0	Amplicon size(bp)	Reference
rbcL	H1f	F: CCACAAACAGAGACTAAAGC	55℃	568	Fofana et al 1997
	Fofana	R: GTAAAATCAAGTCCACCGCG			
	1f	F:ATGTCACCACAAACAGAAAC	55℃	704	Fay et al 1998
	724R	R:TCGCATGTACCTGCAGTAGC			
trnH-psbA	fewPA	F:GTTATGCATGAACGTAATGCTC	55℃	401	CBOL (http://barcoding.si.edu)
	revTH	R:CGCGCATGGTGGATTCACAATCC			
	trnH(GUG)-Saltonstall	F:ACTGCCTTGATCCACTTGGC	55℃	545	Saltonstall 2001
	PsbAr2	R:GTAGTAGGTATCTGGTTTACCGCT			
matK	MF	F:ATTTGCGATCTATTCATTCAAT	58℃	948	Sun et al 2013
	MR	R:TGAGATTCCGCAGGTCATT			
	390F	F:CGATCTATTCATTCAATATTTC	55℃	794	CBOL (http://barcoding.si.edu)
	1326R	R:TCTAGCACACGAAAGTCGAAGT			

	GT1-F	F:CCTATATCCACTTCTCTTTCAGGAGT	55℃	810	This study
	GT1-R	R:CCCTTTGACACCAGAATTGC			
	GT2-F	F:TTTACGATCAAGGTCTTCTGGA	55℃	620	This study
	GT2-R	R:CATATCCAACCAAATCGATGA			
ITS	5a fwd	F:CCTTATCATTTAGAGGAAGGAG	50℃	707	CBOL (http://barcoding.si.edu)
	4 rev	R:TCCTCCGCTTATTGATATGC			
	S2F	F:ATGCGATACTTGGTGTGAAT	56℃	226	CBOL (http://barcoding.si.edu)
	S3R	R:GACGCTTCTCCAGACTACAAT			

2.2.3 Sequence editing, alignment and analysis

Raw sequence was edited using CodonCode Aligner (version 3.7.1) (http://codoncode.com/). MatGAT v2.01 (Campanella, et al., 2003) was used to generate similarity/identity matrix. TaxonGap v2.4.1 (Slabbinck, et al., 2008) was further used for visualization of species separability from one another. Multiple alignment of the DNA sequences were made using ClustalW program and the interspecific and intraspecific divergences of each bar coding region was computed by calculating Kimura 2-parameter (K2P) distances in MEGA5 (Tamura, et al., 2011). K2P(Kimura, 1980), a model used to estimate evolutionary distances, distinguishes between two types of substitutions: transitions, where a purine is replaced by another purine (A<-->G) or a pyrimidine is replaced by another pyrimidine (C<-->T), and transversions, where a purine is replaced by a pyrimidine or a pyrimidine is replaced by a purine (A or G <--> C or T). The assumption of K2P model is that the rate of transitions is different from the rate of transversions where it assumes that transition substitutions (purine-purine or pyrimidine-pyrimidine) can be more frequent than transversion substitutions (purine-pyrimidine). K2P was used for two main reasons; 1) K2P is the most frequently used model and this will help to allow comparison of our results with other DNA bar coding studies regardless of the recent reports indicating this model might not be the best and 2) the K2P model was adopted because it performs best for low value genetic distances or it is the most effective model when genetic distances are low, and is therefore popularly used for species-level analysis (Nei and Kumar, 2000; Herbert et al, 2003). The

discrimination power of each marker was assessed by Wilcoxon signed rank tests and the Wilcoxon two-sample test using an online calculator (http://www.fon.hum.uva.nl/Service/Statistics/Wilcoxon_Test.html). The Wilcoxon signed ranks test was computed as described below

1. For each item in a sample of n items, a difference score, *Di,* between the two-paired values were computed.

2. The set of n absolute differences, |Di| were listed by neglecting the + and - signs.

3. Any absolute difference score of zero was omitted from further analysis, thereby yielding a set of n' nonzero absolute difference scores, where n'≤ n. After removing values with absolute difference scores of zero, n' becomes the actual sample size.

4. Ranks, *Ri* were assigned from 1 to n' to each of the |Di| such that the smallest absolute difference score gets rank 1 and the largest gets rank n'.

5. The symbol + or - were reassigned to each of the n' ranks, *Ri,* depending on whether *Di* was originally positive or negative.

6. The Wilcoxon test statistic, *W,* was computed as the sum of the positive ranks based on the equation below

$$W = \sum_{i=1}^{n'} R_i^{(+)}$$

2.3 Results

2.3.1 Amplification efficiency and sequence information

One of the candidate markers (*rbcL*) was promising both in terms of ease of amplification and sequence quality. The efficiency of amplification across the samples was the highest (96.8%) for *rbcL*, followed by 93.7% in *matK*, and 90.6% in *trnH-psbA* using the universal primers, H1f/Fofana, MF/MR and trnH(GUG)-Saltonstall/PsbAr2 for the three markers respectively. The primers were selected in preference to the other primers listed in Table 2.2 based on the band quality observed. High quality bidirectional sequences were obtained with *rbcL* but the *matK* and *trnH-psbA* required manual editing in two and four species respectively. The non-coding *trnH-psbA* was found to be the best marker both with the number of variable sites (64/507), parsimony informative sites (52/507) and number of singleton sites followed by *matK* and the least in *rbcL* (Table 2.3). The *rbcL* was with the most conserved sites (538/568) followed by *trnH-psbA* (443/507) and the least being *matK* (848/947). The primer combination for amplification of *ITS* region was not good enough for PCR amplification, hence, not considered for further sequence analysis. Thus the potential of the *ITS* region to be used as *Dioscorea* species identification remains to be investigated with different primer combination. The new primers we have designed based on *matK* regions GT1 and GT2 were good in terms of amplification efficiency. However, none of the guinea yam species (*D. rotundata, D. cayenensis, D.abyssinica, D. mangenotiana* and *D. praehensilis*) were identified with these markers.

Table 2.3. Sequence highlights of the DNA barcoding regions.

	Marker		
	rbcL	*matK*	*trnH-psbA*
Total number of species successfully sequenced	21	19	19
Number of species identified	10	12	7
Percentage identification (%)	47.6	63.1	36.8
Sequence alignment length (bp)	568	947	507
Conserved sites (bp)	538	848	443
Variable sites (bp)	30	99	64
Parsimony-informative sites (bp)	28	93	52
Singleton sites (bp)	2	6	12

2.3.2 Interspecific and intraspecific divergence

The *matK* region had the highest level of mean interspecific divergence of 0.0196 (SD 0.0209) compared with the other markers evaluated (Table 2.4) while *rbcL* had the least in both interspecific and intraspecific divergence. The non-coding region, *trnH-psbA* showed high intraspecific divergence 0.0009 (SD 0.0025) than the coding regions, *rbcL* and *matK* although the combination of the two (*rbcL* + *matK*) revealed

the highest intraspecific distance. Likewise the Wilcoxson signed rank test indicated significant variation between the species for *matK* when compared to *rbcL*, and *psbA-trnH* (Table 2.5). The Wilcoxson two samples test indicated highly significant variation of interspecific variation than intraspecific variation for all the markers (Table 2.6).

Table 2.4. Measures of inter-specific and intra-specific divergence of the DNA bar-coding regions used based on Kimura 2-parameter.

Markers	Interspecific distance (K2P mean ± std)	Intraspecific distance (K2P mean ± std)
rbcL	0.0114 ± 0.0073	0.0003 ± 0.0008
matK	0.0196 ± 0.0209	0.0005 ± 0.0018
rbcL + matK	0.0158 ± 0.0134	0.0012 ± 0.0012
trnH-psbA	0.0162 ± 0.0366	0.0009 ± 0.0025

Presented are the Kimura 2-parameter mean ± standard deviations.

Table 2.5. Wilcoxon signed-rank tests of inter-specific divergences among markers.

W+	W-	Relative ranks[1]	Sample size	Result*
rbcL	*trnH-psbA*	W+=9824,W-=2896	159	*rbcL>trnH-psbA*
matK	*rbcL*	W+=9458.5,W-=3261.5	159	*matK>rbcL*
matK	*trnH-psbA*	W+=10,073.5, W-=2806.5	160	*matK>trnH-psbA*

*The p-value is 0. The result is significant at P≤ 0.05

[1]The symbols "W+" and "W-" represent the sum of all of the positive values and the sum of all of the negative values in the Signed Rank column, respectively. Symbol ">" is used if the interspecific divergence for a locus significantly exceeds that of another locus.

Table 2.6. Wilcoxon two-sample test based on interspecific versus intraspecific Kimura 2-distances of the three markers.

Marker	The Wilcoxon two sample test			
rbcL	#A=210	#B=21	W=391	P≤2.542e-12
matK	#A=171	#B=19	W=332.5	P≤7.315e-11
trnH-psbA	#A=171	#B=19	W=926	P≤9.429e-05

2.3.3 Discriminatory power of the markers

The discriminatory power of *rbcL* marker was not good enough although it was the best with regard to its amplification efficiency and sequence quality (Table 2.3). The *rbcL* region clearly defined only 10 of the 21 species (47.6%). The *trnH-psbA* marker was the least promising both in PCR amplification, sequence quality and species identification. The *trnH-psbA* region identified only 7 of 19 sequenced species with percent identification of 36.8. *MatK* performance was best among the three markers evaluated in this study in terms of discriminatory power where 12 out of 19 taxonomic species (63.1%) were defined. The two markers combination (*rbcL+ matK*) helped in identifying additional two more species. This increased the discrimination efficiency to 73.7%. Moreover, the highest number of species with separability values of greater than 0 were observed in *rbcL + matK* (Figure 2.2). The *trnH-psbA* showed separability value =0 for more number of species indicating its poor discriminatory power. Overall, most species were identified based on a combination of the two coding regions (*rbcL+ matK*) except five species (*D.rotundata*, *D. cayenensis*, *D. abyssinica*, *D. praehensilis* and *D. mangenotiana*).

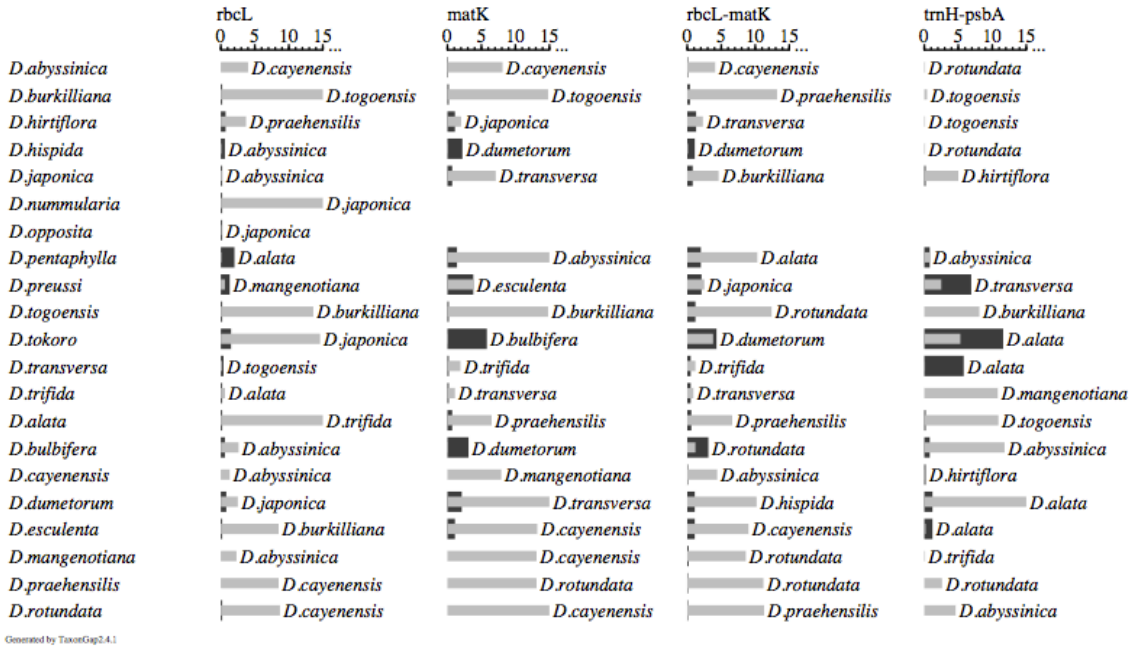


Figure 2.2. The discrimination efficiency of the markers across major cultivated and wild relatives of *Dioscorea* species. The figure indicates the list of species on the left panel, within species heterogeneity (light grey bars) and separability between species (darker bars) for *rbcL*, *matK, rbcL* + *matK* and *trnH-psbA*. The right panel also shows the names of the closely related species based on similarity matrix calculated using Taxongap.

2.4 Discussion

DNA bar coding has been used and proven for efficient species discrimination of flowering plants (Kress, et al., 2005), family Polygonaceae (Song, et al., 2009), family Fabaceae (Gao, et al., 2011) and several other land plant species (Chen, et al., 2010, Fazekas, et al., 2008, Hollingsworth, et al., 2011, Kress and Erickson, 2007). Even though there is standard animal DNA barcode, (i.e. the *cytochrome oxidase 1* (*CO1*) mitochondrial gene that fits the required criteria), finding a plant equivalent has proved difficult (Hollingsworth, et al., 2011). In the current study, the *rbcL*, one of the candidate markers, fits the criteria described by Hollingsworth et al (2011), in terms of universality (ease of amplification and sequencing) and sequence quality but fails to fulfill the ideal marker criteria with regard to species identification having low interspecific divergence and discrimination efficiency with highly conserved sites (Table 2.3 and Table 2.4). Similar reports on poor performance of the *rbcL* marker have been reported by different authors (Clement and Donoghue, 2012, Li, et al., 2014), which limit its potential to use as a universal DNA barcode for plants.

The *trnH-psbA* region showed a higher number of variable sites, parsimony informative sites, singleton sites (Table 2.3) and higher intraspecific divergence (Table 2.4). However, the *trnH-psbA* was the worst for amplification efficiency, sequence quality and species discrimination among the markers assessed. Hence, this marker could not fulfill the required criteria of a desirable DNA barcode. The

increased number of variable sites is related with large number of SNPs observed in three highly divergent species (Figure 2.2) with high separability value.

The *matK* region has been suggested as a candidate for DNA barcoding of plant families, such as Zingiberaceae (Selvaraj, et al., 2008), Fabaceae (Gao, et al., 2011) and as a universal candidate barcode for the whole angiosperms (Yu, et al., 2011). The interspecific distances revealed and larger number of species defined (14 out of 19) by *matK* makes it the best among the markers evaluated. The Wilcoxson rank test similarly indicated significant variation between the species for *matK* when compared to *rbcL* and *psbA-trnH* (Table 2.5). Furthermore, the combination of the two loci (*rbcL + matK*) helped in defining more number of species but none of the markers could identify the five guinea yam species of west African origin.

The inability of all the markers to identify between the five groups (*D.rotundata, D. cayenensis, D. abyssinica, D. praehensilis* and *D. mangenotiana*) could indicate less or recent genetic divergence among the species or unreliability of previous taxonomic classification. A study based on sequencing of three non-coding chloroplast DNA sequences encompassing tRNA genes $trnT_{UGU}$, $trnL_{UAA}$ and $trnF_{GAA}$ similarly could not discriminate these five species from one another (Ramser, et al., 1997). Likewise, a recent observation (Girma, et al., 2014) based on genotyping by sequencing of these species reveals its genetic proximity and significant admixture among one another.

This PhD study suggests the combination of the two single locus-coding regions (*rbcL* and *matK*) as potential multi locus DNA barcoding regions for *Dioscorea* species identification, regardless of difficulty to discriminate some species that possibly have had a recent divergence. However, further study on other chloroplast including nuclear regions and other plastid genes suggested as potential DNA barcode for flowering plants is important to confirm and clearly understand the taxonomy of *Dioscorea* species in general and for plant species that have difficulty for identification like the guinea yams.

2.5 References

Blaxter, M., J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd, et al. 2005. Defining operational taxonomic units using DNA barcode data. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 360: 1935-1943. doi:10.1098/rstb.2005.1725.

Campanella, J.J., L. Bitincka and J. Smalley. 2003. MatGAT: an application that generates similarity/identity matrices using protein or DNA sequences. BMC Bioinformatics 4: 29. doi:10.1186/1471-2105-4-29.

Chen, S., H. Yao, J. Han, C. Liu, J. Song, L. Shi, et al. 2010. Validation of the *ITS2* Region as a Novel DNA Barcode for Identifying Medicinal Plant Species. PLoS ONE 5: e8613. doi:10.1371/journal.pone.0008613.

Clement, W.L. and M.J. Donoghue. 2012. Barcoding success as a function of phylogenetic relatedness in *Viburnum*, a clade of woody angiosperms. BMC evolutionary biology 12: 73. doi:10.1186/1471-2148-12-73.

Cock, P.J., C.J. Fields, N. Goto, M.L. Heuer and P.M. Rice. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38: 1767-1771. doi:10.1093/nar/gkp1137.

Coursey, D.G. 1967. Yams, An account for the Nature, Origins, Cultivation and Utilization of the Useful Members of the Dioscoreaceae. Longmans, Greens and co Ltd., UK, pp230.

Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8: 186-194.

Fazekas, A.J., K.S. Burgess, P.R. Kesanakurti, S.W. Graham, S.G. Newmaster, B.C. Husband, et al. 2008. Multiple Multilocus DNA Barcodes from the Plastid Genome Discriminate Plant Species Equally Well. PLoS ONE 3: e2802. doi:10.1371/journal.pone.0002802.

Gao, T., Z. Sun, H. Yao, J. Song, Y. Zhu, x. Ma, et al. 2011. Identification of Fabaceae Plants Using the DNA Barcode *matK*. Planta Med 77: 92-94.

Gao, T., H. Yao, J. Song, C. Liu, Y. Zhu, X. Ma, et al. 2010. Identification of medicinal plants in the family Fabaceae using a potential DNA barcode *ITS2*. Journal of ethnopharmacology 130: 116-121. doi:10.1016/j.jep.2010.04.026.

Girma, G., K. Hyma, R. Asiedu, S. Mitchell, M. Gedil and C. Spillane. 2014. Next-generation sequencing based genotyping, cytometry and phenotyping for understanding diversity and evolution of guinea yams. Theoretical and Applied Genetics 127: 1783-1794. doi:10.1007/s00122-014-2339-2.

Govaerts, R., P. Wilkin and R.M.K. Saunders. 2007. World Checklist of Dioscoreales. Yams and their allies. The Board of Trustees of the Royal Botanic Gardens, Kew. p. 1-65.

Hebert, P.D. N., Cywinska, A., Ball, S.L. and deWaard, J.R. 2003. Biological identifications through DNA barcodes. Proc. R. Soc. Lond. B 270:313-321

Hollingsworth, P.M., L.L. Forrest, J.L. Spouge, M. Hajibabaei, S. Ratnasingham, M. van der Bank, et al. 2009. A DNA barcode for land plants. Proceedings of the National Academy of Sciences 106: 12794-12797. doi:10.1073/pnas.0905845106.

Hollingsworth, P.M., S.W. Graham and D.P. Little. 2011. Choosing and Using a Plant DNA Barcode. PLoS ONE 6: e19254. doi:10.1371/journal.pone.0019254.

Janzen, D.H., W. Hallwachs, P. Blandin, J.M. Burns, J.-M. Cadiou, I. Chacon, et al. 2009. Integration of DNA barcoding into an ongoing inventory of complex tropical biodiversity. Molecular Ecology Resources 9: 1-26. doi:10.1111/j.1755-0998.2009.02628.x.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution 16: 111-120.

Kress, W.J. and D.L. Erickson. 2007. A Two-Locus Global DNA Barcode for Land Plants: The Coding *rbcL* Gene Complements the Non-Coding *trnH-psbA* spacer Region. PLoS ONE 2: e508. doi:10.1371/journal.pone.0000508.

Kress, W.J., K.J. Wurdack, E.A. Zimmer, L.A. Weigt and D.H. Janzen. 2005. Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences of the United States of America 102: 8369-8374. doi:10.1073/pnas.0503123102.

Lebot, V. 2009. Tropical root and tuber crops: cassava, sweet potato, yams and aroids. CABI Publishers, Wallingford,UK: CABI pp. 413.

Li, Y., Y. Feng, X.-Y. Wang, B. Liu and G.-H. Lv. 2014. Failure of DNA barcoding in discriminating *Calligonum* species. Nordic Journal of Botany: no-no. doi:10.1111/njb.00423.

Little, D.P. 2010. A unified index of sequence quality and contig overlap for DNA barcoding. Bioinformatics 26: 2780-2781.

Miller, S.E. 2007. DNA barcoding and the renaissance of taxonomy. Proceedings of the National Academy of Sciences 104: 4775-4776. doi:10.1073/pnas.0700466104.

Nei, M. and Kumar, S. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York. pp. 333.

Pang, X., C. Liu, L. Shi, R. Liu, D. Liang, H. Li, et al. 2012. Utility of the *trnH–psbA* Intergenic Spacer Region and Its Combinations as Plant DNA Barcodes: A Meta-Analysis. PLoS ONE 7: e48833. doi:10.1371/journal.pone.0048833.

Pereira, F., J. Carneiro and A. Amorim. 2008. Identification of species with DNA-based technology: current progress and challenges. Recent patents on DNA & gene sequences 2: 187-199.

Ramser, J., K. Weising, R. Terauchi, G. Kahl, C. Lopez-Peralta and W. Terhalle. 1997. Molecular marker based taxonomy and phylogeny of Guinea yam (*Dioscorea rotundata - D. cayenensis*). Genome 40: 903-915.

Schaefer, J. and K. Strimmer. 2005. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics 21: 754 - 764.

Selvaraj, D., R.K. Sarma and R. Sathishkumar. 2008. Phylogenetic analysis of chloroplast matK gene from Zingiberaceae for plant DNA barcoding. Bioinformation 3: 24-27.

Slabbinck, B., P. Dawyndt, M. Martens, P. De Vos and B. De Baets. 2008. TaxonGap: a visualization tool for intra- and inter-species variation among individual biomarkers. Bioinformatics 24: 866-867. doi:10.1093/bioinformatics/btn031.

Song, J., H. Yao, Y. Li, X. Li, Y. Lin, C. Liu, et al. 2009. Authentication of the family Polygonaceae in Chinese pharmacopoeia by DNA barcoding technique. Journal of ethnopharmacology 124: 434-439. doi:10.1016/j.jep.2009.05.042.

Sun, X.Q., Y.J. Zhu, J.L. Guo, B. Peng, M.M. Bai and Y.Y. Hang. 2012. DNA barcoding the *Dioscorea* in China, a vital group in the evolution of monocotyledon: use of matK gene for species discrimination. PLoS One 7: e32057. doi:10.1371/journal.pone.0032057.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei and S. Kumar. 2011. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28: 2731-2739. doi:10.1093/molbev/msr121.

Trivedi, U.H., T. Cezard, S. Bridgett, A. Montazam, J. Nichols, M. Blaxter, et al. 2014. Quality control of next-generation sequencing data without a reference. Frontiers in genetics 5: 111. doi:10.3389/fgene.2014.00111.

Yu, J., J.-H. Xue and S.-L. Zhou. 2011. New universal *matK* primers for DNA barcoding angiosperms. Journal of Systematics and Evolution 49: 176-181. doi:10.1111/j.1759-6831.2011.00134.x.

Chapter 3

3. Understanding genomic diversity and relatedness among guinea yams utilizing GBS, cytometry and phenotypic data

3.1 Background and Justification

The native *D. rotundata* Poiret and *D. cayenensis* Lamarck (also referred to as Guinea yams or the *D*. *cayenensis-rotundata* complex) are the most important and most widely cultivated (Mignouna, et al., 2003) among several yam species in West Africa.

The guinea yams were likely domesticated by farmers from wild yams of the section *Enantiophyllum* (Burkill, 1960, Terauchi, et al., 1992, Zannou, et al., 2006). Domestication is still ongoing in Benin (Scarcelli, et al., 2006, Scarcelli, et al., 2006, Zannou, et al., 2006), although the practice is limited to a small number of farmers (Cornet, et al., 2010). A recent study by (Zannou, et al., 2006) indicated that Benin farmers consider the wild yam tubers to be edible after three consecutive planting and harvests and the term "ennoblement" was suggested for yam (Mignouna, 2003) instead of domestication. As part of the domestication process, the farmer's select wild forms for tuber shape and taste which resemble some cultivated varieties in their vegetative parts (Zannou, et al., 2004). In addition, several authors have

reported the direct use of wild yams as a source of food in West Africa (Bahuchet, et al., 1991, Sato, 2001).

Although Africa represents 96% of the total production of yams worldwide (estimated at 40 million tonnes average for the period of 1992 to 2011), no African country is among the top five countries delivering the highest yields (FAOSTAT, 2013). The top five countries producing high yield per area include Japan, Papua New Guinea, Tonga, Jamaica, and Portugal whereas Nigeria, Cote d'Ivoire, Ghana, Benin and Togo are the top five countries in terms of total production.

African farmers face multiple constraints to achieve high yam output. Diseases and storage pests are the major constraints to yam production in West Africa and, over time, these limitations have become more severe (Aidoo, et al., 2011, Amusa, et al., 2003, Baimey, et al., 2006). Breeding for improved varieties in yam is challenging due to the polyploid nature of the crop. Transfer of desirable genes from the secondary genepool of wild relatives to the cultivated primary genepool remains difficult in many crops, including in yams (Spillane and Gepts, 2001). Yet, the wild relatives of yams can harbor desirable genes and genetic diversity that has potential for utilization in breeding efforts to enhance the agronomic performance of yam cultivars. Therefore, understanding the genetic relationship and the biology of yam wild relatives is important for improving cultivated yam species.

To date, there is no clear-cut information on the extent of genetic diversity within and between cultivated guinea yam species and their wild relatives. Genetic diversity in cultivated and wild guinea yams has been investigated using AFLPs, RAPDs, microsatellites and RFLPs (Ramser, et al., 1997, Scarcelli, et al., 2006, Terauchi, et al., 1992). However, these studies could not discriminate some of the wild species from cultivated types, and concluded that the wild and cultivated *Dioscorea* species were very closely related. A recent study on guinea yam collections from Ethiopia using SSR loci (Mengesha, et al., 2013) found no clear distinction between cultivated and wild species.

Morphological characterization studies on cultivars from Benin and Cameroon distinguished individuals and further classified them into *D. rotundata*, *D. cayenensis*, and *D. rotundata x D. cayenensis* groups (Dansi, et al., 1999, Mignouna, et al., 2002). These and other authors suggested the possibility of natural hybridization between different species as a cause of cultivars with heterogeneous morphological traits. However, the difficulty to find reliable and stable morphological traits to discriminate between cultivars was also indicated. *D. abyssinica* Hochst. ex Kunth, *D. praehensilis* Benth, *D. burkilliana* J. Miege, *D. mangenotiana* J. Miege and *D. liebrechtsiana* De Wild were suggested as progenitors of cultivated guinea yam based on shared morphological similarity between plants of wild and cultivated species (Dansi, et al., 1999, Mignouna, et al., 2002, Terauchi, et al., 1992).

Guinea yams, *D. rotundata* and *D. cayenensis*, are polyploid species in which different lines can display different ploidy levels. It has been proposed that *D. rotundata* is a tetraploid with a basic chromosome number of 10 (x = 10) (Dansi, et al., 2001, Gamiette, et al., 1999, Obidiegwu, et al., 2009). Hexaploid and octaploid individuals have been reported in *D. cayenensis* based on DNA flow cytometry, using *Solanum lycopersium* L. (Obidiegwu, et al., 2009) and the tetraploid *D. rotundata* (Dansi, et al., 2001, Gamiette, et al., 1999) as internal standards. However, a study based on segregation patterns of isozyme and microsatellite loci has indicated that *D. rotundata* is diploid, with a chromosome number of 20 (2n=40) (Scarcelli, et al., 2005). Flow cytometry histograms for *D. cayenensis-rotundata* were not distinct from those of its related wild relatives (*D. abyssinica*, *D. mangenotiana*, *D. burkilliana* and *D. praehensilis*) (Gamiette, et al., 1999).

Similar ploidy studies have been performed for two of the other agriculturally most important *Dioscorea* species. *D. trifida* Linnaeus, once thought to be octoploid, is now considered to be an autotetraploid (Bousalem, et al., 2006). Likewise, a study based on the microsatellite segregation analysis of four different progenies has demonstrated that *D. alata* Linnaeus accessions can be diploid, triploid and tetraploid (2n=2x, 3x, 4x), respectively, and not tetraploid, hexaploid and octoploid (2n=4x, 6x, 8x) as previously assumed, with a basic chromosome number of 20 (Arnau, et al., 2009). A study by Nemorin, et al. (2012) further confirmed the autotetraploid nature of the 2n = 80 clones of *D. alata*. However, the extent of

polyploidy is not yet known across guinea yam genepools, which represents an important knowledge gap in understanding the biology and agricultural performance of cultivated guinea yams. For instance, it is possible that ploidy could play an important role in both the morphological and agronomical characteristics of guinea yams.

Flow cytometry, a technique that determines DNA content in a large number of nuclei, alone cannot provide conclusive evidence of ploidy level. Emshwiller (2002) indicated that it can be difficult to distinguish DNA content levels among close ploidy levels. Previously reported as heptaploid (2n=7x=49), *Oxalis tuberosa* Molina, was later found to be actually octoploid (2n=8x=64) using a combination of flow cytometry and molecular evidence. This highlighted the importance of combining both molecular and cytological data in confirming ploidy levels.

The greatest advantage of next generation sequence-based genotyping approaches such as Genotyping By Sequencing (GBS) is reducing "ascertainment bias associated with marker discovery in panels differing from the target population" (Poland and Rife, 2012). The GBS also offers an advantage by simultaneously discovering polymorphisms and obtaining genotypic information across the population of interest. Poland and Rife (2012) have highlighted that GBS represents a fast and inexpensive approach that can enable genotyping of large populations of selection candidates within breeding programs. This can further assist breeders to more efficiently choose genetically diverse parents in breeding programs that employ

both interspecific and intraspecific hybridization. GBS diversity assessment can also provide a means for identifying potential gaps in species collection and further guiding germplasm collecting missions. Taking advantage of the power of Genotyping-by-sequencing approaches, this study aims 1) to increase understanding of genomic diversity and genetic structure of guinea yams and their wild relatives, and 2) to investigate the morphological and ploidy variation within and between cultivated guinea yam species.

3.2 Materials and methods

3.2.1 Plant materials

A total of seven guinea yam species were used for this study. All individual accessions of the two cultivated species *D. rotundata* and *D. cayenensis* including two of the wild species, *D. mangenotiana* and *D. praehensilis* were obtained from International Institute of Tropical Agriculture (IITA) field genebank. The *D. togoensis* accessions were collected from the IITA forest, where they are conserved *in situ*. Accessions of two other wild species, *D. abyssinica* and *D. burkilliana* were kindly supplied by Professor Alexander Dansi from Benin. The accessions of *D. burkilliana* were collected from wild populations while *D. abyssinica* was collected from Northern region of Benin where there is evidence of ongoing domestication of wild yams by farmers (Scarcelli, et al., 2006, Scarcelli, et al., 2006, Zannou, et al., 2006) (Figure 3.1). All of the individual accessions (Table 3.2) were used for genotyping; the two cultivated species (comprising 43 *D. rotundata* and 21 *D.*

cayenensis) were also assessed for morphological variation. The cultivated species including two of the wild species, *D. mangenotiana* and *D. praehensilis* were evaluated for ploidy level using a flow cytometry approach.

