MAKERERE UNIVERSITY

GENE TRANSCRIPT ANALYSIS IN DROUGHT STRESSED 'AAA' AND 'ABB' BANANAS USING NEXT GENERATION SEQUENCING TECHNOLOGIES

BY

NYINE MOSES

BLT, MUK

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL IN PARTIAL FULFILMENT FOR THE REQUIREMENTS OF THE AWARD OF MASTER OF SCIENCE IN MOLECULAR BIOLOGY AND BIOTECHNOLOGY OF MAKERERE UNIVERSITY

AUGUST, 2012

DECLARATION

I **NYINE Moses** hereby declare that the dissertation presented here is original and has never been presented by any student in any institution for an academic award.

Sign: _____ **Date:** 03rd/08/2012

This dissertation is submitted with approval of the supervisors

Sign: _____ **Date:** 03/08/2012

Dr. Ann Nanteza, PhD
Lecturer,
Department of Biomolecular Resources and Biolaboratory Sciences,
School of Biosecurity, Biotechnology and Laboratory Sciences,
College of Veterinary Medicine, Animal Resources and Biosecurity,
Makerere University, Kampala.
E-mail: nantezaa@vetmed.mak.ac.ug

Sign: _____ **Date:** 03/08/2012

Dr. James Lorenzen, PhD
Banana Breeder,
International Institute of Tropical Agriculture,
Namulonge, Uganda.
E-mail: j.lorenzen@cgiar.org

DEDICATION

To my wife, Juliet and my children; Maliza and Hope.

ACKNOWLEDGEMENTS

This work has been made a reality by the financial support from IITA-CIALCA project and IITA-Banana Breeding program, Namulonge, Uganda to which I am so grateful. I am strongly indebted to my supervisors; Dr. James Lorenzen and Dr. Ann Nanteza for their unforgettable endless mentorship and encouragement throughout the course of undertaking this project. I would like to acknowledge the support provided by the IITA-Uganda banana breeding team especially Mr. Michael Batte, Mr. Festo Ssentamu and Mr. Rogers Bazanye during experimental setup, management, data collection in the screenhouse and other logistics that were provided. I also extend my sincere gratitude to the JCVI team; Dr. Christopher D. Town and Mrs. Agnes Chan for their technical support in 454 sequencing and assembly of the reference transcriptome. I am grateful to the CIRAD group for allowing me to use their DH Pahang genome before it was available in the Public domain. I sincerely appreciate the support provided by the bioinformatics group at BecA-ILRI, Nairobi, especially Dr. Etienne de Villiers, Mr. Alan Orth, Mr. Ngara Mtakai and Mr. Nelson Ndegwa for without you I would not have managed to fully analyze the data. I extend my heartfelt gratitude to IITA management for the efficient logistical support that created a conducive environment for undertaking this project and special thanks to Dr. Piet van Asten, Ms. Susan Katebalirwe, Ms. Janet Anyango, Ms. Maria Nanyanzi, Ms. Beatrice Sakwah (IITA-Uganda) and Mrs. Susan Karonga (IITA-Nairobi). I thank my friends Ms. Mercy Kitavi, Mr. Perez Muchunguzi, Mr. Idd Ramathani, Mr. Frances Osingada, Mr. Davis Gimonde for the moral support offered to me. I am very much indebted to my family especially to my dear wife Juliet and my daughters; Maliza and Hope for their patience when I could not provide the expected attention. This work would not have been possible without support from Makerere University, Missouri University DNA core facility and Evrogen laboratory.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABA	Abscisic acid
ABF	Abscisic acid binding factor
ABRE	Abscisic acid responsive element
ABREB	Abscisic acid responsive element binding
AP2	Apetala 2
Apaf-1	Apoptotic protease activating factor 1
BBTV	Banana Bunch Top Virus
BecA	Biosciences east and central Africa
CBL	Calcineurin B-Like
CDK	Cyclin-Dependent Kinase
cDNA	Complementary DNA
CDPKs	Calcium-Dependent Protein Kinases
CDRE	Cachaco Dry Relative Expression
CGIAR	Consultative Group on International Agricultural Research
CLD	Cachaco Leaf Dry
CLW	Cachaco Leaf Well-watered
CIALCA	Consortium for Improving Agriculture-based Livelihoods in Central Africa
CIRAD	Centre de Coopération Internationale en Recherche Agronomique pour le Développement
CRD	Cachaco Root Dry
CRW	Cachaco Root Well-watered
DH	Double Haploid
DREB	Dehydration Responsive Element Binding
DSN	Duplex Specific Nuclease
EAHB	East African highland banana
ERD	Early Responsive to Dehydration
ERF	Ethylene Responsive Factor
FAO	Food and Agriculture Organization
GMGC	Global Musa Genomic Consortium
IITA	International Institute of Tropical Agriculture

ILRI	International Livestock Research Institute
INIBAP	International Network for Improvement of Banana and Plantain
JCVI	J. Craig Venter Institute
JIRCAS	Japan International Research Center for Agricultural Sciences
KAAS	KEGG Automatic Annotation Server
KEGG	Kyoto Encyclopedia of Genes and Genomes
LEA	Late Embryogenesis Abundant
LHCPII	Light-harvesting chlorophyll binding protein II
LSRP	Leaf Senescence Related Protein
MAPKs	Mitogen Activated Protein Kinases
MDRE	Mbwazirume Dry Relative Expression
MLD	Mbwazirume Leaf Dry
MLW	Mbwazirume Leaf Well-watered
MRD	Mbwazirume Root Dry
MRW	Mbwazirume Root Well-watered
NAC	**N**AM, **A**TAF1,2 and **C**UC2
NCED	9-cis epoxycarotenoid dioxygenase
NGS	Next Generation Sequencing
NIP	Nodulin Intrinsic Proteins
PIP	Plasma Intrinsic Proteins
QTL	Quantitative Trait Loci
ROS	Reactive oxygen species
Sfi	*Streptococcus fimbriatus*
SIC-SGP	Senescence Inducible Chloroplast Stay Green Protein
SIP	Small and basic Intrinsic Proteins
SMART	Switching Mechanism at the 5'-end of the RNA Transcript
SNP	Single Nucleotide Polymorphism
SOD	Superoxide dismutase
TIP	Tonoplast Intrinsic Proteins
ZEP	Zeaxanthin epoxidase

ABSTRACT

The main purpose of this study was to determine differences in gene expression and allelic heterozygosity in two banana genotypes, Mbwazirume 'AAA' which is susceptible to drought stress and Cachaco 'ABB' which is tolerant when exposed to drought stress. This was achieved using data from 454 and Illumina sequencing platforms. Drought stress tolerance in bananas has been associated with the B genome but no linkage between gene expression and drought had been illustrated. This has been in part caused by lack of a reference genome/reference transcriptome making it hard to conduct gene expression studies. In this study a reference transcriptome was generated using 454-pyrosequencing technology that comprises of 21201 contigs of lengths ranging between 500-4000 bp. During the course of this project the DH Pahang genome from the CIRAD team was accessed, which together with the reference transcriptome enabled a comprehensive analysis of gene expression and allelic heterozygosity in the two genotypes. Mbwazirume and Cachaco share many genes but what differs is the expression level of these genes. Many genes that are associated with drought tolerance were up-regulated in Cachaco and down-regulated in Mbwazirume. In cases where both genotypes showed down-regulation such as expression of aquaporins, Cachaco maintained a higher level than Mbwazirume. Transcription factors such as MYB44, MYC4, NAC, bZIP, DREB1 and DREB3 were up-regulated in tissues of Cachaco and down-regulated in Mbwazirume. The same trend was observed for genes for antioxidant enzymes and cell cycle regulating proteins. The ability of Cachaco leaves to express high levels of SIC-SGP with reduced expression of LSRP and chlorophyll catabolic enzymes would help it retain more green leaves under dry conditions. Many SNPs were detected in Cachaco, with majority being heterozygous. This was attributed to the presence of two copies of the B genome. It is possible that some of these alleles affect gene activity and efficiency of gene transcription, which results in differences in response to drought stress.

CHAPTER ONE

INTRODUCTION

1.1 Background

Banana (*Musa spp*) is a tropical and sub-tropical perennial giant herb on which many millions of people depend for their livelihood. Banana ranks number eight after wheat, rice, maize, potato, cassava, soybean, and barley as the world's most produced crops (FA0STAT, 2009). Pests and diseases are a threat to banana production in many countries (Jones, 2000). In the tropics and sub-tropics, climate changes have caused unpredictable weather and have made it hard to forecast seasons for agriculture. Drought is one of the major banana production constraints in tropical regions. It has become paramount to breed for plants that can tolerate environmental stresses in order to have sustainable food production (CGIAR, 2003). In bananas, drought tolerance has been associated with the B genome (Stover and Simmonds, 1987; Nelson *et al*., 2006). Therefore, bananas with a B genome are known for their ability to withstand adverse environmental stresses such as drought and high salinity. These include the BB, AB, ABB, AAB, BBB and AABB genotypes. However, many cultivars, such as the East African highland bananas (EAHB-AAA) are sensitive to drought yet many farmers without irrigation commercially grow them. The yields of EAHB-AAA in Uganda are reduced by approximately 20-65% due to insufficient water (Van Asten *et al*., 2011). Physiological studies have been conducted in an effort to elucidate how bananas respond to drought stress. However, it is important to understand the genotype effect on these physiological responses by determining the variation in gene expression. Although there are many gel-based molecular markers that have been used to genotype and characterize bananas, there is still limited knowledge about the molecular mechanisms involved in banana responses to drought stress. This study focused on several important needs using next generation sequencing. The first output was a reference transcriptome from 454 sequencing of cDNA libraries from two

cultivars, ‘AAA’ Mbwazirume and ‘ABB’ Cachaco, since the banana reference genome was not yet available. Secondly SNPs that could be useful in other studies were mined from the common dataset generated from both 454 and Illumina (Solexa) technologies. Also differences in gene expression in these genotypes exposed to different water stress conditions were determined. This work is expected to also facilitate primer design for gene validation based on proteomics data from KU Leuven generated using the same cultivars exposed to water deficit stress to allow them have DNA sequences of their identified genes of interest. Relevant libraries were made from different tissues (root and leaf), of plants exposed to two water levels (well-watered, pF 1.8-2.1, and dry, pF 2.8-3.1). Additional libraries were generated from flower and fruit tissues of same genotypes but under field conditions for purposes of improving single nucleotide polymorphism (SNPs) detection.

1.2 Problem statement

Drought is one of the major production constraints affecting banana crop in tropics and subtropics. Breeding for resistance has been proposed as the only sustainable solution to production constraints. Before breeders attempt gene introgression to improve susceptible varieties, there is need to have a clear understanding on the physiology of different genotypes under different water stress levels and the molecular basis for differences observed. To date there are quite a number of physiological studies that have been conducted in an effort to understand how bananas respond to drought stress. However, there is limited knowledge about the molecular mechanisms involved in banana response to drought stress.

1.3 Research hypothesis

There is no difference in the gene expression profiles of Cachaco ‘ABB’ and Mbwazirume ‘AAA’ bananas under drought stress.

1.4 Broad objective

To determine differences in gene expression and allelic heterozygosity in Mbwazirume 'AAA' and Cachaco 'ABB' bananas exposed to drought stress using data from next generation sequencing technologies.

1.4.1 Specific objectives

i. To generate a reference transcriptome from 454 sequencing of cDNA libraries from two cultivars, Mbwazirume 'AAA' and Cachaco'ABB', improved by Illumina libraries.
ii. To analyze differences in gene expression in the two genotypes exposed to different water stress conditions with respect to tissue, treatment and genotype.
iii. To mine the 454 / Illumina dataset for single nucleotide polymorphisms (SNPs) that could be useful in other studies.

1.5 Significance and Justification

This study generated a reference transcriptome from cDNA libraries from Mbwazirume and Cachaco. This will be very useful for many scientists even with the presence of reference banana genome sequence. Single nucleotide polymorphisms (SNPs) have become commonly used molecular markers. The SNPs generated by this study could be useful in other studies such as genetic diversity of more homogeneous group of bananas (EAHB-AAA). This work adds to the available knowledge and gives insights on gene expression differences in the two genotypes under different water stress conditions, thus expanding molecular understanding for the basis of physiological responses. The key genes that seem to play significant role in drought stress tolerance in banana have been highlighted and this will assist breeders to genetically manipulate them and be able to quickly select the best hybrid genotypes via marker-assisted selection.

CHAPTER TWO

LITERATURE REVIEW

2.1 The origin, domestication history and geographical distribution of bananas

Bananas and plantains, collectively known as bananas, belong to the genus *Musa,* which is divided into five sections: *Australimusa* (2n=20), *Callimusa* (2n=20), *Eumusa* (2n=22) and *Rhodochlamys* (2n=22). The fifth section *Incertae sedis* that contains only three wild species is not well documented (Daniells *et al.,* 2001). The present day cultivated edible bananas are believed to have arisen from intra- and inter-specific hybridization among *Musa acuminata* (A genome) and *Musa balbisiana* (B genome) both belonging to section *Eumusa* (Heuzé and Tran, 2011). The two species are wild diploid bananas endemic in South East Asia, which stretches from India to Papua New Guinea including Malaysia and Indonesia. The present day cultivated bananas are believed to have originated from this area (INIBAP, 1995).

In the area of origin of bananas, it is postulated that some diploids, possibly hybrids acquired the capacity to produce more pulp and progressively became parthenocarpic (INIBAP, 1995). Later, female sterility developed such that even pollinated flowers produce seedless fruits (Simmonds, 1962). As this took place, human intervention accelerated the process of banana evolution and domestication whereby hybrids that were seedless, palatable and had other good traits were selected and grown near human settlements. The wide spread of many popular cultivated bananas could have occurred by traders from Arabia, Persia, India and Indonesia who navigated the Indian Ocean from South East Asia. As they moved, they carried along with them suckers of different varieties with a broad mixture of genomic combinations such as AA, AB, AAA, AAB, ABB, AABB, AAAB, ABBB and delivered them to the coastal areas. Likewise, the Portuguese

and Spaniards between 16th and 19th century carried bananas to all over tropical America (INIBAP, 1995).

East Africa is considered a secondary center of banana genetic diversity harboring a variety of cultivars that are not found elsewhere in the world but greatly believed to have been introduced by Arab traders at East African coast way back in 600 AD. These cultivars are called the East African Highland Bananas, comprising mostly triploid AAA bananas. The question that remains unanswered is, why are they not found in other parts of the world? A number of speculations about this question are made as compiled by Karamura, (1999). This subgroup of bananas is called Lujugira-Mutika (Shepherd, 1957). The accessions in Uganda have been grouped into five clone sets (Nfuuka, Nakitembe, Nakabululu, Musakala and Mbidde) based on analysis of morphological characters (Karamura, 1999).

Banana plants grow with varying degrees of success in diverse climatic conditions, but commercial banana plantations are primarily found in equatorial regions comprising of the humid tropics and subtropics. About 130 countries are known to grow bananas and these include South East Asian nations (Cambodia, Indonesia, Laos, Malaysia, Philippines, Singapore, Thailand and Vietnam). In these areas over 500 different banana varieties are grown alongside other wild uncultivated genotypes. Other banana producing countries include; Ecuador, Costa Rica, Colombia, India, Brazil and China. In West Africa especially Nigeria and Cameroon, large fields of plantain cultivars are maintained (Ortiz and Vuylsteke, 1994) and the same also grow in Latin American. The Caribbean countries mostly grow the Cavendish bananas, which accounted for about 10% of world's production two decades back (Daniells, 1990). Production of Cavendish bananas in Latin America and the Caribbeans in 1990s was estimated to be 56% of which 32% was for export and 24% for local consumption (INIBAP, 1999).

2.2 General overview about banana production constraints

More than 70 million people in 15 countries in Sub-Saharan Africa depend on banana for their livelihood and food supply (FAO, 2010a). However, their livelihood is being threatened by a number of banana production constraints. These constraints are both biotic and abiotic. The biotic constraints include pests and diseases (Jones, 2000). Banana pests include the banana weevil (*Cosmopolites sordidus*) and the parasitic nematodes. Many nematode species have been associated with banana yield decline and amongst them are; *Radophorus similis*, *Helictylenchus multicinctus* and *Pratylenchus Goodye.* These attack and damage the banana root system that leads to toppling of plants hence, yield loss. Thus they are associated with rapid decline in yield. Many bacterial, fungal and viral diseases have been reported to affect banana and plantain and cause varying degrees of yield loss (Jones, 2000). For instance, banana bacterial wilt reduces crop yield by up to 100% (Biruma *et al*., 2007) whereas black sigatoka, a fungal disease that affects the leaves (Pro*Musa,* 2002) reduces yield by 30-50% (Rowe and Rosales, 1996). Of late, banana bunchy top virus (BBTV) transmitted by *Pentalonia nigronervosa* (banana aphid) though first reported in 1889 in many Asian banana growing countries, is reported to affect areas of Rwanda, Burundi and parts of Democratic Republic of Congo including many other banana growing areas. It is said to be more significant on plantains than bananas (Kumar and Hanna, 2010) causing significant yield decline in those areas.

Rainfall is an abiotic factor that impacts on crop production in rain fed agricultural systems. East Africa has two rainfall seasons with an average of 900–1100 mm yr^{-1} in parts of Southwest Uganda, East Rwanda and the western Kagera region in Tanzania. This increases to above 1400 mm yr^{-1} in the high altitude areas close to the Albertine rift, Mt. Elgon, and some patches bordering Lake Victoria. However, with a reported annual rainfall span of 600–2700 mm yr^{-1}, variation between years and sites is considerable (Bouwmeester *et al.*, 2009; Van Asten *et al.,*

2011). Banana requires an average rainfall of 1300 mm yr-1 to give maximum yield. This means that any reductions in the amounts received impacts greatly on the crop's yield. Looking at the last 20 years, drought in 1983/84, 1991/92, 1995/96, 1999/2001, 2004/2005 led to serious famine in various parts of the region (http://weadapt.org/knowledge-base/national-adaptation-planning/climate-changes-in-east-africa).

Due to the food crisis and environmental (abiotic) changes, it is becoming ever more important to breed environmentally stress-tolerant crops (CGIAR, 2003). Plant productivity is greatly affected by environmental stresses, such as drought and high salinity (JIRCAS, 2005). Since bananas and plantains are mostly cultivated in Tropics and Sub-Tropics, taking a global and long-term view, the availability of water is thought to be the most critical limiting factor for photosynthesis on dry land, and hence for agricultural production. Famine caused by drought has scourged humanity down the ages. Water stress causes stomatal closure and has deleterious effects on numerous plant processes, which not only reduce photosynthesis but also damage the photosynthetic machinery of the chloroplasts through a process known as photo-oxidation.

The most productive plant communities are the ones best supplied with water (Öpek *et al*., 2005). However, water supply depends on environmental changes and as seasons have changed, rains are no longer predictable. This calls for high technology investments such as irrigation systems, but water is limited in many areas. With additional investments in technology and adaptation, the effect of climate change on actual agricultural production could be reduced, but because these inputs raise the cost of production, prices could also rise. Similarly, increased irrigation could help farmers cope with droughts and excessive heat, but water shortages and the high cost of irrigation systems limit the potential of irrigation to solve the problem (Cline, 2007). The need

for crop varieties that can tolerate drought stress with maximum yield potential would be very beneficial for rain-fed agriculture.

To achieve the above, breeding strategies have to be geared towards drought tolerant genotypes amongst the breeding priorities. Successful breeding depends on understanding the genetic potentials of different genotypes. Many molecular tools exist to date that can be used to dissect the genome of an organism. However, next generation sequencing technologies are faster, robust, cost effective and high throughput compared to traditional Sanger technology and other gel-based methods (Imelfort and Edward, 2009). Understanding the genetics of bananas especially how genes are expressed in response to drought stress is an important step in molecular breeding strategy of tolerant banana genotypes.

2.3 Plant response to water deficit stress

Water deficit stress hereafter referred to as drought stress can be defined as the exposure of plant to low external water potential. A low plant water content results into several physiological responses that manifest as reduced plant growth and productivity. When plants are exposed to drought stress, abscisic acid (ABA) is produced that plays a central role in signal transduction. It induces stomatal closure, which leads to limited carbon dioxide (CO_2) fixation and reduced $NADP^+$ regeneration by the Calvin Cycle. Thus increasing the reactive oxygen species (ROS) such as: hydrogen peroxide (H_2O_2), superoxide (O_2^-) and hydroxyl (OH) radicals. These ROS cause oxidative stress that plants have to deal with in order to survive (Abedi and Pakniyat, 2010). Plants employ both non-enzymatic and enzymatic strategies to control the effects of oxidative stress resulting from reactive oxygen species (Xu *et al.,* 2008). They respond and adapt to drought stress at both the cellular and molecular levels, for instance by the accumulation of osmolytes and proteins specifically involved in stress tolerance (Shinozaki and Yamaguchi-

Shinozaki, 2007; Fleury *et al.,* 2010). To the farmers, these responses are of great importance since they result in yield reduction for the drought sensitive crops. However, the extent of the effect greatly depends on the atmospheric conditions, type of crop and genotype (genetic makeup) and growth stage amongst others. It is known that most plants are unable to take up any water at pF≈4.2 and this point is referred to as the permanent wilting point (Koorevaar *et al,* 1983). Banana plants respond to drought stress by reducing the leaf area through which transpiration takes place and by slowing down the transpiration rate through stomatal closure. These physiological responses slow down photosynthesis and manifest as reduced leaf emergence and growth rate accompanied by delayed fruiting under field conditions (Robinson and Saúco, 2010).

Drought stress results into significant changes in gene expression. While some genes are up-regulated, others are down-regulated and a number of proteins and RNAs are produced that play significant roles in biosynthesis reactions. For instance, genes for the synthesis of the hormone abscisic acid (ABA) are up-regulated, leading to increased levels of this hormone (Xiong and Zhu, 2003; Moumeni *et al*., 2011). Abscisic acid transported into leaves induces stomatal closure in the early stages of drought stress. But ABA also forms a part of the signaling chain for the control of other genes responsive to drought stress (Öpik *et al*., 2005). Although some stress responsive genes are regulated by ABA, others are induced by stress only or a combination of stress and ABA (Luo *et al.,* 1992).

Studies on various plants have reported several stress responsive genes. Ideally one would think of improving stress sensitive plants by introducing some of these genes into plants of interest to confer tolerance to drought stress. However, plant tolerance to drought stress is a multifaceted process that involves many genes coding for products that are involved in diverse pathways. That

is to say drought tolerance is a quantitative trait (Fluery *et al*., 2010). This is a major drawback to plant breeders given the limited accessible germplasm with diverse genetic diversity. A solution to this problem is to transform a plant with drought stress-response transcription factors that can efficiently regulate many downstream genes involved in protecting the plant against drought stress (Hardy, 2010). Therefore, a lot of literature is available on the expression profiles of transcription factors and several other genes involved in stress responses. These include: dehydration responsive element binding (DREB) transcription factor which binds DRE/CRT elements and are induced by both cold and drought stress (Zhou *et al.,* 2010), stress sensor proteins such as calcineurin B-like (CBL) proteins together with calmodulin act as sensors for calcium concentration changes during drought stress, abscisic acid responsive elements (ABRE) and abscisic acid responsive element binding factor (AREB/ABF) are *cis*-acting elements at the promoter region induced by ABA (Cheong *et al.,* 2003; Hardy, 2010), apetala2/ethylene responsive factor (AP2/ERF) is a transcription factor involved in ABA independent pathways during drought stress response (Hardy, 2010) , NAC proteins are induced by multiple stresses including drought (Fujita *et al*., 2004), zinc finger proteins, C-repeat binding factor (CBF), early responsive to dehydration (ERD) and late embryogenesis abundant (LEA) (Hu *et al.,* 2006, Hardy, 2010). Other key genes that confer drought stress tolerance include those coding for antioxidant enzymes such superoxide dismutase, guaiacol peroxidase, catalase, ascorbate peroxidase and enzymes involved in the biosynthesis of non-enzymatic antioxidants such as tocopherol cyclase involved in tocopherol (vitamin E) synthesis (Vidi *et al*., 2006). A study in *Brassica napus* (oilseed rape) showed up-regulation of antioxidant enzymes such as superoxide dismutase and guaiacol peroxidase in drought resistant cultivars (Abedi and Pakniyat, 2010) and this was consistent with other studies on poplar (Xiao *et al*., 2008), Sunflower (Gunes *et al*., 2008) and cowpea (Manivannan *et al*., 2007). In cowpea plants, up-regulation of cystatin in

leaves of tolerant cultivars was indicated to confer drought stress tolerance by blocking cysteine proteinase activities. Cystatins are proteinaceous reversible inhibitors of cysteine proteinases, such as papain and cathepsin (Diop *et al.,* 2004)

2.4 Dissection of banana genome

The most effective way of dissecting and fully understanding the genome of an organism is by having its entire genome sequenced. To date a number of plants have had their genome fully sequenced and the data is available to the public, for example *Arabidopsis thaliana* (http://www.arabidopsis.org), a model plant for dicots, and *Oryza sativa,* the first monocot and cereal plant to have its genome fully sequenced and annotated (Zhao *et al*., 2004; (http://rise.genomics.org.cn/)). Banana has been considered to be an ideal model for understanding genomic evolution in relation to biotic and abiotic stresses, among the polyploid and vegetatively propagated crops (http://www.musagenomics.org/). To this effect, Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD) conducted a full genome-sequencing project for banana in 2009. However, this reference genome is yet to be made available for public use. Nevertheless, molecular understanding of banana at the transcriptome level is not limited by availability of a reference genome sequence. Next generation sequencing such as 454 sequencing technology offers an opportunity for *de novo* assembly of reads to generate a reference transcriptome, against which gene expression studies can be referenced. This is because unlike other next generation sequencing technologies, the 454 technologies gives longer reads, ranging on average from 400 to 600 bp thus making *de novo* assembly of reads into contigs much easier (Margulies *et al*., 2005). Indeed, the search for agronomically important genes such as those involved in disease resistance and drought tolerance can be done by cDNA sequence analysis. Tremendous research has been carried out in this direction. By using next generation sequencing, it is now clear that banana like other high

eukaryotes contains a large amount of repetitive DNA sequences that are known not to code for anything. Such sequences constitute more than 50% of the so far sequenced genomic libraries of *Musa acuminata* and the most frequent repeats are Ty1/copia and Ty3/gypsy retroelements (Hřibová *et al*., 2010). Similarly, Cheung and Town, (2007), did a large scale BAC end sequencing which showed a great potential to anchor a small proportion of the genome of *Musa acuminata* to the genomes of *Oryza sativa* and possibly *Arabidopsis thaliana*. This means that even in the absence of reference genome it is still possible to get information from partial sequences because of macro- and micro-synteny. This can be detected by carrying out BLASTN match to fully sequenced and annotated species. Lescot *et al*., (2008), described the utility of comparative analyses between distantly related monocot species such as Oryza and *Musa* in improving the understanding of monocot genome evolution. They revealed the presence of micro-synteny regions that have persisted in the course of evolution between order Poales and Zingiberales over a long period. Sequencing and deciphering of gene function in banana is very important in understanding how genes interact with each other and the environment to make a plant thrive or die for example under conditions of water deficit (INIBAP, 2005). It should be noted that plant tolerance to drought stress is a multifaceted process that involves synergistic effects of different genes that encode proteins involved in different metabolic pathways.

2.5 Molecular analysis of gene expression

When organisms are exposed to stressful conditions such as water deficit, they respond by switching ON some genes and switching OFF others or up-regulation and down-regulation of different genes. Transcript abundance can be measured to gain an understanding on how different genes participate in response to a given condition. Scientists have often employed microarray technology in the study of gene expression (Schena *et al*., 1995), but more recently this is being accomplished by RNA-sequencing technologies.

2.5.1 Micro-array technology

Gene expression analysis by microarray technology involves isolating mRNA and converting it to cDNA by enzyme reverse transcriptase. Then the fluorescently labeled cDNA is hybridized to the chip to reveal the expression level for each gene, identifiable by its known location on the chip. Generally, two predominant microarray technologies are commonly used that is the cDNA microarray technology and the oligonucleotide microarray technology. The most commonly used oligonucleotide microarray technology is the Affimetrix Gene Chip. However, there are other technologies like Affymetrix 23 000 ATH1 GeneChip and 22 000 Agilent oligoarrays that have been employed on *Arabidopsis* and rice respectively (Shinozaki and Yamaguchi-Shinozaki, 2007).

Microarray technologies have found wide use in the scientific research. They have been used to identify clusters of co-expressed genes (Eisen *et al*., 1998) and allowed inferences about biological processes implicated in plant development and environmental responses (Wullschleger and Difazio, 2003). The value of microarrays to elucidate the genetic control of gene expression variation was demonstrated in mice and *Drosophila melanogaster* (Karp *et al*., 2000; Eaves *et al*., 2002; Wayne and McIntyre, 2002). Microarrays have been used to determine gene expression levels in segregating populations and identify genomic regions (gene expression QTLs, or eQTLs) explaining transcript variation in co-regulated genes (Brem *et al.*, 2002; Schadt *et al*., 2003; Yvert *et al*., 2003). A number of stress-inducible genes have been identified using microarray analysis in various plant species, such as *Arabidopsis* and rice (Seki *et al.*, 2002*a*; Seki *et al*., 2002*b*).