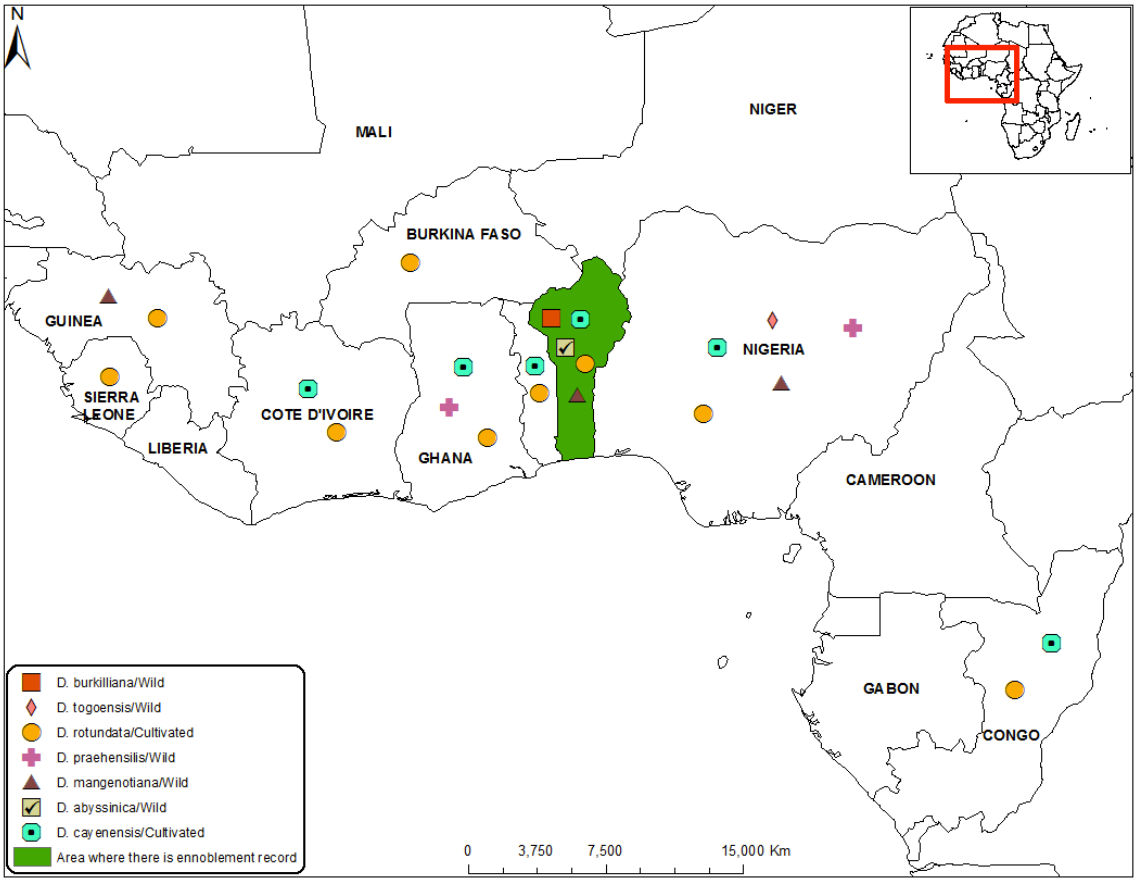


Figure 3.1. Map indicating collection sites for wild and cultivated guinea yam species used in this study. Benin is shaded in green as this is the region where there is evidence of ongoing domestication of wild yams by farmers, via a farmer-driven selection process called ennoblement.

3.2.2 Phenotyping of yam accessions

All individuals of the cultivated species within the IITA field genebank were assessed in 2012 for intra-specific and interspecific morphological variation. The materials were planted following standard procedures (Dumet and Ogunsola, 2008) as routine field genebank regeneration during the main growing season at the IITA experimental plot, Ibadan (Latitude: 7°30'8''N; Longitude: 3°54'38''E), Nigeria. Data

was collected from three individuals planted and labeled as A, B and C per accession. Fourteen yam morphological descriptors (IPGRI/IITA, 1997) were used. The descriptors consisted of stem color, vigor, presence and absence of barky patches and waxiness, leaf shape, leaf color, distance between lobes, sex, number of inflorescences, flower color, tuber flesh color (observed on the upper, middle and lower part), and tuber beneath skin color (Table 3.1).

3.2.3 Ploidy analysis

The ploidy level of the cultivated species was analyzed using a flow cytometry approach. Two of the wild species, *D. mangenotiana* and *D. praehensilis*, whose genome size (611 Mbp) is similar to that of the cultivated *D. rotundata* (Hamon, et al., 1992) were also analyzed using a diploid *D. rotundata* accession (TDr 1673, 2x=2n=40) as standard. Ploidy analysis was perfomed using previously described protocols (Babil, et al., 2010). Young leaves were collected from individual plants. A leaf blade of approximately $5mm^2$ was chopped to homogenize the tissue by adding 500µl ice cold OTTO I buffer (0.1M citric acid monohydrate 0.5% Tween 20). The homogenate was filtered through a 50-µm-pore size nylon filter into a plastic tube. The cell suspension was incubated for 5 minutes at room temperature. The nuclear DNA was stained by adding 2ml of OTTO II buffer (0.4M Na_2PO_4 supplemented with 4 µg/ml of DAPI – 4,6-diamidino-2-phenylindole) and 1 µl/ml mercaptoethanol to each tube. Relative fluorescence intensity was measured to determine the ploidy by using the standard as internal reference. The flow cytometer was adjusted so that

the peak representing the G1 nuclei of the diploid standard (TDr 1673) was set at channel 50.

3.2.4 Yam DNA samples

A total of 95 yam accessions comprising the two cultivated species including five of its wild relatives were genotyped (Table 3.2; Figure 3.1). Leaf samples were collected and lyophilized. DNA was extracted using a Qiagen-DNeasy plant mini kit (QIAGEN GmbH). Samples were quantified using a Nanodrop ND-1000 spectrophotometer (Thermo Scientific). For further quality and quantity assessments, 1 μL (100ng) DNA of all samples was run on 1% w/v agarose gel along with 500ng of two λ *Hin*dIII size standards per gel. A trial digestion was done for ten randomly selected DNA samples using 1U of *Hin*dIII, which were run on a 1% w/v agarose gel along with the λ *Hin*dIII size standards. Two different concentrations of DNA (100ng and 500ng) were used for each digest. The restriction enzyme digested better at the lower DNA concentration (100ng).

Table 3.1. Morphological descriptors used for characterization of the two cultivated species, *D. rotundata* and *D. cayenensis*, from IITA genebank collection.

Morphological descriptors	Parameters used
Sex	1=female, 2=male, 3=monoecious and 4= no flowering
Stem color	1=green, 2=brownish green, 3=purple and 4=dark green
Vigorousity	1=low, 2= intermediate and 3=high
Leaf color	1=green, 2=yellowish green and 3=dark green
Distance between lobes	1=no measurable distance, 2=intermediate and very distant
Flower color	0=not available, 1=white and 2=yellowish
Number of inflorescence	0=none, 1=≤10, 2=11-29 and 3=≥30
Presence of barky patches on stem	0=absent and 1=present
Presence of waxiness on stem	0=absent and 1=present
Leaf shape	1=ovate,2=cordate, 3=sagittate and 4=hastate
Flesh color of upper part of tuber	2=creamy white, 3= yellow and 5=purplish white
Flesh color of middle part of tuber	2=creamy white, 3= yellow and 5=purplish white
Flesh color of lower part of tuber	2=creamy white, 3= yellow and 5=purplish white
Tuber beneath skin color	2=creamy white, 3= yellow and 5=purplish white

3.2.5 GBS libraries and sequencing

GBS libraries were prepared and analyzed at the Institute for Genomic Diversity (IGD) at Cornell University, following (Elshire, et al., 2011). PstI enzyme was used for digestion and for creating a library containing 96 unique barcodes (95 uniquely named samples and one negative control containing no DNA). The GBS library was sequenced on a single Illumina HiSeq lane. A total of 118,383,523 100bp reads were generated and used for SNP calling.

3.2.6 Phenotypic data analysis

Multiple Correspondence Analysis (MCA) was performed for the categorical phenotypic data with FactoMineR package (Lê, et al., 2008) using R software (R Core Team, 2013) to detect the underlying pattern and structures in a data set.

3.2.7 Analysis of GBS data

A modified version of the non-reference Genotyping By Sequencing SNP calling pipeline UNEAK (http://www.maizegenetics.net/gbs-bioinformatics), as implemented in Tassel Version 3.0.160 (Lu, et al., 2013), was used for SNP calling (see supplementary materials for XML configuration files and barcode keyfile). A total of 6,371 SNPs were identified. A filtered dataset was created using VCFtools version v0.1.10 (Danecek, et al., 2011) by first filtering genotypes with quality scores less than 98 (--GQ 98), and then removing SNP loci with more than 90% missing data (--geno 0.1). A total of 2,215 SNP loci remained after filtering. Multi-

dimensional Scaling analysis (MDS) was conducted using PLINK version v1.07 (Purcell, et al., 2007). The individual db-8, originally identified as *D. burkilliana* based on morphology, was found to be a potential mis-identified *D. mangenotiana* based on its pattern of heterozygosity and genetic similarity and was treated as *D. mangenotiana* for further analyses.

Nucleotide distances (substitution rates per site) between and within groups were calculated using MEGA version 5 (Tamura, et al., 2011). A Maximum Parsimony (MP) analysis was carried out on the 2,215 SNPs using PHYLIP. Five hundred sets of weights were generated by bootstrapping (seqboot). From these 500 replicates, 25,370 MP trees representing 23,588 topologies were generated (dnapars). Trees were re-rooted at the longest branch using Newick tools v.1.6 (Junier and Zdobnov, 2010) and visualized with Densitree v2.1.10 (Bouckaert, 2010). A cluster analysis using weighted correlation network was performed on genotypes in R (R Core Team, 2013, R Development Core Team, 2010) using the R package WGCNA (Langfelder and Horvath, 2008).

The proportion of heterozygous SNPs for each individual was calculated as the number of heterozygous SNPs divided by the total number of genotyped SNPs for that individual. Pairwise comparisons of allele frequencies and the proportion of private alleles were calculated between groups (as defined using phylogenetic and MDS analysis), using loci that were genotyped in both groups. The pairwise allele frequencies, distribution of minor allele frequencies and proportion of shared

(present in both populations) versus private (present in one group or the other) alleles are shown in Figure 3.5.

3.3 Results

3.3.1 Morphological diversity among cultivated yam species

*P*henotypic descriptors have been extensively used for plant genetic resources management and conservation (Zamir, 2013). Apart from tuber flesh color, none of the phenotypic descriptors used in this study could distinguish the two cultivated species from each other, although some descriptor traits correlated with ploidy level (Figure 3.2 and 3.4). Morphological traits associated with increased ploidy levels in *D. rotundata* included presence of barky patches, absence of waxiness on stem, and dark green leaf color. The yellow color of tuber flesh observed in *D. cayenensis* was absent in *D. rotundata*. *D. rotundata* was the most phenotypically diverse species in terms of flowering pattern (male, female, monoecious, and non-flowering). In *D. cayenensis*, only male or non-flowering accessions were observed. Some traits including stem color, leaf color, leaf shape, absence and presence of barky patches and waxiness, showed variation in *D. rotundata* but not in *D. cayenensis* (Figure 3.3 and 3.4).

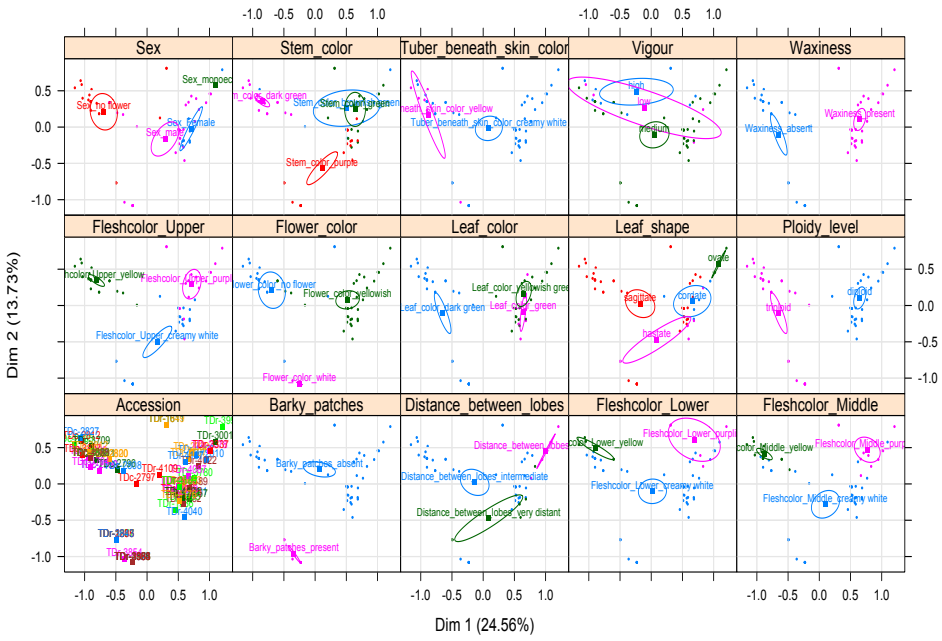


Figure 3.2. MCA performed using the plotellipses function in R, which draws confidence ellipses around the categories of all the categorical variables used. The plotted confidence ellipses around the categories of several variables are to see whether the categories of a categorical variable are significantly different from each other and indicate a) female sex type, creamy white tuber beneath skin and flesh color, brownish green stem, vigour, no waxiness, dark green leaf, cordate leaf, diploid, no barky paches and intermdiate distance between leaf lobes (plot ellipses with blue line); b) monoecious, green stem, medium vigourosity, yellow flesh tuber, yellowish flower, yellowish green leaf, ovate leaf and distance between lobes of very distance (dark green line); c) male, dark green stem, yellow tuber beneath skin, low vigorosity, no waxiness, purple flesh tuber, white flower, green leaf, hastate leaf shape, triploid, stem with barky paches and no distance between lobes (pink line); and ellipses with red line indicate non-flowering and purple stem and sagittate leaf (red line).

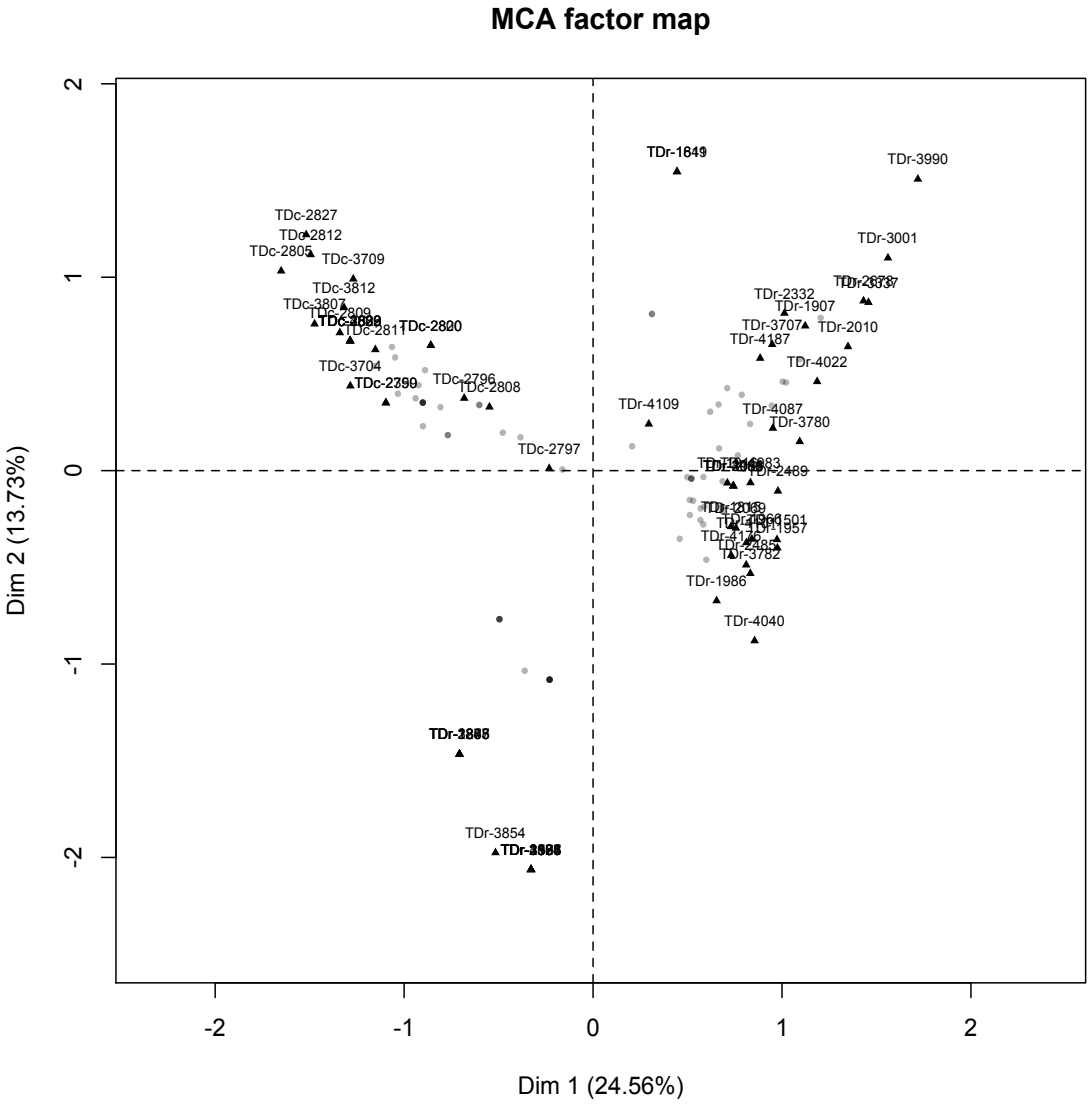


Figure 3.3. MCA showing a) the distribution of individual accessions of *D.rotundata* and *D. cayenensis* (dark grey triangle) and morphological variables (light grey circle).

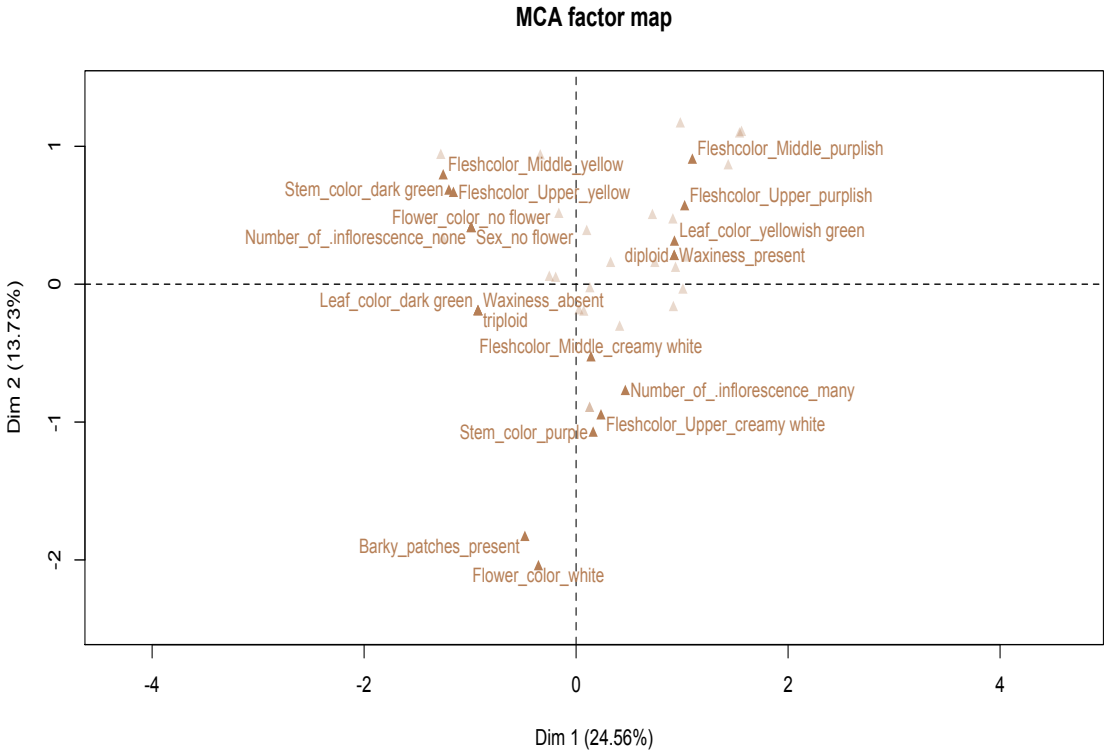


Figure 3.4. MCA showing the top 20 categories of morphological traits contributing the most to the variation (dark grey triangle) including the least contributing variables (light grey triangle).

3.3.2 Ploidy variation across different species of guinea yams

The within-species ploidy level was constant amongst accessions of *D. cayenensis* (3x, N=21), *D. praehensilis* (2x, N=7), and *D. mangenotiana* (3x, N=5). In contrast, both diploid (74.4%) and triploid (25.6%) accessions were observed for *D. rotundata* (Table 3.2). The coefficient of observed variation was below 5% in all flow cytometry histograms, indicating the reliability of ploidy measurements. The different ploidy level accessions within a given species displayed differing phenotypes. For instance, triploid *D. rotundata* individuals all had distinct features, which were absent in the diploid accessions (i.e., thick dark green leaves, stems with barky patches and no waxiness). Moreover, all triploid (3x) individuals were either male or consistently non-flowering. All female flowering plants (N=8) as well as the monoecious (N=1), non-flowering (N=4), and remaining male accessions (N=17) were diploid (data not shown).

Table 3.2. Ploidy diversity and levels of cultivated and wild yam species in Africa.

Species	Accession number	Cultivated or wild?	Country of origin	Ploidy level
*D.rotundata**	TDr 1673	Cultivated	Togo	2x
D. cayenensis	TDc-2359	Cultivated	Nigeria	3x
D. cayenensis	TDc-2790	Cultivated	Togo	3x
D. cayenensis	TDc-2796	Cultivated	Togo	3x
D. cayenensis	TDc-2797	Cultivated	Ghana	3x
D. cayenensis	TDc-2800	Cultivated	Nigeria	3x
D. cayenensis	TDc-2805	Cultivated	Togo	3x
D. cayenensis	TDc-2806	Cultivated	Benin	3x
D. cayenensis	TDc-2808	Cultivated	Cote d'voire	3x
D. cayenensis	TDc-2809	Cultivated	Benin	3x
D. cayenensis	TDc-2811	Cultivated	Togo	3x
D. cayenensis	TDc-2812	Cultivated	Nigeria	3x
D. cayenensis	TDc-2820	Cultivated	Ghana	3x
D. cayenensis	TDc-2822	Cultivated	Ghana	3x
D. cayenensis	TDc-2823	Cultivated	Ghana	3x
D. cayenensis	TDc-2827	Cultivated	Ghana	3x

D. cayenensis	TDc-3704	Cultivated	Congo	3x
D. cayenensis	TDc-3709	Cultivated	Congo	3x
D. cayenensis	TDc-3807	Cultivated	Nigeria	3x
D. cayenensis	TDc-3812	Cultivated	Nigeria	3x
D. cayenensis	TDc-3839	Cultivated	Nigeria	3x
D. cayenensis	TDc-4089	Cultivated	Benin	3x
D. rotundata	TDr-1501	Cultivated	Togo	2x
D. rotundata	TDr-1591	Cultivated	Togo	3x
D. rotundata	TDr-1611	Cultivated	Togo	2x
D. rotundata	TDr-1815	Cultivated	Togo	2x
D. rotundata	TDr-1849	Cultivated	Togo	2x
D. rotundata	TDr-1877	Cultivated	Cote d'voire	3x
D. rotundata	TDr-1888	Cultivated	Cote d'voire	3x
D. rotundata	TDr-1907	Cultivated	Ghana	2x
D. rotundata	TDr-1916	Cultivated	Cote d'voire	2x
D. rotundata	TDr-1957	Cultivated	Ghana	2x
D. rotundata	TDr-1966	Cultivated	Ghana	2x
D. rotundata	TDr-1983	Cultivated	Cote d'voire	2x
D. rotundata	TDr-1986	Cultivated	Ghana	2x

D. rotundata	TDr-2008	Cultivated	Ghana	2x
D. rotundata	TDr-2010	Cultivated	Cote d'Ivoire	2x
D. rotundata	TDr-2069	Cultivated	Nigeria	2x
D. rotundata	TDr-2178	Cultivated	Nigeria	3x
D. rotundata	TDr-2205	Cultivated	Nigeria	3x
D. rotundata	TDr-2332	Cultivated	Nigeria	2x
D. rotundata	TDr-2485	Cultivated	Togo	2x
D. rotundata	TDr-2489	Cultivated	Togo	2x
D. rotundata	TDr-2527	Cultivated	Togo	3x
D. rotundata	TDr-2678	Cultivated	Cote d'voire	2x
D. rotundata	TDr-3001	Cultivated	Nigeria	2x
D. rotundata	TDr-3337	Cultivated	Ghana	2x
D. rotundata	TDr-3707	Cultivated	Congo	2x
D. rotundata	TDr-3780	Cultivated	Benin	2x
D. rotundata	TDr-3782	Cultivated	Benin	2x
D. rotundata	TDr-3843	Cultivated	Benin	3x
D. rotundata	TDr-3854	Cultivated	Benin	3x
D. rotundata	TDr-3866	Cultivated	Benin	3x
D. rotundata	TDr-3983	Cultivated	Nigeria	3x

D. rotundata	TDr-3985	Cultivated	Nigeria	2x
D. rotundata	TDr-3990	Cultivated	Benin	2x
D. rotundata	TDr-4022	Cultivated	Nigeria	2x
D. rotundata	TDr-4040	Cultivated	Nigeria	2x
D. rotundata	TDr-4087	Cultivated	Burkina Faso	2x
D. rotundata	TDr-4101	Cultivated	Benin	2x
D. rotundata	TDr-4109	Cultivated	Burkina Faso	2x
D. rotundata	TDr-4176	Cultivated	Guinea	2x
D. rotundata	TDr-4181	Cultivated	Guinea	2x
D. rotundata	TDr-4184	Cultivated	Sierra Leone	3x
D. rotundata	TDr-4187	Cultivated	Sierra Leone	2x
D. praehensilis	TDp 3022	Wild	Ghana	2x
D. praehensilis	TDp 3025	Wild	Ghana	2x
D. praehensilis	Dp IITA-1	Wild	Nigeria	2x
D. praehensilis	Dp IITA-2	Wild	Nigeria	2x
D. praehensilis	Dp IITA-a	Wild	Nigeria	2x
D. praehensilis	Dp IITA-b	Wild	Nigeria	2x
D. praehensilis	Dp IITA-c	Wild	Nigeria	2x

D. mangenotiana	TDm 2938	Wild	Nigeria	3x
D. mangenotiana	TDm 3051	Wild	Nigeria	3x
D. mangenotiana	TDm 3053	Wild	Nigeria	3x
D. mangenotiana	TDm 3054	Wild	Nigeria	3x
D. mangenotiana	TDm 3803	Wild	Guinea	3x
D. abyssinica	da-1	Wild	Benin	NA
D. abyssinica	da-3	Wild	Benin	NA
D.burkilliana	db-7	Wild	Benin	NA
D.burkilliana	db-10	Wild	Benin	NA
D.burkilliana	db-6	Wild	Benin	NA
D.burkilliana	db-11	Wild	Benin	NA
D.burkilliana	db-5	Wild	Benin	NA
D.burkilliana	db-2	Wild	Benin	NA
D.burkilliana **	db-8	Wild	Benin	NA
D. togensis	dt-iita-3	Wild	Nigeria	NA
D. togensis	dt-iita-2	Wild	Nigeria	NA
D. togensis	dt-iita-4	Wild	Nigeria	NA

D. togensis	dt-ii	Wild	Nigeria	NA
D. togensis	dt-i	Wild	Nigeria	NA
D. togensis	dt-5	Wild	Nigeria	NA

* an individual used as a standard for ploidy analysis

** misidentified individual

The genetic clustering analysis showed admixture of some individuals across different ploidy groups. Two 2x *D. rotundata* accessions were admixed with 3x *D. rotundata*, while two 2x accessions were admixed with 3x *D. cayenensis* groups (TDr 4187, TDr 3337, TDr 3990 and TDr 4087), which exhibited high heterozygosity. Ploidy variation was also associated with incorporation of alleles from wild germplasm into the *D. cayenensis* – *D. rotundata* complex. *D. cayenensis* (3x) harboured alleles from the wild species *D. burkilliana*, whereas 3x *D. rotundata* contained *D. togoensis* alleles, potentially indicating allo-polyploid origins of these 3x cultivated accessions (Girma et al, 2014). However, some 3x individuals of *D. rotundata* (TDr 3983, TDr 1888 and TDr 3854) did not have high levels of heterozygosity, indicating autopolyploidy or hybridization between closely related individuals as possible routes to polyploidy. The reduced heterozygosity in *D. burkilliana, D. togoensis* and *D. abyssinica* suggests that these are diploid.

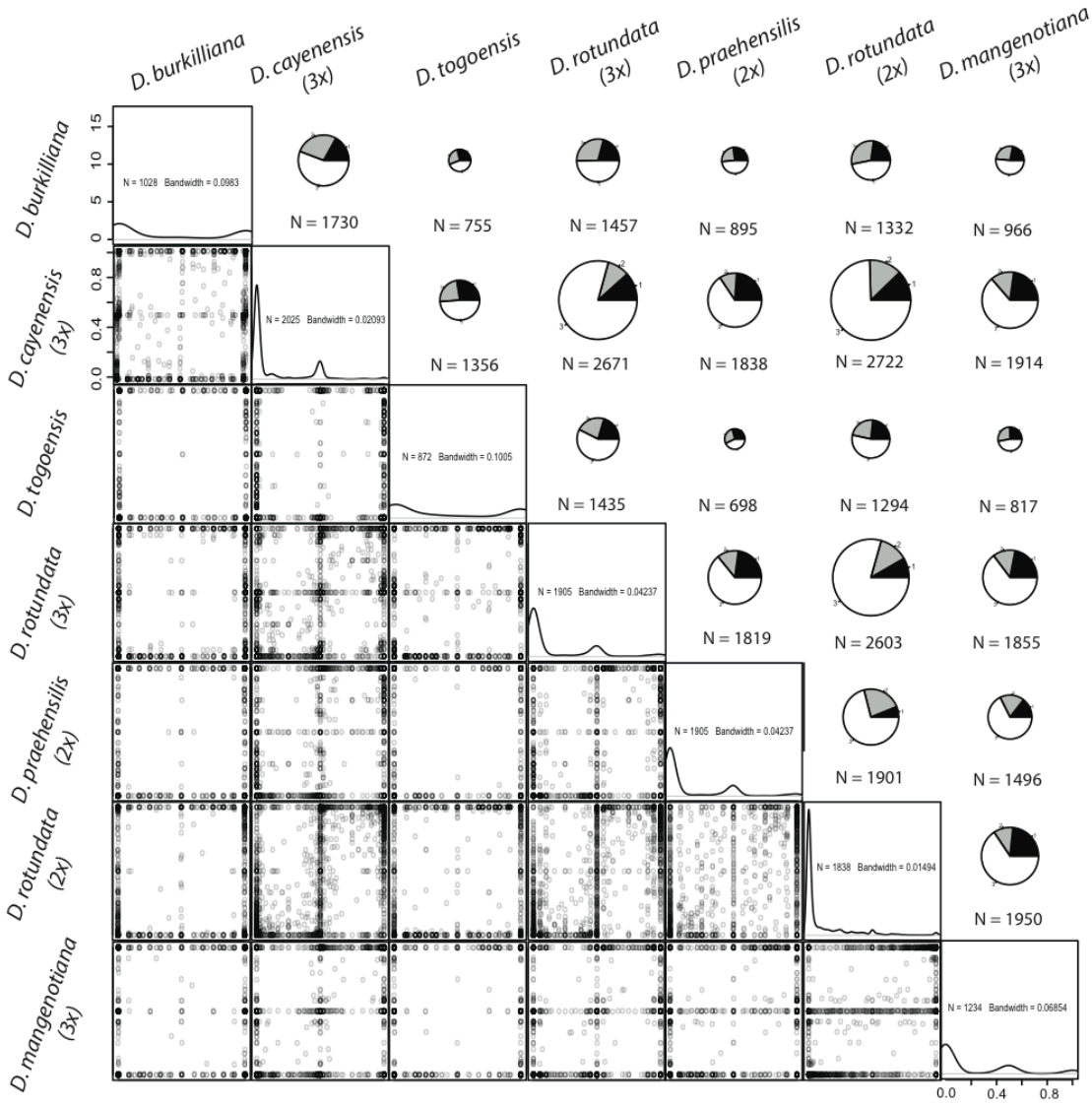


Figure 3.5. Frequency and proportion of private alleles. The lower diagonal area contains plots of pair wise allele frequencies (major and minor) between groups. In each box, the points around the edges represent alleles that are fixed in one population or the other, while points in the middle are segregating in both. The lower diagonal area of the Figure also shows plots of minor allele frequencies for each group (all on the same scale, 0.0 to 1.0) - peaks at ~50% (0.5) can be seen in groups that are 3x. The upper diagonal area of the Figure contains pie charts

indicating the proportion of shared and private alleles (major and minor). Shared alleles represented in white, while private alleles specific to the x-axis group are in black, and private alleles specific to the y-axis group are in grey.

3.3.3 Genetic diversity patterns and genetic structure of yams

The maximum parsimony analysis distinguished *D. burkilliana* from *D. togoensis*, but also distinguished *D. praehensilis* and *D. mangenotiana* from the cultivated *D. cayenensis* and *D. rotundata* (Girma et al, 2014). Two *D. abyssinica* individuals appeared to be closely related to *D. rotundata* (2x). The mean group differences in substitution rate per site (Table 3.3) indicated that the wild guinea yams *D. togoensis* and *D. burkilliana* are the most distant among wild populations from the cultivated species. Conversely, the analysis indicated that *D. mangenotiana*, *D. praehensilis* and *D. abyssinica* wild species are genetically closer to the cultivated species, *D. cayenensis* and *D. rotundata* (Girma et al, 2014). The number of base substitutions per site (from averaging over all sequence pairs between groups) ranged from 0.03 between *D. abyssinica* and (2x) *D. rotundata* to 1.15 between *D. cayenensis* and *D. togoensis*.

The heterozygosity levels among individuals varied between 10 to 20% and appeared to be correlated with ploidy levels (Girma et al, 2014). For instance, the triploid *D. mangenotiana and D. cayenensis* had a higher proportion of heterozygous sites than the diploid *D. rotundata* and *D. praehensilis*. The wild yam species formed some discrete genetic groupings (Girma et al, 2014), with *D. burkilliana* and *D.*

togoensis being quite distinct from the other species. All other accessions clustered into three groups predominantly composed of the cultivated *D. rotundata* and *D. cayenensis* diploids and triploids (Girma et al, 2014). However, accessions from the wild species *D. praehensilis*, *D. mangenotiana* and *D. abyssinica* clustered together with the cultivated species.

Table 3.3. Estimates of evolutionary divergence over sequence pairs between groups. Lower left diagonal: average number of base substitutions per site over all sequence pairs between groups. Upper right diagonal: standard error estimate(s) are shown above the diagonal.

	abyssinica	*burkilliana*	*mangenotiana*	*praehensilis*	*togoensis*	*cayenensis*	*rotundata_2x*	*rotundata_3x*
abyssinica		0.1058505	0.0098201	0.0139287	0.1537379	0.0090677	0.0032199	0.0054628
burkilliana	0.8032981		0.0918105	0.0986323	0.1265432	0.0597261	0.0899553	0.1023735
mangenotiana	0.0582308	0.7092618		0.0165598	0.1549088	0.0182714	0.0091955	0.0140782
praehensilis	0.1165054	0.8175970	0.1354147		0.1510850	0.0138823	0.0102911	0.0131464
togoensis	0.9549081	0.8918185	1.0429453	1.0149002		0.1929068	0.1369335	0.0689549
cayenensis	0.0524431	0.5150288	0.1452489	0.1193613	1.1551302		0.0059603	0.0099503
rotundata_2x	0.0304090	0.8032775	0.0792767	0.1174252	0.9331154	0.0584111		0.0039185
rotundata_3x	0.0425668	0.8452944	0.1233333	0.1244797	0.5777506	0.0900501	0.0449827	

3.4 Discussion

3.4.1 Identification of novel SNPs using Genotyping-By-Sequencing (GBS)

Genotyping by Sequencing (GBS) is increasingly used for genetic diversity analyses, gene identification, and plant breeding. GBS has been applied to wheat genomic selection (Poland, et al., 2012), analysis of switchgrass genomic diversity (Lu, et al., 2013), development of genetic maps in barley and wheat (Poland, et al., 2012), and genome wide association studies in sorghum (Morris, et al., 2013). Here we demonstrate that GBS is an effective tool for analysis of guinea yam genomic diversity, regardless of the complexity of guinea yams in terms of ploidy level, genome size, and the current lack of a reference genome.

3.4.2 Recent origins of cultivated yams from wild ancestors such as D. burkilliana

The low genetic divergence between the two cultivated species, *D. rotundata* and *D. cayenensis* (Table 3.3), confirms previous studies (using RFLP analysis) which suggested that these two species display a recent evolutionary divergence (Terauchi, et al., 1992). The clear separation of *D. togoensis* and *D. burkilliana* illustrates the isolation of these species from the *rotundata-cayenensis* complex. The relatively lower divergence (Table 3.3) and higher allele sharing (Figure 3.5; Girma et al, 2014) between *D. cayenensis* and *D. burkilliana* substantiates earlier suggestions (Onyilagha and Lowe, 1986, Ramser, et al., 1997, Terauchi, et al., 1992) that *D. burkilliana* could be the possible ancestor of *D. cayenensis.* However, *D.*

togoensis seems to contribute more to *D. rotundata* (based on allele sharing) than to *D. cayenensis*, in contrast to previous reports (Ramser, et al., 1997, Terauchi, et al., 1992). The minimal differentiation and closer similarity of *D. mangenotiana*, *D. praehensilis* and *D. abyssinica* (Table 3.3) with the *rotundata-cayenensis* complex indicates that these wild relatives are either of recent divergence or variants of the cultivated species.

3.4.3 Population genetic structure of the cultivated guinea yams and its wild relatives likely reflects ongoing domestication practices or past hybridization events

The wild relatives of yam display distinct clustering based on multi-dimensional scaling, maximum parsimony and genotype clustering (Girma et al, 2014). However, some of the wild relatives showed some genetic admixture with cultivated forms. The close genetic relationship between the wild and cultivated species could also be due to the difficulty to phenotypically differentiate the cultivated species from the wild species, or gene flow occurring via interspecific hybridization between wild and cultivated species (Cornet, et al., 2010, Scarcelli, et al., 2006, Scarcelli, et al., 2006).

Our study also shows evidence of admixture in *D. abyssinica* (Girma et al, 2014). The spontaneous formation of hybrids between wild and cultivated yams demonstrated by Scarcelli, et al. (2006) suggests a mechanism for naturally occurring genetic

admixture between cultivated and wild relatives. In contrast, Dansi, et al. (1999) found no evidence for the deliberate use of *D. togoensis* plants for domestication purposes. There is also no report of farmers harvesting *D. burkilliana* for food or for domestication purposes, although farmers do recognize both of these species as wild (Dansi, et al. (1999). Ethnobotanical evidence suggests that gene flow between these two wild species cultivated yams is minimal. This is supported by our data, which showed little genetic contribution of these two wild species to the cultivated gene pool (*rotundata-cayenensis* complex). The increased heterozygosity levels we found in some 2x accessions for *D. rotundata* also supports a role for admixture arising from interspecific hybridization.

Breeding (crossing) experiments conducted at IITA have confirmed the sexual compatibility within cultivated, and between cultivated yams and their wild relatives. Interspecific crossing studies were conducted with the objective to transfer traits from wild relatives to cultivated lines. Cultivated species (e.g. *D. rotundata x D. cayenensis)* and wild and cultivated species (e.g. *D. rotundata* x *D. praehensilis* and *D. rotundata* x *D. togenesis*) have been successfully crossed (Akoroda, 1985); Robert Asiedu, personal communication).

3.4.4 Morphological descriptors lack resolving power to differentiate the two cultivated yam species.

The taxonomy of *D. rotundata* and *D. cayenensis* has been under investigation and scientific debate for decades. Some taxonomists have considered *D. rotundata* as

subspecies of *D. cayenensis,* indicated as *D. cayenensis subsp. rotundata* (Poiret) J. Miège 1968 (White Guinea Yam) whereas Terauchi, et al. (1992) suggested that 'yellow yam', *D. cayenensis* should be treated as a variety of *D. rotundata*, denoted as *D. rotundata var. x 'cayenen- sis' (*on the basis of its nuclear ribosomal DNA characteristics). On the other hand, Hamon and Toure (1990) observed several intermediate forms, and proposed to treat the two species as the *D. cayenensis-rotundata* species complex.

In this study, we have observed yellow tuber flesh color in some parts of the tuber in all *D. cayenensis* accessions investigated (Figure 3.2 and 3.4). However, as a classifier, the yellow tuber flesh color is ambiguous in some accessions of *D. cayenensis* even though it is the most commonly used approach for classifying the two species as either yellow or white yams. Illustrating the challenges of using morphological descriptors, none of the morphological descriptors we used were distinct for the two yam species highlighting the difficulty to distinguish the two species using such criteria. However, our analysis determined that some of the morphological traits are correlated with ploidy level. The presence of barky patches, absence of waxiness, and dark green leaf color are closely related with *3x D. rotundata.* The darker green color could be associated with chlorophyll content, which can increase as ploidy level increases (Yildiz, 2013).

3.4.5 Ploidy variation in guinea yams due to auto- and allo-polyploidy.

The pattern of allele sharing where *D. cayenensis* harboured *D. burkilliana* alleles, 3x *D. rotundata* harboured *D. togoensis* alleles and a few 3x *D. rotundata* showed reduced heterozygosity, suggest that the polyploidization process in guinea yams likely involves both allo-polyploidy and auto-polyploidy. Moreover, the increased heterozygosity in some 2x *D. rotundata* accessions highlights the presence of gene flow between closely related species. Additionally, increased ploidy levels and heterozygosity in *D. cayenensis* and allele sharing between the two cultivated species indicate that *D. cayenensis* arose from *D. rotundata* but not vice versa. Our results also confirm Terauchi, et al. (1992) earlier suggestion to consider *D. cayenensis* as a subspecies of *D. rotundata*.

3.4.6 GBS data will be most powerful when combined with reference genome

Despite the lack of a reference genome, the UNEAK pipeline was successfully used to call large number of SNPs in switchgrass (Lu, et al., 2013), which were further validated using maize GBS data. While we have utilized the GBS data in the absence of a reference genome for yam, we recognize that GBS is most powerful when the reference genome is available. Access to a reference genome for yam would help for identifying more SNPs and avoiding potential bias associated with the conservative SNP calling employed in the UNEAK pipeline. The GBS data generated in this study (and made publicly available) will be compatible with the yam reference genome

when the genome sequence is released, and will allow further assessment of molecular diversity in yam.

Specifically, the limitation of identifying bi-allelic SNPs that differ at only one base pair within a 64 base pair tag may lead to biases when estimating true rates of divergence within or among species due to mis-identified or unobserved loci, especially when divergence rates are high. This problem may also be exacerbated by low sample sizes for some species. With a reference genome, these biases can be significantly reduced. The GBS raw sequence data generated in this study will be reanalyzed in the future using a reference-sequence based pipeline for calling SNPs once the genome sequence of *Dioscorea* becomes available (Tamiru, et al., 2013).

3.4.7 Implications for guinea yam conservation and improvement programs

We advocate the wider use of GBS (even in species lacking a reference genome), as it can help generating genotypic information across the whole population of interest (including germplasm collections) at a much lower cost per data point. Similarly, GBS could be used for further understanding of genetic relationship studies of other species within the genus *Dioscorea*. GBS is cost-effective and has major potential for characterization of the yam genebank collection maintained at IITA (and other yam germplasm collections), as it can assess the extent and distribution of genetic diversity in the collections. Such knowledge is necessary for improved management

of the genebank, either through identifying duplicates or guiding the need for further germplasm collection. The close genetic similarity of some wild yams with the cultivated forms and sexual compatibility between species provides an opportunity for yam improvement through incorporation of genes and traits from wild relatives. The use of wild relatives in yam breeding programs can allow the tapping of important traits present in the wild genetic pool and that were not yet captured in domesticated germplasm. Variation in ploidy within and between species is a challenge but also an opportunity for managing both intraspecific and interspecific hybridization in breeding programs. Overall, the use of GBS combined with a better understanding of ploidy relationships among species is essential for improving understanding of genetic relationships between wild and cultivated forms of guinea yams, which is critical for understanding the evolution, domestication and ongoing use of guinea yams as an important staple food crop.

3.5. References

Aidoo, R., F. Nimoh, J.-E. Bakang, K. Yankyera, S. Fialor and R. Abaidoo. 2011. Economics of small scale seed yam production in Ghana: Implications for commercialization Journal of sustainable development in Africa 13: 7.

Akoroda, M.O. 1985. Pollination management for controlled hybridization of white yam. Scientia Horticulturae. 25: 201-209.

Amusa, N., A. Adegbite, S. Mohammed and R. Baiyewu. 2003. Yam diseases and its management in Nigeria. African Journal of Biotechnology 2: 497-502.

Arnau, G., A. Nemorin, E. Maledon and K. Abraham. 2009. Revision of ploidy status of *Dioscorea alata* L. (Dioscoreaceae) by cytogenetic and microsatellite segregation analysis. Theor Appl Genet 118: 1239-1249. doi:10.1007/s00122-009-0977-6.

Babil, P.K., K. Irie, H. Shiwachi, Y. Tun, H. Toyohara and H. Fujimaki. 2010. Ploidy variation and their effects on Leaf and stoma traits of water yam (*Dioscorea alata* L.) collected in Myanmar. Trop.Agr.Develop. 54: 132-139.

Bahuchet, S., D. McKey and I. de Garine. 1991. Wild yams revisited: Is independence from agriculture possible for rain forest hunter-gatherers? Human Ecology 19: 213-243.

Baimey, H., D. Coyne and N. Labuschagne. 2006. Effect of fertilizer application on yam nematode (*Scutellonema bradys*) multiplication and consequent damage

to yam (*Dioscorea* spp.) under field and storage conditions in Benin. International Journal of Pest Management 52: 63-70.

Bouckaert, R.R. 2010. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics 26: 1372-1373. doi:10.1093/bioinformatics/btq110.

Bousalem, M., G. Arnau, I. Hochu, R. Arnolin, V. Viader, S. Santoni, et al. 2006. Microsatellite segregation analysis and cytogenetic evidence for tetrasomic inheritance in the American yam *Dioscorea trifida* and a new basic chromosome number in the Dioscoreae. Theor Appl Genet 113: 439-451. doi:10.1007/s00122-006-0309-z.

Burkill, I.H. 1960. The organography and the evolution of Dioscoreaceae, the family of the Yams. Journal of the Linnean Society of London, Botany 56: 319-412. doi:10.1111/j.1095-8339.1960.tb02508.x.

Cornet, D., M. Deu, M. Baco, A. Agbangla, M. Duval and J. Noyer. 2010. Impact of farmer selection on yam genetic diversity. Conservation Genetics 11: 2255-2265.

Coursey, D.G. 1967. Yams, An account for the Nature, Origins, Cultivation and Utilization of the Useful Members of the Dioscoreaceae. Longmans, Greens and co Ltd., UK, pp.230.

Danecek, P., A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, et al. 2011. The variant call format and VCFtools. Bioinformatics 27: 2156-2158. doi:10.1093/bioinformatics/btr330.

Dansi, A., H.D. Mignouna, M. Pillay and S. Zok. 2001. Ploidy variation in the cultivated yams (*Dioscorea cayenensis-Dioscorea rotundata* complex) from Cameroon as determined by flow cytometry. Euphytica 119: 301-307. doi:10.1023/A:1017510402681.

Dansi, A., H.D. Mignouna, Zoundjih, eacute, J. kpon, A. Sangare, et al. 1999. Morphological diversity, cultivar groups and possible descent in the cultivated yams (*Dioscorea cayenensis/D. rotundata*) complex in Benin Republic. Genetic Resources and Crop Evolution 46: 371-388.

Deschamps, S., V. Llaca and G.D. May. 2012. Genotyping-by-Sequencing in Plants. Biology 1: 460-483.

Dumet, D. and D. Ogunsola. 2008. Regeneration guidelines: yams, Crop Specific Regeneration Guidelines. In: M. E. Dulloo, I. Thormann, M. A. Jorge and J. Hanson, editors, CGIAR System-wide Genetic Resource Programme. Rome.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, et al. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6: e19379. doi:10.1371/journal.pone.0019379.

Emshwiller, E. 2002. Ploidy Levels among Species in the '*Oxalis tuberosa* Alliance' as Inferred by Flow Cytometry. Annals of Botany 89: 741-753. doi:10.1093/aob/mcf135.

FAOSTAT. 2013. Statistical data base. FAO, Rome,Italy. Date accessed February 2014

Gamiette, F., F. Bakry and G. Ano. 1999. Ploidy determination of some yam species (*Dioscorea* spp.) by flow cytometry and conventional chromosomes counting. Genetic Resources and Crop Evolution 46: 19-27. doi:10.1023/A:1008649130567.

Girma, G., K. Hyma, R. Asiedu, S. Mitchell, M. Gedil and C. Spillane. 2014. Next-generation sequencing based genotyping, cytometry and phenotyping for understanding diversity and evolution of guinea yams. Theoretical and Applied Genetics 127: 1783-1794. doi:10.1007/s00122-014-2339-2.

Govaerts, R., P. Wilkin and R.M.K. Saunders. 2007. World Checklist of Dioscoreales. Yams and their allies. The Board of Trustees of the Royal Botanic Gardens, Kew. p. 1-65.

Hamon, P., J.-P. Brizard, J. Zoundjihékpon, C. Duperray and A. Borgel. 1992. Etude des index d' ADN de huit ignames (*Dioscorea* sp.) par cytométrie en flux. Canadian Journal of Botany 70: 996-1000.

Hamon, P. and B. Toure. 1990. Characterization of traditional yam varieties belonging to the *Dioscorea cayenensis-rotundata* complex by their isozymic patterns. Euphytica 46: 101-107.

IPGRI/IITA. 1997. Descriptors for Yam (*Dioscorea* spp.). International Institute of Tropical Agriculture, Ibadan, Nigeria/International Plant Genetic Resources Institute, Rome, Italy.

Junier, T. and E.M. Zdobnov. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. Bioinformatics 26: 1669-1670. doi:10.1093/bioinformatics/btq243.

Langfelder, P. and S. Horvath. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9: 559.

Lê, S., J. Josse and F. Husson. 2008. FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software. Journal of Statistical Software 25: 1-18.

Lebot, V. 2009. Tropical root and tuber crops: cassava, sweet potato, yams and aroids. CABI Publishers, Wallingford,UK: CABI pp. 413.

Lu, F., A.E. Lipka, J. Glaubitz, R. Elshire, J.H. Cherney, M.D. Casler, et al. 2013. Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. PLoS Genet 9: e1003215. doi:10.1371/journal.pgen.1003215.

Mengesha, W., S. Demissew, M. Fay, R. Smith, I. Nordal and P. Wilkin. 2013. Genetic diversity and population structure of Guinea yams and their wild relatives in South and South West Ethiopia as revealed by microsatellite markers. Genetic Resources and Crop Evolution 60: 529-541. doi:10.1007/s10722-012-9856-0.

Mignouna, H., M. Abang and R. Asiedu. 2007. Yams. In: C. Kole, editor Genome mapping and molecular breeding Pulses, Sugar and Tuber Crops. Springer, Heidelberg, Berlin, New York, Tokyo. p. 271–296.

Mignouna, H.D., M.M. Abang and S.A. Fagbemi. 2003. A comparative assessment of molecular marker assays (AFLP, RAPD and SSR) for white yam (*Dioscorea rotundata*) germplasm characterization. Annals of Applied Biology 142: 269-276. doi:10.1111/j.1744-7348.2003.tb00250.x.

Mignouna, H.D., A. Dansi and S. Zok. 2002. Morphological and isozymic diversity of the cultivated yams (*Dioscorea cayenensis/Dioscorea rotundata* complex) of Cameroon. Genetic Resources and Crop Evolution 49: 21-29. doi:10.1023/A:1013805813522.

Morris, G.P., P. Ramu, S.P. Deshpande, C.T. Hash, T. Shah, H.D. Upadhyaya, et al. 2013. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proceedings of the National Academy of Sciences of the United States of America 110: 453-458. doi:10.1073/pnas.1215985110.

Nemorin, A., K. Abraham, J. David and G. Arnau. 2012. Inheritance pattern of tetraploid *Dioscorea alata* and evidence of double reduction using microsatellite marker segregation analysis. Mol Breeding 30: 1657-1667. doi:10.1007/s11032-012-9749-0.

Obidiegwu, J., J. Loureiro, E. Ene-Obong, E. Rodriguez, M. Kolesnikova-Allen, C. Santos, et al. 2009. Ploidy level studies on the *Dioscorea cayenensis/Dioscore rotundata* complex core set. Euphytica 169: 319-326.

Onyilagha, J.C. and J. Lowe. 1986. Studies on the relationship of *Dioscorea cayenensis* and *Dioscorea rotundata* cultivars. Euphytica 35: 733-739. doi:10.1007/BF00028581.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, et al. 2012. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. Plant Gen. 5: 103-113. doi:10.3835/plantgenome2012.06.0006.

Poland, J.A., P.J. Brown, M.E. Sorrells and J.-L. Jannink. 2012. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. PLoS ONE 7: e32253. doi:10.1371/journal.pone.0032253.

Poland, J.A. and T.W. Rife. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Gen. 5: 92-102. doi:10.3835/plantgenome2012.05.0005.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81: 559-575. doi:10.1086/519795.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria.URL http://www.R-project.org/.

Ramser, J., K. Weising, R. Terauchi, G. Kahl, C. Lopez-Peralta and W. Terhalle. 1997. Molecular marker based taxonomy and phylogeny of Guinea yam (*Dioscorea rotundata - D. cayenensis*). Genome 40: 903-915.

Sato, H. 2001. The potential of edible wild yams and yam-like plants as a staple food resource in the African tropical rain forest. African study monographs. Supplementary issue. 26: 123-134.

Scarcelli, N., M. Couderc, M. Baco, J. Egah and Y. Vigouroux. 2013. Clonal diversity and estimation of relative clone age: application to agrobiodiversity of yam (*Dioscorea rotundata*). BMC Plant Biology 13: 178.

Scarcelli, N., O. Dainou, C. Agbangla, S. Tostain and J.L. Pham. 2005. Segregation patterns of isozyme loci and microsatellite markers show the diploidy of African yam *Dioscorea rotundata* (2n = 40). Theor Appl Genet 111: 226-232. doi:10.1007/s00122-005-2003-y.

Scarcelli, N., S. Tostain, C. Mariac, C. Agbangla, O. Da, J. Berthaud, et al. 2006. Genetic Nature of Yams (*Dioscorea* sp.) Domesticated by Farmers in Benin (West Africa). Genetic Resources and Crop Evolution 53: 121-130. doi:10.1007/s10722-004-1950-5.

Scarcelli, N., S. Tostain, Y. Vigouroux, C. Agbangla, O. DaÏNou and J.L. Pham. 2006. Farmers' use of wild relative and sexual reproduction in a vegetatively propagated crop. The case of yam in Benin. Molecular Ecology 15: 2421-2431. doi:10.1111/j.1365-294X.2006.02958.x.

Spillane, C. and P. Gepts. 2001. Evolutionary and genetic perspectives on the dynamics of crop genepools. In: D. Cooper, C. Spillane and T. Hodgkin, editors, Broadening the Genetic Base of Crop Production. CABI, Wallingford UK. p. 25-70.

Spindel, J., M. Wright, C. Chen, J. Cobb, J. Gage, S. Harrington, et al. 2013. Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theoretical and Applied Genetics 126: 2699-2716.

Tamiru, M., S. Natsume, H. Takagi, P.K. Babil, S. Yamanaka, A. Lopez-Montes, et al. 2013. Whole genome sequencing of Guinea yam (*Dioscorea rotundata*). First Global Conference on Yam. International Institute of Tropical Agriculture(IITA), Accra, Ghana.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei and S. Kumar. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-2739. doi:10.1093/molbev/msr121.

Terauchi, R., V.A. Chikaleke, G. Thottappilly and S.K. Hahn. 1992. Origin and phylogeny of Guinea yams as revealed by RFLP analysis of chloroplast DNA and nuclear ribosomal DNA. Theoretical and Applied Genetics 83: 743-751. doi:10.1007/BF00226693.

Tortoe, C., P.-T. Johnson, L. Abbey, E. Baidoo, D. Anang, S.G. Acquaah, et al. 2012. Sensory properties of pre-treated balst-chilled (*Dioscorea rotundata*) as a convenience food product. African Journal of Food Science and Technology 3: 59-65.

Yildiz, M. 2013. Plant Responses at Different Ploidy Levels In:Current Progress in Biological Research.InTech.

Zamir, D. 2013. Where Have All the Crop Phenotypes Gone? PLoS biology 11: e1001595.

Zannou, A., A. Ahanchede, P. Struik, P. Richard, J. Zoundjihekpon, R. Tossou, et al. 2004. Yam and cowpea diversity management by farmers in the Guinea-Sudan transition zone of Benin. NJAS 52: 393-420.

Zannou, A., P. Richards and P.C. Struik. 2006. Knowledge on yam variety development: insights from farmers' and researchers' practices. Knowledge Management for Development Journal 2: 30-39.

Chapter 4

4. Morphological, SSR and ploidy analysis of aerial tuber producing accessions of *D.alata* L. for its potential utilization as planting material

4.1 Background and Justification

Dioscorea alata L. commonly called water yam is among the top ten staple and agriculturally important yam species (Lebot, 2009). Its wide adaptation and cultivation makes it one of the most important *Dioscorea* species used for food. Moreover it has relatively better agronomic flexibility than other cultivated yam species due to its ease of propagation and high multiplication ratio from vegetative propagules (Petro, et al., 2011). Tubers from this species are well known for their high nutritional content (Siqueira, et al., 2012).

Poor reproductive development in yams (*Dioscorea* spp.) has been reported to be associated with the polyploid nature of the crop (Egesi, et al., 2002). The increased ploidy level is generally considered to enable polyploid genotypes to grow in a wide range of environments and ploidy increases are used as sources of variability for crop improvement (Jan, et al., 1988, Yildiz, 2013). However, the effects of increased ploidy level are not currently predictable, as in some cases polyploid plants can have less performance (the ability to grow across wider range of environments and

having desirable traits) than diploids indicating poor fertility (Allario, et al., 2011). Determining the ploidy levels and its effect on phenotypic performance is therefore important for the rational utilization of yam germplasm (and ploidy manipulations) in breeding program.

Due to poor production and germination of botanical seeds, yam production is generally restricted to clonal propagation, mainly from underground tubers. The tuber propagules that are set-aside for next season planting accounts for 30% of the total harvest (Kabeya, et al., 2013), which otherwise would have been used for consumption. Moreover, a significant percentage of the tuber propagule material is lost due to poor post-harvest management practices. The transportation cost of this bulk material is also another challenge to yam production. Matsumoto, et al. (2010) further indicated the limited productivity of water yam in which only a single or few tubers per plant is obtained and suggested the need for clonal propagation through vine cuttings.

Vine cuttings are an excellent option for yam propagation because they can provide disease-free materials (once healthy mother plant that provides the cuttings is selected and well inspected) and allow for several cuttings per plant, as well as saving yam tubers for harvest. However, vine cuttings have limitations in relation to current requirements for trained manpower, special media, and at least 150 days to produce the minitubers. Vine cuttings that consider the aerial tuber producing

potential of a variety could have significant potential for boosting yam production as an alternative propagating material.

Aerial tubers also called bulbils develop from an accessory bud found in the abaxial side to the axillary bud next to leaf petiole in the leaf axils. Aerial tubers serve as a means of vegetative reproduction and for dispersal of plants (Walck, et al., 2010). Among the main cultivated yams *D. bulbifera*, *D. japonica*, *D. opposita* and *D. pentaphylla* bear bulbils (Okagami and Tanno, 1991). Some *Dioscorea* species such as *D. polystachya* (Raz, 2002) in North America and *D. bulbifera* rely almost entirely on aerial tubers for reproduction. *D. alata* is also known to have aerial tubers in some accessions. However, its utilization for propagation or yam production and the variation among aerial tuber producing varieties is not yet reported.

Utilizing aerial tubers as a planting material could offer huge advantages for farmers by saving the underground harvest for consumption, which is usually used as seed. Making use of aerial tubers could minimize transportation costs, encourages large-scale yam production, as there is no need to store the consumable underground tuber for the purpose of next season planting.

Several studies have been reported on genetic diversity study of *D. alata* (Egesi, et al., 2006, Lebot, et al., 1998, Malapa, et al., 2005, Obidiegwu, et al., 2009, Siqueira, et al., 2012) but no report has been made in particular to aerial tuber producing

varieties regardless of its potential utilization. The aims of this study were therefore to assess the morphological, ploidy level and genetic variability across aerial tuber and non-aerial tuber producing accessions of the *D. alata* germplasm collection.

4.2 Materials and methods

4.2.1 Plant material, Experimental layout and data collection

All *D.alata* materials, originally comprising 813 accessions under International Institute of Tropical Agriculture (IITA) field genebank, were assessed for aerial tuber production over two consecutive years (2011and 2012). This has identified 121 accessions as aerial tuber producing accessions. The materials were planted following standard procedures (Dumet and Ogunsola, 2008) as routine field genebank regeneration during the main growing season at the IITA experimental plot, Ibadan (Latitude: 7°30'8''N; Longitude: 3°54'38''E), Nigeria. A total of 21 morphological descriptors for yam (Table 4.1) obtained from International Plant Genetic Resources Institute (IPGRI) and the International Institute of Tropical Agriculture (IITA) (IPGRI/IITA, 1997) were used to characterize the aerial tuber producing accessions (n=121) plus accessions with out aerial tubers (n=18).

4.2.2 Determination of ploidy level

Young leaves were collected from a total of 139 individuals representing both aerial tuber producing (n=121) and accessions with out aerial tubers (n=18). Leaf blade of

approximately 5mm^2 was chopped to homogenize the tissue for flow cytometry analysis using Ploidy Analyzer (Partec GmbH, Germany). The homogenate was filtered with 50μm pore size nylon filter. Extracted nuclei were stained with DAPI (4, 6-diamino-s phenylinode). The flow cytometry analyses were conducted at the rate of 5-20 nuclei per second. The intensity of fluorescence was measured to determine the ploidy. A diploid *D. alata* accession (TDa 1375, 2x=2n=40 was used as a standard. The amount of DNA in a sampled plant was determined in relation to that of the standard to identify the ploidy level. The flow cytometer was adjusted so that the peak representing the G1 nuclei of the standard was set at channel 50.

Table 4.1. Morphological descriptors used to characterize aerial and non-aerial tuber producing *D. alata* accessions.

Morphological traits	Parametric representation
Presence of aerial tuber	0=absent and 1=present
Leaf shape	1=cordate, 2=sagittate and 3=hastate
Wing color	1=green, 2=green with purple edges and 3=purple
Petiole color	1=green with purple base, 2=green with purple leaf junction, 3= purplish green with purple at both ends and 4=green
Petiole wing color	1=green, 2=green with purple edges and 3=purple
Young leaf color	1=green, 2=pale green, 3=yellowish, 4=purplish green and 5=purple
Tuber shape	1=round, 2=oval, 3=ovaloblong, 5=flattened and 6=irregular
Tendency of tuber to branch	0=no branch, 3=slightly branched, 5=branched
Spiny roots on tuber surface	0=no, 3=few and 7=many
Roots on tuber surface	0=no, 3=few and 7=many
Place of roots on tuber	1=lower, 2=middle, 3=upper and 4=entire tuber

Prickly appearance on tuber	0=no and 1=yes
Wrinkles on tuber surface	0=no, 3=few and 7=many
Presence of cracks	0=absent and 1=present
Flesh color of upper part of the tuber	1=white, 2=creamy white, 3=yellow, 4=purplish, 5=purplish white, 6=creamy, 7=brownish white and 8=deep purple
Flesh color of middle part of the tuber	1=white, 2=creamy white, 3=yellow, 4=purplish, 5=purplish white, 6=creamy, 7=brownish white and 8=deep purple
Flesh color of lower part of the tuber	1=white, 2=creamy white, 3=yellow, 4=purplish, 5=purplish white, 6=creamy, 7=brownish white and 8=deep purple
Tuber beneath skin color	1=white, 2=creamy white, 3=yellow, 4=purplish, 5=purplish white, 6=creamy, 7=brownish white and 8=deep purple
Sprout count	Total sprout count per accession
Number of aerial tuber	Count
Number of underground tuber	Count

4.2.3 Molecular characterization and fragment analysis

DNA was extracted from lyophilized leaf samples of individual accessions with aerial tuber (n=121) and without aerial tuber (n=6) using Qiagen-DNeasy plant mini kit. A total of eight SSR primers were used for genotyping of the *D. alata* accessions (Table 4.2). The polymerase chain reaction was conducted on Veriti 96 well thermal cycler (Applied Biosystems, USA). PCR amplification was carried out in a 20 µl reaction mixture containing 25ng DNA, 0.8mM of the dNTPs, 3mM of Mgcl2, labeled forward and unlabelled reverse primers, each 10µM, 0.9x Taq buffer and 1 unit of Taq Polymerase. The forward primer was 5'-labeled with one of the fluorochromes, 6-FAM. The amplification program was 3 min initial denaturation at 94°C, 35 cycles of 45 sec denaturation at 94°C, 30 sec primer annealing at 55°C, and 1min extension at 72°C; with a final extension of 10 min at 72°C. The amplified DNA fragments were separated by electrophoresis on 1.5% w/v agarose to check the amplification product. Fragment analysis through capillary electrophoresis using Genetic Analyzer (ABI 3130xl) was performed. The GeneMapper software version 4.0 developed by Applied Biosystems was used for sizing and genotyping microsatellite data generated through capillary electrophoresis.

Table 4.2. List of SSR primers, number of alleles scored, expected fragment size range and Polymorphic Information Content (PIC).