The proteins identified by microarray analysis and thought to function in abiotic stress tolerance include molecules such as chaperones, late embryogenesis abundant (LEA) proteins, osmotin,

antifreeze proteins, mRNA-binding proteins, key enzymes for osmolyte biosynthesis, water channel proteins, sugar and proline transporters, detoxification enzymes, and various proteases. Also regulatory proteins such as protein kinases, protein phosphatases, enzymes involved in phospholipid metabolism, and other signaling molecules such as calmodulin-binding protein are involved in further regulation of signal transduction and stress-responsive gene expression (Shinozaki *et al.*, 2003).

Studies on drought-inducible genes in *Arabidopsis* and rice using microarrays indicated that more than half of the drought-inducible genes were also induced by high salinity and abscisic acid (ABA) treatments, implicating significant cross-talk between the drought, high salinity, and ABA response pathways as compared to only 10% of them that were also induced by cold stress (Shinozaki and Yamaguchi-Shinozaki, 2007). Similarly, studies to determine the role of proline, K/Na ratio and chlorophyll in wheat genotypes indicated that genotypes with higher proline, K/Na ratio and chlorophyll contents had higher grain yield (Khan *et al*., 2009). Chlorophyll is used to trap the solar energy used in photosynthesis hence, a key factor in crop yield. Increased concentration of sodium in the plant under conditions of high salinity disturbs many metabolic activities because it is known to be a toxic element (Akram *et al*., 2007). Plants that are able to restrict sodium in their roots and avoid its translocation are said to be tolerant to high salinity. Proline accumulation in plants under conditions of drought and salinity stress is a key strategy in ensuring tolerance to these stress conditions. It has been found out that this amino acid is involved in prevention of programmed cell death (apoptosis). Proline and glycinebetaine have been implicated in inducing antioxidant defense gene expression and suppression of cell death in cultured tobacco cells under salt stress (Banu *et al*., 2009). While in animals, prevention of programmed cell death results into cancers, in plants it is a survival strategy under conditions that lead to accelerated apoptosis such as drought and high salinity stress.

Despite the significant contributions micro-array technologies have made to gene expression studies, the technologies are quite laborious, the precision is not optimal due to cross hybridization of closely related genes and the technology is limited to the number of reference genes that can be spotted on the chip. Therefore, next generation sequencing technologies are evolving at a faster rate to try and close the loopholes in micro-arrays.

2.5.2 The role of next generation sequencing technologies in research

Sanger sequencing has been greatly utilized by scientist and up to date it still remains the gold standard because of its ability to produce longer reads with minimum errors (Sanger and Coulson, 1975). With this method genes can be analyzed at nucleotide level. However, it is very expensive with low throughput. The advent of next generation sequencing technologies also referred to as second and third generation sequencing or 2G and 3G technologies have revolutionized biological research. They have taken center stage because of high throughput and reduced cost and time of sequence generation. These methods produce millions to billions of relatively short reads, usually at the expense of read accuracy. They eliminate the use of *in vivo* bacterial cloning stage of the Sanger methodology by using either 'emulsion PCR' (Roche 454, Applied Biosystems-SOLiD) or 'bridge PCR' (Illumina) for target amplification to generate a polony thus speeding up sample preparation (Shendure and Ji, 2008).

The first commercial next generation sequencing system was produced by 454 technologies and commercialized by Roche. It works on the principle of pyrophosphate detection that was earlier described by Nyren and Lundin, (1985). To date several other companies including Illumina/Solexa (Turcatti *et al.,* 2008), Applied Biosystems, Helicos Biosciences and Pacific Biosciences have joined the competition (Imelfort and Edwards, 2009). The platforms offer a variety of experimental approaches for characterizing a transcriptome, including single-end and

paired-end cDNA sequencing, tag profiling, methylation assays, small RNA sequencing, sample tagging ("barcoding") to permit small subsample identification following multiplex sequencing, and splice variant analyses (Wall *et al*., 2009). Hřobiva *et al*., (2010) used low-depth 454 sequencing to thoroughly characterize the repetitive part of banana (*Musa acuminata* cv. 'Calcutta 4') genome such as the LTR-retrotransposons, DNA transposons, LINE-like elements, tandem repeats and simple sequence repeats. They found out that the LTR-retrotransposons were the most abundant repetitive sequences than the DNA transposons.

Sequencing by 454-technology and *de novo* assembly of the transcriptome was done to identify genes and to reconstruct the metabolic pathways involved in the production of biofuel precursors in *D. tertiolecta.* This is a non-model microalgae species involved in biosynthesis and catabolism of fatty acids, triacylglycrols, and starch to produce biofuel (Rasmani-Yasdi *et al*., 2011). Cassava (*Manihot esculanta*) genome was sequenced by 454 sequencing technology using the whole genome shortgun strategy (http://www.phytozome.net/cassava#D1). Its availability to the public has sparked off several research projects geared towards improving the crop. For organisms whose reference genome sequence is available, next generation sequencing technologies are being used in resequencing, and genetic variations important for molecular breeding can now be detected and analyzed effectively (www.bgisequence.com). The main challenge of utilizing these new technologies was lack of bioinformatics tools to handle the data in areas of sequence quality scoring, alignment, assembly and data release (Shendure and Ji, 2008) but currently many scientists in bioinformatics and computational biology are committed to developing tools that allow *de novo* assembly of sequence reads, discovery of sequence variations such as SNP and INDEL detection. It is now possible to sequence the genome of organisms at a lower cost, generate information on gene expression profiles even in organisms whose genome has not been fully sequenced.

2.5.3 Trends in bioinformatics

When the basic structure of DNA, was discovered by Watson and Crick in 1953, it was elucidated to contain two chains that composed of four repeating nucleotides. The order of which determined the characteristics of an organism. This discovery stimulated scientists to find out the order of the nucleotides of the DNA and the variation that exists between different organisms. A significant breakthrough was achieved when Fredrick Sanger in 1975 developed the chain termination method using the dideoxynucleotides (Sanger and Coulson, 1975). After gel electrophoresis, the nucleotides would be ordered manually but this was very laborious exercise and time consuming. The introduction of dye-terminator technology using fluorescent-labeled dideoxynucleotides was a great improvement that allowed sequencing in a single reaction tube other than the four parallel reactions. This technology lent itself to automation increasing the speed of sequencing hence, the DNA sequence databases started to grow rapidly. Since then various sequencing technologies have continued to emerge with increased throughput, robustness, less time requirement and reduced cost such as the next generation sequencing technologies.

As more and more sequences were generated, challenges in handling large sequence datasets and getting meaningful information about them became evident. Therefore, various algorithms and statistical tools had to be developed to allow alignment, assembly, polymorphism detection and visualization of alignment and structural variants within the sequences and this led to the development of bioinformatics field (Magi *et al.,* 2010). Basic local alignment search tool (BLAST) which was modified into BLAT (Kent, 2002) is a traditional alignment tool that is widely used for aligning sequences but it is limited in speed and accuracy when handling millions of short reads. The new tools that have been developed have the ability to quickly and efficiently align billions of short reads, align non-unique reads and reads that do not match exactly the

reference genome (Magi *et al*., 2010). These tools are compatible with different sequencing platforms such as Bowtie (Langmead *et al.,* 2009) for Illumina, BWA (Li and Durbin, 2010) for Illumina, SOLiD and 454, SOAP2 (Li *et al.,* 2009) for Illumina and Maq (Li *et al,* 2008) for Illumina and SOLiD. They put into consideration issues regarding sequence length of references and reads, quality of reads with respect to sequencing errors, sequence similarity to a reference sequence (SNPs, micro-indels), and the number of reads to be aligned with the reference (Malhis *et al.*, 2009). The earlier developed alignment tools include; Exonerate (Slater and Birney, 2005), MUMer (Delcher *et al.,* 2002; Kurtz *et al.,* 2004) and MUMmerGPU (Schatz *et al.,* 2007) among others. However, these tools have inherent problems such as stalling during program execution when reads are many, production of large number of reads that are misaligned and generation of large number of mismatches that would be considered sequencing errors which result into false positive SNPs. To solve the above problems, Slider algorithm was developed that uses not only the most probable base, but also all possible bases with a probability above a certain base probability threshold (baseMinPrb) provided by the Illumina probability files in order to generate all possible reads with probability above a certain read probability threshold (read_0_MinPrb). This reduced the level of alignment approximation needed and improved accuracy thus reducing the number of misaligned reads and the number of base mismatches (Malhis *et al.,* 2009).

In situations where the reference genome of an organism does not exist, *de novo* assembly of reads into contiguous sequences (contigs) and contigs into scaffolds is done. Newbler provides the tools used for *de novo* assembly of 454-reads. The algorithms that have been used in *de novo* assembly of bacterial genome are based on de Bruijn graphs and these include; Velvet (Zerbino and Birney, 2008), SOAPdenovo (Li *et al.,* 2008) and Edena (Hernandez *et al*., 2008). Next generation sequencing technologies though they are of high throughput, they are prone to errors. The high number of reads generated and availability of tools that can resolve such errors during

assembly normally compensate this. Verified consensus assembly by K-mer extension (VCAKE) is a modified algorithm of simple K-mer that overcomes error by using high depth coverage (Jeck *et al.,* 2007). Similarly, the short sequence assembly by K-mer search and 3' read extension (SSAKE) is another tool for aggressively assembling millions of short nucleotide sequences by progressively searching through a prefix tree for the longest possible overlap between any two sequences. The short sequence assembly by K-mer search and 3' read extension is designed to help leverage the information from short sequence reads by stringently assembling them into contiguous sequences that can be used to characterize novel sequencing targets (Warren *et al.,* 2007). The VCAKE and SSAKE assembly process are almost identical at the beginning but VCAKE uses, two multi-FASTA files to separately populate *bin* and *set* hash tables from a pool of reads. Great divergence from the SSAKE method occurs during extension of the seed sequences from *set* (Jeck *et al.*, 2007).

Following successful development of alignment or assembly tools for short reads user-friendly tools had to be developed for visualization of alignments or assemblies. Among the viewing tools are: Eagle View (Huang and Marth, 2008), Maq View (Li *et al.,* 2008), Map View (Bao *et al.,* 2009) and Tablet (Milne *et al.,* 2010).

Detection of sequence variants such as single nucleotide polymorphisms (SNPs), insertion and deletion (INDEL) is easily done after alignment of reads with the reference sequence. All the available tools for SNP and INDEL discovery involve data preparation after which each nucleotide is called under a Bayesian framework (Magi *et al.,* 2010). Some of the programs that provide variant detection tools include Maq (Li et al., 2008) and SSAHA2 (Ning *et al.,* 2001).

With advancement in computer technology and wide use of Internet, many bioinformatics software tools have been developed that incorporate sequence alignment, assembly, trimming,

mapping, translation, polymorphism detection, generation of restriction maps, gene expression profiling and primer designing. Many of the software programs are web-based and open source so that different users can modify them based on their requirements for example galaxy (http://galaxy.psu.edu/). Some of these web-based software tools are limited in their functionality. For instance Clustalw (http://www.genome.jp/tools/clustalw/) is useful in multiple alignment of sequences, Expasy (http://web.expasy.org/translate/) is used as sequence translation tool, SNP finder (http://snpsfinder.lanl.gov/) for detection of SNPs, webcutter (http://rna.lundberg.gu.se/cutter2/) for generating restriction maps, Primer 3 (http://frodo.wi.mit.edu/) for primer designing. However, there are other patented (restricted access) software that are very key in analysis of next generation sequence data. These are maintained and upgraded to cope up with the dynamics of sequencing technologies. Such software include; Newbler and CLC Bio (CLC genomic workbench). They have many tools with different capabilities to handle data from various next generation sequencing platforms. They are user friendly compatible with graphical user interface programs.

In functional genomics scientists are much interested in knowing what type of proteins do the sequences generated code for, are they structural protein, transcriptional factor or enzymes, where in the cell are the proteins located, what are their functions, if enzymes which metabolic pathways are they involved in and how do they interact with each other? The process of answering these questions is what is referred to as annotation of the sequences. Therefore, an efficient functional annotation of DNA sequences is a major requirement in biological research. An integrated and biologist-oriented solution that is based on gene ontology vocabulary is available known as Blast2GO (www.blast2go.com). It is different from the traditional Blast in that it combines various annotation strategies and tools controlling type and intensity of annotation, it has numerous graphical features such as the interactive GO-graph visualization for

gene-set function profiling or descriptive charts, it has general sequence management features and has high throughput capabilities (Götz *et al* 2008). It allows InterProscan to be done and assignment of enzyme codes to the sequences using gene ontology terms. It is also linked to Kyoto Encyclopedia of Genes and Genomes (KEGG) that allow metabolic pathways map retrieval.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Study location

The study was conducted in five locations. Screenhouse drought trials, extraction, purification and stabilization of total RNA was done at International Institute of Tropical Agriculture (IITA) research station at Namulonge, located 30 km North of Kampala in Uganda at 32° 27' E longitude and 0° 32' N latitude. Complementary DNA (cDNA) synthesis was done at Evrogen, Russia as instructed by J. Craig Venter Institute (JCVI) and sequencing by 454 technology to generate a reference trascriptome was done at JCVI in USA while cDNA synthesis and subsequent sequencing by Illumina was done at the DNA Core Facility laboratory of the University of Missouri in the USA. Data analyses requiring high computing power was done at Biosciences east and central Africa/International Livestock Research Institute (BecA-ILRI), Nairobi.

3.2 Experimental design

The study involved two banana genotypes that is: Mbwazirume "AAA" and Cachaco "ABB". Young plantlets of approximately same age and uniform height were planted in buckets containing 30 kg of thoroughly mixed soil with manure (in the ratio of 3:1 respectively). The soil was fully saturated with water prior to planting. The experiment had two treatments; well-watered (pF = 1.8-2.1) and dry (pF = 2.8-3.1) and plants were randomly assigned to these treatments. The experiment had two replications with two plants per genotype. Buckets with plants were covered with transparent polythene bags up to the lower pseudostem to ensure that water loss from the soil was only through the leaves by transpiration. Tensiometers filled with water were inserted into the soil to a depth of 15 cm in representative buckets to monitor the

changes in soil moisture content due to transpiration. The treatments were initiated three weeks later after planting but the dry treatment stabilized after 10 weeks from the time of planting. The biomass data were taken on weekly basis and the minimum and maximum bucket weights adjusted according to the adopted plant growth model developed by IITA to maintain the pF within the range. Plants were maintained in a screenhouse for four months to ensure stability of all treatments before sampling. Also field-grown samples from these genotypes were included to provide a greater range of tissues for RNA extraction and analysis.

3.3 Samples

Root and leaf samples were obtained from both genotypes in each treatment. Fruit and flower samples were obtained from the same genotypes under field conditions. This was because bananas take long time to flower under screenhouse conditions. All samples were taken on a bright sunny day after 1:00 pm to ensure that optimal gene expression was captured. Samples were aseptically taken ensuring that no contamination with exogenous ribonucleases (RNases) occurred and quickly frozen in liquid nitrogen.

3.4 RNA isolation

Frozen samples were crushed into powder in liquid nitrogen without allowing them to thaw. Approximately 0.1 g of the powder were transferred into a 1.5 ml eppendorf tube pre-cooled in liquid nitrogen. Without allowing the sample to warm even slightly, the PureLink plant RNA extraction reagent from Invitrogen (Cat No. 12322-012) was added and mixed thoroughly with the frozen powder by vortexing. The process of RNA isolation was done following the small scale RNA isolation protocol from Invitrogen revised on 14th December 2005 (appendix I). The extracted RNA was run on a 1.2% agarose gel in 1x Tris Acetate Ethylenediaminetetraacetic acid (1x TAE) buffer to check its quality and concentration. Cleanup of the extracted RNA was done

using RNeasy mini kit (50) from QIAGEN (Cat No. 74904). During cleanup, on column digestion of contaminating DNA was done using DNase I enzyme from QIAGEN (Cat No. 79254). For stability of RNA, all samples were transferred to GenTegra tubes (GTR5025-S), dried by high speedvac centrifugation for approximately 4 hours at 30^{o}C, and stored at -20^{o}C before complementary DNA (cDNA) library construction and sequencing. For samples intended for 454 pyrosequencing, approximately equal concentration of total RNA from different tissues and treatments for a particular genotype were drawn and pooled together to form single composite samples for each genotype. For Illumina sequencing, RNA from different tissues, genotypes and treatment were treated as independent samples during library construction and sequencing.

3.5 Sample preparation for 454 sequencing

3.5.1 Complementary DNA synthesis

The pooled RNA from Mbwazirume and Cachaco tissues was used for complementary DNA (cDNA) synthesis following the switching mechanism at the 5' end of RNA transcript (SMART) approach (Zhu *et al*., 2001). Normalization of cDNA involved denaturation/reassociation, treatment with duplex-specific nuclease (DSN), (Shagin *et al*., 2002) and amplification of normalized fraction by PCR. The sequences of primers used are shown in the appendix II. To synthesize the first cDNA strand, a primer annealing mixture (5 µl) containing 0.3 µg of total RNA, 10 pmol SMART-SfiIA oligonucleotide and 10 pmol CDS-SfiIB-GC T23 primer was heated and maintained at 72^{o}C for 2 min and cooled on ice for 2 min. First-strand cDNA synthesis was then initiated by mixing the annealed primer-RNA with Reverse Transcriptase to a final volume of 10 µl, containing 1x First-Strand Buffer (50 mM Tris-HCl (pH 8.3); 75 mM KCl;

6 mM $MgCl_2$), 2 mM DTT and 1 mM of each dNTP. The reaction mixture was incubated at 42^{o}C for 2 h in an air incubator and then cooled on ice.

To prepare the double stranded cDNA (ds cDNA), the first-strand cDNA was diluted 5 times with TE buffer, heated at 70^{o}C for 7 min and used for amplification by Long-Distance PCR (Barnes, 1994). The 50 µl PCR reaction mixture contained 1µl diluted first-strand cDNA, 1 x Encyclo reaction buffer (Evrogen), 200 µM dNTPs, 0.3 µM SMART PCR primer and 1 x Encyclo Polymerase mix (Evrogen). Nineteen (19) PCR cycles were performed in MJ Research PTC-200 DNA thermocycler. Each cycle included denaturation at 95^{o}C for 7 s, annealing at 65^{o}C for 20 s and extension at 72^{o}C for 3 min. Amplified cDNA PCR products were purified using QIAquick PCR Purification Kit (QIAGEN, CA) and concentrated by ethanol precipitation. DNA pellet was diluted by milli-Q water to a final cDNA concentration 50 ng/ul.

3.5.2 Complementary DNA (cDNA) normalization and library construction

Complementary DNA normalization involved three stages that included; hybridization, duplex-specific nuclease (DSN) treatment and two PCR reactions.

3.5.2.1 Hybridization

Hybridization reaction contained 3 µl of purified ds cDNA and 1 µl of 4x hybridization buffer (200 mM HEPES-HCl, pH 8.0; 2 M NaCl). The reaction mixture was overlaid with one drop of mineral oil and incubated at 98^{o}C for 3 min followed by 68^{o}C for 5 h.

3.5.2.2 Duplex-Specific Nuclease treatment

Preheated 3.5 µl milli-Q water, 1 µl of 5x DNAse buffer (500 mM Tris-HCl, pH 8.0; 50 mM MgCl2, 10 mM DTT) and 0.5 µl DSN enzyme were added to the hybridization reaction at 68^{o}C.

The mixture was further incubated at 67°C for 20 min. On completion of DSN treatment, DSN-enzymes were inactivated by heating at 97 °C for 5 min.

3.5.2.3 Polymerase chain reactions (PCRs)

Complementary DNA samples were diluted by adding 30 µl milli-Q water and used for PCR amplification. The first PCR reaction (50 µl) contained 1 µl of diluted cDNA, 1x Encyclo reaction buffer (Evrogen), 200 µM dNTPs, 0.3 µM SMART PCR primer and 1x Encyclo Polymerase mix (Evrogen). Eighteen (18) PCR cycles were performed in a thermocycler and involved denaturation at 95°C for 7 s, annealing at 65°C for 20 s and extension at 72°C for 3 min. To perform the second PCR, 1 µl from ds cDNA preparation (non-normalized cDNA) and 1 µl of cDNA sample from the first PCR (normalized cDNA) was diluted by adding 19 µl of milli-Q water and used for PCR amplification. The PCR reaction (50 µl) contained 1 µl 20x diluted cDNA (first-PCR), 1x Encyclo buffer (Evrogen), 200 µM dNTP mix, 0.2 µM, CDS-SfiBI T19-454 primer, 0.2 µM SfiIA PCR primer and 1x Encyclo polymerase mix (Evrogen). Three PCR cycles were performed in MJ Research PTC-200 DNA thermoycler for every sample. Each cycle included denaturation at 95°C for 7 s, annealing at 50°C for 20 s and extension at 72°C for 3 min. An additional 11 PCR cycles were performed on every sample and each cycle included denaturation at 95°C for 7 s, annealing at 63°C for 20 s and extension at 72°C for 3 min. Agarose gel electrophoresis was carried out on cDNA before and after normalization to check whether the process was effective. The normalized cDNA was used for library construction following the protocol from Roche 454 sequencing technologies.

3.5.3 454 sequencing

After library construction, sequencing of clones was performed according to the optimized conditions of 454 sequencer in the JCVI laboratory, USA.

3.6 Sample preparation for Illumina sequencing

Libraries were constructed following the manufacturer's protocol with reagents supplied in Illumina's TruSeq RNA sample preparation kit (#RS-930-2001). Briefly, the poly-A containing mRNAs were purified from total RNA. The mRNA was fragmented and used to generate double-stranded cDNA. The bar code containing adaptors were ligated to the ends of cDNAs to allow multiplex sequencing.

3.6.1 Complementary DNA synthesis

Total RNA (2 ug) was first incubated in a thermocycler for 5 minutes at 65^{o}C in a total volume of 50 µl in a 96-well PCR plate to linearize the RNA. The plate was removed and incubated an additional 5 min at room temperature allowing RNA to bind to the poly-T oligo-attached magnetic beads. Beads were washed by placing the PCR plate on the magnetic stand at room temperature for 5 min and the supernatant was discarded. Bead Washing Buffer (200 µl) was added and returned to the magnetic stand for 5 min. Supernatant was removed and discarded. The plate was removed from the magnetic stand and Elution Buffer (50 µl) added to each well. The plate was incubated at 80^{o}C for 2 min and then placed at room temperature. RNA bound to beads by adding 50 µl of Bead Binding Buffer and incubating for 5 min at room temperature. The beads were washed as previously described.

3.6.1.1 First strand cDNA synthesis

First strand cDNA synthesis was performed by adding elute, prime and fragment Mix (19.5 µl) to each well. The mixture was incubated for 8 min at 94^{o}C. The plate was placed on the magnetic stand at room temperature for 5 min. From the plate, 17 µl of the fragmented and primed RNA were transferred to a new PCR plate. First strand Master Mix and Superscript II mix (8 µl) were

added to each well and gently mixed. Incubation was performed in a thermocycler under the following program, 25°C for 10 min followed by 42°C for 50 min and finally at 70°C for 15 min.

3.6.1.2 Second strand cDNA synthesis

Second strand cDNA synthesis was performed by adding 25 µl of second strand Master Mix to each well. Mixture was incubated at 16°C for 1 hr. Aline PCR Clean beads (90 µl) were added to each well containing 50 µl of ds cDNA. The plate was incubated at room temperature for 15 min and placed on the magnetic stand for 5 min. The supernatant (135 µl) was removed and discarded. Each well was washed by adding of 200 µl of 80% ethanol and incubated at room temperature for 30 s before removing the supernatant. Wash steps were repeated once and the plate was allowed to dry on magnetic stand for 15 min. Re-suspension Buffer (52.5 µl) was added to each well. The plate was returned to the magnetic stand at room temperature for 5 min and 50 µl of supernatant was transferred to a new PCR plate.

3.6.2 Complementary DNA Library construction

3.6.2.1 Repair of overhangs and addition of adaptors

Fragment overhang ends were converted to blunt ends by the addition of the End Repair Mix (40 µl) to each well and incubated at 30°C for 30 min. Aline PCR clean beads (160 µl) was added to each well containing 100 µl of End Repair Mix. The plate was incubated at room temperature for 15 min. Supernatant (127.5 µl) was removed and discarded. Each well was washed with 80% ethanol as previously described. The dried pellet was resuspended in 20 µl of resuspension buffer and 15 µl were transferred to a new PCR plate. The 3' ends of the fragments were adenylated with the addition of A-tailing mix (12.5 µl) to each well and then incubated at 37°C for 30 min. DNA Ligase Mix (2.5 µl) and a single RNA adapter mix (2.5 µl) were added to each well and then incubated at 37°C for 10 min. The ligation reaction was stopped by adding 5 µl of

Stop Ligase Mix. Aline PCR Clean beads (42 µl) was added to each well. The plate was incubated at room temperature for 15 min. Supernatant (79.5 µl) was removed and discarded. Each well was washed with 80% ethanol as previously described. The dried pellet was resuspended in 52.5 µl resuspension buffer and 50 µl were transferred to a new PCR plate. Aline PCR clean beads (50 µl) was added to each well. The plate was incubated at room temperature for 15 min. Supernatant (95 µl) was removed and discarded. Each was washed with 80% ethanol as previously described. The dried pellet was re-suspended in 22.5 µl resuspension buffer and 20 µl were transferred to a new PCR plate.

3.6.2.2 Polymerase chain reaction (PCR) and product purification

Complementary DNA fragments were enriched by adding PCR primer cocktail (5 µl) and 25 µl PCR master mix to each well. Polymerase chain reaction amplification was performed as follows: initial denaturation at 98^{o}C for 30 s followed by 15 cycles of denturation at 98^{o}C for 10 s, annealing at 60^{o}C 30 s and extension at 72^{o}C for 30 s then a final extension at 72^{o}C for 5 min. The amplified cDNA constructs were purified by adding 50 µl of Aline PCR clean beads to each well. The plate was incubated at room temperature for 15 min. Supernatant (95 µl) was removed and discarded. Each well was washed with 80% ethanol as previously described. The dried pellet was re-suspended in 32.5 µl of resuspension buffer, incubated at room temperature for 2 min, and then placed on the magnetic stand for 5 min. Supernatant (30 µl) was transferred to a low binding microcentrifuge tube for storage.

3.6.3 Illumina sequencing

The final constructs of each purified library were evaluated using the BioAnalyzer 2100 automated electrophoresis system, quantified with the Qubit flourometer using the quant-iT HS dsDNA reagent kit (Invitrogen), and diluted according to Illumina's standard sequencing

protocol for sequencing on the HiSeq 2000. All bar-coded libraries from different tissues of the same genotype were pooled together for multiplex sequencing. In total two lanes were used, each with six different bar-coded libraries; i.e. one lane for Mbwazirume and the other for Cachaco. The libraries were diluted to 10 nM, and sequencing done as per optimized conditions at the DNA Core Facility laboratory University of Missouri in the USA.

3.7 Data analysis

A mixed assembly of Cachaco and Mbwazirume 454 reads was carried out using Newbler software v. 2.6 (20110517-1502) to form large contigs (≥500 bp long) that represented the gene transcript sequences and formed the reference transcriptome on which downstream analyses were based. The reference transcriptome was annotated and the expressed genes identified based on information from non-redundant databases. Annotation was done automatically using KEGG and blastx with blast2go (Götz *et al.,* 2008). Pathway analysis was also carried out to determine the average number of transcripts involved in different metabolic pathways. In the course of undertaking this study, sequencing and annotation of a double haploid banana (DH) Pahang genome (AA), hereafter referred to as reference genome (RG) was completed and availed to IITA after signing the material transfer agreement with CIRAD. The reference transcriptome was mapped to the reference genome to determine the extent to which the reference transcriptome could be useful and the percentage coverage of the genes available in the reference genome.