Primer	Sequences (5'-3')	Size range (bp)	Number of alleles	PIC
YSR24	F: GGTGTTGTTGGGTTTCATTGTC	110-145	13	0.36
	R: TCCCTCTTCTCATTTCACTCCC			
YSR33	F: ACCATGGGATGAAGGGAAGG	163-234	8	0.33
	R: GCATATGGTGCATGGGAGC			
YSR36	F: CCTTACCACCGGACTCCTC	118-177	10	0.50
	R: TGCAGCAATACACCGGAAC			
YSR53	F: CTCATAAGCAGAGCCTTCTCTC	333-350	6	0.50
	R: TACAGTCCCTGTTTGAGCATAG			
YSR65	F: ACAAATGCACGCTCTGAAGG	145-184	7	0.41
	R: GGGCAGTAGAATTTGGTGCG			
YSR66	F: ATATTGACTGACCACCAGATCA	207-260	5	0.46
	R: GAAGAGTCTTGGATTTCTACCA			
YSR74	F: TGGTGTTTGAGAATGGAGGATTG	450-520	4	0.42
	R: ACTTGATCTTTGTCTTGATGGC			
YSR75	F: TCGCTCAACCTAATCCTCTATT	308-355	5	0.50
	R: TCAAACCAGCCAAAACATC			
Average			7.25	0.43

4.2.4 Morphological data analysis

Multiple Correspondence Analysis (MCA) was performed with a set of 21 categorical phenotypic variables using FactoMineR package (Lê, et al., 2008) in R software (R Core Team, 2013) to detect the structure of individuals and its correlation with aerial tuber production.

4.2.4 Molecular data analysis

To evaluate the performance of SSR markers for the study, the polymorphic information content of each marker was calculated with the formula described by Roldán-Ruiz, et al. (2000) PICi =2*fi*(1-*fi*), where PICi is the polymorphic information content of marker 'i', *f*i is the frequency of the marker alleles which were present and 1-*f*i represents the frequency of marker alleles which were absent.

Genetic diversity was estimated by Shannon diversity index (Lewontin, 1972) as

$$H' = -\Sigma\ pi \log pi$$

where *p*i is the frequency of a given allele for each population. Shannon diversity index was used to measure the total diversity (Hsp) as well as the mean intra-population diversity (Hpopn). The proportion of diversity between populations was then calculated as (1-Hpop/Hsp). The population was defined based on geographic location where the materials were originally obtained.

Genetic diversity based on the presence of more than one allele in a given loci within a population as Number of Polymorphic Loci (NPL), Percent Polymorphic Loci (PPL), Nei's gene diversity (H), Shannon's information index (I) and population differentiation as GST was calculated using POPGENE version 1.32 software (Yeh, et al., 1997).

A neighbor-joining tree was generated for 127 individual accessions based on 58 loci obtained from 8 SSR markers using DARwin software version 5.0 (Perrier and Jacquemoud-Collet, 2006). To further look at the patterns of variation among individual accessions, a three-dimensional Principal Coordinate Analysis (PCoA) was performed based on Jaccard's coefficient (Jaccard, 1908). Jaccard's coefficient was calculated using PAST software version 1.18. (Hammer, et al., 2001). Using the first three axis PCoA were constructed with STATISTICA software version 6.0 (Statistica Stat Soft, 2001).

4.3 Results

4.3.1 Morphological variation within yams for aerial tuber production

Multiple correspondence analyses indicated variable projections across different dimensions and non-aerial tuber producing accessions were explained by different dimension from accessions with aerial tubers (Figure 4.2). The two dimension of the MCA explained about 23% of the total variance. The individual's scatterplot

appeared homogeneous where there is no particular group of individuals except for non-aerial tuber producing accessions. The individuals with non-aerial tubers tended to cluster and form a distinct group. A similar pattern of distribution was observed for some phenotypic variables (including leaf shape, leaf color, wing color and tuber flesh color) with the pattern of aerial tuber production. Moreover, with the confidence ellipse the above mentioned variables are both much linked to the second dimension (Figure 4.2). A group of accessions with sagittate leaf shape, low to high anthocyanin pigmentation (pale green to purple leaf and wing) and creamy white to deep purple flesh tuber color was mostly associated with aerial tuber production except in a few cases (Figure 4.2). Accessions with cordate leaf shape were all aerial tuber producing and showed green with purple edge to completely green wing color, pale green to purplish green leaf and mostly creamy white with few yellowish flesh tuber color. The extent of aerial tuber production increases as the leaf shape changes from hastate (no aerial tuber), to sagittate (mostly aerial tuber producing with few exceptions) to cordate were aerial tuber is always present. Similarly, the mean aerial tuber count per sprout is highest for cordate followed by sagittate and none in hastate leaf types (Figure 4.3). However, accessions with hastate leaf shape displayed a higher mean number of underground tubers per sprout (Figure 4.4).

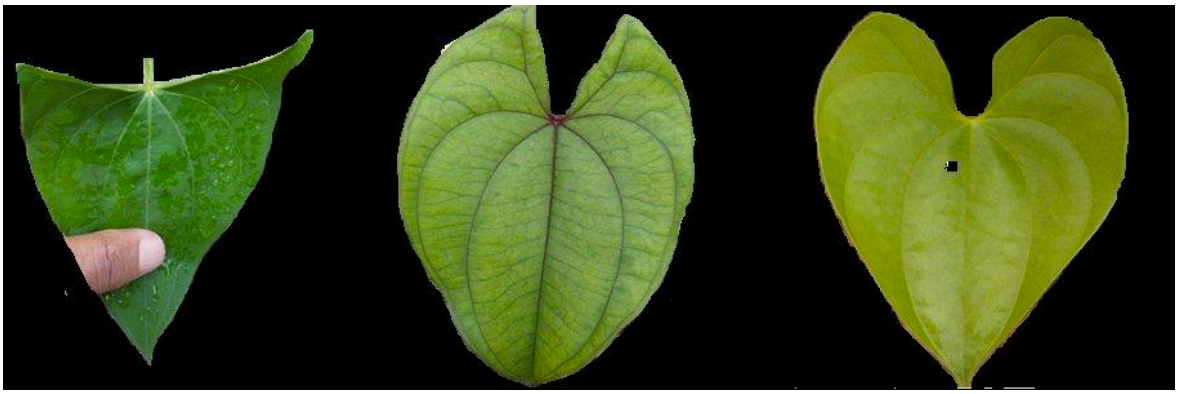


Figure 4.1a. From left to right; hastate, sagittate and cordate leaf shapes representing accessions with different extent of aerial tuber production.

Figure 4.1b. Aerial tuber 'primordium' developing from axillary buds.

Figure 4.1c. Aerial and underground tubers harvested from a single aerial tuber.

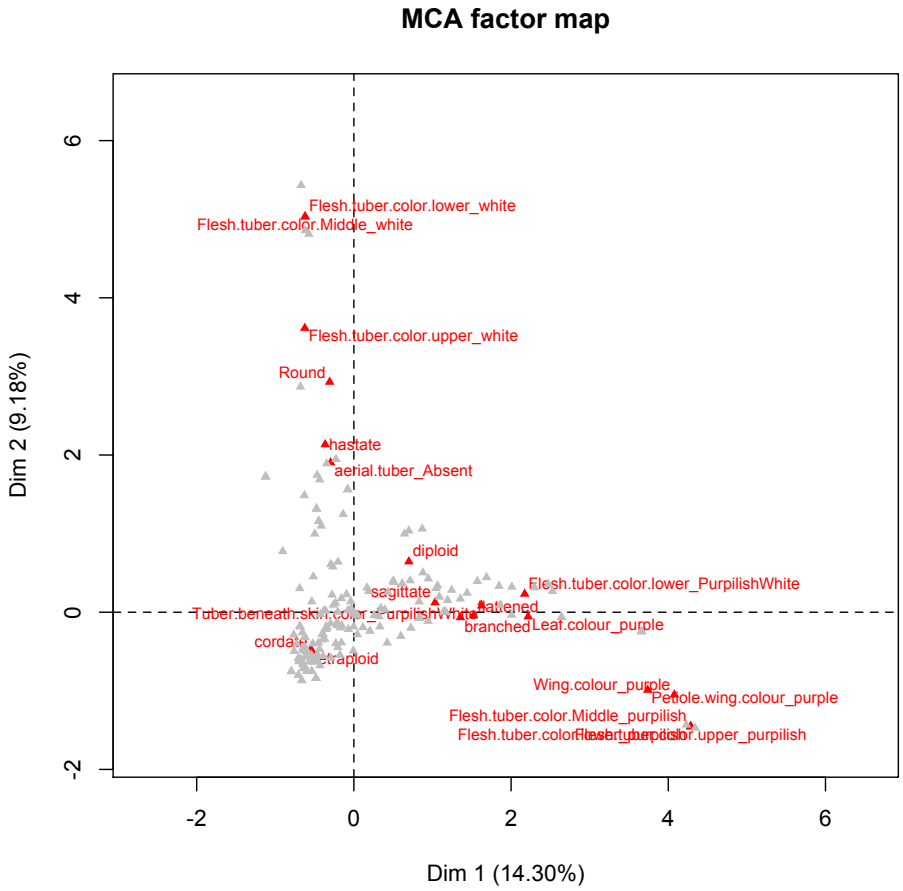
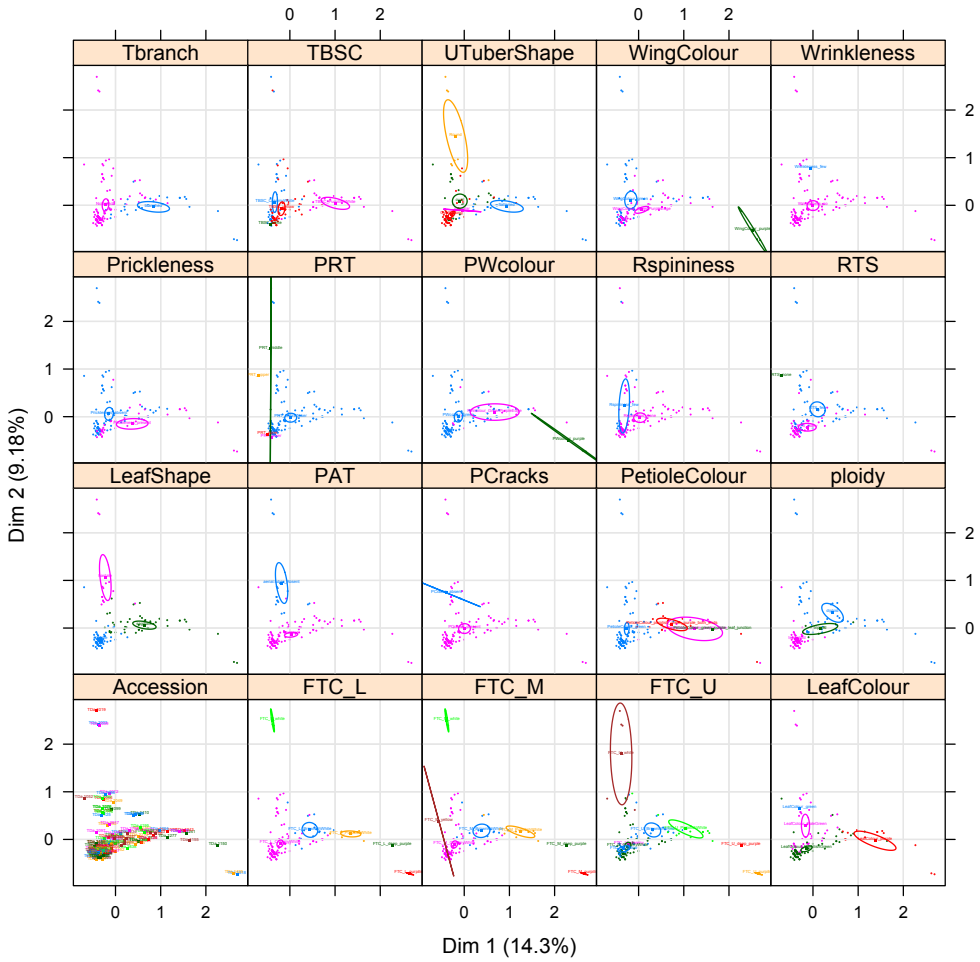


Figure 4.2a. MCA showing 20 most discriminant variables (red triangle) and individual accessions (grey triangle) of *D. alata* based on 21 categorical morphological descriptors. The 20 most discriminant variables include white flesh tuber color at lower part or upper part or middle part, purplish white flesh color at lower part, purple flesh tuber color at lower part or middle part or upper part) purplish-white tuber beneath skin color, diploid, tetraploid, purple color leaf or wing or petiole wing, tuber shape (round, flattened and branched), absence of aerial tuber and leaf shape (hastate, sagittate and cordate).



Key

FTC_M: Flesh Tuber Color_Middle part; **FTC_L**: Flesh Tuber Color_Lower part; **FTC_U**: Flesh Tuber Color_Upper part; **TBSC**: Tuber Beneath Skin Color; **PWcolor:** Petiole Wing color; **PCracks:** Presence of Cracks; **PRT**: Presence of Roots on Tuber surface; **Tbranch**: Tuber branching; **UTuberShape:** Underground Tuber Shape; **PAT:** Presence of Aerial Tuber; **Rspininess:** Root spininess; **RTS:** The number of Roots on Tuber Surface.

Figure 4.2b. MCA showing confidence ellipses around the categories of all the categorical variables used. The plotted confidence ellipses around the categories of variables are to see whether the categories of a categorical variable are significantly different from each other and indicate a) branched tuber, flattened tuber shape, creamy white TBSC, green wing, few wrinkles, no prickliness, roots on entire tuber,

green PWcolor, few root spinnines, few roots on tuber surface, cordate leaf, no aerial tuber, no cracks on tuber, green petiole, diploid, brownish green FTC and green leaf (plot ellipses with blue line); b) no tuber branching, purplish TBSC, irregular tuber, green with purple edge wing, no wrinkles, prickle on tuber, roots on lower tuber, green with purple edge PWcolor, no root spinnines , many roots on tuber surface, hastate leaf, presence of aerial tuber, presence of cracks, green with purple base petiole, tetraploid, creamy white FTC and pale green leaf (plot ellipses with pink line); c) white TBSC, oval tuber, purple wing, roots on middle part of tuber, purple PW, no roots on tuber surface, sagittate leaf, green with purple at both junction of petiole, triploid, deep purple FTC and purplish green leaf (plot ellipses with dark green line); yellow TBSC, oval oblong tuber, no roots on tuber, petiole with purplish green with purple at both ends, purple FTC and leaf (plot ellipses with red line); round tuber, roots on upper part of tuber and purplish white FTC (plot ellipses with yellow line); white lower and middle part FTC and purplish white upper tuber FTC (plot ellipses with green line); and yellow FTC (plot ellipses with dark red line).

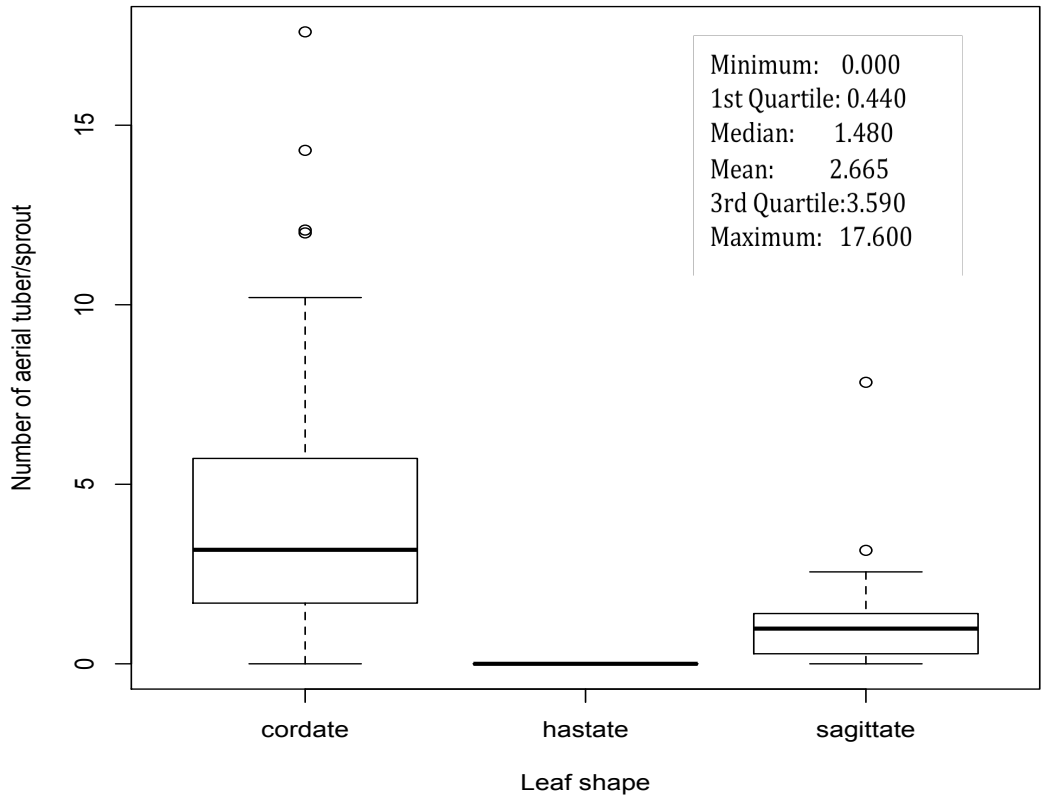


Figure 4.3. Number of aerial tubers per sprout across different group of *D. alata* accessions with different leaf shape. The average/median and spread/inter-quartile range of number of aerial tuber for cordate leaf shape accessions is much larger than sagittate group. The five numbers are the minimum score, lower quartile,

median, upper quartile and the maximum score. The small circles (^{0}) on the boxplot explain outliers.

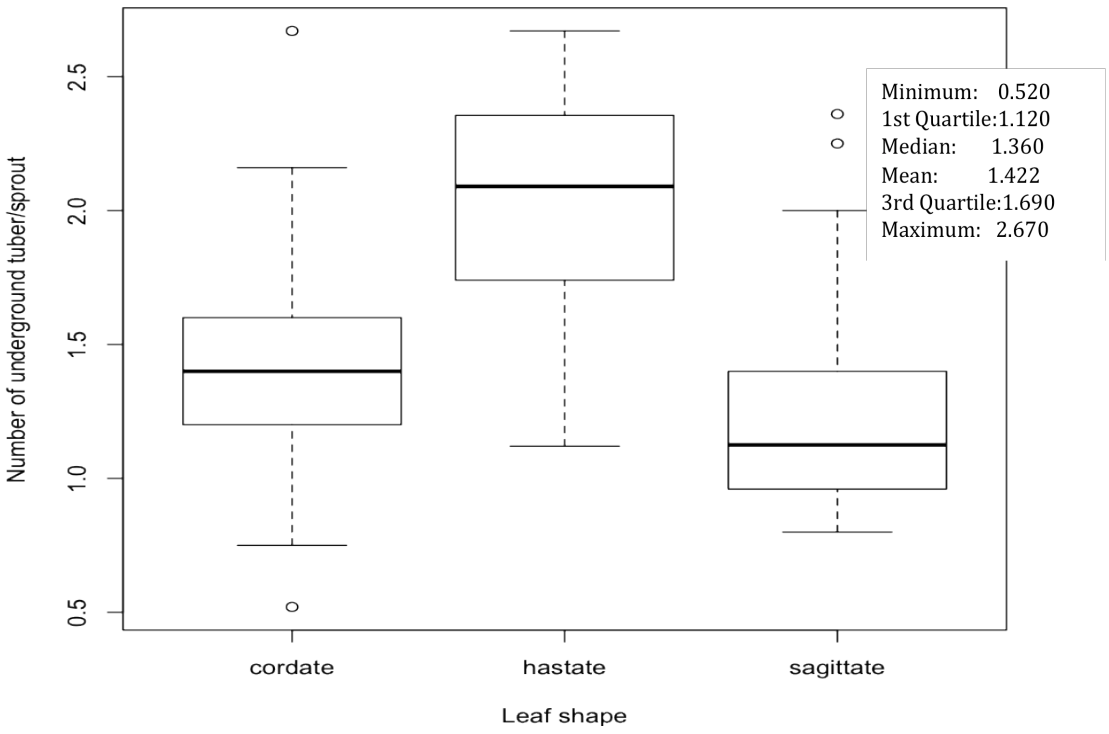


Figure 4.4. Mean number of underground tubers per sprout across different group of *D. alata* accessions with different leaf shape. The average number of underground tuber in hastate is larger than both cordate and sagittate groups. The inter-quartile range is similar for all the three groups. The five numbers are the minimum score, lower quartile, median, upper quartile and the maximum score. The small circles (^{0}) on the boxplot explain outlier.

4.3.2 Ploidy variation of aerial tuber producing yams

Ploidy analysis showed that 53% of the accessions were tetraploid followed by 43% diploid and 4% triploid. The three groups of accessions defined based on phenotypic variables showed different patterns of ploidy. The first group of accessions with hastate leaf shape, no aerial tubers and without anthocyanin pigmentation was correlated with a diploid ploidy (n=15). The second group with sagittate leaf shape, both aerial and non- aerial tuber producing and different extent of pigmentation was mainly found to be diploid (n=44) with the exception of three accessions that were triploid (n=3). The third group with cordate leaf shape, always aerial tuber producing and with anthocyanin pigmentation were both tetraploid (n=74) and triploid (n=3) (Table 4.3). Overall, increases in the number of aerial tubers were observed with increased ploidy level, and as the leaf shape gets progressively roundish (Figure 4.1 and 4.5).

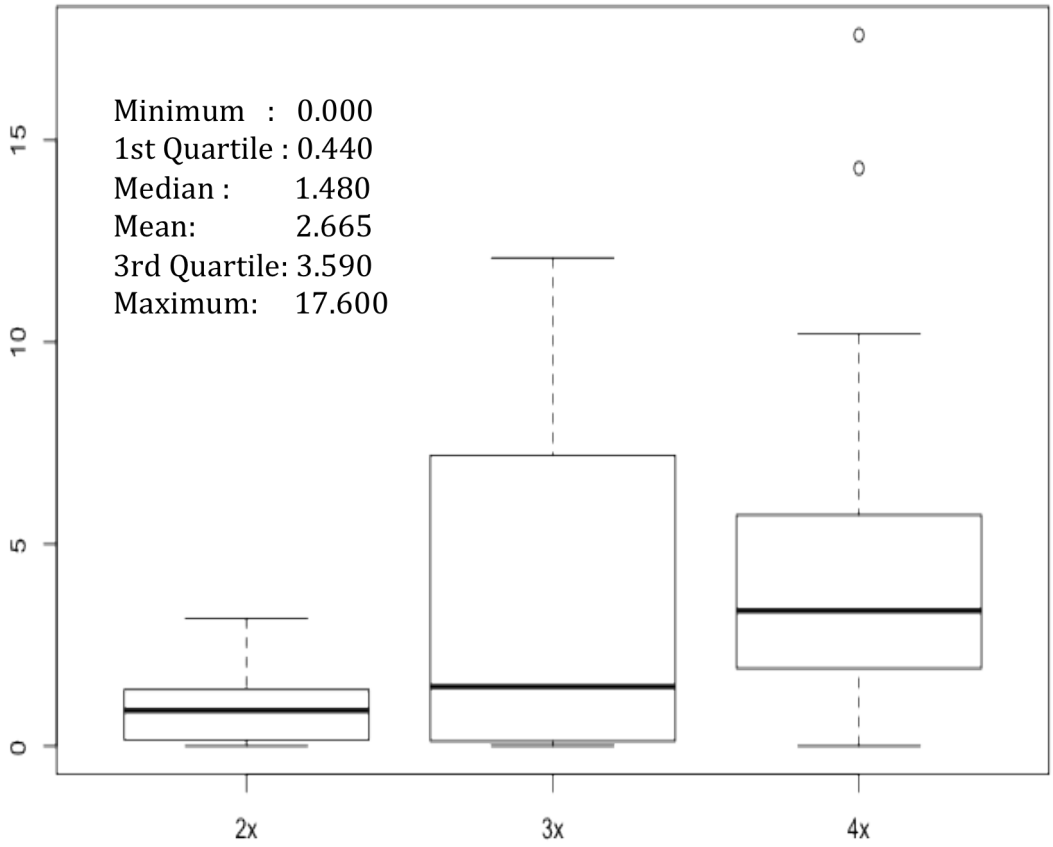


Figure 4.5. Mean number of aerial tubers per sprout across *D. alata* accessions of different ploidy levels (2x=40, 3x= 60 and 4x=80) with number of individual (n) =59, 6 and 74 individuals respectively. The average number of aerial tuber for tetraploid is larger than both triploid and diploid groups. The inter-quartile range is wider in triploid followed by tetraploid and diploid. The five numbers are the minimum score, lower quartile, median, upper quartile and the maximum score. The small circles (^{0}) on the boxplot explain outlier.

Table 4.3. Ploidy level, presence of aerial tubers with respective leaf shape of 139 *D. alata* accessions.

Accession	Ploidy	Leaf shape	Aerial tuber	Accession	Ploidy	Leaf shape	Aerial tuber
WAB 56-104[a]	2x						
TDa 1375[b]	2x	Hastate	Absent	TDa 1170	2x	Sagittate	Present
TDa 1003	2x	Hastate	Absent	TDa 1189	4x	Cordate	Present
TDa 1012	2x	Hastate	Absent	TDa 1192	2x	Hastate	Absent
TDa 1018	4x	Cordate	Present	TDa 1202	4x	Cordate	Present
TDa 1019	2x	Hastate	Absent	TDa 1207	2x	Sagittate	Present
TDa 1023	4x	Cordate	Present	TDa 1211	4x	Cordate	Present
TDa 1036	4x	Cordate	Present	TDa 1220	4x	Cordate	Present
TDa 1037	4x	Cordate	Present	TDa 1222	4x	Cordate	Present
TDa 1038	4x	Cordate	Present	TDa 1224	2x	Sagittate	Present
TDa 1039	2x	Hastate	Absent	TDa 1230	4x	Cordate	Present
TDa 1044	4x	Cordate	Present	TDa 1234	2x	Sagittate	Present

TDa 1046	2x	Hastate	Absent	TDa 1236	2x	Sagittate	Present
TDa 1049	2x	Hastate	Absent	TDa 1237	4x	Cordate	Present
TDa 1050	2x	Hastate	Absent	TDa 1241	2x	Sagittate	Present
TDa 1052	4x	Cordate	Present	TDa 1243	2x	Sagittate	Present
TDa 1066	2x	Hastate	Absent	TDa 1261	4x	Cordate	Present
TDa 1073	4x	Cordate	Present	TDa 1263	2x	Sagittate	Present
TDa 1076	2x	Hastate	Absent	TDa 1267	2x	Sagittate	Present
TDa 1079	2x	Sagittate	Present	TDa 1275	2x	Sagittate	Present
TDa 1090	2x	Sagittate	Absent	TDa 1276	2x	Sagittate	Present
TDa 1091	4x	Cordate	Present	TDa 1277	2x	Sagittate	Present
TDa 1094	2x	Sagittate	Present	TDa 1280	4x	Cordate	Present
TDa 1096	2x	Sagittate	Present	TDa 1284	4x	Cordate	Present
TDa 1097	4x	Cordate	Present	TDa 1285	4x	Cordate	Present
TDa 1099	2x	Sagittate	Present	TDa 1313	4x	Cordate	Present
TDa 1104	2x	Hastate	Absent	TDa 1318	2x	Sagittate	Present
TDa 1107	4x	Cordate	Present	TDa 1319	4x	Cordate	Present

TDa 1113	2x	Hastate	Absent	TDa 1338	4x	Cordate	Present
TDa 1114	4x	Cordate	Present	TDa 1340	4x	Cordate	Present
TDa 1125	2x	Hastate	Absent	TDa 1341	2x	Sagittate	Present
TDa 1137	4x	Cordate	Present	TDa 1344	2x	Sagittate	Present
TDa 1150	2x	Sagittate	Present	TDa 1353	3x	Sagittate	Present
TDa 1151	2x	Sagittate	Present	TDa 1354	4x	Cordate	Present
TDa 1152	4x	Cordate	Present	TDa 1361	4x	Cordate	Present
TDa 1154	4x	Cordate	Present	TDa 1362	2x	Sagittate	Present
TDa 1155	2x	Sagittate	Present	TDa 1364	4x	Cordate	Present
TDa 1156	4x	Cordate	Present	TDa 3277	3x	Sagittate	Present
TDa 1162	2x	Sagittate	Present	TDa 3744	4x	Cordate	Present
TDa 1168	4x	Cordate	Present	TDa 3752	2x	Sagittate	Present
TDa 1390	4x	Cordate	Present	TDa 1385	4x	Cordate	Present
TDa 1394	2x	Sagittate	Present	TDa 3762	4x	Cordate	Present
TDa 1396	2x	Sagittate	Present	TDa 1386	2x	Hastate	Absent
TDa 1410	2x	Sagittate	Present	TDa 3789	4x	Cordate	Present

TDa 1419	4x	Cordate	Present	TDa 3798	4x	Cordate	Present
TDa 1425	2x	Sagittate	Present	TDa 3838	4x	Cordate	Present
TDa 1426	4x	Cordate	Present	TDa 3903	3x	Cordate	Present
TDa 1428	2x	Sagittate	Present	TDa 3911	4x	Cordate	Present
TDa 1438	4x	Cordate	Present	TDa 3262	4x	Cordate	Present
TDa 1440	4x	Cordate	Present	TDa 3268	4x	Cordate	Present
TDa 1441	4x	Cordate	Present	TDa 3271	4x	Cordate	Present
TDa 1446	4x	Cordate	Present	TDa 3272	2x	Sagittate	Present
TDa 1451	4x	Cordate	Present	TDa 3917	2x	Sagittate	Absent
TDa 1452	2x	Sagittate	Present	TDa 3920	4x	Cordate	Present
TDa 1458	4x	Cordate	Present	TDa 3925	4x	Cordate	Present
TDa 1465	2x	Sagittate	Absent	TDa 3926	4x	Cordate	Present
TDa 1469	4x	Cordate	Present	TDa 4046	4x	Cordate	Present
TDa 1471	4x	Cordate	Present	TDa 4049	2x	Sagittate	Present
TDa 2847	3x	Cordate	Present	TDa 4062	4x	Cordate	Present
TDa 2849	4x	Cordate	Present	TDa 4125	2x	Sagittate	Present

TDa 2871	2x	Sagittate	Present	TDa 4129	4x	Cordate	Present
TDa 2874	4x	Cordate	Present	TDa 4134	2x	Sagittate	Present
TDa 2883	4x	Cordate	Present	TDa 4194	2x	Sagittate	Present
TDa 3131	3x	Sagittate	Present	TDa 4195	2x	Sagittate	Present
TDa 3146	4x	Cordate	Present				
TDa 3157	4x	Cordate	Present				
TDa 3161	2x	Sagittate	Present				
TDa 3163	4x	Cordate	Present				
TDa 3186	4x	Cordate	Present				
TDa 3189	2x	Hastate	Absent				
TDa 3199	4x	Cordate	Present				
TDa 3201	4x	Cordate	Present				
TDa 3204	3x	Cordate	Present				
TDa 3213	2x	Sagittate	Present				
TDa 3225	2x	Sagittate	Present				
TDa 3229	4x	Cordate	Present				

TDa 3246	4x	Cordate	Present				
TDa 3258	4x	Cordate	Present				

[a]*Oryza sativa* variety used as an external standard

[b]*D. alata* accession used as internal standard

4.3.3 SSR Polymorphism across yam accessions

The analysis of genetic diversity using SSRs involved a total of 58 alleles scored with an average of 7.25 alleles per primer with the highest allele number (13) in YSR24 and the lowest (4) with primer YSR74. The PIC values of the polymorphic SSR markers ranged from 0.33 to 0.50 with average 0.43. The highest PIC value (PIC=0.50) was obtained in 3 of the 8 markers; YSR36, YSR53 and YSR75 and hence highly informative. The remaining SSRs were reasonably informative and had a PIC value of ≥ 0.33 (Table 4.2).

4.3.4 Genetic diversity and its partitioning across yam populations

Shannon's diversity index (H), for populations of different geographic origin (Table 4.4), ranges from 0.52 to 0.61 with mean genetic variation for population (Hpopn=0.57) and mean genetic variation for entire data (Hsp=0.61). The genetic variation within population (Hpopn/Hsp= 0.98) was greater than genetic variation between populations (1-Hpopn/Hsp= 0.02). The genetic diversity indexes (Table 4.5) showed that the populations from Togo and Nigeria have the highest genetic diversity, whereas the lowest was within the populations from Sierra Leone. In addition, population differentiation as Gst value of 0.36 between population and Gst

value of 0.052 within population representing different pattern of aerial tuber production and leaf shape was obtained.

Table 4.4. Shannon's diversity index (H) within and among *D. alata* populations.

Population	H
Benin	0.57
Cote d'Ivoire	0.59
Ghana	0.52
Nigeria	0.61
Sierra Leone	0.56
Togo	0.58
Hpopn	0.57
Hsp	0.61
Hpopn/Hsp	0.98
1-Hpopn/Hsp	0.02

Hpopn = mean genetic variation for population, Hsp = mean genetic variation for the entire data, Hpopn/Hsp = proportion of genetic variations within populations and 1-Hpopn/Hsp= proportion of genetic variations between different populations.