Reads from Illumina sequencing platform were trimmed using CLC genomic workbench 5.1 tools basing on the quality control results generated by galaxy. For gene expression profiling, trimmed reads from the respective genotype tissues under the two treatments were separately mapped to the reference transcriptome and reference genome using RNA-Seq analysis under CLC high throughput sequencing tools. To ensure comparability the following conditions were

used when mapping reads to the references. The minimum length fraction was set at 0.8, minimum similarity fraction at 0.9 and number of hits for a read at 10. For the case of the reference genome, the exon discovery parameters were set as follow; the required relative expression level was set at 0.2, minimum number of reads was 10, minimum length was 50 and the gene expression values were calculated and reported as reads per kilo base of exon model value (RPKM). To demonstrate the difference in response to drought stress in the two genotypes based on differences in gene expression, two experiments were designed using RNA-Seq sample map files. Experiment 1 (CLRD vs MLRD) compared the two genotypes under drought stress and it had two groups where group 1 had Cachaco root dry (CRD) and Cachaco leaf dry (CLD) samples whereas group 2 contained Mbwazirume root dry (MRD) and Mbwazirume lead dry (MLD) samples. Experiment 2 (CLRW vs MLRW) compared the two genotypes under well-watered conditions and also had two groups i.e group 1 had Cachaco root well-watered (CRW) and Cachcaco leaf (CLW) samples whereas group 2 had Mbwazirume root well-watered (MRW) and Mbwazirume leaf well-watered (MLW) samples. Quality control was performed on the two experiments. Due to variations between samples, $\log_{10}$ transformation of expression values was done so that the data follow a Gaussian distribution. An unpaired t-test analysis was carried out on transformed expression values to determine if a significant difference existed in gene expression levels between Cachaco and Mbwazirume tissues under well-watered and dry treatment in respect to the individual genes at 95% confidence level. Results were displayed as volcano plots with genes whose expression was significantly different at $P < 0.05$ presented as red dots. Finally, a comparison between Cachaco and Mbwazirume tissue specific gene expressions was done to identify the novel genes tightly linked to drought stress responses in two genotypes. The expression of genes in tissues under dry treatment was divided by the expression in tissues under well-watered conditions to get the relative expression. This was used to indicate

whether the gene was up-regulated or down-regulated. When the relative expression value was 1.0, it meant that there was no difference in expression for the tissue under dry treatment as compared to that under well-watered conditions. A gene was considered to be up-regulated or down-regulated only when the relative expression value was >1.0 or <1.0 respectively.

Single nucleotide polymorphisms (SNPs) were called for Mbwazirume and Cachaco using the SNP detection tool in CLC genomic workbench 5.1. To increase the validity of the detected SNPs, duplicate reads from trimmed 454 and Illumina reads were removed before mapping onto the references. Mapping was done under high stringency conditions to eliminate false positive SNPs. That is, the mismatch cost was set at 1, insertion/deletion was 3, length fraction was 0.9 and similarity was 0.95. During SNP detection, the following parameters were set, window length was 11, quality of central base was 30, quality of surrounding bases was 25, minimum coverage was 10, variant frequency was 30% and ploidy level was set at 3 since the two genotypes are triploids. The total number of SNPs detected in Mbwazirume and Cachaco were reported. These were further categorized as homozygous or heterozygous SNPs. Only SNPs whose frequency was ≥ 95% were considered homozygous and heterozygous SNPs were assumed to be loci with variant base frequency ranging between 30-70%. The average distance between SNPs in the two genotypes was calculated based on the reference genome.

CHAPTER FOUR

RESULTS

4.1 Effects of drought stress on banana plants

Following the four months of drought stress in screenhouse, significant differences in physiological responses were observed. Plants under dry treatment (pF 2.8-3.1) showed remarkable reduction in vigor and growth rate (Figure 1)

Figure 1: Physiological effects of drought stress on Mbwazirume 'AAA' and Cachaco 'ABB'. Images A and B are well-watered Mbwazirume and Cachaco plants while images C and D are the same genotypes under dry treatment respectively.

At the time of sampling, both Mbwazirume and Cachaco showed reduced plant height and leaf emergence under drought stress. There was reduction in leaf area with Mbwazirume showing leaf folding and loss of green color of leaves. The root mass of Mbwazirume was reduced under drought stress as compared to Cachaco under stress. Mbwazirume showed the least weekly weight gain under dry treatment but under well-watered treatment it had the most weekly weight gain (Figure 2). It was at this point when treatments and genotypes were showing clear differences that samples were taken for total RNA extraction.

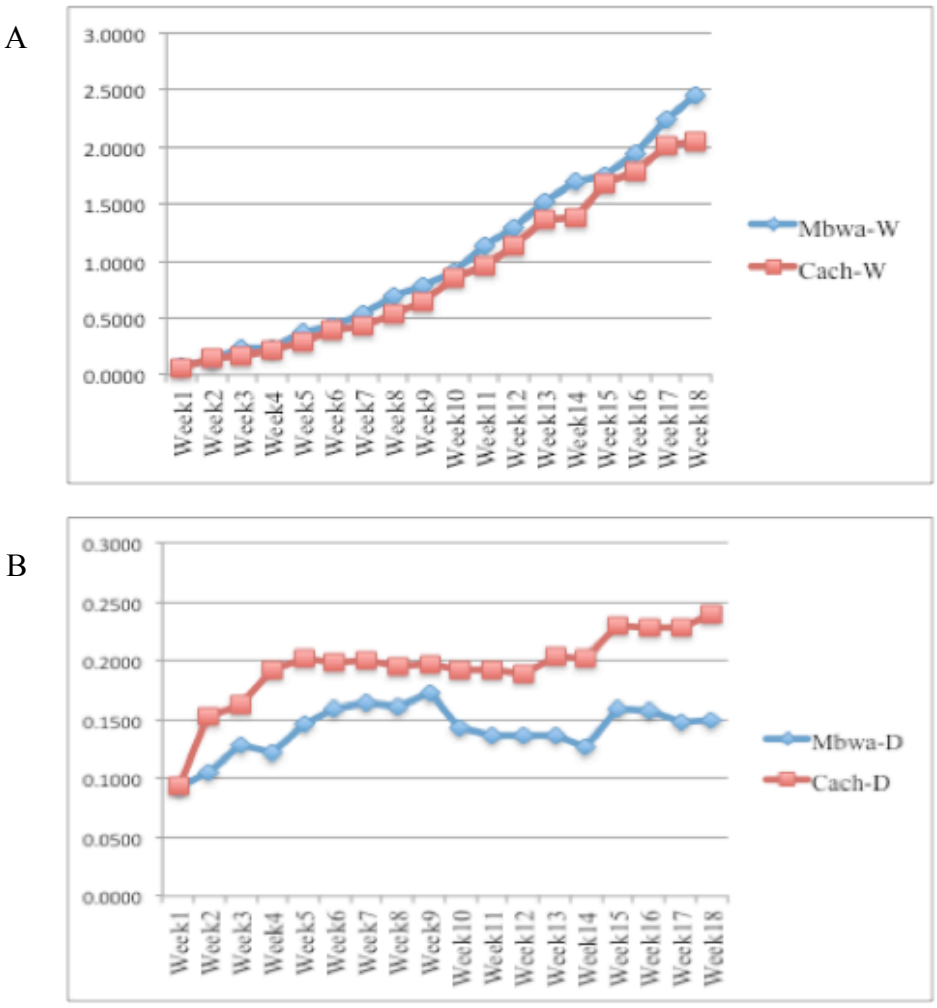

Figure 2: Weekly weight gains for Mbwazirume and Cachaco under well watered and drought stressed conditions. Where A is Mbwa & Cach well watered and B is Mbwa & Cach drought stressed.

4.2 Sequencing and *de novo* assembly of 454 reads

In this study, the 454 GS FLX platform was used to generate the reference transcriptome from cDNA libraries of Mbwazirume 'AAA', and Cachaco 'ABB' drought stressed and control tissues (Figure 1). The rationale of choosing 454 technology was based on its ability to generate longer reads which make *de novo* assembly much easier and as such does not require paired end reads.

Sequencing of the two libraries yielded 656100 (201 Mb) and 571161 (172 Mb) reads with an average read length of 307 bp and 301 bp for Mbwazirume and Cachaco respectively. Mixed assembly of Mbwazirume and Cachaco reads using Newbler v.2.6 (20110517-1502) software yielded >25250 contiguous sequences (contigs) out of which 21201 (84%) were large contigs with lengths ranging between 500-4143 bp. These formed the reference transcriptome onto which downstream analyses were initially based. The remaining 4049 (16%) contigs were too short and were excluded from the reference transcriptome. Although the large contig lengths ranged from 500-4143 bp, the majority of them were between 500 bp and 1000 bp long giving an average contig length of 734 bp based on 3 to 900 reads per contig. The distributions of contigs lengths that constitute the reference transcriptome are summarized in figure 3.

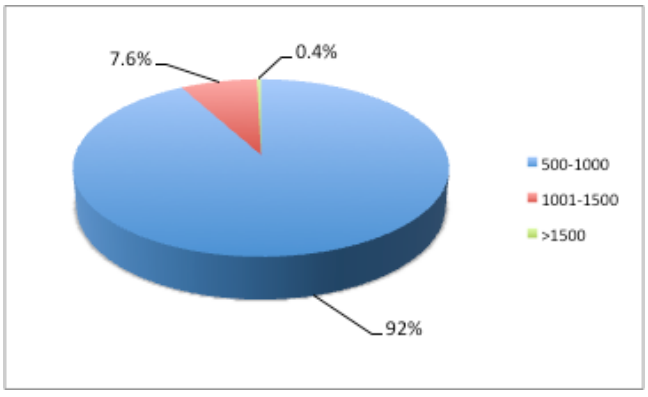

Figure 3: Distribution of large contigs lengths in base pairs that constitute the reference transcriptome generated from sequencing of cDNA libraries of Mbwazirume and Cachaco drought stressed and control tissues using 454 sequencing technology.

4.2.1 Mapping of 454 reads to the reference transcriptome

To identify the number of large contigs that were unique to Mbwazirume and Cachaco, mapping of the respective 454 reads to the reference transcriptome under moderate stringent conditions using CLC genomic workbench 5.1 mapping tool was done and generated the results summarized in Tables 1.

Table 1: Summary statistics for the distribution of Mbwazirume and Cachaco reads after mapping to a reference transcriptome

	Count		Average length		Total bases	
	Mbwa	**Cach**	**Mbwa**	**Cach**	**Mbwa**	**Cach**
Reads	656,100	571,161	307.21	301.53	201,559,538	172,221,640
Matched	**505,462**	**441,407**	**317.44**	**313.96**	**160,451,558**	**138,582,300**
Not matched	150,638	129,754	272.89	259.25	41,107,980	33,639,340
References	21201	21201	734	734	15,577,909	15,577,909

The number of reads that matched to the reference transcriptome (RT) for the two libraries was not different. Mbwazirume had 77% as compared to Cachaco with 77.3% of matched reads. Some of the unmatched reads are likely to be part of the short contigs (<500 bases) that were excluded from the reference transcriptome, or possibly were rare reads not represented in the alignments. When Mbwazirume reads were mapped to the RT, 2016 (9.5%) contigs had no Mbwazirume reads mapped to them suggesting that these contigs were unique to Cachaco. Similarly, when Cachaco reads were subjected to the above analysis 928 (4.4%) contigs had no Cachaco reads mapped to them suggesting that these contigs were unique to Mbwazirume. Therefore, the shared contigs were 18257 (86.1%) in both libraries.

4.2.2 Automatic annotation of the reference transcriptome

To understand the biological functions embedded in sequences comprising the reference transcriptome, further analysis was done. All large contigs were submitted to Kyoto

Encyclopedia of Genes and Genomes (KEGG) Automatic Annotation Server (KAAS). Out of the 21201 contigs submitted only 3798 (17.9%) where annotated. The 17.9% of contigs that were annotated were distributed among 281 protein families involved in metabolism, genetic information processing, environment information processing and cellular processes. The number of large contigs that matched the KEGG biological processes ranged between 2 and 1000 contigs with an average of 2 large contigs for a particular process. Also the number of contigs that matched a particular biological molecule ranged from 1 to 10 with an average of 2 contigs. Representative pathways generated by KEGG are shown in Figure 4.

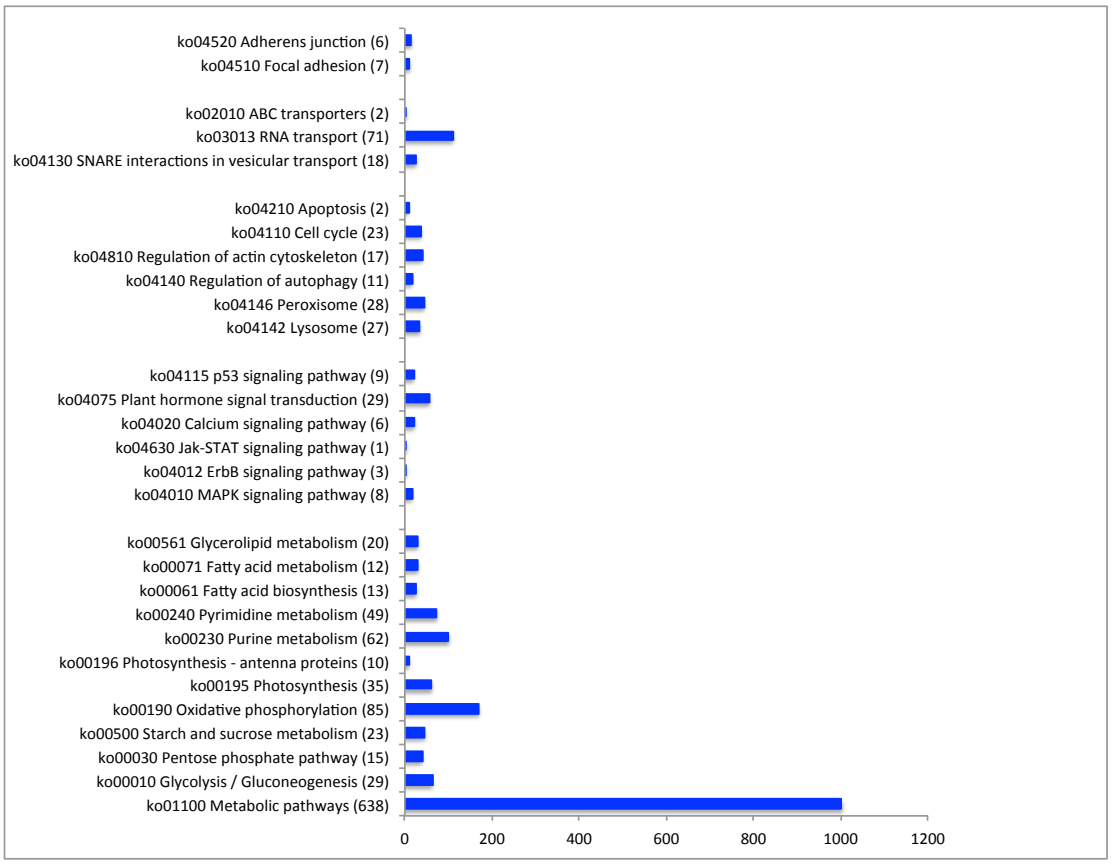

Figure 4: Representative KEGG ontology terms assigned to the large contigs with more emphasis on those involved in stress response. The values in brackets represent the number of enzymes/proteins molecules matching the large contigs.