Table 4.5. Genetic Variation Statistics for 6 populations of different geographic origin of *D. alata* accessions.

Population	NPL	PPL	H	I
Benin	41	70.69	0.20	0.31
Cote d'Ivoire	23	39.66	0.15	0.22
Ghana	44	75.86	0.20	0.32
Nigeria	49	84.48	0.22	0.34
Sierra Leone	18	31.03	0.11	0.17
Togo	52	89.66	0.22	0.34
Overall	55.00	94.83	0.22	0.35

NPL= number of polymorphic loci; PPL= percent polymorphic loci; H= Nei's gene diversity and I= Shannon's information index

4.3.5 Cluster Analysis

The neighbor-joining tree generated using SSR markers revealed a clear distinction of accessions according to their pattern of aerial tuber production (Figure 4.6). The groups with cordate leaf shape accessions, also correlated with consistent aerial tuber production, form a distinct cluster group. All the accessions without aerial tubers were found within the accessions with sagittate leaf shape cluster group.

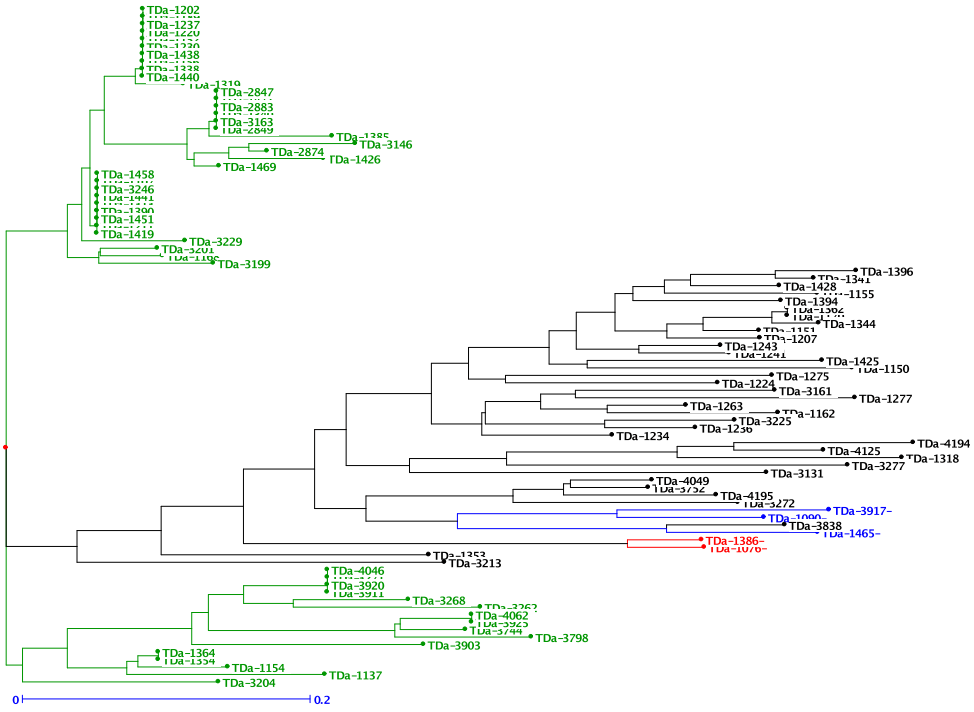


Figure 4.6. Rooted neighbor-joining tree generated for *D. alata* accessions using 58 alleles of 8 SSR markers with midpoint rooting method. It shows different cluster group of accessions producing aerial tubers (green and black) and individuals without aerial tubers (blue and red). Individuals with more than 50% missing data were excluded.

4.3.6 Principal Coordinate Analysis (PCoA) using SSR data

The first 3 coordinates of the PCoA having eigen values of 7.60, 2.69 and 1.58 with variance of 34.28%, 12.16% and 7.12% respectively were used to illustrate the grouping of individuals using three coordinates. The PCoA revealed a distinct clustering of the accessions into different groups mainly according to the pattern of

aerial tuber production, ploidy levels and leaf shape (Figure 4.7). The groups include a) a distinct group of accessions with aerial tuber, all tetraploid and cordate leaf shape and b) a group of mainly diploids having sagittate leaf shape with aerial tuber including non-aerial tuber producing accessions (both sagittate and hastate leaf shape) with the triploids admixed.

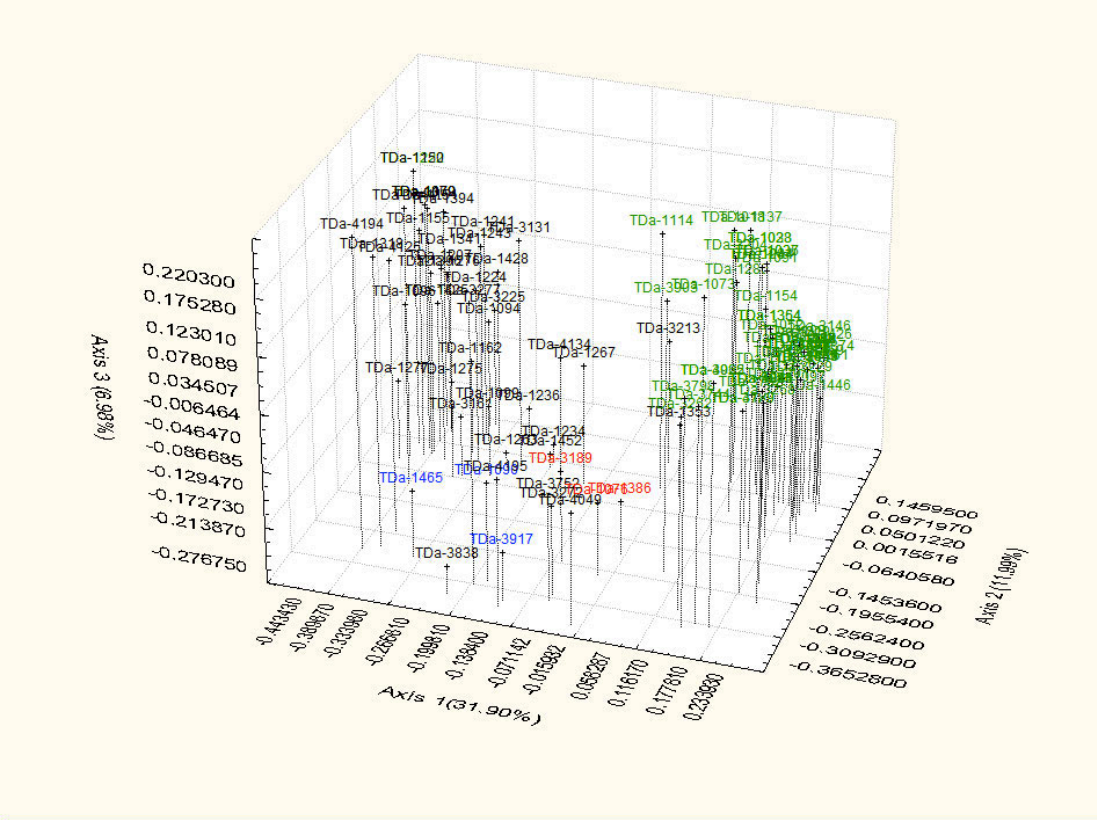


Figure 4.7. A PCoA indicating genetic relationships among 127 individuals of aerial tuber producing (green and black) and non-aerial tuber producing (blue and red) *D. alata* accessions inferred from Jaccard's similarity matrix based on 58 alleles of 8 SSR markers. The percent variation explained by each axis is indicated inside parentheses.

4.4 Discussion

4.4.1 Morphological and ploidy variation across aerial tuber producing D. alata accessions

The MCA shows that the shape, flesh tuber color of underground tubers; shape and color of leaves; stem wing color, and absence and presence of aerial tubers were among the variables with high contribution to variability within the species as the most discriminant traits. Likewise, an assessment of morphological variation among accessions of *D. alata* from Malaysia based on 47 morphological variables indicated that shape, size and flesh color of underground tubers; shape and color of aerial tubers; position, shape, size and vein color of the leaves; petiole color; shoot growth rate; and number of days for shoots to germinate were among the characters contributing largely to the species variability (Hasan, et al., 2008). The current observation indicated that the aerial tuber production feature of accessions was associated with phenotypic variables such as leaf shape, anthocyanin pigmentation and ploidy level where all accessions with hastate leaf shape were projected on the same MCA dimension with absence of aerial tubers, no anthocyanin pigmentation (white flesh tuber color) (Figure 4.2) and all diploid (Table 4.3). Whereas the sagittate and cordate leaf shape accessions were explained by different dimensions of MCA and associated with presence of aerial tuber, anthocyanin pigmentation and higher ploidy levels. The leaf shape variation was further observed to have

correlation with ploidy level. An increase in ploidy level among *D. alata* accessions as the leaf shape gets roundish were similarly reported (Babil, et al., 2010).

A study based on microsatellite segregation analysis in four different progenies has demonstrated that *D. alata* accessions are diploid, triploid and tetraploid (2n = 2x, 3x, 4x), respectively, and not tetraploid, hexaploid and octaploid, as previously assumed (Arnau, et al., 2009). Likewise, three ploidy levels (2x=40, 3x=60 and 4x=80) were observed among *D. alata* accessions investigated in the current study. In addition, variation in leaf shape, pattern of anthocyanin pigmentation (green, purplish green, purple color of leaf, wing and petiole) and aerial tuber production were observed as the ploidy level changes. An increase in aerial tuber production as the ploidy level increase in *D. alata* accessions is an advantage of polyploidy in yam improvement. The increased in ploidy level was reported to have several advantages in crop plant through enabling polyploid genotypes to resist against biotic and abiotic stress factors, consequently to grow in the wide range of environments and as sources of variability for plant improvement, although it cannot be always anticipated (Jan, et al., 1988, Yildiz, 2013).

4.4.2 Genetic diversity and Population differentiation

SSR markers have been widely used in genetic diversity assessments of different plant species including yams (Mignouna, et al., 2003, Obidiegwu, et al., 2009).

According to (Paál, et al., 2013) highly informative marker has the PIC ≥ 0.5 whereas, reasonably informative marker has the value 0.5 > PIC ≥ 0.25. The markers used in this study (Table 4.2), with mean (PIC=0.43), were therefore all informative indicating their usefulness for genetic diversity assessment of *D. alata* accessions. Moreover, the markers were useful in revealing the pattern of aerial tuber production in addition to showing variability among individual accessions.

D. alata is generally the most diverse *Dioscorea* species growing across different tropical regions of the world. Few recent studies indicate high morphological and molecular diversity among different collections of *D. alata* (Dansi, et al., 2013, Obidiegwu, et al., 2009, Siqueira, et al., 2012). Similarly, the Shannon's diversity index in this study shows high genetic variability among individual accessions (Table 4.4). Both neighbor joining tree and PCoA show distinction of accessions according to the pattern of aerial tuber production. The differentiation (Gst=0.36) between population representing different pattern of aerial tuber production and leaf shape and Gst value of 0.052 within each population showed high differentiation between and moderate genetic differentiation within population according to Wright's classification (Wright, 1978) respectively. The individual accession based neighbor joining tree and PCoA reveals no distinction according to the geographic origin. This is further supported by a study on *D. alata* accessions of IITA genebank originally collected from 9 West African countries indicating no relatedness of the accessions and their geographical area of collection (Obidiegwu,

et al., 2009). Similarly the current study also shows no accession relatedness according to the geographic origin (data not shown). This could be due to possible exchange of planting materials across different countries.

4.4.3 Significance of the study in yam breeding

In addition to saving the consumable underground tuber, the use of aerial tuber as planting material could have significant contribution in solving the problem with yam planting material. Moreover, the mini tuber that is produced through vine cutting and aerial tubers are the same in terms of origin as both are formed at the base of axillary buds. Hence, vine cutting technique for mini tuber production (Kikuno, et al., 2007) that considers the aerial tuber production potential of the mother plant would be more effective and efficient. Development of improved varieties with aerial tuber is therefore important to consider in yam breeding strategies. In addition to its help in understanding the genetic diversity of aerial tuber producing accessions, the current morphological descriptors, SSR markers and ploidy variation can also be used for efficient screening of aerial tuber producing accessions of *D. alata*.

4.5. References

Allario, T., J. Brumos, J.M. Colmenero-Flores, F. Tadeo, Y. Froelicher, M. Talon, et al. 2011. Large changes in anatomy and physiology between diploid Rangpur lime (*Citrus limonia*) and its autotetraploid are not associated with large changes in leaf gene expression. Journal of Experimental Botany 62: 2507-2519. doi:10.1093/jxb/erq467.

Arnau, G., A. Nemorin, E. Maledon and K. Abraham. 2009. Revision of ploidy status of *Dioscorea alata* L. (Dioscoreaceae) by cytogenetic and microsatellite segregation analysis. Theor Appl Genet 118: 1239-1249. doi:10.1007/s00122-009-0977-6.

Babil, P.K., K. Irie, H. Shiwachi, Y. Tun, H. Toyohara and H. Fujimaki. 2010. Ploidy variation and their effects on Leaf and stoma traits of water yam (*Dioscorea alata* L.) collected in Myanmar. Trop.Agr.Develop. 54: 132-139.

Dansi, A., H. Dantsey-Barry, I. Dossou-Aminon, E. N'Kpenu, A. Agre, Y. Sunu, et al. 2013. Varietal diversity and genetic erosion of cultivated yams (*Dioscorea cayenensis* Poir-*D. roundata* Lam complex and *D. alata* L.) in Togo. International Journal of Biodiversity and Conservation 5: 223-239. doi:10.5897/IJBC12.131.

Dumet, D. and D. Ogunsola. 2008. Regeneration guidelines: yams, Crop Specific Regeneration Guidelines. In: M. E. Dulloo, I. Thormann, M. A. Jorge and J. Hanson, editors, CGIAR System-wide Genetic Resource Programme. Rome.

Egesi, C.N., R. Asiedu, G. Ude, S. Ogunyemi and J.K. Egunjobi. 2006. AFLP marker diversity in water yam (*Dioscorea alata* L.). Plant Genetic Resources 4: 181-187. doi:doi:10.1079/PGR2006121.

Egesi, C.N., M. Pillay, R. Asiedu and J.K. Egunjobi. 2002. Ploidy analysis in water yam, Dioscorea alata L. germplasm. Euphytica 128: 225-230. doi:10.1023/A:1020868218902.

Govaerts, R., P. Wilkin and R.M.K. Saunders. 2007. World Checklist of Dioscoreales. Yams and their allies. The Board of Trustees of the Royal Botanic Gardens, Kew. p. 1-65.

Hammer, O., D.A.T. Harper and P.D. Ryan. 2001. PAST: Paleontological statistics software package for education and data analysis. Palaeontologia electronica 4: 9.

Hasan, S.M.Z., A.A. Ngadin, R.M. Shah and N. Mohamad. 2008. Morphological variability of greater yam (*Dioscorea alata* L.) in Malaysia. Plant Genetic Resources 6: 52-61. doi:doi:10.1017/S1479262108920050.

IPGRI/IITA. 1997. Descriptors for Yam (*Dioscorea* spp.). International Institute of Tropical Agriculture, Ibadan, Nigeria/International Plant Genetic Resources Institute, Rome, Italy.

Jaccard, P. 1908. Nouvelles recherches Sur la distribution florale. Bull.Soc.Vaud.Sci.Nat 44: 223-270.

Jan, C.C., J.M. Chandler and S.A. Wagner. 1988. Induced tetraploidy and trisomic production of Helianthus annuus L. Genome 30: 647-651. doi:10.1139/g88-109.

Kabeya, M.J., U.C. Kabeya, B.D. Bekele and H. Kikuno. 2013. Vine cuttings technology in food yam (*Discorea rotundata*) production. Asian Journal of Plant Science and Research 3: 107-111.

Kikuno, H., R. Matsumoto, H. Shiwachi, H. Toyohara and R. Asiedu. 2007. Comparative effects of explants sources and age of plant on rooting, shooting and tuber formation of vine cuttings from yams (*Dioscorea* spp.). Japanese Journal of Tropical Agriculture 51.

Lê, S., J. Josse and F. Husson. 2008. FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software. Journal of Statistical Software 25: 1-18.

Lebot, V. 2009. Tropical root and tuber crops: cassava, sweet potato, yams and aroids. CABI Publishers, Wallingford,UK: CABI pp. 413.

Lebot, V., B. Trilles, J.L. Noyer and J. Modesto. 1998. Genetic relationships between *Dioscorea alata* L. cultivars. Genetic Resources and Crop Evolution 45: 499-509. doi:10.1023/A:1008603303314.

Lewontin, R.C. 1972. The apportionment of human diversity. Evol.Biol. 6: 381-398.

Malapa, R., G. Arnau, J.L. Noyer and V. Lebot. 2005. Genetic Diversity of the Greater Yam (*Dioscorea alata* L.) and Relatedness to *D. nummularia* Lam. and *D.*

transversa Br. as Revealed with AFLP Markers. Genetic Resources and Crop Evolution 52: 919-929. doi:10.1007/s10722-003-6122-5.

Matsumoto, R., H. Shiwachi, H. Kikuno, R. Irie, H. Toyohara, A. Komamine, et al. 2010. Characterization of sprouting and shoot formation processes of rooted cuttings of water yam (*Dioscorea alata* L.). Trop.Agr.Develop. 54: 107-112.

Mignouna, H.D., M.M. Abang and S.A. Fagbemi. 2003. A comparative assessment of molecular marker assays (AFLP, RAPD and SSR) for white yam (*Dioscorea rotundata*) germplasm characterization. Annals of Applied Biology 142: 269-276. doi:10.1111/j.1744-7348.2003.tb00250.x.

Obidiegwu, J., M. Kolesnikova-Allen, C. Muoneke and R. Asiedu. 2009. SSR markers reveal diversity in Guinea yam (*Dioscorea cayenensis/D.rotundata*) core set. African Journal of Biotechnology 8: 2730-2739.

Obidiegwu, J.E., R. Asiedu, E.E. Ene-Obong, C.O. Mouneke and M. Kolesnikova-Allen. 2009. Genetic characterization of some water yam (*Dioscorea alata* L.) in West Africa with simple sequence repeats. Journal of Food, Agriculture & Environment 7: 132-136.

Okagami, N. and N. Tanno. 1991. Dormancy in *Dioscorea*: Comparison of Dormant Characters in Bulbils of a Northern Species (*D. opposita*) and a Southern Species (*D. bulbifera* var. vera). Journal of plant physiology 138: 559-565. doi:http://dx.doi.org/10.1016/S0176-1617(11)80241-5.

Paál, D., J. Kopernický, J. Gasper, D. Vašíček, K. Vašíčková, M. Bauerová, et al. 2013. Microsatellite analysis of the slovak carniolan honey bee (*Apis mellifera* carnica). Journal of Microbiology, Biotechnology and Food Sciences 2: 1517-1525.

Perrier, J. and J. Jacquemoud-Collet. 2006. DARwin software http://darwin.cirad.fr/darwin.

Petro, D., T. Onyeka, S. Etienne and S. Rubens. 2011. An interspecific genetic map of water yam (*Dioscorea alata* L.) based on AFLP markers and QTL analysis for anthracnose resistance. Euphytica 179: 405-416.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria.URL http://www.R-project.org/.

Raz, L.F.o.N.A.E.C.e. 2002. Dioscoreaceae R. Brown: Yam Family. Flora of North America North of MexicoMagnoliophyta: Liliidae: Liliales and Orchidales. p. 479-485

Roldán-Ruiz, I., J. Dendauw, E. Van Bockstaele, A. Depicker and M. De Loose. 2000. AFLP markers reveal high polymorphic rates in ryegrasses (*Lolium* spp.). Mol Breeding 6: 125-134. doi:10.1023/A:1009680614564.

Siqueira, M.V.B.M., G. Dequigiovanni, M.A. Corazon-Guivin, J.C. Feltran and E.A. Veasey. 2012. DNA fingerprinting of water yam (*Dioscorea alata*) cultivars in Brazil based on microsatellite markers. Horticultura Brasileira 30: 653-659.

Statistica Stat Soft, I. 2001. STATISTICA (data analysis software system) Version 6.0. www.statsoft.com.

Walck, J.L., M.S. Cofer and S.N. Hidayati. 2010. Understanding the germination of bulbils from an ecological perspective: a case study on Chinese yam (*Dioscorea polystachya*). Ann Bot 106: 945-955. doi:10.1093/aob/mcq189.

Wright S. 1978. Evolution and the genetics of populations. V. 4. Variability within and among natural populations. Chicago:University of Chicago Press; Pp.580.

Yeh, F.C., R.-C. Yang and T. Boyle. 1997. Population genetic analysis of codominant markers and qualitative traits. . Belgian J. Bot. 129: 157.

Yildiz, M. 2013. Plant Responses at Different Ploidy Levels In:Current Progress in Biological Research.InTech.

Chapter 5

5 Phenotyping and SuperSAGE analysis for determination of flowering and sex-related genes in *D. rotundata* (Poiret)

5.1 Background and Justification

Dioecy is one of the major characteristics of the genus *Dioscorea* (Terauchi and Kahl, 1999), making the synchronization of flowering time very difficult. It is also common to find non-flowering types in *D. rotundata*. Many cultivars flower only rarely. In addition, when they do flower they seldom set fertile seeds (Lebot, 2009). *D. rotundata* is mostly dioecious, with separate male and female plants, although a few lines have been identified as monoecious.

Flowers of *D. rotundata are* numerous and are usually borne on spikes. They are actinomorphic, small and inconspicuous, and are pollinated by small insects, including thrips, that are attracted by floral scent (Terauchi and Kahl, 1999). In the context of yam pollinators, Lebot (2009) indicated that night flying insects are responsible for pollination which don't require visual attraction, as yam flowers are insignificant in color but sweetly scented. The staminate flowers are sessile (stalk less) and one to six inflorescences are formed per internode mostly pointing downwards. Male plants have six stamens. In pistillate flowers, usually one or two

inflorescences are formed per internode pointing downwards. The perianth consists of three green sepals and three yellowish green petals. The sepals and petals of pistillate flowers resemble those in staminate flowers but are lobed above the ovary. The ovary is inferior and trilocular, i.e the female flowers have three carpels with each containing two ovules. After fertilization, the inferior ovary develops into a capsule with three wings each containing two seeds.

The sex chromosomes of 40 angiosperm species have been reported so far (Aryal and Ming, 2014, Ming et al., 2011). Heteromorphic sex chromosomes were identified in 20 plant species, while in the remaining 20 plant species sex chromosomes cannot be distinguished at the cytological level and hence are homomorphic. Among the species with homomorphic sex chromosomes, the XY system is found in 15 species, and the ZW system is found in five species. Of the 20 species with heteromorphic sex chromosomes, the XY system is found in 19 species and the ZW system is found in one species. The sex chromosomes of almost all *Dioscorea* species are not yet identified, except for *D. tokoro* where heterogametic sex (XY) for male and homogametic sex (XX) for female is suggested based on AFLP markers that showed only heterozygotes in the male parent that had tight linkages with the sex of its progeny (Terauchi and Kahl, 1999). The small size and large numbers of *Dioscorea* chromosomes made identification of the sex determining chromosomes difficult at cytological level (Dansi, et al., 2000, Martin, 1966). In addition, the genes for flowering and sex differentiation in yams are largely not yet known.

Flowering development in angiosperms involves gene expression in meristems that leads to conversion from vegetative meristems to flowering meristems in response to environmental signals. Multiple pathways that respond to different environmental and developmental signals are also known to control the flowering process (Simpson and Dean, 2002). Molecular and genetic analyses in eudicot plants such as *Arabidopsis thaliana* and *Antirrhinum majus* have identified several genes that specify floral organ identity (Ma and dePamphilis, 2000). This led to grouping the genes into three classes; A, B, and C based on floral organ identity that the genes specify in the developing flower. Hence, the ABC model was proposed (Coen and Meyerowitz, 1991). According to this model the class A genes are required for sepals and petals development, class B genes are required in petals and stamens development and class C genes are required in stamens and carpels development. In addition, class A and C genes are mutually antagonistic to each other, and in the absence of one the other expands to occupy the entire flower. Recent studies indicates that there are additional classes of genes important for ovule and all the organ development, namely D and E gene classes respectively (Figure 5.1), which led to a modified ABCDE model of flowering (Su, et al., 2013). As master regulators of floral organ identity, plant MADS-box genes that are known to encode transcription factors are at the heart of the classic ABC model for floral development (Heijmans, et al., 2012).

Several factors underlie the regulation of sex differentiation in angiosperms including genetic regulation of sex determination, which refers to DNA polymorphisms in the loci that determine the development of male and female characteristics (Spigler, et al., 2008), epigenetic and environmental regulation (Jaligot, et al., 2011), and physiological regulation by phytohormones (Acosta, et al., 2009).

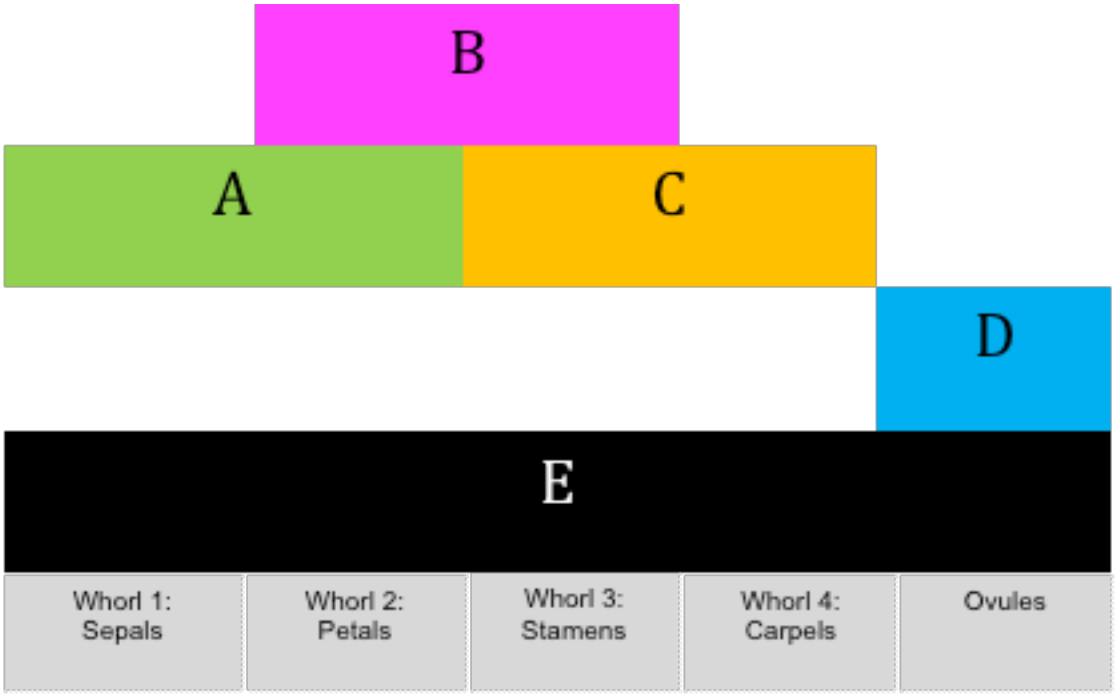


Figure 5.1. The ABCDE gene model of flower development according to Su, et al. (2013). The figure indicates different gene classes in flower development including class A genes (green box) required for sepals and petals development, class B genes (purple box) required in petals and stamens development, class C genes (yellow box) required in stamens and carpels development, class D genes (light blue box)

required in ovule development and class E genes (black box) required for all the organ development.

The needs for efficient improvement of yam require a better understanding of the molecular and genetic mechanisms of flowering. The ongoing whole genome sequencing project of *D. rotundata* (Tamiru, et al., 2013) is expected to provide opportunity for genetics and genomics studies aimed at dissecting the regulation of important processes including flowering and sex determination in yam. The identification of flowering genes has immense significance in scaling up utilization of the available yam germplasm in breeding programs. However, a toolbox for functional genomics (i.e. knockouts and overexpression lines) in yam has yet to be developed that can be used in a high-throughput manner to dissect functions of individual genes.

Next generation sequencing based high throughput SuperSAGE (Serial Analysis of Gene Expression) that involves sequencing of longer fragments and simultaneous analysis of multiple samples by using indexing (barcoding) was suggested as a reliable protocol for tag based gene expression profiling (Matsumura, et al., 2010). SuperSAGE is a variant of the Serial Analysis of Gene Expression (SAGE) expression profiling technology, in which 26-bp tags are extracted from cDNA using the type III restriction endonuclease EcoP15I, generates the longest distance so far reported between the recognition and cleavage sites. SAGE uses the restriction endonuclease,

BsMFI to generate 15bp tag, which is too short to identify the gene of origin; The use of a longer tag size in SuperSAGE allows a secure tag-to-gene annotation by homology searches against genome, transcript, or expressed sequence tag sequences. The technique also has an advantage over other techniques based on next generation sequencing (such as DGE-TAG that provides a relatively short tag reads (21-bp) which sometimes create tag-to-gene annotation more difficult) and RNA-Seq that requires a large amount of sequence reads to fully cover the dynamic range and to provide a truly quantitative gene expression profiling (Matsumura, et al. (2010).

This study represents the first attempt to identify candidate genes differentially regulated in relation to flowering and sex in *D. rotundata* using expressed tag sequences generated by the high throughput SuperSAGE technique. In addition, phenotypic variations related to flowering pattern were assessed across the IITA yam germplasm collection.

5.2 Materials and Methods

5.2.1 Morphological characterization

All accessions of *D. rotundata* (N=1938) planted for routine field maintenance under IITA genebank were characterized based on selected morphological traits and assessed for their flowering pattern and other morphological traits for two consecutive years (2010 and 2011). Twelve phenotypic yam descriptors jointly

developed by the International Plant Genetic Resources Institute and the International Institute of Tropical Agriculture (IPGRI/IITA, 1997) were used (Table 1).

5.2.2 RNA extraction

Tubers of seven accessions representing male (N=2), female (N=2) and monoecious (N=3) (Table 2) flowering types selected based on their consistency of flowering and sex type over the two years were planted in pots under screen house conditions for sampling at the appropriate growth stages (Figure 5.2). Total RNA was extracted using the Qiagen RNeasy plant mini kit according to the manufacturer's protocol (Qiagen, Venlo, the Netherlands). On-column DNAase treatment was performed to remove contaminating DNA.

5.2.3 Library preparation and sequencing

SuperScript II double-strand cDNA synthesis kit was employed for cDNA synthesis using the biotinylated adapter-oligo (dT) primer (5'-bio-CTGATCTAGAGGTACCGGATCC-CAGCAGTTTTTTTTTTTTTTTT-3'). Synthesized cDNA was purified using Qiagen PCR purification Kit. The library was prepared following the protocols of Matsumura, et al. (2010). Briefly, the NlaIII, anchoring enzyme was used to digest the total population of transcripts so that short fragments are isolated. The fragments (NlaIII-digested cDNA) were bound to streptavidin-coated magnetic beads (Dynabeads streptavidin M-270), and non-biotinylated cDNA fragments were removed by washing. Adapter 2 was ligated to digested cDNA fragments bound to the magnetic beads. The type III restriction enzyme EcoP151 was used for digestion of adapter 2-cDNA after washing. Adapter2-26bp fragments were further ligated to adapter 1 (that are specific for each of the samples). The adapters prepared following procedure described by (Matsumura, et al., 2010) were used. The adapter2-tag-adapter1 ligates was amplified using PhusionHigh polymerase and GEX primers (5'-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3' and 5'-CAAGCAGAAGACGGCATACGATCT-3'). The amplification program was 98^0C for 1min, 10 cycles at 98^0C for 35 sec, and 60^0C for 30 sec. The PCR product comprising 8 tubes per sample was pooled and concentrated using Qiagen MinElute reaction purification kit. The amplification product was run on an 8% non-denaturing polyacrylamide gel. After staining with SYBR green (Takkara Bio), the band around 123-125-bp size was cut out from the gel, and DNA purified after its elution from the

gel pieces. The purified PCR product from each sample was analyzed for its quantity and quality on an Agilent Bioanalyzer 2100.

As a next step the PCR product was cloned using Invitrogen: - zeroblunt Topo PCR cloning kit for sequencing and later transformation using one shot chemical transformation protocol. Colony PCR was done by using 2x colony PCR mixture and purified using QIAGEN PCR purification kit. Purified and mixed PCR products were applied to Illumina Genome Analyzer II for sequencing reactions. Samples were bulk-sequenced on a single lane of an Illumina Genome Analyzer II, which generated single-end 35-bp long single reads. Of these, the first 4-bp corresponds to an index sequence and the final 5-bp to an adapter sequence that were ligated to each fragment during preparation of supeSAGE libraries.

Table 5.1. Flower related and other phenotypic traits used for characterization of *D. rotundata* accessions.