Further annotation was done by blastx via NCBI to non-redundant database using Blast2GO tools (Götz *et al.,* 2008) with the threshold E-value set at 10^{-6}. Out of the 21201 contigs blasted, 18146 (85.6%) contigs were assigned gene ontology terms that corresponded to different plant species including monocots and dicots. There were more hits on *Oryza sativa* compared to other monocots like *Zea mays* and *Sorghum bicolor* (Figure 5). The remaining 3055 (14.4%) contigs could not be assigned any gene ontology term either by KAAS or Blast2GO because they did not contain an open reading frame (ORF). This was confirmed by translating some of these sequences with Expasy translation tool.

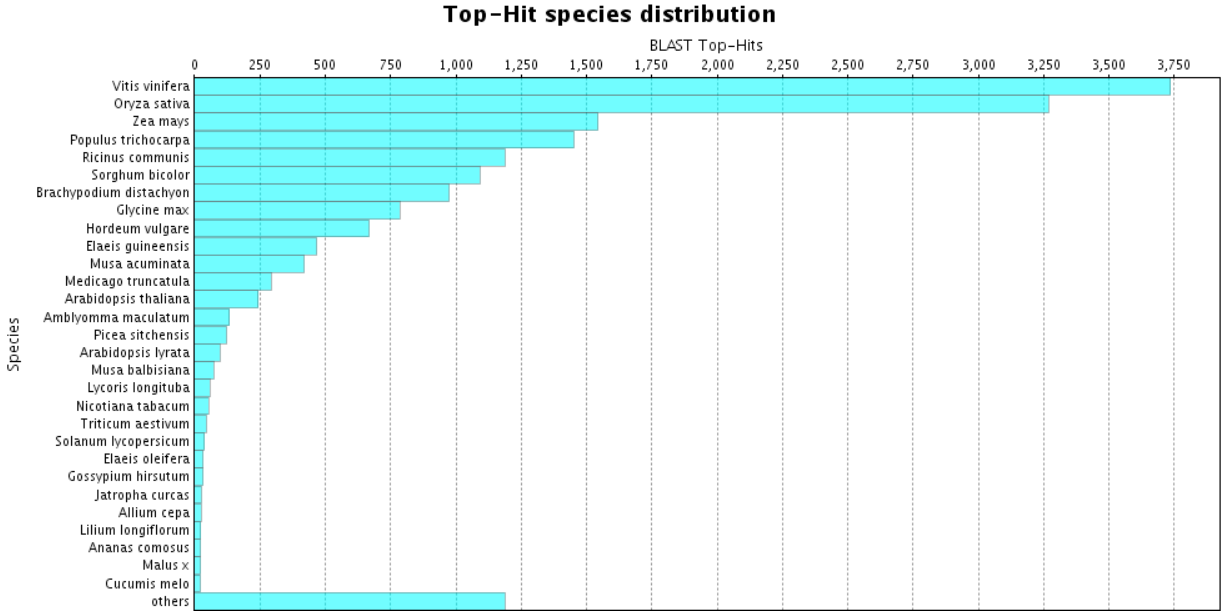

Figure 5: Species distribution of blast hits of 18146 contigs obtained with blast2go

4.3 General comparison of total gene expression in Cachaco and Mbwazirume

The comparison of all genes expressed in the different tissues and a hierarchical clustering of samples based on $\log_{10}$ transformed expression values was done. It was observed that Cachaco leaf well-watered (CLW) clustered together with Cachaco leaf dry (CLD) tissue, while Cachaco root dry (CRD) clustered with Cachaco root well-watered (CRW) tissue (Figure 6)

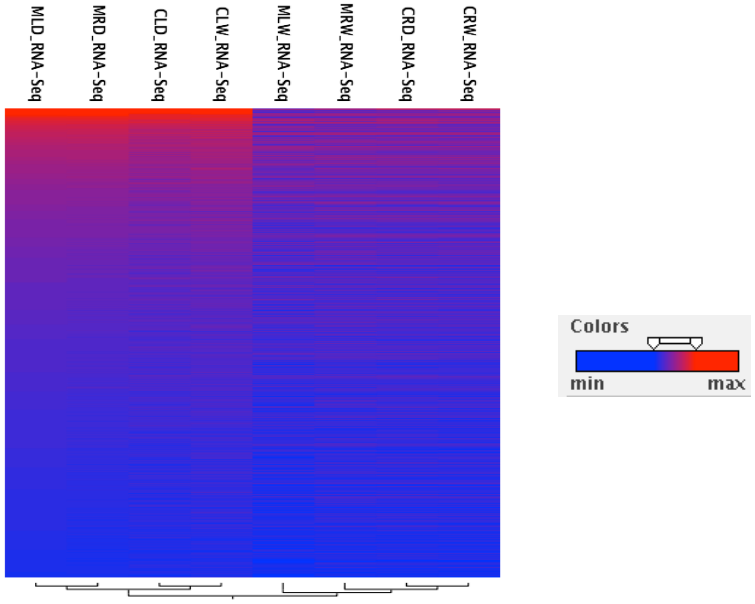

Figure 6: Hierarchical clustering of different samples based on $\log_{10}$ transformation of gene expression values. Red color indicates maximum expression and blue minimum expression.

When unpaired t-test statistic was performed on experiment 1 (CLRD vs MLRD) and experiment 2 (CLRW vs MLRW), there was significant difference in gene expression within and between groups at $P < 0.05$. There were more genes that significantly varied in expression under experiment 1 compared to experiment 2 among the two genotypes as presented by the volcano plots in figure 7 (A and B) below respectively.

A B

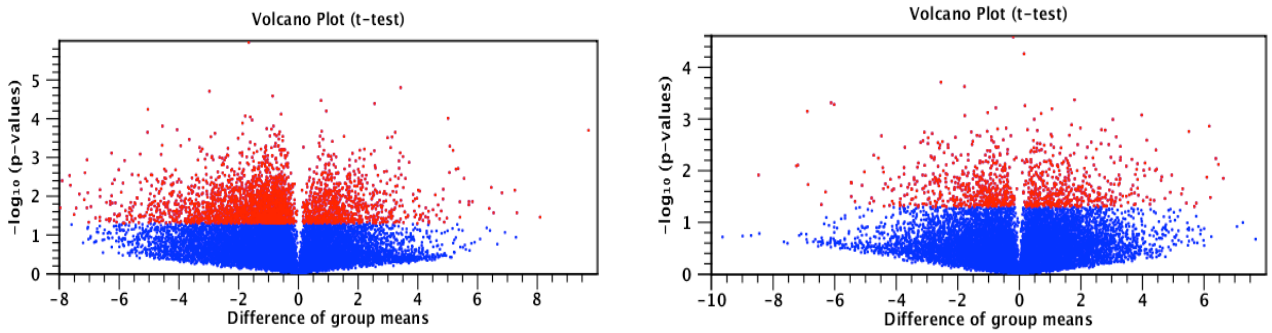

Figure 7: Volcano plots for the t-test analysis on $\log_{10}$ transformed gene expression values showing significant differences between Cachaco and Mbwazirume expressions. Where A is for experiment 1 and B is for experiment 2. The red dots represent genes whose expressions were significantly different within and between genotypes.

4.4 Validation of reference transcriptome using DH Pahang genome

In the course of undertaking this project, sequencing and annotation of DH Pahang genome 'AA' was completed by Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD). Although the genome is not yet available in the public domain, the International Institute of Tropical Agriculture (IITA) signed a material transfer agreement to get a copy of the annotated reference genome. The genome assembly that is about 473Mbp contains 36542 genes. When the reference transcriptome was mapped to the Pahang genome using CLC genomic workbench 5.1, 14225 out of 21201 454-large contigs (67.1%) matched the Pahang genome at moderately stringent conditions. That is, the mismatch cost was set at 2, insertion/deletion cost at 3, length fraction at 0.5 and similarity at 0.5. The matched contigs mapped to 12305 (33.7%) of the putative genes in the Pahang genome. With the availability of a reference genome, it was not necessary for us to improve the reference transcriptome using Illumina reads as planned earlier. To demonstrate the relevance and reliability of the reference transcriptome, a comparison was done on the expression profile of some genes. Many of these genes showed a similar trend in expression profile based on the reference transcriptome and reference genome in respect to treatments, genotypes and tissues. However, some genes showed

tissue-specific expression variation with respect to the reference used (Figure 8). Therefore, the Pahang genome was used in expression analysis since it was more comprehensive than the reference transcriptome.

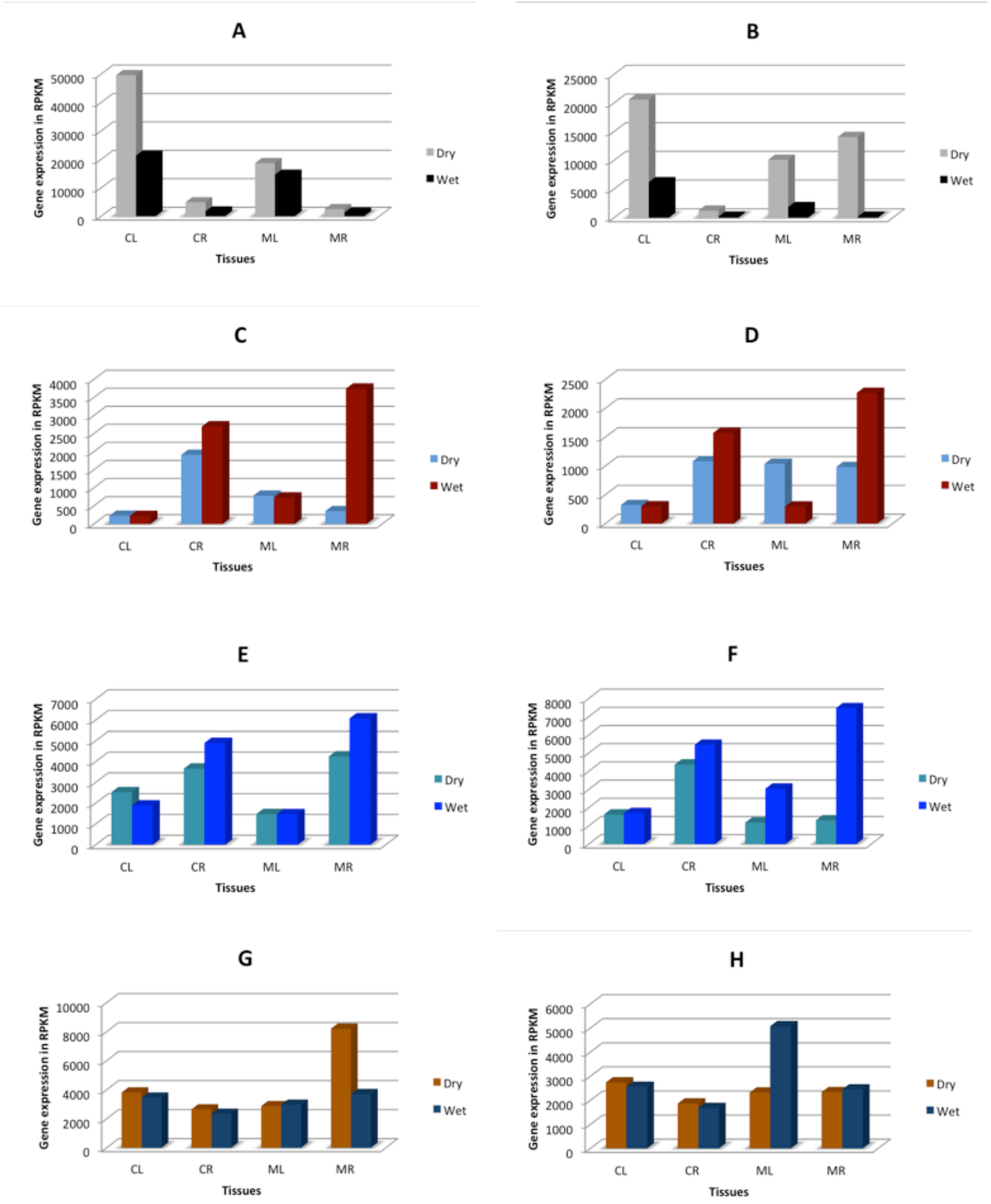

Figure 8: Graphical comparison of gene expression profiles in reads per kilo base of exon model value (RPKM) based on the reference transcriptome (RT) and DH Pahang genome (RG). Graphs A, C, E and G are based on RT whereas B, D, F and H are based on RG. Where A & B are for acidic endochitinase, C & D are for lipoxygenase, E & F are for aquaporin, G & H are for Catalase. CL=Cachaco Leaf, CR=Cachaco Root, ML=Mbwazirume Leaf and MR=Mbwazirume Root. Acidic endochitinase and Lipoxygenase show consistent expression with RT and RG.

4.5 Gene expression patterns in Cachaco and Mbwazirume tissues

To understand the molecular basis for difference in physiological response to drought stress observed between Cachaco 'ABB' and Mbwazirume 'AAA', an expression analysis was done and the genes that have been reported to confer drought stress tolerance in various plant species examined. The genes considered include those coding for transcription factors, metabolic enzymes, antioxidant enzymes, Signal transduction molecules, channel proteins and cell cycle regulating proteins.

4.5.1 Expression patterns of some of the transcription factors

Transcription factors are biological molecules mostly proteins, that bind and interact specifically with certain motifs at the promoter or enhancer regions of a gene. The binding of such factors allows further binding of other transcription factors leading to formation of a complex, which attracts the RNA polymerase to bind the promoter region and initiates transcription. Therefore, various transcription factors are very important in regulation of gene expression. Of utmost importance are those transcription factors whose expression is regulated by environmental stresses such as drought and salinity. The expression profiles of such transcription factors are given in Table 2. The well-watered treatment was used as the reference expression to determine whether the gene was up-regulated (UR) or down-regulated (DR) under drought stress condition.

Table 2: Summary of relative expression patterns of transcription factors in Cachaco and Mbwazirume leaf and root tissues under drought stress

Gene name	Cachaco 'ABB'		Mbwazirume 'AAA'	
	Leaf	Root	Leaf	Root
DREB transcription factors	3.153 (1.303)	2.126 (**0.904**)	0.095 (**1.146**)	0.483 (0.939)
MYB transcription factor, putative	2.042	1.688	1.195	0.965
MYB transcription factor	0.988 (**1.188**)	1.439 (1.450)	1.466 (**0.908**)	2.159 (2.182)
MYC transcription factor	2.354	1.262	0.482	1.314
B3 domain containing protein, putative	1.308	1.087	1.515	0.495
Ethylene responsive transcription factor	1.280	1.570	0.951	0.877
Early responsive to dehydration stress related protein	0.760	1.541	0.709	0.483
bZIP transcription factor domain containing proteins	1.252 (1.010)	1.145 (**0.893**)	0.662 (0.952)	0.712 (**1.056**)
AP2-like ethylene responsive transcription factor	1.017	1.372	0.705	0.413
NAC domain containing protein	1.638 (1.047)	2.835 (2.170)	0.400 (0.871)	0.490 (**1.416**)

Values in parentheses indicate relative expression based on reference transcriptome, where bolded values indicate deviation from the trend observed with the reference genome.