Morphological descriptors	Parameters used
Sex	1=female, 2=male, 3=monoecious and 4= no flowering
Inflorescence position	1=pointing upward and 2= pointing downward
Average length of inflorescence	≤5cm=short, 6-15=intermediate and ≥ 16cm=long
Number of inflorescence per plant	<10=few, 11-29=medium and ≥ 30=many
Number of inflorescence per internode	Count
Flower color	0=not available, 1=white and 2=yellowish
Stem color	1=green, 2=brownish green, 3=purplish green and 4=purple
Vigour	3=low, 5= intermediate and 7=high
Leaf color	1=green, 2=yellowish green and 3=dark green
Leaf shape	1=hastate, 2=sagitate and 3=cordate
Presence of barky patches on stem	0=absent and 1=present
Presence of waxiness on stem	0=absent and 1=present

5.2.4 Data analysis

Phenotypic analysis: Multiple Correspondence Analyses (MCA) was performed for the phenotypic data using FactoMineR package (Lê, et al., 2008) in R software (R Core Team, 2013) to detect the underlying structure of morphological variables and its correlation with flowering patterns in the data set.

5.2.5 SuperSAGE data analysis

The reads in fastq format were sorted to their corresponding samples using the 4-bp index sequencing using a perl script (Appendix A), after which the reads were trimmed to 26-bp following removal the 4-bp index and 5-bp adapter sequences.

The R package edgeR (Robinson, et al., 2010) was used for examining differential expression of replicated count data of tags across different pairs of flowering patterns (male vs female, male vs monoecious and female vs monoecious). The edger software was initially developed for serial analysis of gene expression and it is so far the only software for differential expression of SAGE data, which can account for biological variability when there are few replicate samples. Tags with very low counts were filtered out and tags that were expressed in at least one sample from each flowering group were considered for the analysis. Additionally, since two replicate samples were sequenced from each flowering group, tags that expressed in both replicates were considered. The dispersion parameter for each tag, a measure of the degree of inter-library variation for that tag was estimated using the Negative Binomial (NB) model which gives an idea of overall variability across the genome

for the dataset. NB is used for modeling count variables, usually for over-dispersed count data, which is when the variance exceeds the mean. Trimmed Mean of M values (TMM) normalization (Robinson and Oshlack, 2010) was implemented to ensure that technical effects have less impact on the results. The TMM method estimates scale factors between samples with the aim to ensure the relative distributions of genes between samples are comparable. Defining Y_{gk} as the observed count for gene g in library k summarized from the raw reads, μ_{gk} as the true and unknown expression level (number of transcripts), L_g as the length of gene g and N_k as total number of reads for library k and S_k as the total RNA output of a sample was used a framework for more explanation for the requirement normalization (Robinson and Oshlack, 2010). The model for the expected value of Y_{gk} was indicated as:

$$E[Y_{gk}] = \frac{\mu_{gk} L_g}{S_k} N_k$$

$$\text{where } S_k = \sum_{g=1}^{G} \mu_{gk} L_g;$$

The relative RNA production of two samples, fk = Sk/Sk', essentially a global fold change was determined using a weighted trimmed mean of the log expression ratios. For sequencing data, the gene-wise log-fold- changes was defined as:

$$M_g = \log_2 \frac{Y_{gk}/N_k}{Y_{gk'}/N_{k'}}$$

and absolute expression levels:

$$A_g = \frac{1}{2} \log_2 \left(Y_{gk}/N_k \bullet Y_{gk'}/N_{k'} \right) \text{ for } Y_{g\bullet} \neq 0$$

After estimating the dispersion the function exacTest conducts tagwise tests using the exact negative binomial test proposed by Robinson and Smyth (2007). Benjamini and Hochberg's (1995) algorithm is used to control the False Discovery Rate (FDR). The p-value was calculated using the Fisher Exact test and the "BH" method by Benjamini Hochberg was used for correction of P-values for multiple comparisons. The test results for the most significant tags are displayed by the topTags function. The counts per million for the tags that edgeR has identified as the most differentially expressed were also listed.

For annotation of SuperSAGE tags, the selected tags were first aligned to the draft *Dioscorea rotundata* scaffold sequence, followed by extraction of the upstream 2000-bp sequences. The 2000-bp sequences were finally used as queries for blasting against the non-redundant (nr) database of the National Center for Biotechnology Information (NCBI) and the Universal Protein Resource (UniProt).

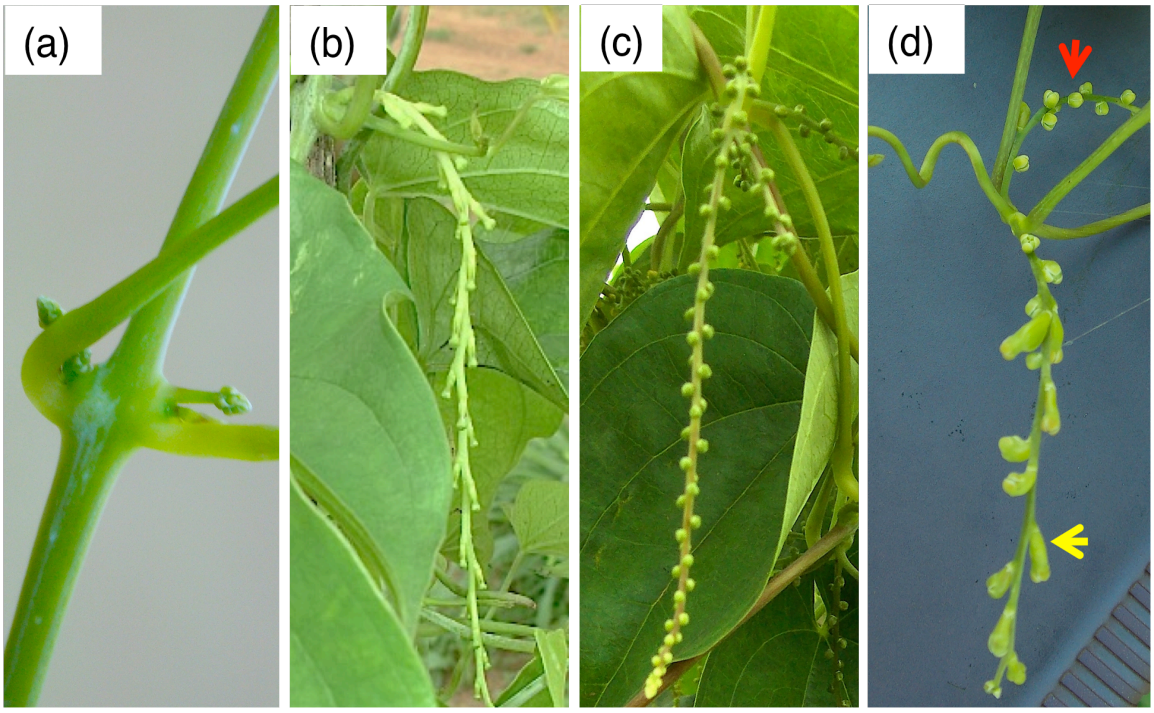


Figure 5.2. Flowering variation in *D. rotundata* a) female flower at early stage, also indicating a stage we have collected samples for total RNA extraction, b) female c) male and d) monoecious inflorescence.

5.3 Results

5.3.1 Morphological variation of D. rotundata genebank collection

The flowering patterns of 1938 *D. rotundata* accessions collected primarily from the main yam growing regions of West and Central Africa were assessed over two consecutive growing seasons in 2010 and 2011. *D. rotundata* accessions are easily distinguished based on their flower type as female, male, or monoecious (the production of unisexual female and male flowers on the same plant) (Figure 1). In 2010, 996 of the accessions (51.4%) failed to flower, while 745 (38.4%), 170 (8.8%), and 27 (1.4%) were male, female, and monoecious, respectively (Figure

5.3a). In the 2011 growing season, 939 (48.5%) were male, followed by 630 (32.5%) non-flowering, 287 (14.8%) female, and 82 (4.2%) monoecious type accessions (Figure 5.3b). Most accessions were consistent over the two seasons with respect to flowering, while some were not. About 326 (16.8%) male, 169 (8.7%) female and 43 (2.2%) monoecious accessions failed to flower in one of the two seasons, hence the discrepancy observed in the proportion of accessions with different sexes between the seasons (Figure 5.3). Overall, the majority of *D. rotundata* accessions maintained at IITA were either male or non-flowering, while the female accessions represented less than 15% of the collection. The result further revealed that the monoecious types are very rare in *D. rotundata*.

In addition to sex, we collected morphological data over the same period using 14 selected traits (Table 5.1). The categorical data was then subjected to MCA, which grouped the accessions into three major clusters that mainly reflected their sex, suggesting an overall correlation between sex and the selected morphological traits (Figure 5.4). The clustering included a distinct group consisting of non-flowering accessions that were distinguished by traits such as purplish green stem, non-waxy stem, stem with non-barky patches, dark green leaf color, and hastate leaf shape. A second cluster made entirely of male accessions was correlated with purplish green stem, presence of barky patches, non-waxy stem, dark green leaf, and sagittate leaf shape. The third group was composed of a mixture of male, female and monoecious flowering accessions that were identified by waxiness, absence of barky patches, either green, brownish green or purple stem color, and pale green or green leaf color as distinct traits.

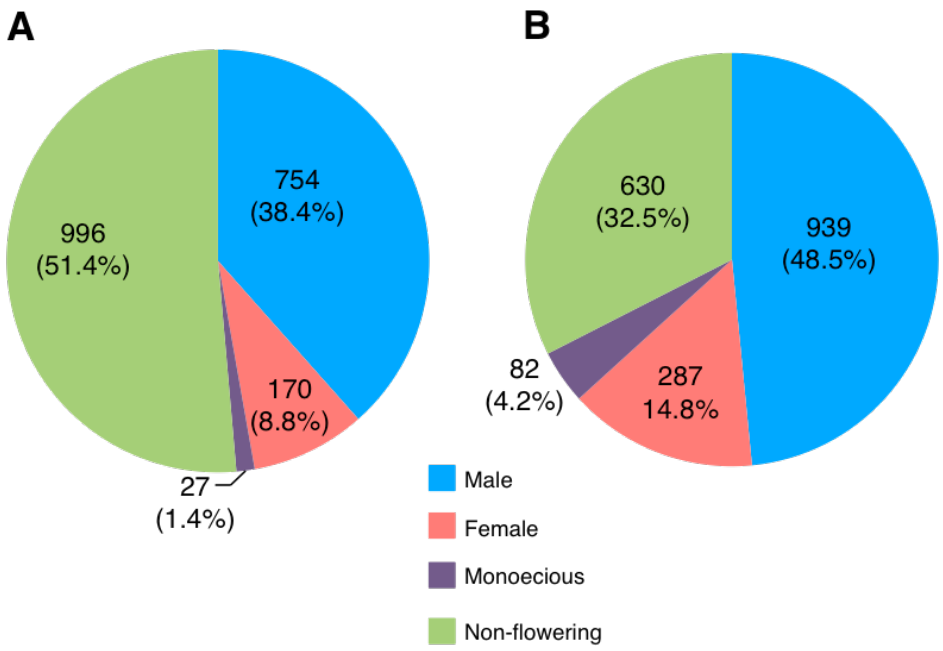


Figure 5.3. Sex distribution in yam (*D. rotundata*) accessions. The proportion of male, female, monoecious, and non-flowering accessions among 1938 genbank accessions in 2010 (A) and 2011 (B) growing seasons.

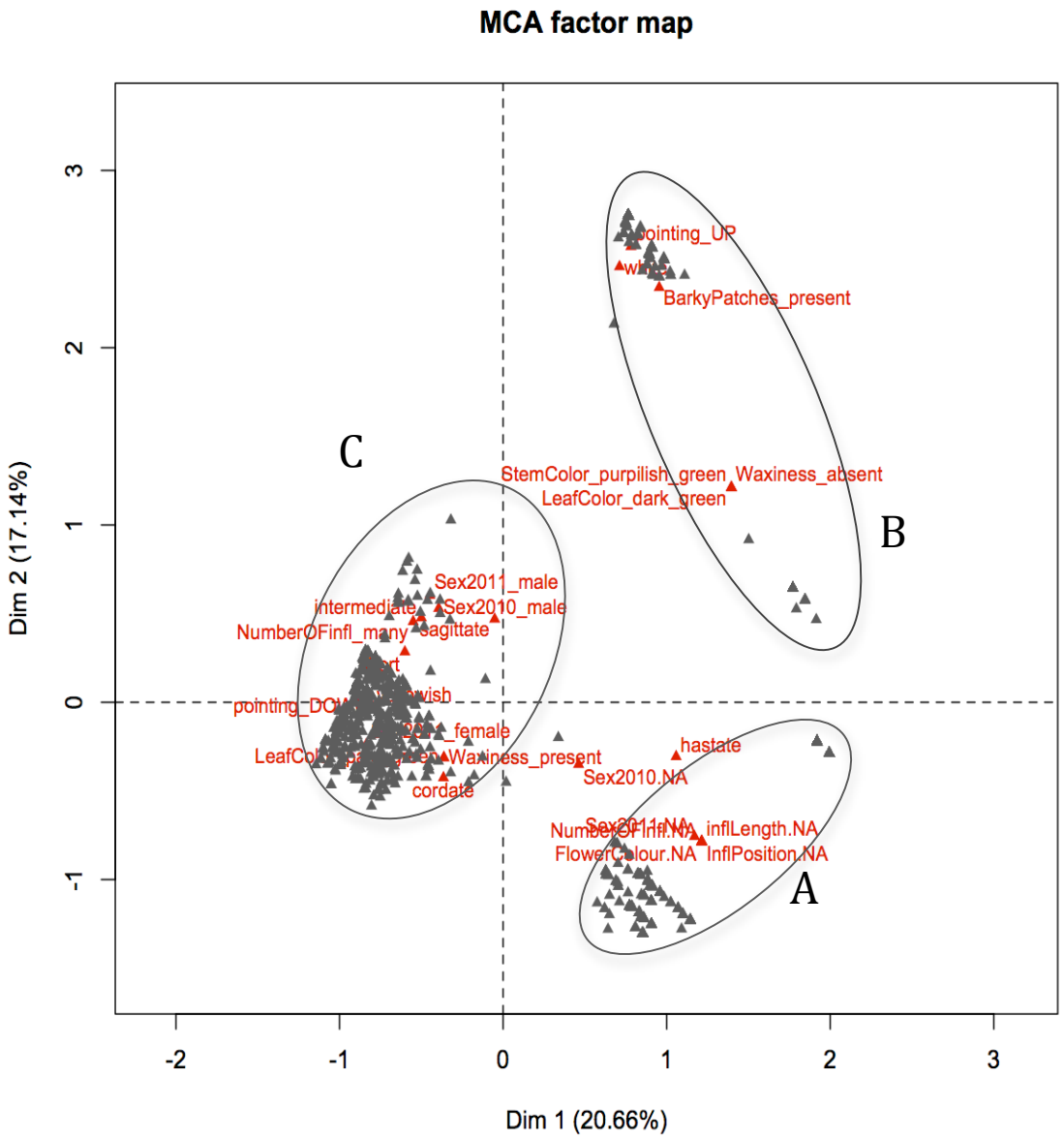


Figure 5.4 Multiple Correspondence Analysis (MCA) of sex type and phenotypic traits in yam (*D. rotundata*). The pattern of relationship between individual plants (grey triangles) and the 20 most discriminant morphological traits (red triangles) are provided. The circles with grey lines represent the three main cluster: Cluster A = non-flowering accessions; Cluster B = male accessions; Cluster C = male, female, and monoecious accessions.

5.3.2 Tags generated from the SuperSAGE library

The SuperSAGE libraries were multiplexed and sequenced on a single lane of an Illumina genome analyzer II, generating a total of 8,332,523 quality tags extracted from the initial sequence reads (Table 5.2) through Quality Control (QC). Sequence reads were selected according to their quality in FASTAQ format. Specifically, short reads in which more than 10% of nucleotides had PHRED quality score of less than 30 were excluded from analysis. The tags were sorted based on Adapter1 sequence, which is specific to each sample. All the singleton tags were excluded, whereas non-singleton tags were used for further analysis. A total of 20,236 unique tags were identified and used for differential expression analysis. Of these, 6,335 tags that were ten or less in number in each sample were excluded, while the remaining 13,901 that were more abundant were further considered for constructing a venn diagram (Figure 5.5). Accordingly, 43% of tags were shared among the three flowering groups, whereas others were specific to male (1855), female (1648) and monoecious (765). The remaining 19.0% were shared between male vs female, 2.7% between male vs monoecious, and 4.5% between female vs monoecious groups.

Table 5.2. Summary of tags generated for the different flowering groups by SuperSAGE analysis.

Sample	Sex	Total tags	Unique tags	Non-singleton
TDr3631	Male	1,251,361	17,773	16,602
TDr2965	Male	1,460,689	18,234	17,408
TDr4087	Female	1,049,552	18,620	17,853
TDr1679	Female	560,257	17,534	15,887
TDr4162	Monoecious	1,209,229	18,032	17,146
TDr1506	Monoecious	1,309,189	18,746	18,177
TDr1819	Monoecious	1,492,246	18,918	18,403

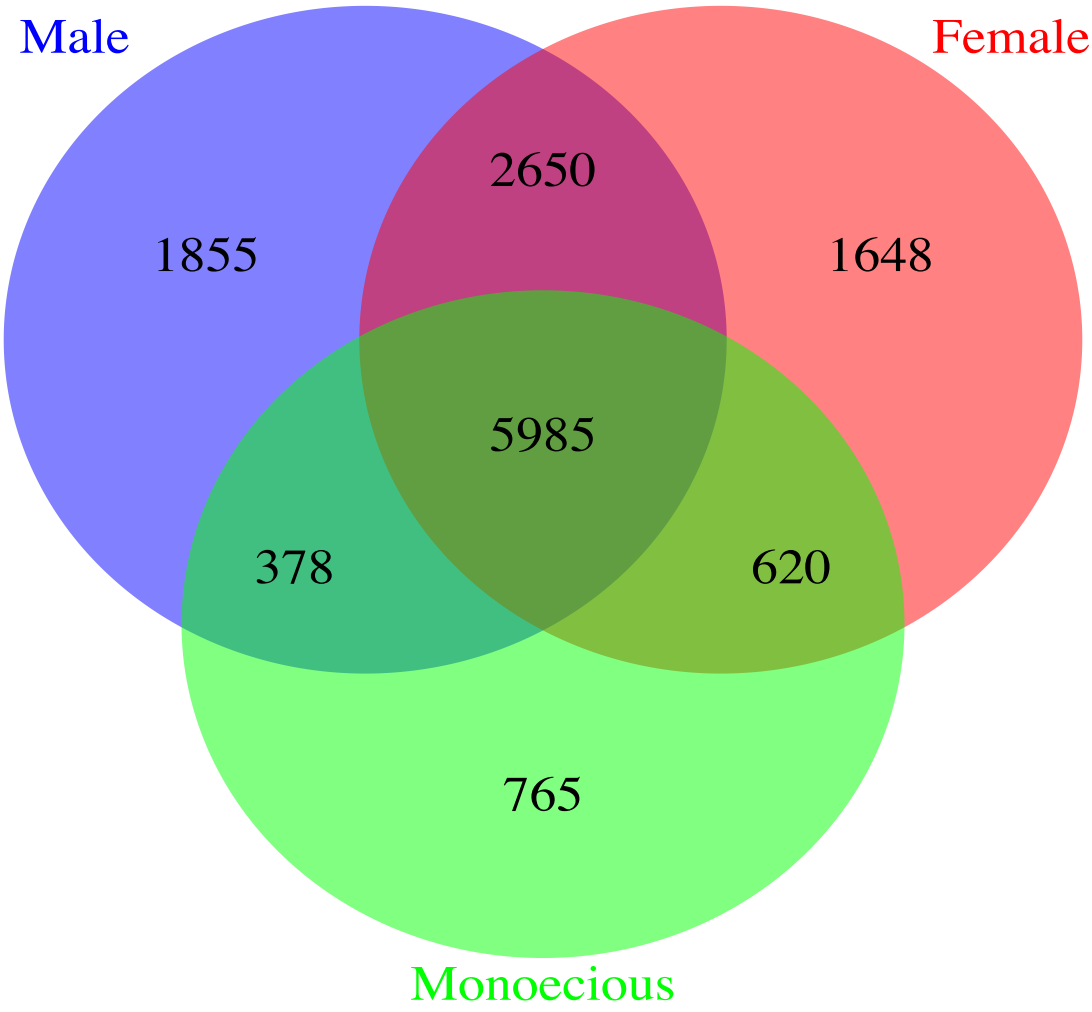


Figure 5.5. Venn diagram showing unique tags, as well as tags shared among male, female and monoecious flower groups.

5.3.3. Differential gene expression among flowering groups

The fold change of differential expression, and gene abundance (count per million) of all the singleton tags (N= 20,236) was compared (Figure 5.6) across different flower groups. In the current study a total of 100 tags/genes were differentially expressed with p and FDR values less than 0.01. The differential gene expression estimated based on the number of SuperSAGE tags varied across male vs. female, male vs. monoecious and female vs. monoecious groups with 13, 67 and 20 tags resectively (Appendix B). Of the 13 genes highly expressed in male and female groups, five were highly expressed in male while the remaining eight were expressed in female sex type. Likewise, the male vs monoecious group comparison revealed that 25 tags were highly represented in male, and 42 were abundant in monoecious sex type. Between the female vs monoecious flower group 11 and nine tags were expressed highly in female and monoecious sex type, respectively. The tag abundance indicated as average logCPM (counts per million) ranges from a minimum of 6.36 to 11.88 among all differentially expressed tags.

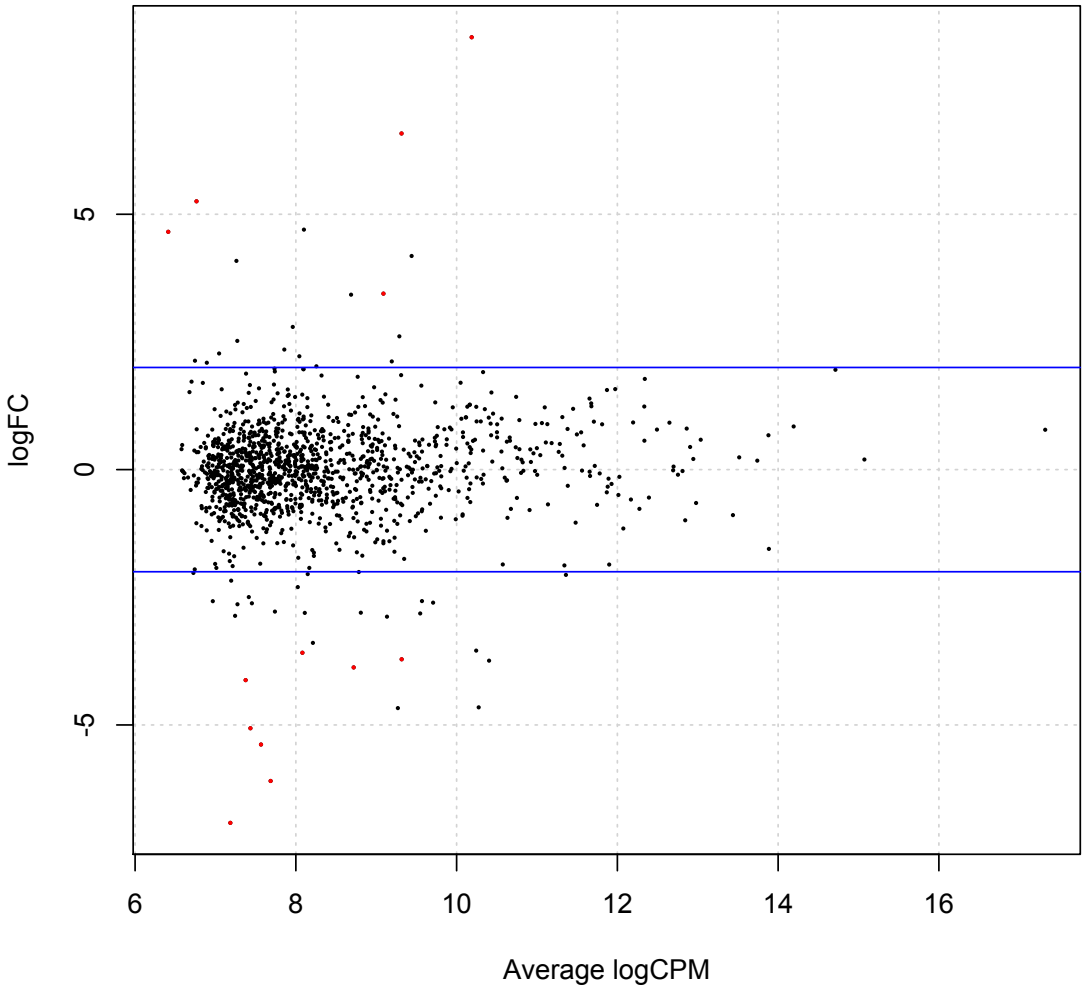


Figure 5.6a. Differentially expressed tags and abundance across male vs. female flower group. Differentially expressed tags are represented by red dots. Fold change values for male vs. female group are plotted against average log expression values (standardized read counts). The logFC indicates the fold changes of differential expression whereas logCPM indicate count per million or tag/gene abundance. The horizontal blue lines represent 4-fold changes.

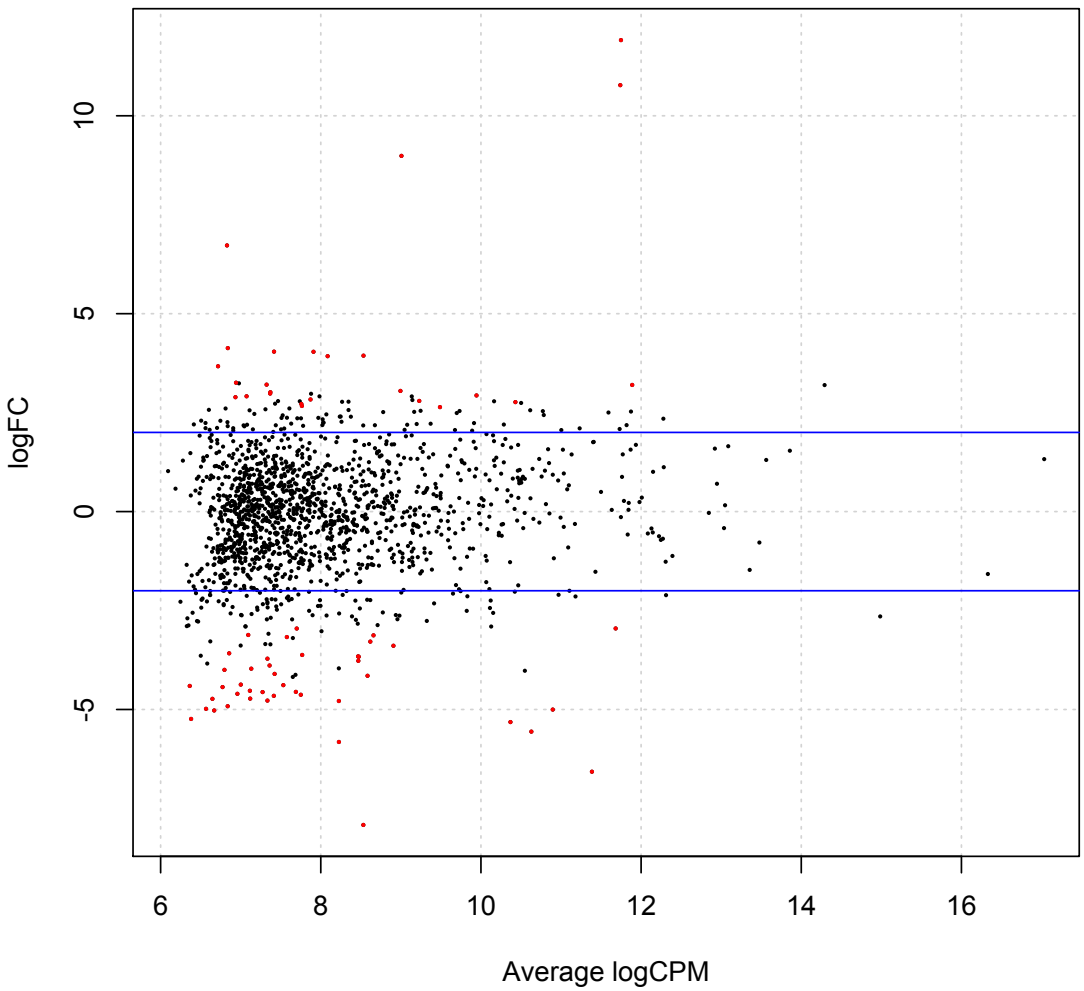


Figure 5.6b. Differentially expressed tags and abundance among male vs. monoecious group. Differentially expressed tags are represented by red dots. Fold change values for male vs. monoecious group are plotted against average log expression values (standardized read counts). The logFC indicates the fold changes of differential expression whereas logCPM indicate count per million or tag/gene abundance. The horizontal blue lines represent 4-fold changes.

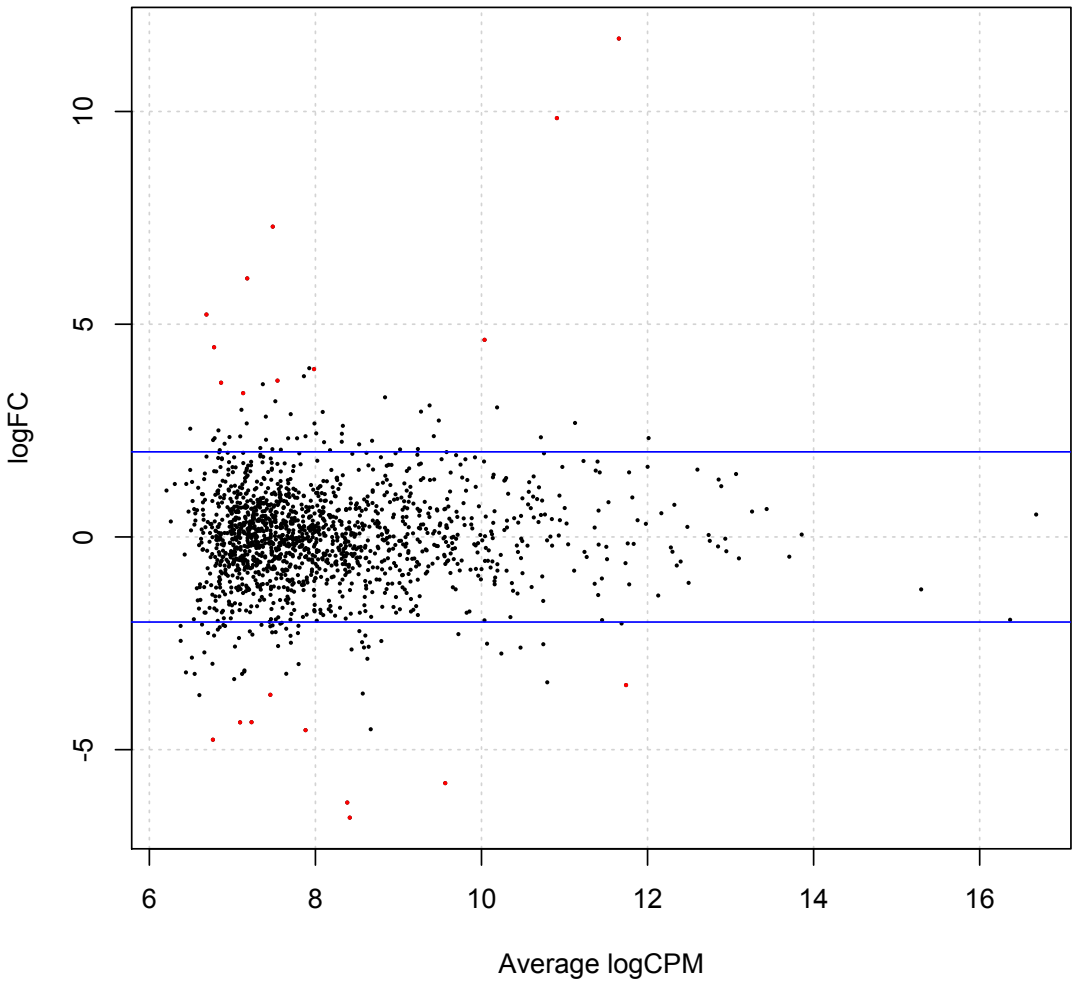


Figure 5.6c. Differentially expressed tags and abundance among female vs. monoecious group. Differentially expressed tags are represented by red dots. Fold change values for female vs. monoecious group are plotted against average log expression values (standardized read counts). The logFC indicates the fold changes of differential expression whereas logCPM indicate count per million or tag/gene abundance. The horizontal blue lines represent 4-fold changes.

5.3.4 Gene annotation, tag-to-gene

Among all (100) differentially expressed tags used for further analysis 88 were unique. Tag annotation was therefore carried out for a total of 88 tags against the NCBI and UniProt databases after aligning the tags to the draft *D. rotundata* scaffold sequence and collecting 2000-bp upstream regions. Each tag was aligned to *D. rotundata* assembled scaffolds using BLASTN. From BLASTN result, about 2-kb sequence upstream of the each tag was extracted, which were then used for tag annotation. About 98.86% of the tags were successfully annotated, while 15 (17.04 %) did not match sequences available in databases, and the remaining one tag (1.13%) could not be aligned to the draft *D. rotundata* sequence. Sequence alignment from some of the eight (9.09%) tags showed high e-values, suggesting possible sequence hits by chance or random alignment. About 14(15.90%) of the tags corresponded to either proteins of unknown function, unnamed, uncharacterized or hypothetical proteins (Table 5.3). The genes identified using 19 (21.59%) of the tags consisiting of 17 unique genes corresponded to those that have been reported for a role in flower development and/or to be expressed in flowers in multiple species (Table 5.4). The remaining genes were not reported for involvement or expressed in flower and flower development.

Among the 17 genes reported for expression or involvement in flower and flower development four were highly expressed in female, eight in male and six in monoecius sex types (Table 5.4). The numbers of genes differentially expressed were 16 in male, 15 in monoecious and four in female sex types. Of the total genes only one was differentially expressed across all the flowering groups, whereas two

in female vs monoecious; three in male vs female and 14 in male vs monoecious groups alone.

Table 5.3. List of differentially expressed tags, result of tag annotation, and candidate genes reported for expression reported for involvement in flower and flower development.

Tag sequence	E-value	Putative protein	Expression and involvement in flower and flower development
CATGAAAATTACCATCACCAAAAAAA	NoHit	NoHit	-
CATGAAACCCCCTCGGGCGAAGTTTC	N/A	N/A	-
CATGAAAGTGTTGAAAGTTAAAAAAA	0.0000001	No conserved domain	-
CATGAACCTTGTGTTTGTATTTAAAA	1E-161	Serine--glyoxylate aminotransaminase	-
CATGAACTACGGCCCTGGTGCCGCCG	0	Pectinesterase inhibitor	Yes
CATGAAGATTGTCATTCCCTGAATTG	6E-88	Trichome birefringence-like 23	Yes
CATGAAGGGAACAAAAGAAATAAAAA	6.3	Uncharacterized protein	-
CATGAAGGTAGGGATGATTTTTTAAA	N/A	N/A	-
CATGAATCACTGTGTAACTGATGCAT	5E-13	Glutamine amidotransferases class-II (GATase).	-
CATGACACACCGGACATTACTGGACT	0.0001	No conserved domain	-
CATGACCACCGTCCTGCTGTCTTAAT	2E-47	Predicted protein/No conserved domains	-

CATGACTACATCTGGTCCTATGAATA	2E-75	NAC domain protein	Yes
CATGACTGTTATGATAAAAAAAAAAA	3E-32	Transposase-associated domain	-
CATGAGAAGCTGCTTCTTGGGTGGGA	6E-74	Serine-glycine hydroxymethyltransferase (SHMT)	-
CATGAGGTTCTAGGGTTTGGTTATTT	N/A	N/A	-
CATGAGTAATAAAGTTAACTTCTCGT	1E-15	Mannose specific lectin	-
CATGATGATCAGGGTAGCATATGAGC	0	ABC transporter F family member 1	-
CATGATTAATTTGAAAAAAAAAAAAA	2E-20	Exostosin family protein	-
CATGATTAATTTGAAGACTGCTCAGT	5E-103	Transferase family protein	Yes
CATGATTGGCTTTGCTGCGTCTCTGC	0.001	Photosystem II, 22 kDa Protein	-
CATGCAACAACAAGCTCGCAAGGCTG	0.045	WRKY transcription factor 47-1	-
CATGCAACGCCAAGGAGATTTTCGTC	1E-94	Pathogenesis-related thaumatin family protein	-
CATGCAAGTTCTACAGAGAATAAAAA	2E-75	NAC domain protein	Yes
CATGCACACAATCATCATCATCATCA	N/A	N/A	-
CATGCAGATCTTCGTGAAGACCCTGA	0	TolB protein-related isoform 1	-
CATGCAGCCACTTGCCCTGTTTCCTT	0.000008	Vacuolar-processing enzyme (VPE)	Yes
CATGCATCCATCGCTGGCCTTGTTTT	2E-36	Major intrinsic protein (MIP) superfamily.	-
CATGCATGCGTGGATGGGTGGACGTA	6E-95	Probable aquaporin TIP1-1 (AQPs)/Major intrinsic protein (MIP) superfamily.	Yes

CATGCATGCTGTCAGTTTTGGGAGTC	6E-41	T-complex 1 subunit eta	-
CATGCATGGATTTATTTGTTGGAAAG	5E-30	Mannan endo-1,4-beta-mannosidase 5-like	-
CATGCATTTGTGTGTTATGTAATAAG	3E-18	PR10 protein	-
CATGCCAAGAAGTTTAGTGCTTGGAT	2E-41	Glutathione-S transferase	Yes
CATGCCACGTAATTACGTATTATTAT	7.2	Unnamed protein product	-
CATGCCACTGAAGCTGCAAACAAAAT	7E-118	Lanatoside 15-O-acetylesterase	-
CATGCCGACGCCTTTGTCCACGCCAC	4E-37	Zinc finger family protein	Yes
CATGCCGGAGGCGGTGATGTGGCTCA	2E-28	Unnamed protein product	-
CATGCCTAGAGCTTCTTCTTGAAGTA	5E-89	LRR receptor-like kinase family protein	-
CATGCCTTCGCTTGGTTGTGAAAAAA	1E-64	NAC-like protein 13	-
CATGCGCCCATCGGCTCGCTATGATG	N/A	N/A	-
CATGCGCCGGCGCGCACATTGGCCTA	N/A	N/A	-
CATGCGCGCGCGCGTCACGCCGCGCC	N/A	N/A	-
CATGCGCGCGCGTCACGCCGCGCCGT	N/A	N/A	-
CATGCGCTTGTACTGCAACTTATAAA	1E-110	No lysine kinase 4 isoform 1	-
CATGCGGCCTTTGCTGCGCTTGAGCT	9E-43	Alternative transcript type 3	-
CATGCGTGCGCGGAGTGTCGGATCGG	2E-61	Myb/SANT-like DNA-binding domain	-
CATGCGTGGATGGGTGGACGTAGTTT	6E-95	Probable aquaporin TIP1-1 (AQPs)/Major intrinsic	Yes

		protein (MIP) superfamily.	
CATGCTCCGGCCCGCTGATGGGAATG	0	Transketolase (TK)	Yes
CATGCTGAGCGGTTTCCGATCCCGAA	0	Catalase 2	-
CATGCTTCCAGAACTCCGTTTTCGGG	0.9	Hypothetical protein	-
CATGCTTGCTTGTTTGATCCTTATAT	8E-22	Probable aldehyde oxidase 2-like	-
CATGCTTGTATCAGTTAGCTACTGCG	N/A	N/A	-
CATGGAAGATACAACCCTCAGCTTTC	0.000002	LINE-1 reverse transcriptase like	-
CATGGAATTAAGCACCTAAGTTTGCT	5E-11	Tuber lectin 1	-
CATGGACGGATGATCCGTTTAGGAGA	0.13	No conserved domain	-
CATGGAGAATGGAGCGGTTGCGGGGA	1E-41	No conserved domains	-
CATGGAGTGGTCTTTGCATTTAGTAT	N/A	N/A	-
CATGGATCAAGTGAGCAGCAACTTCA	2E-79	Stay green protein	-
CATGGATTGTTTGTTTGGGGATGAAT	N/A	N/A	-
CATGGCCGACCTTCCACGATCGGTCA	1E-85	Acyl: COA ligase	-
CATGGCCGCGCCAATGCTCGGAGGCA	N/A	N/A	-
CATGGCGACTCGTAAGGTGCAGGGGC	6E-80	Peptidase S8 family domain	-
CATGGCGCTACTGGTGCTGGGGGCTG	8E-46	Anthranilate synthase 01	-
CATGGCGGTGGCGGTGGCCGTCGAGG	1E-39	Hypothetical protein	-

CATGGCTCGTGTTAAAGTAACTAATA	1E-10	AAA family ATPase	-
CATGGCTGAGGACGGCGAAGGTGGTG	8E-83	Malonyltransferase	Yes
CATGGCTTGCAGCAGCTGCGGCGGAT	1E-30	Chaperone protein dnaJ-related	-
CATGGGAGAGGGTGACCGATCCTCGT	8E-171	No conserved domain	-
CATGGGCTTGCTTGCCTTCGCTGCTG	0.004	Glucan 1,3-beta-glucosidase	-
CATGGGGATCCCCAATAGCATCTCCA	0	Lipoxygenase (LOX)	Yes
CATGGGGGGAAGTGGGTTTGTGGGCT	0.0007	Protein of unknown function (DUF1677)	-
CATGGGTGGCGCGACTTCTCGCTGTG	1E-11	D6-type cyclin isoform 3	-
CATGGGTGTCCCTTCCCAAAGGTAAG	2E-57	40S ribosomal S4 (RPS4)	Yes
CATGGTTTATGTAGGGACAAGGATCA	3E-155	Aspartic proteinase nepenthesin-2 precursor	Yes
CATGTACATTGTGCGCTAGTTGCTCA	0.31	CTP synthase 1 isoform X2	-
CATGTAGAGTAGTTGTATTCAGTGGT	0.22	Uncharacterized protein	-
CATGTATCTCGTCCCGCCCAGGCCAA	2E-53	40S ribosomal protein S8 isoform 3	-
CATGTCAGATCGAGATTGGCTTGCAT	N/A	N/A	-
CATGTCCAAATGCCACTGGATATGTA	2E-29	GDSL esterase/lipase APG	Yes
CATGTGAAGAAGTGATGAGTTTGTTT	N/A	N/A	-
CATGTGACTTAACCGCAACCAGGGAA	1E-32	DnaJ-like protein	Yes
CATGTGAGTTGCTCCCGTTGACCTGC	N/A	N/A	-

CATGTGCGCTGCCTCAAATTTGCAAG	7E-16	Transcription factor PIF3	Yes
CATGTGCTCGCAGGCCGTCGAAAAAA	0.000000009	Histidine kinase 3	-
CATGTGGATGATGATAATGAATGCAT	2E-69	Primary amine oxidase	-
CATGTGGCCTAGCGGTACCCCGAGTT	4	Protein of unknown function	-
CATGTGTGTTGTGTATTCTGTTTTTC	5E-24	Stem-specific TSJT1	-
CATGTTCTCAATGTTTGGATTCTTTG	6E-144	Chloroplast chlorophyll a/b binding protein	-
CATGTTGCTAGCTCAGCGGTTGGGTT	8E-53	Long-chain fatty acid CoA synthetase	Yes

Yes = gene involvement/ expression in flower; - = no report on the expression or involvement of the gene in flower or flower development; NoHit= Tag that could aligned to the draft *D. rotundata* sequence, and N/A= did not match any sequences in the dat base used.

Table 5.4. Summary on differentially expressed genes reported to express in flower across different flower sex types.

Tag sequence	Male	Female	Monoecious	Putative protein	Source
CATGAACTACGGCCCTGGTGCCGCCG	+	0	-	Pectinesterase inhibitor	*Lycoris aurea*
CATGAAGATTGTCATTCCCTGAATTG	+	0	-	Trichome birefringence-like 23	*Theobroma cacao*
CATGACTACATCTGGTCCTATGAATA	-	0	+	NAC domain protein	*Theobroma cacao*
CATGATTAATTTGAAGACTGCTCAGT	-	0	+	Transferase family protein	*Populus trichocarpa*
CATGCAAGTTCTACAGAGAATAAAAA	-	0	+	NAC domain protein	*Theobroma cacao*
CATGCAGCCACTTGCCCTGTTTCCTT	-	0	+	VPE	*Gossypium arboreum*
CATGCATGCGTGGATGGGTGGACGTA	+	+	-	AQPs	*Oryza sativa Japonica group*
CATGCCAAGAAGTTTAGTGCTTGGAT	+	0	-	Glutathione-S transferase	*Hyacinthus orientalis*
CATGCCGACGCCTTTGTCCACGCCAC	+	0	-	Zinc finger family protein	*Populus trichocarpa*
CATGCGTGGATGGGTGGACGTAGTTT	+	+	-	AQPs	*Oryza sativa Japonica group*
CATGCTCCGGCCCGCTGATGGGAATG	-	+	0	TK	*Camellia sinensis*
CATGGCTGAGGACGGCGAAGGTGGTG	+	0	-	Malonyltransferase	*Iris x hollandica*
CATGGGGATCCCCAATAGCATCTCCA	+	0	-	LOX	*Malus domestica*
CATGGGTGTCCCTTCCCAAAGGTAAG	0	+	-	RPS4	*Gossypium arboreum*

CATGGTTTATGTAGGGACAAGGATCA	-	0	+	Aspartic proteinase nepenthesin-2 precursor	*Zea mays*
CATGTCCAAATGCCACTGGATATGTA	+	0	-	GDSL esterase/lipase APG	*Glycine soja*
CATGTGACTTAACCGCAACCAGGGAA	-	0	+	DnaJ-like protein	*Glycine max*
CATGTGCGCTGCCTCAAATTTGCAAG	-	0	+	PIF3	*Medicago truncatula*
CATGTTGCTAGCTCAGCGGTTGGGTT	-	+	0	Long-chain fatty acid CoA synthetase	*Cucumis sativus*

+=highly expressed tags; - = poorly expressed tags and 0= no expression (zero expression)

5.4 Discussion

5.4.1 Flowering vs morphological variation

The dioecious and monoecious flowering patterns of *D. rotundata* were previously reported (Dansi, et al., 1999, Hamadina, et al., 2009, Hamon and Toure, 1990). Consistent with these observations, our finding similarly indicates the presence of dioecious (separate male and female flower) and monoecious (consisting of female and male flower on the same plant) *D. rotundata* accessions. The monoecious plants were predominantly male, as most of the inflorescences consisted of male flowers only. Female flowers rarely formed on the same inflorescences with male flowers. This phenomenon in a *D. rotundata* cultivar was described as trimonoecious (Hamadina, et al. 2009).

Overall, our observation across all accessions in the IITA genebank indicated more male than female accessions, and that monoecious plants and male flowers are more numerous compared to female ones. Non-flowering accessions were the most frequent followed by male flowering accessions (Figure 5.3). The intensity of flowering appeared to vary according to sex. An observation on materials collected from Benin similarly showed rare flowering in female plants that produce a limited number of flowers, whereas the male flowering cultivars, whenever they flower, produce flowers in abundance (Dansi, et al., 1999). Female plants have longer inflorescences and generally appear vigorous in most cases (data not shown).

Morphological characters that could be used distinguish the different flowering groups are important particularly for breeders to easily select germplasm and cultivars prior to flowering for breeding experiments. The current study revealed three distinct groups including a) non-flowering accessions associated with hastate leaf shape, deep green leaf colour, purplish green stem colour with reduced barky patches and no waxiness, b) Male flowering accessions with upward pointing inflorescence, white flower, with sagittate shape and dark green colour, purplish stem colour with barky patches and no waxiness, and c) a group consisting male, female, monoecious and rarely flowering accessions which has general distinct features such as pale green leaf colour, downward pointing inflorescence, yellowish flower, stem with no barky patches and waxiness absent.

5.4.2 The expression and putative role of differentially expressed hypothetical genes in flowering in yams

In this study, several genes differentially expressed in relation to flowering were identified in *D. rotundata* (Table 5.3 and Figure 5.6). But only a few of these genes have been reported for their expression across flower organs or involvement in flower development (Table 5.4). The majority of these genes were reported only for their expression and only a few of them were characterized for their involvement in flower development. We have categorized the genes into six major groups.

The first group includes those reported only for their expression in flowers, which consisted of;

1. *Pectinesterase inhibitor* expressed in flowers of *Arabidopsis thaliana* (Micheli, et al., 1998),
2. *Malonyltransferase*, expressed in flowers of *Salvia splendens* (Suzuki, et al., 2004),
3. *A glutathione S-transferase-like* gene for its high expression in flowers of *Cucurbita maxima* Duch. cv. Ebisu.(Momose, et al., 2013) and
4. *AQPs/MIP* (Alexandersson, et al., 2005).

The second group was identified for their organ-specific expression and includes;

1. *VPE*, its increase in *Citrus sinensis* L. during flower development and highest levels in flowers at anthesis and petals from flowers at this stage were observed (Alonso and Granell, 1995) and
2. *Trichome birefringence-like 23*, expressed at mature pollen stage (Wang, et al., 2008) and in petal, sepal, pedicel, stamen, pollen, and petal differentiation and expansion stages of *Arabidopsis thaliana* (Schmid, et al., 2005, Wang, et al., 2008).

The third group was reported for their involvement in conversion of vegetative to reproductive phase and includes;

1. The *long chain fatty acid CoA sythetase* that is confirmed in *Solanum lycopersicum* for high transcription levels of the gene in the anther and

petal, preferentially in the sites subject to epidermal fusion (Smirnova, et al., 2013). The gene was reported to be involved in flower development and its deficiency impairs fertility and floral morphology and

2. *GDSL esterase/lipase APG*, suggested for its potential involvement in flowering (Ling, 2008).

The fourth group was indicated for their role in flower color development and includes;

1. A *glutathione S-transferase-like* gene, reported for flower color intensity in *Dianthus caryophyllus* L.(Momose, et al., 2013).

The fifth group encompasses those involved in flowering time, photoperiod and scenescence regulation and includes;

1. *NAC* domain containing protein, which was proposed for playing an important role in the coordination of cold response and flowering time (Yoo, et al., 2007),
2. *Transferase family protein*, that regulates flowering time via the flowering repressor FLOWERING LOCUS C in *Arabidopsis thaliana* (Wang, et al., 2012),
3. *Zinc finger family protein*, regulating flowering time and abiotic stress tolerance in *Chrysanthemum morifolium* (Yang, et al., 2014) where transgenic lines with suppressed expression of the gene flowered

earlier than wild-type plants and showed decreased tolerance to freezing and drought stresses,

4. An *LOX* gene known to regulate cell death related to flower senescence and flower opening (Liu and Han, 2010). A dramatic increase in *LOX* gene in response to senescence was observed in *Rosa hybrida* cv. Kardinal (Fukuchi-Mizutani, et al., 2000),
5. The *RPS4*, suggested for its important role in regulating flowering time. A delay in flowering time was showed by silencing of the genes encoding *RPS4* and rhodanese in *Glycine max* (Ai-Hua, et al., 2014) and
6. Transcription factor *PIF3*, suggested to play an important role in the control of flowering through clock-independent regulation of CO and FT gene expression and causes early flowering in *Arabidopsis thaliana* (Oda, et al., 2004).

The sixth group are those with sex-specific expression and includes;

1. The *DnaJ-like protein*, detected predominantly in male flower of *Salix bakko* (Futamura, et al., 1999) and
2. a *TK* gene in which transgenic plants in cucumber showed a higher ratio of female flower and yield relative to the wild type plants. Moreover, the decrease in net photosynthetic rates and carboxylation efficiency were less in transgenic plants than that in wild type during low temperature and low light intensity (Bi, et al., 2013). The up regulation of *TK* gene in the current study and the vigourosity of

female plant in general (personal observation) could indicate its involvement in female flower development in *D. rotundata*.

In the current study we have successfully identified genes that are differentially expressed in different *D. rotundata* flowers, including those previously reported for a role in flowering and flower development. However, the exact role of these genes, if any, in sex determination is not known. It is also possible that these genes are downstream of the candidate gene/genes that control sex in yams, the identification of which requires techniques such as map-base cloning or association studies. Verification or detailed analysis of the role of the candidate genes identified here also requires additional studies that involve the use of allelic variants, gene-knockout/knockdown, and/or over-expression lines.

5.4.3 Significance of the study for yam Improvement

The results from the assessment of the morphological characteristics will have an important role for the efficient utilization of yam germplasm by IITA and other yam breeding programs. The current study is the first of its kind in attempting to identify candidate flowering related genes in yams. The results reported will contribute to understanding the flowering biology of yam. The identification of candidate genes in this study could be used as important inputs for other research projects aiming to overcome the erratic to non-flowering, poor fruit setting and seed germination of this crop species, and thereby make development of varieties through recombination (breeding) much easier. Once identified and confirmed with further functional experiments, sex and flowering candidate genes can be incorporated into

cultivars with inconsistent flowering to non-flowering to induce regular flowering cultivars hence, tackling one of the challenges in yam breeding. The findings in this study open an avenue for further genetic and genomic studies of genes implicated in yam flowering. In addition to the genetic factors, there will also be a need to understand the environmental and epigenetic factors underling sex and flowering regulation in addition to the genetic factors.

5.5. References

Acosta, I.F., H. Laparra, S.P. Romero, E. Schmelz, M. Hamberg, J.P. Mottinger, et al. 2009. tasselseed1 is a lipoxygenase affecting jasmonic acid signaling in sex determination of maize. Science 323: 262-265. doi:10.1126/science.1164645.

Ai-Hua, S., C. Yin-Hua, S. Zhi-Hui, Z. Xiao-Juan, W. Xue-Jun, Q. De-Zheng, et al. 2014. Identification of photoperiod-regulated gene in soybean and functional analysis in Nicotiana benthamiana. Journal of genetics 93: 43-51.

Alexandersson, E., L. Fraysse, S. Sjovall-Larsen, S. Gustavsson, M. Fellert, M. Karlsson, et al. 2005. Whole gene family expression and drought stress regulation of aquaporins. Plant Mol Biol 59: 469-484. doi:10.1007/s11103-005-0352-1.

Alonso, J.M. and A. Granell. 1995. A putative vacuolar processing protease is regulated by ethylene and also during fruit ripening in Citrus fruit. Plant Physiol 109: 541-547.

Aryal, R. and R. Ming. 2014. Sex determination in flowering plants: Papaya as a model system. Plant science : an international journal of experimental plant biology 217-218: 56-62. doi:10.1016/j.plantsci.2013.10.018.

Benjamini, Y and Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1. (1995), pp. 289-300.

Bi, H., M. Wang, X. Dong and X. Ai. 2013. Cloning and expression analysis of transketolase gene in Cucumis sativus L. Plant physiology and biochemistry : PPB / Societe francaise de physiologie vegetale 70: 512-521. doi:10.1016/j.plaphy.2013.06.017.

Coen, E.S. and E.M. Meyerowitz. 1991. The war of the whorls: genetic interactions controlling flower development. Nature 353: 31-37.

Dansi, A., H.D. Mignouna, Zoundjih, eacute, J. kpon, A. Sangare, et al. 1999. Morphological diversity, cultivar groups and possible descent in the cultivated yams (*Dioscorea cayenensis/D. rotundata*) complex in Benin Republic. Genetic Resources and Crop Evolution 46: 371-388.

Dansi, A., M. Pillay, H. Mignouna, O. Daïnou, F. Mondeil and K. Moutaïrou. 2000. Ploidy level of the cultivated yams (*Dioscorea cayenensis/D. rotundata* complex) from Benin Republic as determined by chromosome counting and flow cytometry. African Crop Science Journal 8: 355-364.

Fukuchi-Mizutani, M., K. Ishiguro, T. Nakayama, Y. Utsunomiya, Y. Tanaka, T. Kusumi, et al. 2000. Molecular and functional characterization of a rose lipoxygenase cDNA related to flower senescence. Plant science : an international journal of experimental plant biology 160: 129-137.

Futamura, N., N. Ishiiminami, N. Hayashida and K. Shinohara. 1999. Expression of DnaJ homologs and Hsp70 in the Japanese willow (Salix gilgiana Seemen). Plant & cell physiology 40: 524-531.

Gowda, M., C. Jantasuriyarat, R.A. Dean and G.L. Wang. 2004. Robust-LongSAGE (RL-SAGE): a substantially improved LongSAGE method for gene discovery and transcriptome analysis. Plant Physiol 134: 890-897. doi:10.1104/pp.103.034496.

Hamadina, E.I., P.Q. CRAUFURD and R. ASIEDU. 2009. Flowering intensity in white yam (*Dioscorea rotundata*). The Journal of Agricultural Science 147: 469-477. doi:doi:10.1017/S0021859609008697.

Hamon, P. and B. Toure. 1990. Characterization of traditional yam varieties belonging to the *Dioscorea cayenensis-rotundata* complex by their isozymic patterns. Euphytica 46: 101-107.

Heijmans, K., P. Morel and M. Vandenbussche. 2012. MADS-box Genes and Floral Development: the Dark Side. Journal of Experimental Botany. doi:10.1093/jxb/ers233.

Hossain, M.Z. and M. Fujita. 2002. Purification of a phi-type glutathione S-transferase from pumpkin flowers, and molecular cloning of its cDNA. Bioscience, biotechnology, and biochemistry 66: 2068-2076. doi:10.1271/bbb.66.2068.

IPGRI/IITA. 1997. Descriptors for Yam (*Dioscorea* spp.). International Institute of Tropical Agriculture, Ibadan, Nigeria/International Plant Genetic Resources Institute, Rome, Italy.

Jaligot, E., S. Adler, É. Debladis, T. Beulé, F. Richaud, P. Ilbert, et al. 2011. Epigenetic imbalance and the floral developmental abnormality of the in vitro-regenerated oil palm *Elaeis guineensis*. Annals of Botany. doi:10.1093/aob/mcq266.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution 16: 111-120.

Lê, S., J. Josse and F. Husson. 2008. FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software. Journal of Statistical Software 25: 1-18.

Lebot, V. 2009. Tropical root and tuber crops: cassava, sweet potato, yams and aroids. CABI Publishers, Wallingford,UK: CABI pp. 413.

Ling, H. 2008. Sequence analysis of GDSL lipase gene family in Arabidopsis thaliana. Pakistan journal of biological sciences: PJBS 11: 763-767.

Liu, S. and B. Han. 2010. Differential expression pattern of an acidic 9/13-lipoxygenase in flower opening and senescence and in leaf response to phloem feeders in the tea plant. BMC Plant Biology 10: 1-15. doi:10.1186/1471-2229-10-228.

Ma, H. and C. dePamphilis. 2000. The ABCs of Floral Evolution. Cell 101: 5-8. doi:http://dx.doi.org/10.1016/S0092-8674(00)80618-2.

Marioni, J.C., C.E. Mason, S.M. Mane, M. Stephens and Y. Gilad. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18: 1509-1517. doi:10.1101/gr.079558.108.

Martin, F. 1966. Sex Ratio and Sex Determination in *Dioscorea*. Journal of Heredity 57: 95-99.

Matsumura, H., K. Yoshida, S. Luo, E. Kimura, T. Fujibe, Z. Albertyn, et al. 2010. High-Throughput SuperSAGE for Digital Gene Expression Analysis of Multiple Samples Using Next Generation Sequencing. PLoS ONE 5: e12010. doi:10.1371/journal.pone.0012010.

Micheli, F., C. Holliger, R. Goldberg and L. Richard. 1998. Characterization of the pectin methylesterase-like gene AtPME3: a new member of a gene family comprising at least 12 genes in Arabidopsis thaliana. Gene 220: 13-20.

Mignouna, H., M. Abang and R. Asiedu. 2007. Yams. In: C. Kole, editor Genome mapping and molecular breeding Pulses, Sugar and Tuber Crops. Springer, Heidelberg, Berlin, New York, Tokyo. p. 271–296.

Ming, R., A. Bendahmane and S.S. Renner. 2011. Sex chromosomes in land plants. Annu Rev Plant Biol 62: 485-514. doi:10.1146/annurev-arplant-042110-103914.

Momose, M., Y. Itoh, N. Umemoto, M. Nakayama and Y. Ozeki. 2013. Reverted glutathione S-transferase-like genes that influence flower color intensity of carnation (Dianthus caryophyllus L.) originated from excision of a

transposable element. Breeding science 63: 435-440. doi:10.1270/jsbbs.63.435.

Nielsen, K.L., A.L. Hogh and J. Emmersen. 2006. DeepSAGE--digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. Nucleic acids research 34: e133. doi:10.1093/nar/gkl714.

Oda, A., S. Fujiwara, H. Kamada, G. Coupland and T. Mizoguchi. 2004. Antisense suppression of the Arabidopsis PIF3 gene does not affect circadian rhythms but causes early flowering and increases FT expression. FEBS letters 557: 259-264.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria.URL http://www.R-project.org/.

Robinson, M.D and Smyth, G.K. 2007. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23(21):2881-7.

Robinson, M.D. and A. Oshlack. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11: R25. doi:10.1186/gb-2010-11-3-r25.

Robinson, M.D., D.J. McCarthy and G.K. Smyth. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-140. doi:10.1093/bioinformatics/btp616.

Saha, S., A.B. Sparks, C. Rago, V. Akmaev, C.J. Wang, B. Vogelstein, et al. 2002. Using the transcriptome to annotate the genome. Nat Biotech 20: 508-512. doi:http://www.nature.com/nbt/journal/v20/n5/suppinfo/nbt0502-508_S1.html.

Scarcelli, N., A. Barnaud, W. Eiserhardt, U.A. Treier, M. Seveno, A. d'Anfray, et al. 2011. A Set of 100 Chloroplast DNA Primer Pairs to Study Population Genetics and Phylogeny in Monocotyledons. PLoS ONE 6: e19954. doi:10.1371/journal.pone.0019954.

Schmid, M., T.S. Davison, S.R. Henz, U.J. Pape, M. Demar, M. Vingron, et al. 2005. A gene expression map of Arabidopsis thaliana development. Nat Genet 37: 501-506. doi:10.1038/ng1543.

Simpson, G.G. and C. Dean. 2002. Arabidopsis, the Rosetta stone of flowering time? Science 296: 285-289. doi:10.1126/science.296.5566.285.

Smirnova, A., J. Leide and M. Riederer. 2013. Deficiency in a very-long-chain fatty acid beta-ketoacyl-coenzyme a synthase of tomato impairs microgametogenesis and causes floral organ fusion. Plant Physiol 161: 196-209. doi:10.1104/pp.112.206656.

Spigler, R.B., K.S. Lewers, D.S. Main and T.L. Ashman. 2008. Genetic mapping of sex determination in a wild strawberry, *Fragaria virginiana*, reveals earliest form of sex chromosome. Heredity 101: 507-517. doi:10.1038/hdy.2008.100.

Sterk, P., H. Booij, G.A. Schellekens, A. Van Kammen and S.C. De Vries. 1991. Cell-specific expression of the carrot *EP2 lipid transfer protein* gene. Plant Cell 3: 907-921. doi:10.1105/tpc.3.9.907.

Su, C.-l., W.-C. Chen, A.-Y. Lee, C.-Y. Chen, Y.-C.A. Chang, Y.-T. Chao, et al. 2013. A Modified ABCDE Model of Flowering in Orchids Based on Gene Expression Profiling Studies of the Moth Orchid *Phalaenopsis aphrodite*. PLoS ONE 8: e80462. doi:10.1371/journal.pone.0080462.

Suzuki, H., S. Sawada, K. Watanabe, S. Nagae, M.A. Yamaguchi, T. Nakayama, et al. 2004. Identification and characterization of a novel anthocyanin malonyltransferase from scarlet sage (Salvia splendens) flowers: an enzyme that is phylogenetically separated from other anthocyanin acyltransferases. Plant J 38: 994-1003. doi:10.1111/j.1365-313X.2004.02101.x.

Tamiru, M., S. Natsume, H. Takagi, P.K. Babil, S. Yamanaka, A. Lopez-Montes, et al. 2013. Whole genome sequencing of Guinea yam (*Dioscorea rotundata*). First Global Conference on Yam. International Institute of Tropical Agriculture(IITA), Accra, Ghana.

Terauchi, R. and G. Kahl. 1999. Mapping of the *Dioscorea tokoro* genome: AFLP markers linked to sex. Genome 42: 752-762. doi:10.1139/g99-001.

Terauchi, R. and G. Kahl. 1999. Sex determination in *Dioscorea tokoro*, a wild yam species. In: C. Ainsworth, editor Sex Determination in Plants. BIOS, Oxford OX4 1RE, UK. Pp.244.

Wang, B., S.H. Jin, H.Q. Hu, Y.G. Sun, Y.W. Wang, P. Han, et al. 2012. UGT87A2, an Arabidopsis glycosyltransferase, regulates flowering time via FLOWERING LOCUS C. The New phytologist 194: 666-675. doi:10.1111/j.1469-8137.2012.04107.x.

Wang, Y., W.Z. Zhang, L.F. Song, J.J. Zou, Z. Su and W.H. Wu. 2008. Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in Arabidopsis. Plant Physiol 148: 1201-1211. doi:10.1104/pp.108.126375.

Velculescu, V., L. Zhang, B. Vogelstein and K. Kinzler. 1995. Serial analysis of gene expression. Science 270: 484 - 487.

Yang, Y., C. Ma, Y. Xu, Q. Wei, M. Imtiaz, H. Lan, et al. 2014. A Zinc Finger Protein Regulates Flowering Time and Abiotic Stress Tolerance in Chrysanthemum by Modulating Gibberellin Biosynthesis. Plant Cell 26: 2038-2054. doi:10.1105/tpc.114.124867.