Many MYB transcription factor isoforms were expressed but the one with significant increase under drought stress was MYB44. It increased in both Cachaco and Mbwazirume tissues with Cachaco presenting the highest expression. The putative MYB2 trancription factor was down-regulated in Cachaco and up-regulated in Mbwazirume tissues. Among the MYC transcription factor isoforms, MYC4 showed increased expression in both leaf and root tissues of Cachaco while reduced expression was observed in Mbwazirume leaf and an increase in root tissues (Table 3). This may suggest that MYB44 and MYC4 isoforms have a significant contribution in response to drought stress amongst the MYB and MYC gene families. The up-regulation of MYB2 in Mbwazirume may require synergistic effects of other factors.

Table 3: Summary of the relative expression values for MYB44, MYB2 and MYC4 transcription factors in Mbwazirume and Cachaco leaf and root tissue under drought stress

Gene name	**MYB44, Putative**		**MYB44**		**MYB2**		**MYC4**	
Tissue	Leaf	Root	Leaf	Root	Leaf	Root	Leaf	Root
CDRE	3.748	2.265	1.094	2.233	0.401	0.333	2.780	1.601
MDRE	1.568	1.217	5.113	2.450	5.798	2.575	0.457	1.710

CDRE = Cachaco dry relative expression and MDRE = Mbwazirume dry relative expression

4.5.2 Expression profile of some of the antioxidant enzymes

During drought stress, plants minimize water loss by lowering the rate of transpiration. This is achieved by limiting stomatal conductance, which is effected by stomatal closure due to response of guard cells to abscisic acid. In C_3 plants such as banana, stomatal closure increases photorespiration, which is associated with oxidative stress, increased reactive oxygen species (Noctor *et al.,* 2002). Increased reactive oxygen species damages the photosynthetic apparatus and increases cell deaths. Antioxidant enzymes such as superoxide dismutase, catalase, ascorbate peroxidase, guaiacol peroxidase, glutathione reductase and polyphenol oxidase remove these reactive oxygen species and prevent cell deaths due to oxidative stress (Noctor *et al.,* 2002). The expression profiles of these enzymes were examined in the two banana genotypes (Table 4).

Table 4: Summary of relative expression profile of antioxidant enzymes in Cachaco and Mbwazirume leaf and root tissues under drought stress

Gene name	Cachaco 'ABB'		Mbwazirume 'AAA'	
	Leaf	Root	Leaf	Root
Catalase	1.073 (1.103)	1.110 (1.126)	0.458 (0.967)	0.955 (**2.219**)
L-ascorbate peroxidase	0.838^{++} (0.823)	0.911^{++} (0.947)	0.722 (**1.044**)	0.481 (0.633)
Superoxide dismutase	0.713	1.074	0.597	1.178
Glutathione reductase	0.900^{++}	0.665^{++}	0.533	0.297
Probable phospholipid hydroperoxide glutathione peroxidase	0.967^{++}	1.162	0.915	0.712
Polyphenol oxidase, chloroplastic	1.440	1.529	0.086	0.026

++ - Means that even if there was down-regulation, the expression was higher than that of comparable tissue in the other genotype. Values in parentheses indicate relative expression based on RT

The enzyme catalase had two isoforms expressed that is catalase-2 and catalase-A but the former likely has a significant contribution in antioxidation activity due to its higher expression levels. Although there was a general down-regulation of L-ascorbate peroxidase under drought stress, there were differences in expression of different isoforms. Stromal L-ascorbate peroxidase decreased in the leaf and root of both genotypes but the thylakoid-bound L-ascorbate peroxidase increased in Cachaco leaf. Comparing the expression levels in both genotypes, Cachaco had higher expression levels compared to Mbwazirume even though both showed a decreasing trend. This was also observed in the expression of glutathione reductase. This probably gives Cachaco an additional survival advantage under drought stress. Three isoforms of superoxide dismutase were expressed in different patterns. Superoxide dismutase [Cu-Zn] increased in roots and decreased in leaves for both genotypes. Superoxide dismutase [Fe] was down-regulated in Cachaco leaf but up-regulated in Cachaco root, Mbwazirume leaf and root tissues while superoxide dismutase [Mn] showed a decreasing trend with stress in both genotypes (data presented in appendix III-B). Polyphenol oxidase increased with stress in Cachaco but was reduced in Mbwazirume tissues under stress.

4.5.3 Expression profile of some of the signal transduction molecules

Plants respond to environmental changes through a number of chemical signaling pathways. Drought stress is associated with increase in abscisic acid (ABA). At transcriptome level, an increase in ABA can be examined by looking at the expression levels of enzymes that are involved in its biosynthesis such as 9-*cis* epoxycarotenoid dioxygenase (NCED), zeaxanthin epoxidase (ZEP), and ABA 8'hydroxylase, as well as receptors that bind ABA and allow downstream regulation of ABA-sensitive genes. Other signal transduction molecules include calcineurin B-like proteins (CBL), calmodulin and calcium-dependent protein kinase (CDPK). The relative expression of these signal transduction molecules are summarized below (Table 5).

Table 5: Summary of relative expression profile of some of the signal transduction molecules

Gene name	Cachaco 'ABB'		Mbwazirume 'AAA'	
	Leaf	Root	Leaf	Root
9-cis expoxycarotenoid dioxygenase (NCED)	1.438	3.218	0.378	0.653
Zeaxanthin epoxidase (ZEP)	1.474	1.601	0.886	2.191
Calcineurin B-Like protein (CBL)	1.079	1.108	0.883	0.858
ABA receptors	1.602	1.102	0.559	0.891
Calmodulin	0.794	1.201	0.539	0.427
Calcium-dependent protein kinase (CDPK)	1.096	1.366	0.536	0.481

In this study, an increase in the expression of NCED_(chloroplastic), ZEP_(chloroplastic) and ABA 8' hydroxylase was noted in the leaf and root tissues of Cachaco under drought stress whereas a down-regulation was observed in Mbwazirume tissues except ZEP_(Chloroplastic) which increased in the root tissues. Abscisic acid 8' hydroxylase was much higher in roots while NCED_(chloroplastic) and ZEP_(chloroplastic) were more expressed in leaves. With increased biosynthesis of ABA, there was an increased expression of abscisic acid receptors in leaf and root tissues of Cachaco as compared to Mbwazirume. The isoforms of such receptors include PYR1, PYL2, PYL5, PYL6, PYL7, PYL8 and PYL10 amongst which PYL8 was the most highly expressed.

Calmodulins and calcineurin B-like proteins act as calcium sensors and their expression is induced by multiple stresses including drought stress. Changes in calcium concentration play a very important role in signal transduction that regulates transcription of certain genes. Drought stress increased the expression of CBL proteins especially CBL-1 and CBL-3 in Cachaco tissues whereas a decrease was observed in Mbwazirume tissues. Other isoforms expressed included CBL-4, CBL-6, CBL-7 and CBL-9. However, CBL-9 showed an increased expression in both genotypes but its levels were not as high as those of CBL-1 and CBL-3. The expression of calmodulin increased in Cachaco root only and decrease was observed in Cachaco leaf and in both tissues of Mbwazirume but the expression level in Cachaco leaf was higher than that of Mbwazirume leaf and root under drought stress.

Calcium dependent protein kinases (CDPKs) are serine/threonine kinases that are critical for plant abiotic stress signaling; a number of isoforms occur in different plants. In this study there was a general increase in expression of CDPKs in the leaf and root tissues of Cachaco and a decrease in the tissues of Mbwazirume under drought stress. Sixteen main isoforms of CDPKs were expressed in the two genotypes at varying levels. Those with high expression levels included CDPK-2, CDPK-3, CDPK-13 and CDPK-28. The highest expression was recorded in Cachaco roots for CDPK-3 and CDPK-28 while in Cachaco leaf it was CDPK-13 and CDPK-28 (for details see appendix III-C).

4.5.4 Expression profile of channel proteins

Channel/transporter proteins located within the cell membrane or tonoplast are very crucial in moving materials in and out of the cell. Water is an essential requirement for proper plant growth, which must be taken in by plant cells and this occurs in channels called aquaporins. The expression of various aquaporins was examined and generally there was down-regulation of

aquaporins with stress in both Cachaco and Mbwazirume. The major classes of aquaporins expressed in banana included Plasma Intrinsic Proteins (PIP), Tonoplast Intrinsic Proteins (TIP), Small and basic Intrinsic Proteins (SIP) and the Nodulin Intrinsic Proteins (NIP). Under drought stress, Mbwazirume leaf tissue showed up-regulation of NIP, TIP and SIP. Only PIP was up-regulated in Cachaco leaf tissues. In the root tissues, all classes of aquaporins were down-regulated under stress in both genotypes. When the four classes of aquaporins were compared in both leaf and root tissues under drought stress, Cachaco had a higher expression than Mbwazirume (Figure 8 A & B). It appears like PIP and TIP are the main aquaporin classes in banana plants given their high expression levels in both well-watered and dry treatments.

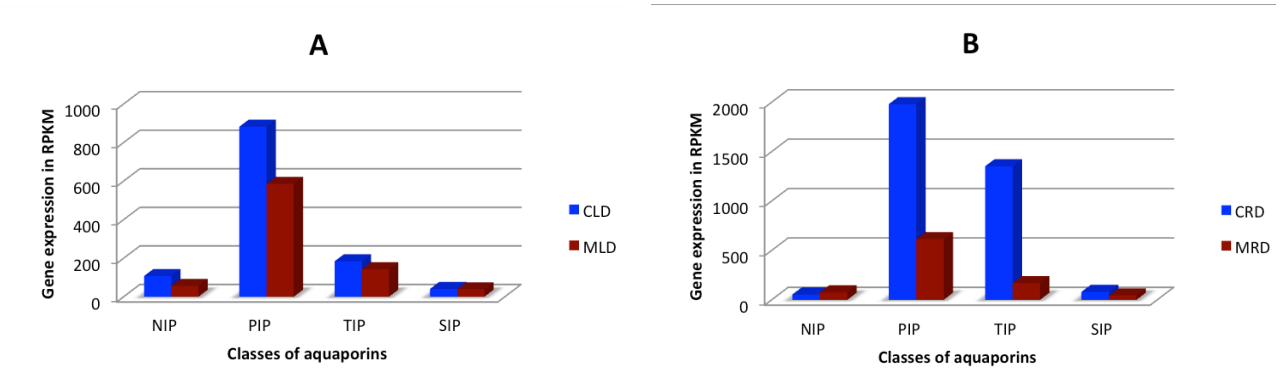

Figure 9: Expression levels of different classes of aquaporins in the leaves and roots of Mbwazirume and Cachaco under drought stress. Where A is expression in leaves and B is expression in roots. In both A and B there was a significant increase in expression levels in Cachaco than Mbwazirume especially for PIP and TIP.

Many other channels/transporters exist within the cell membrane such as potassium channels/transporters, magnesium transporters, sodium channels/transporters and amino acid transporters. Under drought stress, potassium channels/transporters were up-regulated in the both leaf and root tissues of Cachaco. Mbwazirume had these channels up-regulated in roots but down-regulated in leaves. The same trend of expression was noted for magnesium transporters in both genotypes.

4.5.5 Expression profile of some of the cell cycle regulating proteins

In this study the effect of drought stress on the expression of some of the key cell cycle control genes was determined. These include genes encoding the mitogen-activated protein kinases (MAPKs) that are involved in cellular programs such as proliferation, differentiation, movement and death, the cyclin-dependent kinases (CDKs) that in the presence of cyclins drive the cell through the cell cycle, senescence proteins that are produced in response to cellular ageing and cytochrome c which plays a role in apoptosis (programmed cell death). There was an up-regulation of MAPKs in Cachaco root and Mbwazirume leaf under stress while a down-regulation was noted in Cachaco leaf and Mbwazirume root. Nine isoforms of these MAPKs were expressed at different levels within the tissues of the two genotypes. Cyclin-dependent kinase expression increased in both stressed leaf and root tissues of Cachaco but were reduced in Mbwazirume tissues, while programmed cell death protein genes were up-regulated in Cachaco tissues and down-regulated in Mbwazirume tissues. Up-regulation of leaf senescence related protein was observed in stressed Mbwazirume tissues, while in stressed Cachaco tissues down-regulation was recorded. The expression of putative senescence associated protein was increased in stressed Cachaco root but reduced in stressed Cachaco leaf and leaf and root tissues of Mbwazirume. However, the expression of senescence inducible chloroplast stay green protein increased in stressed Cachaco leaf but was significantly reduced in stressed Mbwazirume leaf tissue. In Cachaco leaf tissues no significant change in the expression of enzymes involved in chlorophyll catabolism was observed under stress, but a significant increase was observed in Mbwazirume leaf. Examples of such enzymes include chlorophyllase-2 (chloroplastic) and chlorophyllide-a oxygenase (chloroplastic). Cytochrome c, a protein that is an integral part of the apoptosome was down-regulated in stressed tissues of both genotypes (appendix III).

4.5.6 Expression of other genes that are reported to be up-regulated under drought stress

In several plant studies different genes have been reported to play a role in stress response either by being up-regulated or down-regulated. In this study we examined the expression of such genes (Table 6).