Yoo, S.Y., Y. Kim, S.Y. Kim, J.S. Lee and J.H. Ahn. 2007. Control of Flowering Time and Cold Response by a *NAC-Domain* Protein in *Arabidopsis*. PLoS ONE 2: e642. doi:10.1371/journal.pone.0000642.

Chapter 6

6. Conclusions and recommendations

In this PhD study, molecular tools were applied for identification of major *Dioscorea* species (chapter 2), next generation sequencing based genotyping techniques were used to study the genetic diversity, relationship and evolution of guinea yams (chapter 3), molecular, morphological and ploidy variation across *D. alata* accessions producing aerial tubers, potential and alternative planting material were analyzed (chapter 4) and novel candidate genes involving in flowering and sex determination were identified using the high-throughput SuperSAGE expression profiling technique (chapter 5).

In chapter 2, DNA barcode regions were evaluated for *Dioscorea* species identification using a single locus *rbcL, matK*, combination of *rbcL* and *matK*, the non-coding intergenic spacer, trnH-psbA of chloroplast regions and the nuclear *ITS* region that showed different performance among markers. The *matK* locus was better for species identification where of the total number of species sequenced, 63.1% (12 out of 19 species) were discriminated in addition to revealing high interspecific variation with mean number of nucleotide substitutions per site of 0.0196 between different DNA sequences with standard deviation (SD=0.0209) among markers assessed. The combination of the two coding regions (*rbcL* + *matK*)

was the best with regard to species discrimination where 73.7% (16 out of 21) species could be defined. The *rbcL* exhibited good PCR amplification efficiency and sequence quality. However, its species discriminatory power was relatively poor with 47.6% (10 species discriminated out of 21) . Similarly, *trnH-psbA* region was also found to be the poorest with regard to discrimination efficiency of only 38.8% (7 species discriminated out of 19). None of the markers used could identify the five closely related guinea yams (*D. rotundata, D. cayenensis, D. abyssinica*, *D.mangenotina* and *D. praehensilis*). The results indicate that a two-locus combination, *rbcL* + *matK* can be utilized as a multi-locus DNA barcode region for *Dioscorea* species identification. Further studies on chloroplast and nuclear regions are worthwhile to develop universal DNA barcodes to confirm and clearly understand the taxonomy of *Dioscorea* species in general and for species that have difficulty for identification like the guinea yams.

In chapter 3, the study utilizing GBS combined with morphological and biological data provided a powerful tool for testing hypotheses that the cultivated guinea yams are cosely related and evolved or arise from the wild relatives through domestication. The study also demonstrated GBS as an effective tool for analysis of guinea yam genomic diversity, regardless of the complexity of guinea yams in terms of ploidy level, genome size, and the current lack of a reference genome.

None of the morphological descriptors we have used in this study could distinguish the two cultivated species from each other, except for tuber flesh color as the only trait. However, some descriptor traits were correlated with ploidy level. *D. rotundata* showed more diversity in terms of flowering pattern consisting of male, female, monoecious, and non-flowering. Moreover, some traits including stem color, leaf color, leaf shape, absence and presence of barky patches and waxiness, showed variation in *D. rotundata* but not in *D. cayenensis.*

The ploidy level in guinea yams is highly correlated with species identity. A single ploidy level was observed across *D. cayenensis* (3x, N=21), *D. praehensilis* (2x, N=7), and *D. mangenotiana* (3x, N=5) accessions, whereas both diploid and triploid accessions were observed in *D. rotundata* (N=32 and N=11, respectively). The triploid *D. rotundata* individuals all had distinct features, which were absent in the diploid accessions. Moreover, all triploid (3x) individuals were either male or consistently non-flowering.

The genetic diversity analysis revealed low divergence between the two cultivated species, and minimal differentiation and closer similarity of *D. mangenotiana*, *D. praehensilis* and *D. abyssinica* with the *rotundata-cayenensis* complex. This suggests that these wild relatives are either of recent divergence or variants of the cultivated species. *D. togoensis* and *D. burkilliana* were the most genetically distant species from the *rotundata-cayenensis* complex among the wild species studied. The

relatively lower divergence and higher allele sharing between *D. cayenensis* and *D. burkilliana* indicates that *D. burkilliana* could be the possible ancestor of *D. cayenensis.* Similarly, the current investigation showed that *D. cayenensis* arose from *D. rotundata* but not vice versa.

The study result further suggested that the polyploidization process in guinea yams involves potentially both allo-polyploidy, polyploids that arise due to the hybridization of two distinct species and auto-polyploidy, polyploids that arise within a species, based on the pattern of allele sharing where *D. cayenensis* harboured *D. burkilliana* alleles, 3x *D. rotundata* harboured *D. togoensis* alleles and few 3x *D. rotundata* showed reduced heterozygosity. Moreover, an increase in heterozygosity of diploid indviduals in *D. rotundata* where two 2x accessions were admixed with 3x *D. cayenensis* and two 2x *D. rotundata* accessions were admixed with 3x *D. rotundata* and reduced heterozygosity of triploid individuals in *D. rotundata* where three 3x *D. rotundata* accessions were clustered with 2x *D. rotundata* groups, supports a role for admixture arising from interspecific hybridization. The reduced heterozygosity in *D. burkilliana, D. togoensis* and *D. abyssinica* suggests these may be diploid.

The GBS data generated can further be mined once the yam reference genome becomes available, and will allow further assessment of diversity and origin of

polyploid accessions in yam. The GBS can be used widely (even in species lacking a reference genome), as it can generate genotypic information across the whole population of interest at a much lower cost per data point or low per sample cost which can be achieved by multiplexing samples. Similarly, GBS can be used for further understanding of genetic relationship studies of other species within the genus *Dioscorea.* Hence, GBS can be applied for all accessions in the IITA genebank for better management through identifying duplicates or guiding the need for further germplasm collection.

The close genetic similarity of some wild yams with the cultivated forms and sexual compatibility between species provides an opportunity for yam improvement through incorporation of genes and traits from wild relatives. Variation in ploidy within and between species can be considered as an opportunity for managing both intraspecific and interspecific hybridization in breeding programs.

In chapter 4, we have characterized aerial tuber producing *D. alata* accessions using ploidy level, phenotypic and genotypic data. The aerial tuber production pattern of accessions were associated with some phenotypic variables and identified different groups consisting of: 1) non aerial tuber producing accessions, with hastate leaf shape, less or no anthocyanin pigmentation and diploid (n=15), 2) mostly aerial tuber producing accessions, different extent of anthocyanin pigmentation, sagittate leaf shape, mainly diploid (n=44) and few triploid (n=3) and 3) all individuals producing aerial tuber, cordate leaf shape, intermediate anthocyanin pigmentation

and mostly tetraploid (n= 74) and few triploid (n=3). An increase in aerial tuber production was found associated with increased ploidy level. Tetraploid individuals produce more aerial tuber per sprout than both triploid and diploid, indicating the advantage of polyploidy in *D. alata*. SSR analysis based on neighbor-joining tree and PCO similarly reveals distinct groups according to the pattern of aerial tuber formation, leaf shape and anthocyanin pigmentation.

In addition to saving the consumable underground tuber, the use of aerial tubers as planting material could make a significant contribution to solving the problem with yam planting material. Vine cutting techniques for mini tuber production and development of improved varieties that consider the aerial tuber production potential of the cultivar could be more effective and efficient. Further evaluation and research on its advantages is therefore important to clearly understand the advantage of using varieties producing both aerial and underground tubers.

In addition to identifying candidate genes involved in flowering and sex determination, Chapter 5 has indicated the variation on flowering pattern and morphological traits across *D. rotundata* genebank accessions. The flowering pattern showed variability across years and inconsistency in flowering in some accessions. The overall observation across all *D. rotundata* accessions conserved in the IITA genebank was that there are more male followed by non-flowering, female and monoecious accessions in respective descending order. In general, the intensity

of flowering appeared to be a function of sex, with male flowers more numerous compared to female ones. The study result using morphological data that identified different group of accessions will be important particularly to breeders for easier selection of germplasm and cultivars prior to flowering. However, the current study is not exhaustive in terms of number of years and location to make final conclusion on flowering pattern and its association between other morphological traits. Hence, evaluation of accessions for flowering across different years, multiple location with additional descriptors is important.

Several candidate genes differentially regulated in relation to flowering and sex differentiation were also identified for the first time using high throughput SuperSAGE technique. Some of the genes were reported only for their expression and the remaining were characterized for their involvement in flower development and sex differentiation. Among the genes identified seven were reported for their expression only in flowers, four genes for their organ specific expression, two genes for involvement in conversion of vegetative to reproductive phase, four genes for having role in flower color development, three genes for involvement in flowering time and photoperiod regulation and two genes were reported for sex specific expression in different flowering plants.

Overall, the findings of the current PhD study can be used as important inputs for further ongoing studies aiming to overcome the challenge of poor fertility in this species. Once identified and confirmed with further experiments, candidate genes

related to flowering and sex differentiation can be incorporated to cultivars with poor flowering. As a way forward, more experiments are required to understand other factors such as environmental and epigenetic factors underling sex and flowering regulation in addition to the “hard wired” genetic factors.

Appendix A: A perl script used for sorting reads according to their corresponding samples based on 4-bp index.

Samples were bulk-sequenced and generated 35-bp long single reads. The first 4-bp corresponds to an index sequence and the final 5-bp to an adapter sequence that were ligated to each fragment during preparation of SuperSAGE libraries. The reads in fastq format were sorted to their corresponding samples using the 4-bp index sequence with a perl script described below, after which the reads were trimmed to 26-bp following removal of the 4-bp index and 5-bp adapter sequences.

```
#!/usr/bin/perl -w

use strict;
use strict;
use File::Basename;
use Getopt::Long;

my($Command) = basename($0);

#---------------------------------------------------------
sub usage() {
        my($msg) = @_;

        if ($msg) {
                print STDERR "$msg\n";
        }
        print STDERR "usage: $Command OPTIONs\n";
        print STDERR "  -i <sample index filename(.txt)>\n";
        print STDERR "  -r <read sequence (fasta,fastq, gz)>\n";
        print STDERR "  -p <Prefix>\n";
        print STDERR "  -t <threshold length of tag (26:default,";
        print STDERR " 26-30, 99,: no threshold)>\n";
        print STDERR "  sample index filename(.txt) and read sequence";
```

```
        print STDERR " (.fasta) should be input. please.....\n";
        exit(0);
}
#-------------------------------------------------------

# constant
my $index_file;
my @read_sequences = ();
my $prefix = "sage";
my $threshold_tag_length = 26;
my $Help = "";

if (scalar(@ARGV) == 0) { &usage(); }


GetOptions(
        'index_file=s'              => \$index_file,        #index file
        'read_sequence=s{,}'        => \@read_sequences,
        'prefix=s'                      => \$prefix,#prefix for output file
        'threshold_tag_length=i'    => \$threshold_tag_length,
        'help'                      => \$Help,
) or &usage();
&usage() if $Help;
&usage() unless defined $index_file;
&usage() unless defined $read_sequences[0];

print "prefix: $prefix, length:$threshold_tag_length\n";

#check
if($threshold_tag_length == 99) {
        $threshold_tag_length = "22,";

} else {
        $threshold_tag_length -= 4;
}

open IN1, $index_file,
        or die "cannot open the index file: $index_file\n";
open OUTPUT1, ">Summary_${prefix}_tag.xls"
        or die "cannot create the file: Summary_${prefix}_tag.xls";
```

```
print OUTPUT1 "sample code\tsample name\tindex seq\t";
print OUTPUT1 "Number of total tags\tNumber of unique tags\t";
print OUTPUT1 "Number of non-singleton tags\n";

while(<IN1>) {

	chomp;
	my @list = split '\t';
	my $index_seq = $list[2];
	print OUTPUT1 "$_\t";

#input "sequence file name".
	#open IN2, $read_sequence or die "cannot open the file: $read_sequence\n";

	my %count = ();
	my $total_tag_no = 0;

	foreach my $fq (@read_sequences) {
		my $fastq_fp = &open_gz_file($fq);

		while(my $line = <$fastq_fp>) {
			chomp $line;

			if ($line =~ /^$index_seq(.*)/) {
				my $seq1 ="$1\n" if defined $1;

				if ($seq1 =~ /(^\w{22,}CATG)/ ) {
					my $seq = $1 if defined $1;
					my $revcom = reverse $seq;
					$revcom =~ tr/ACGT/TGCA/;

					if($revcom =~
/(^CATG\w{$threshold_tag_length})/ ) {
						$total_tag_no++;
						$count{$1}++ if defined $1;

					} else {
						next;
					}
				} else {
```

```
                                        next;
                                }
                        } else {
                                next;
                        }
                }
                close ($fastq_fp);
        }

        open OUTPUT2, ">${list[1]}_${index_seq}.txt"
                or die "cannot create the file: ${list[1]}_${index_seq}.txt\n";
        foreach my $tag_seq (sort { $count{$b} <=> $count{$a} } keys %count){
#               print OUTPUT2 "$tag_seq\t\t$count{$tag_seq}\n";
                print OUTPUT2 "$tag_seq\t$count{$tag_seq}\n";
                #print  "$tag_seq\t\t$count{$tag_seq}\n";
        }
        close OUTPUT2,
        print "sample: $list[1]\n";
        print OUTPUT1 "$total_tag_no\t"; #total_number of tags"
        print "Number of total tags = $total_tag_no\n";

        my $keynum = keys %count;
        print OUTPUT1 "$keynum\t";
        print "Number of unique tags = $keynum\n";

        my $morethan_two = 0;
        my $key = '';
        my $value = '';

        while ( ($key, $value) = each %count ) {
                if($value > 1){
                        $morethan_two++;
                }
  }
        print OUTPUT1 "$morethan_two\n";
        print "Number of non-singleton tags = $morethan_two\n\n";
}
close IN1;
close OUTPUT1;
```

```
#-------------------------------------------------------------
sub open_gz_file() {
	my ($file) = @_;

	die "not found file, $file\n"
		if ( ! -e $file);

	my $file_fp;
	if ($file =~ /\.gz$/) {
		open($file_fp, "gunzip -c $file | ");
	} else {
		open($file_fp, $file);
	}

	return $file_fp;
}
```

Appendix B: List of differentially expressed or absent and present tags across different flower sex type.

List of differentially expressed tags between male and female flower group (p and FDR values< 0.01)

Tag sequence	Male		Female		logFC	logCPM	PValue	FDR
	TDr3631	TDr2965	TDr4087	TDr1679				
Tags highly expressed in female								
CATGCTTGTATCAGTTAGCTACTGCG	1.353286	2.8598295	436.849336	131.219015	-6.917439	7.184788	4.07E-09	1.81E-06
CATGATGATCAGGGTAGCATATGAGC	1.353286	9.532765	418.168278	388.511201	-6.097024	7.686176	7.77E-09	2.60E-06
CATGCGCCGGCGCGCACATTGGCCTA	1.353286	15.252424	354.940085	380.792436	-5.383311	7.566473	4.01E-07	8.93E-05
CATGGCGGTGGCGGTGGCCGTCGAGG	12.17957	7.626212	319.014975	347.344452	-5.064668	7.434685	9.59E-08	2.56E-05
CATGCTCCGGCCCGCTGATGGGAATG	16.239427	19.0655299	336.259028	285.594327	-4.122178	7.375365	2.86E-06	4.26E-04
CATGCCGGAGGCGGTGATGTGGCTCA	79.843851	27.6450184	945.548891	625.220013	-3.873643	8.719361	2.37E-05	3.17E-03
CATGACCACCGTCCTGCTGTCTTAAT	116.382562	63.8695253	681.140082	1685.26382	-3.714033	9.316391	4.16E-05	4.64E-03
CATGTTGCTAGCTCAGCGGTTGGGTT	18.945998	62.9162488	528.817617	465.698857	-3.586247	8.081478	7.98E-05	8.20E-03
Tags highly expressed in male								
CATGTATCTCGTCCCGCCCAGGCCAA	1151.646048	831.2571056	106.338325	74.614734	3.445767	9.089593	4.08E-05	4.64E-03
CATGCTTGCTTGTTTGATCCTTATAT	171.867272	142.0381981	8.622026	2.572922	4.656184	6.412369	2.21E-06	3.69E-04
CATGTGGCCTAGCGGTACCCCGAGTT	227.351981	180.169258	1.437004	10.291687	5.253815	6.763942	1.57E-06	3.00E-04
CATGCGGCCTTTGCTGCGCTTGAGCT	1169.238761	1335.540373	12.93304	12.864609	6.581826	9.314686	3.07E-11	2.05E-08

CATGGAGAATGGAGCGGTTGCGGGGA	3136.916028	1497.597377	1.437004	12.864609	8.466053	10.187486	6.23E-12	8.33E-09

logFC= Fold change of differential expression; log2CPM(counts per million); FDR=false discovery rate

Table 5.3b. List of differentially expressed tags between male and monoecious flower group (p and FDR values < 0.01)

Tag sequence	Male		Monoecious			logFC	logCPM	PValue	FDR
	TDr3631	TDr2965	TDr4162	TDr1506	TDr1819				
Tags highly expressed in monoecious									
CATGCATGGATTTATTTGTTGGAAAG	4.07375	0.98691	695.2469	232.82413	905.15505	-7.917333	8.531268	3.55775E-10	1.11429E-07
CATGCAGCCACTTGCCCTGTTTCCTT	92.33843	0.98691	3980.60452	4237.60699	58.64879	-6.57243	11.38605	5.48429E-06	0.000390382
CATGAAAATTACCATCACCAAAAAAA	13.57918	3.94763	604.23276	416.79676	448.01525	-5.820776	8.226582	1.07163E-08	2.79695E-06
CATGTGTGTTGTGTATTCTGTTTTTC	32.59004	76.97883	2951.63911	1095.52067	752.01368	-5.559596	10.629406	7.67023E-08	1.50145E-05
CATGCCGGAGGCGGTGATGTGGCTCA	80.11717	28.62033	2548.39591	1588.19315	352.30819	-5.317651	10.368261	3.85937E-08	8.63396E-06
CATGATTAATTTGAAGACTGCTCAGT	2.71584	3.94763	131.46487	139.27872	127.74366	-5.238995	6.380364	1.35371E-07	2.35546E-05
CATGCAACAACAAGCTCGCAAGGCTG	8.14751	1.97382	249.0248	95.6242	143.25539	-5.025161	6.668981	1.87871E-06	0.000196137
CATGGCTCGTGTTAAAGTAACTAATA	46.16922	147.0493	3761.91777	1215.05091	351.49641	-5.002643	10.896803	1.57487E-06	0.000189711
CATGCACACAATCATCATCATCATCA	5.43167	3.94763	304.64455	22.86666	126.8312	-4.983778	6.566814	6.47491E-05	0.002534928
CATGATTAATTTGAAAAAAAAAAAAA	8.14751	3.94763	240.1762	104.97874	203.4774	-4.917233	6.83756	8.85811E-07	0.000126107
CATGCCACTGAAGCTGCAAACAAAAT	14.9371	19.73816	653.53209	195.40596	602.22009	-4.785866	8.225952	1.2176E-06	0.000158897
CATGACTACATCTGGTCCTATGAATA	10.86335	7.89526	304.64455	96.66359	375.01888	-4.775766	7.333691	2.75898E-06	0.000227398
CATGATGATCAGGGTAGCATATGAGC	1.35792	9.86908	152.95432	161.10598	164.24184	-4.731398	6.646782	1.70695E-06	0.000190935

CATGAATCACTGTGTAACTGATGCAT	10.86335	5.92145	241.44029	163.18477	261.8745	-4.72594	7.118258	5.42974E-07	8.50298E-05
CATGTCAGATCGAGATTGGCTTGCAT	19.01085	2.96072	274.3065	134.08175	410.60461	-4.65546	7.413784	1.45437E-05	0.000797164
CATGAGGTTCTAGGGTTTGGTTATTT	14.9371	12.8298	359.00022	93.54541	584.88346	-4.628467	7.750803	1.12113E-05	0.000650253
CATGTGGATGATGATAATGAATGCAT	8.14751	7.89526	209.83816	107.05752	277.38623	-4.604371	6.95856	2.12504E-06	0.000207988
CATGACTGTTATGATAAAAAAAAAAA	10.86335	9.86908	164.33109	169.42113	406.95479	-4.559807	7.273393	2.48793E-06	0.000220773
CATGGTTTATGTAGGGACAAGGATCA	20.36877	7.89526	305.90864	185.01203	500.93762	-4.555749	7.689098	3.65787E-06	0.000272773
CATGCAAGTTCTACAGAGAATAAAAA	12.22126	6.90836	259.13748	127.84539	275.56132	-4.527821	7.114796	2.53762E-06	0.000220773
CATGCCTTCGCTTGGTTGTGAAAAAA	6.78959	8.88217	145.36981	112.25449	261.8745	-4.433404	6.773489	3.15085E-06	0.000246711
CATGAAGGTAGGGATGATTTTTTAAA	2.71584	8.88217	84.69371	178.77567	125.00629	-4.405424	6.363874	8.14507E-06	0.000531466
CATGTAGAGTAGTTGTATTCAGTGGT	16.29502	11.8429	385.54601	63.403	437.0658	-4.384735	7.532149	4.66897E-05	0.002031004
CATGCTGAGCGGTTTCCGATCCCGAA	9.50543	9.86908	341.30302	29.10302	239.97558	-4.371812	7.001266	0.000146982	0.004342914
CATGCAACGCCAAGGAGATTTTCGTC	31.23212	37.50251	702.83141	265.04532	870.48177	-4.148954	8.583893	8.83877E-06	0.000553661
CATGCATTTGTGTGTTATGTAATAAG	13.57918	17.76435	542.29258	73.79693	200.74003	-4.101298	7.423602	0.000119322	0.003710718
CATGCGCTTGTACTGCAACTTATAAA	14.9371	6.90836	262.92974	62.36361	198.00267	-3.999229	6.797128	7.32042E-05	0.00272947
CATGGATCAAGTGAGCAGCAACTTCA	20.36877	7.89526	313.49315	66.52118	281.9485	-3.969616	7.131933	0.000121547	0.003710718
CATGCCTAGAGCTTCTTCTTGAAGTA	23.08461	11.8429	334.9826	95.6242	344.90787	-3.889145	7.360483	6.67774E-05	0.002550572
CATGGAGTGGTCTTTGCATTTAGTAT	54.31673	27.63343	492.99326	171.49991	12.8247	-3.771313	8.468567	0.000184343	0.005136224
CATGTGAAGAAGTGATGAGTTTGTTT	16.29502	21.71198	257.8734	89.38784	409.69216	-3.716976	7.333902	0.000120343	0.003710718
CATGAACCTTGTGTTTGTATTTAAAA	35.30587	52.30613	476.56015	405.36344	796.57294	-3.667364	8.474021	2.26919E-05	0.001110487
CATGTGACTTAACCGCAACCAGGGAA	35.30587	52.30613	706.62367	272.32108	689.81574	-3.658841	8.465669	4.22612E-05	0.001946499

CATGTACATTGTGCGCTAGTTGCTCA	28.51628	26.64652	542.29258	137.19993	343.99542	-3.622496	7.767419	9.61449E-05	0.003273105
CATGGACGGATGATCCGTTTAGGAGA	14.9371	14.80362	232.59169	166.30295	139.60557	-3.580773	6.855933	3.62607E-05	0.001720734
CATGTGCTCGCAGGCCGTCGAAAAAA	66.53799	75.99192	470.23972	1205.69637	570.28418	-3.390052	8.907099	9.94603E-05	0.003313933
CATGCCACGTAATTACGTATTATTAT	28.51628	95.73008	697.77507	389.77254	741.82566	-3.285626	8.618409	0.00026221	0.006953694
CATGAAGGGAACAAAAGAAATAAAAA	39.37963	25.65961	323.60583	183.97264	370.4566	-3.169938	7.576542	0.00019694	0.005317389
CATGTGAGTTGCTCCCGTTGACCTGC	42.09546	99.67771	868.42658	566.46942	432.50352	-3.126958	8.658226	0.000294767	0.007327071
CATGAAAGTGTTGAAAGTTAAAAAAA	31.23212	16.77744	269.25016	155.90901	198.91512	-3.117511	7.093912	0.000280951	0.00721261
CATGTGCGCTGCCTCAAATTTGCAAG	32.59004	48.35849	298.32412	236.9817	411.51706	-2.955018	7.699538	0.000358452	0.008635936
CATGCATCCATCGCTGGCCTTGTTTT	662.66405	638.5295	5291.46095	5192.80955	633.44491	-2.953129	11.682123	0.00017087	0.004865131
Tags highly expressed in male									
CATGGGCTTGCTTGCCTTCGCTGCTG	1452.9724	1433.97737	160.53883	374.18163	159.67957	2.639757	9.488498	0.000323604	0.007918184
CATGATTGGCTTTGCTGCGTCTCTGC	537.73558	334.56182	37.92256	59.24543	107.66965	2.668732	7.764325	0.00038076	0.009034388
CATGCCGACGCCTTTGTCCACGCCAC	535.01974	338.50946	36.65847	98.74238	62.95937	2.718642	7.757258	0.000266425	0.006953694
CATGGCGCTACTGGTGCTGGGGGCTG	3861.91916	1783.34281	427.26082	346.11801	470.82662	2.766213	10.430225	0.000121944	0.003710718
CATGAACTACGGCCCTGGTGCCGCCG	1462.47782	997.76402	133.99304	224.50898	172.45394	2.795519	9.22934	8.34093E-05	0.002968613
CATGCCAAGAAGTTTAGTGCTTGGAT	516.00889	443.12171	94.8064	32.2212	74.82128	2.83248	7.871619	0.000158571	0.004598571
CATGGCCGACCTTCCACGATCGGTCA	357.13247	144.08857	41.71481	14.55151	44.71028	2.890448	6.934412	0.000292947	0.007327071
CATGCATGCTGTCAGTTTTGGGAGTC	199.61397	353.31308	24.01762	42.61513	42.88537	2.913153	7.073353	8.54104E-05	0.002972281
CATGGCGACTCGTAAGGTGCAGGGGC	2660.16161	1450.75481	194.66913	382.49678	229.02613	2.934209	9.944481	6.301E-05	0.002530095
CATGGGGGGAAGTGGGTTTGTGGGCT	249.85694	435.22644	41.71481	57.16664	31.02346	2.981548	7.367735	5.31185E-05	0.002248205

CATGGGGATCCCCAATAGCATCTCCA	316.39492	373.05124	31.60213	75.87572	20.074	3.015814	7.370545	0.000123217	0.003710718
CATGGCCGCGCCAATGCTCGGAGGCA	1887.5062	264.49135	159.27474	80.03329	150.55502	3.048083	8.994564	0.000421619	0.009854547
CATGGCTGAGGACGGCGAAGGTGGTG	10480.41211	5806.96686	667.43702	1097.59946	894.2056	3.199293	11.888449	1.00002E-05	0.000602321
CATGCGTGCGCGGAGTGTCGGATCGG	353.05871	327.65347	7.58451	67.56057	34.67328	3.20679	7.32334	0.000186951	0.005136224
CATGTCCAAATGCCACTGGATATGTA	274.29946	248.70082	10.11268	45.73331	25.54873	3.255969	6.941567	4.37718E-05	0.001958475
CATGTTCTCAATGTTTGGATTCTTTG	232.204	229.94957	1.26409	32.2212	20.074	3.672247	6.716986	7.72985E-05	0.002815104
CATGGGTGGCGCGACTTCTCGCTGTG	376.14332	850.71472	7.58451	78.9939	33.76082	3.926434	8.08569	2.08352E-05	0.001052515
CATGAAGATTGTCATTCCCTGAATTG	666.7378	1011.58073	10.11268	112.25449	41.06046	3.939125	8.533282	1.93536E-05	0.001010259
CATGAGTAATAAAGTTAACTTCTCGT	673.52739	419.43591	3.79226	78.9939	16.42418	4.037258	7.909061	6.01838E-05	0.002480204
CATGACACACCGGACATTACTGGACT	483.41885	291.13787	2.52817	42.61513	24.63628	4.040496	7.416115	1.47623E-05	0.000797164
CATGGGAGAGGGTGACCGATCCTCGT	381.57499	138.16713	2.52817	22.86666	18.24909	4.130609	6.839231	6.08062E-06	0.000414011
CATGGCTTGCAGCAGCTGCGGCGGAT	406.01752	146.06239	2.52817	3.11818	1.82491	6.724606	6.828897	5.31871E-12	2.08228E-09
CATGCGGCCTTTGCTGCGCTTGAGCT	1173.24126	1382.65815	1.26409	1.03939	4.56227	8.986724	9.007302	3.74054E-19	1.95256E-16
CATGCGTGGATGGGTGGACGTAGTTT	9312.60252	7764.00549	2.52817	3.11818	8.21209	10.774049	11.739794	2.63219E-25	2.061E-22
CATGCATGCGTGGATGGGTGGACGTA	9566.53321	7622.87764	1.26409	3.11818	1.82491	11.912257	11.748633	4.01965E-28	6.29477E-25

Table 5.3c. List of differentially expressed tags between female and monoecious flower group (p and FDR values < 0.01)

Tag sequence	Female		Monoecious			logFC	logCPM	PValue	FDR
	TDr4087	TDr1679	TDr4162	TDr1506	TDr1819				
Tags highly expressed in monoecious									
CATGTGGCCTAGCGGTACCCCGAGTT	1.38633	11.76841	473.58011	629.85788	569.88356	-6.597639	8.413803	1.35E-08	3.59E-06
CATGGAGAATGGAGCGGTTGCGGGGA	1.38633	14.71051	281.4419	914.95144	438.37197	-6.242613	8.383513	5.09E-07	1.16E-04
CATGCAGATCTTCGTGAAGACCCTGA	2.77267	44.13154	137.87596	415.48519	1164.12112	-5.786349	9.563035	1.13E-05	1.31E-03
CATGCTTGCTTGTTTGATCCTTATAT	8.31803	2.94210	182.66661	179.01224	152.94311	-4.766111	6.765532	2.36E-06	3.76E-04
CATGGATTGTTTGTTTGGGGATGAAT	19.40874	11.76841	474.93320	287.30359	367.25829	-4.540991	7.881118	1.82E-06	3.22E-04
CATGAGAAGCTGCTTCTTGGGTGGGA	6.93169	14.71051	121.77774	332.60916	188.01286	-4.357664	7.092531	2.41E-05	2.26E-03
CATGAATCACTGTGTAACTGATGCAT	18.02241	2.94210	258.43943	173.48717	279.58390	-4.353841	7.230451	1.15E-05	1.31E-03
CATGGAAGATACAACCCTCAGCTTTC	20.79508	20.59471	327.44682	171.27714	320.49861	-3.71042	7.456918	5.40E-05	4.31E-03
CATGCATCCATCGCTGGCCTTGTTTT	471.35533	491.33116	664.01822	520.64907	4946.78414	-3.481407	11.740796	3.27E-05	2.90E-03
Tags highly expressed in female									
CATGGGTGTCCCTTCCCAAAGGTAAG	228.74597	379.53125	20.29629	25.41531	40.91471	3.380477	7.129343	1.54E-05	1.53E-03
CATGCGCGCGCGTCACGCCGCGCCGT	280.04052	232.42611	25.70863	9.94512	26.30231	3.625009	6.863656	6.76E-06	9.80E-04
CATGCGCGCGCGCGTCACGCCGCGCC	548.99033	276.55765	62.24195	7.73509	27.27647	3.673826	7.543062	5.30E-05	4.31E-03
CATGAAACCCCCTCGGGCGAAGTTTC	619.69363	526.63639	58.1827	24.31030	29.22479	3.944665	7.983351	7.59E-07	1.51E-04
CATGGAATTAAGCACCTAAGTTTGCT	346.58480	158.87354	1.35308	27.62534	4.8708	4.458883	6.779363	1.29E-05	1.37E-03
CATGCTTCCAGAACTCCGTTTTCGGG	4382.21826	558.99952	37.88641	207.74259	52.60463	4.632833	10.037003	9.33E-06	1.24E-03

CATGCGCCCATCGGCTCGCTATGATG	163.58802	329.51550	6.76543	6.63008	5.84496	5.227412	6.687503	2.93E-09	9.34E-07
CATGGCGGTGGCGGTGGCCGTCGAGG	307.76730	397.18387	2.70617	3.31504	8.76743	6.076854	7.178614	4.61E-11	1.84E-08
CATGGCTTGCAGCAGCTGCGGCGGAT	651.57943	226.54191	2.70617	3.31504	1.94832	7.293527	7.48623	8.95E-13	4.75E-10
CATGCGTGGATGGGTGGACGTAGTTT	3005.58342	6587.36804	2.70617	3.31504	8.76743	9.844801	10.908245	5.04E-20	4.02E-17
CATGCATGCGTGGATGGGTGGACGTA	3318.89608	12795.20483	1.35308	3.31504	1.94832	11.713857	11.654607	4.88E-22	7.78E-19