Table 6: Summary of relative expression profiles of some of the genes that have been reported in various plant studies to be up-regulated in drought tolerant plants during drought stress

Gene name	Cachaco 'ABB'		Mbwazirume 'AAA'	
	Leaf	Root	Leaf	Root
Calchone synthase	0.338	1.733	0.172	0.477
Calchone flavonone isomerase, putative	0.528	0.951	2.103	0.925
Tocopherol cyclase chloroplastic, putative	2.122	2.080	0.603	2.529
Multicystatin, putative	0.719	0.855	**105.924**^	**172.666**^
Cysteine protease	1.072	0.994	0.905	0.863
Cysteine protease inhibitor	1.991	2.200	0.516	0.911
Patatin	0.694	1.034	**0.131**^{V}	**0.179**^{V}
Epoxide hydrolase	1.236	1.371	0.972	0.972
Sucrose synthase	0.332	0.520	0.318	0.161
Glyceraldehyde-3-phosphate dehydrogenase	0782	0.936	1.664	1.004
Phosphoenolpyruvate carboxylase	0.690	0.515	0.572	0.120
Acidic endochitinase	3.308	**6.257**^	5.426	**71.302**^
Fructose bisphosphate aldolase, cytoplasmic	1.397	0.807	2.944	1.153
Betaine aldehyde dehydrogenase, chloroplastic	0.971	1.072	0.240	0.384
Lipoxygenase	1.076	0.687	3.500	0.433
Dehydrin	**7.185**^	1.250	**0.004**^{V}	**0.010**^{V}
Osmotin-like protein	5.076	1.053	**0.029**^{V}	**0.024**^{V}
ZF-HD homeobox protein	1.554	0.935	**12.909**^	3.876
Putative Histone H1	2.142	1.766	0.647	4.008
Putative Transducin beta-like protein	1.467	1.035	0.650	0.922

Where ^ - there was a huge increase in expression under drought stress, ^{V} - there was a huge decrease in expression under drought stress

4.6 SNP detection

Single nucleotide polymorphisms (SNPs) are common variations found in plant genomes. These occur by single base substitution and these can be used as markers to differentiate even closely related organism. To detect SNPs, trimmed Cachaco and Mbwazirume 454 and Illumina reads without duplicates were separately mapped to the reference transcriptome and reference genome under high stringent conditions. Based on the reference transcriptome, 100,053 SNPs were detected in Mbwazirume against 179,788 SNPs in Cachaco. However, based on reference genome 372,284 SNPs were detected in Mbwazirume against 949,761 SNPs in Cachaco. The trend in the number of SNPs detected was consistent between reference genome and reference transcriptome. Further analysis on detected SNPs was based on the reference genome. To classify the SNPs, some assumption were made; SNP variants with frequencies ≥ 95% were confidently reported as homozygous while those with frequencies approximately 30-70% were reported as heterozygous provided at least two alleles occurred at a locus, while SNPs with frequencies of 70.1-94.9 with one allele called at the respective loci could not be confidently classified as either homozygous or heterozygous. This is because the algorithm set for SNP calling has several filters. One of them is minimum SNP variant frequency, if an allele has a frequency below the set threshold even though it is a SNP, it will not be called. For example, if the reference has a ‘T’ and the variants are ‘G’, ‘C’, and ‘T’ with frequencies of 70%, 20%, and 10% respectively, given that the minimum variant frequency is 30%, only a ‘G’ will be called and reported as a SNP. Alleles ‘C’ and ‘T’ are not reported because their frequencies are below 30%. The second filter is on the quality of central base and surrounding bases. For a base to be considered as a SNP variant and be included in the calculation of allele frequency it must satisfy this inclusion criterion or else it is not reported. Lowering the stringency of these parameters results into calling many SNPs and the number of false positive SNPs also increases. Based on the above assumptions, the

total number of SNPs detected in each genotype, the number of SNPs under the respective categories and the distance between SNPs are summarized (Table 7).

Table 7: Summary of SNPs detected and the average distance between them in Mbwazirume and Cachaco

Genotype	Homozygous (≥ 95%)	Heterozygous (30-70%)	Unclassified (70.1-94.9%)	Total	Av. Distance btn SNPs
Mbwazirume	168,125	135,150	69,009	**372,284**	945
Cachaco	271,788	398,289	278,684	**949,761**	371

CHAPTER FIVE

DISCUSSION

5.1 The reference transcriptome

The ability of some banana varieties to tolerate drought stress has been linked to the presence of the B genome (Stover and Simmonds, 1987; Nelson *et al*., 2006) based on the observed differences in physiological responses. No detailed study has been conducted to examine differences in gene expression as well as genome variations that could possibly account for these differences. Limited knowledge in this area has been partly caused by lack of banana reference genome. For an organism whose reference genome sequence is not yet publically available or has never been sequenced, next generation sequencing and *de novo* assembly of gene transcripts offer greater opportunities in studying the functional part of the genome. The high throughput and relatively low cost of sequencing with next generation sequencing (NGS) platforms have made significant revolution in transcriptomics research. The choice of platform to use is dependent on the objectives of the study. In this current study a reference transcriptome consisting of 21201 large contigs (≥500 bp) was generated by *de novo* assembly of reads from next generation sequencing platform (454 GS FLX platform) while waiting for the availability of banana reference genome. Contigs below 500 bp were excluded from the reference transcriptome. In *Musa acuminata,* sequencing of two BAC clones revealed the average exon length as 234 bp with the average maximum exon length as 1682 bp and an average number of exons as 6.3 (Aert *et al.,* 2004). Therefore, it is likely that many gene transcripts in banana like other eukaryotic organisms on average have a length of approximately 1.0 kb. The algorithm for calculating gene expression report expression levels in reads per kilo base of exon model (RPKM) (Mortazavi *et al*, 2008). The aim was to generate insights about the best explanation for the differences in

physiological responses to drought stress observed between Cachaco 'ABB' and Mbwazirume 'EAHB-AAA' at transcript level.

In the course of undertaking this study the CIRAD team with support from the Global *Musa* Genomic Consortium (GMGC) was able to finish the sequencing and annotation of the DH Pahang genome 'AA' banana that was availed to me even though it was not yet available in the public domain. This gave me an opportunity to validate the reference transcriptome that was a mixed assembly of Cachaco and Mbwazirume reads and gain a better understanding on the expression profiles of genes in the banana genome. When the reference transcriptome (21201 contigs) was mapped to the reference genome, only 67.1% of the contigs mapped to the reference genome representing 33.7% of the total genes present in the DH Pahang genome. The possible explanation for this discrepancy is that the reference transcriptome was generated from cDNA which contains only the coding regions of the genome while the reference genome contains both exons and introns. Also the reference transcriptome contains both A- and B-genomes yet the reference genome is made up of only A-genome. These factors in addition to the mapping algorithms used lower the possibility of the contigs getting hit targets to the reference genome.

The low percentage of gene representation in the reference transcriptome is attributed to the fact that in living organisms genes are divided into two categories; constitutive genes, which are always expressed to maintain life and the inducible genes, which are only expressed under a given set of conditions. Expression of some genes is tissue-specific or stage-specific, but also among those expressed, some are expressed at a low level thus reducing their probability of being represented in the reference transcriptome. In the genome, some genes are silenced or may have become non-functional (pseudogenes), all these reduce the number of genes present in any given cDNA library. The factors that induce the expression of some genes could be internal or external

such as environmental stresses. Drought stress is an abiotic factor that evokes the expression of many genes involved in different pathways as the plant tries to adapt to the stress. Some of these genes are constitutively expressed but under drought stress, they are either up-regulated or down-regulated. When blastx to non-redundant databases via NCBI using blast2go (Götz *et al.,* 2008) was done the genes in the reference transcriptome were identified. Blastx allows the translation of nucleotide sequences into amino acid sequences, which are then used as queries to search non-redundant protein databases such as SWISSPROT and InterPro. Base code degeneracy is a very big issue when the nucleotide sequences are used as queries because different organisms have preferential code words they use to code for different amino acids and this lowers the percentage similarity between nucleotide sequences. Therefore, when annotating *de novo* assembled sequences for a species whose reference genome is not yet available, a better understanding of the sequences is gained by aligning the query sequences to sequences that are fully annotated, and with blastx, chances of finding hits are increased.

Among the monocot plants, banana gene sequences are much closer to those of *Oryza sativa* because the highest number of hits was observed here then followed by *Zea mays* as shown in Figure 4. These results supplement the findings of Lescot *et al.,* (2008) and Arango *et al*., (2011) which indicated that microsynteny exists between *Musa acuminata* genome and *Oryza sativa* genome. Not only microsynteny exists between *Musa acuminata* but also there is high sequence similarity. The reference transcriptome was a mixed assembly of reads from the A and B genomes and this may suggest that not only *Musa acuminata* shows this microsynteny but also *Musa balbisiana* (B genome).

The second reason why the reference transcriptome had low percentage gene representation is that many genes in the banana genome are duplicated as observed in the DH Pahang genome and

this is further increased by the polyploid nature of the plant. The expression levels of the different gene copies in bananas are not the same. Transcription factors and RNA polymerase easily access some copies than others and thus the expression of such copies is always higher. Accessibility of DNA by transcription factors and RNA polymerase depends on the extent of histone acetylation at the promoter regions. When the activity of histone acetyltransferase is high at the promoter region, the rate of transcription of the gene is increased (Choi *et al.,* 2004). Also there are some chances of losing low-abundance gene transcripts in the total mRNA during library construction. In this study, the switching mechanism at the 5' end of the RNA transcript (SMART) approach was used (Zhu *et al*., 2001). One limitation of this technique is that once premature termination of cDNA synthesis occurs, the incorporation of *Sfi*IA restriction site does not occur. Prematurely terminated cDNAs are eliminated during cloning step and such transcripts cannot be represented in the cDNA library. The difference in gene copy-number observed between the reference transcriptome and reference genome possibly accounts for some differences that were observed between the expression patterns compared basing on reference transcriptome and reference genome. However, the reference transcriptome still remains valid and useful in the absence of a reference genome but in the presence of a reference genome a comprehensive analysis of gene expression can be achieved.

The relationship between Cachaco 'ABB' and Mbwazirume 'AAA' is that both are edible triploid varieties of banana that belong to the genus *Musa*. Edible bananas are believed to have arisen from intra- and inter-specific hybridization among *Musa acuminata* 'AA' subspecies and *Musa balbisiana* (BB) (De Langhe *et al.,* 2010, Heuzé and Tran, 2011). A certain degree of homology in gene sequences is expected despite different evolutionary backgrounds. Cachaco contains one copy of the 'A' genome whose genes could be homologous to the 'A' genome in Mbwazirume. The evolution of edible 'AAB' and 'ABB' bananas are believed to have passed

through an intermediary stage of edible 'AB'. Backcrossing of the hybrid 'AB' to edible 'AA' and 'BB' could have resulted into the present day edible 'AAB' and 'ABB' (De Langhe *et al.,* 2010). It is also likely that backcrossing of hybrid 'AB' and edible 'AA' could have resulted into triploid 'AAA' and 'AAB with chromosome segments transferred between the B and the A genomes. The high percentage of shared contigs between the two cultivars shows that random segregation of the A and B genome chromosomes during meiosis most likely generated gametes carrying a variable proportion of recombinant A^b and B^a genome chromosomes (the superscripts indicate B genome alleles in an A genome background, and vice versa). Therefore, it is likely that Mbwazirume exists as AAA while Cachaco as AB^aB. The presence of 'A' and possibility of 'B^a' could probably explain the high percentage of shared contigs observed and edible quality of this 'ABB' banana. It is also likely that *acuminata* and *balbisiana* are similar enough that homologs form common contigs. Although drought tolerance has been associated with the B-genome (Stover and Simmonds, 1987), the underlying mechanisms that give banana plants with B genome a survival advantage and to remain productive under drought stress have not been fully described at gene expression level. The reference transcriptome and the reference genome helped in understanding the expression profiles of genes especially those that are involved in drought stress response.

5.2 Effect of drought stress on gene expression

Drought stress tolerance is a quantitative trait that is controlled by many genes located at different loci thus referred to as quantitative trait loci (QTL). When banana plants are subjected to drought stress, their ability to survive and remain productive depends on the expression levels of genes and synergistic interaction of gene products that neutralize the detrimental effects of drought stress. Response to drought stress is a multifaceted process that requires many

interactions between different cellular pathways to make the plant survive and remain productive. In this study the expression profiles of transcription factors, metabolic enzymes, antioxidant enzymes, signal transduction molecules, channel proteins and cell cycle regulating proteins were examined. The findings indicated that majority of these genes were up-regulated in Cachaco and down-regulated in Mbwazirume under drought stress. In situations where down-regulation of certain genes was noted in both genotypes, the expression levels in Cachaco were significantly higher than those of Mbwazirume. This could probably explain why there are differences in physiological responses to drought stress between the two genotypes. Although differences in tissue expression profiles were observed, the trend of expression was the same in most cases that is when there was up-regulation in roots there was also up-regulation in leaves.

5.2.1 Transcription factors

Food security is one of the major concerns of the rapid growing world population. In 2010 the hungry population of the world was estimated to be 925 million people with majority (239 millions) living in Sub-Saharan Africa (FAO, 2010b). Many factors lead to food insecurity among which drought stress is one of them. Breeding for drought tolerant crop varieties has been proposed as the most sustainable solution to address the decreasing food production (CGIAR, 2003). Due to the complex nature of the trait, partial sterility of cultivated bananas and their polyploid nature, improvement via conventional breeding is difficult. In both Mbwazirume and Cachaco, the majority of the genes that have been reported to confer drought tolerance in various crops *Oryza sativa* (Moumeni *et al*., 2011), Arabidopsis (Cheong *et al*., 2003; Nakashima *et al*., 2009), Alfalfa (Luo *et al*., 1992), and woody plants (Liu *et al*., 2011) are present. What differs are the expression levels of these genes under drought stress conditions, and possibly allele-specific effects. Transcription factors are very important in regulation of gene expression. Transforming drought susceptible plants with transcription factors that are more efficient in

regulating downstream genes that confer tolerance is considered a good strategy in molecular breeding (Pardo, 2010, Hardy, 2010). Of utmost importance are those transcription factors whose expression is regulated by drought stress. Several transcription factors are induced by drought stress. In *Oryza sativa* and *Arabidopsis thaliana* the *cis*-acting elements that have been identified include MYB2, MYC2, DREB2, ABRE, NAC, bZIP and AP2/ERF all of which are transcription factors induced by abscisic acid, drought stress or high salinity. Up-regulation of these genes has been observed in drought tolerant varieties (Shinozaki and Yamaguchi-Shinozaki, 2007; Hardy, 2010). In the current study it was MYB44 and MYC4 which were up-regulated in the drought tolerant Cachaco whereas MYB2 was up-regulated in the drought sensitive Mbwazirume and down-regulated in Cachaco. The former two genes occur in multiple copies within the banana genome indicating possible requirement for more tissue- or stage-specific regulation of their products. This may suggest that banana's MYB44 and MYC4 genes could be considered in plant transformation as a way of improving drought tolerance by regulating gene expression. This could be further investigated using paralog- and allele-specific differences in expression to identify which paralogs would be good candidates for transformation.

5.2.2 Antioxidants

During drought stress, plants minimize water loss by lowering the rate of transpiration. This is achieved by limiting stomatal conductance, which is effected by stomatal closure. In C_3 plants such as banana, stomatal closure increases photorespiration, which is associated with oxidative stress due to increased reactive oxygen species (ROS) (Noctor *et al.*, 2002). Increased ROS damages the photosynthetic apparatus and increases cell deaths. Plants respond to oxidative burst by increasing the expression of both enzymatic and non-enzymatic antioxidants that help to remove ROS (Abedi and Pakniyat, 2010). In Cachaco there was significant increase in catalase

and polyphenol oxidase. Conversely, the expression of glutathione reductase, superoxide dismutase, L-ascorbate peroxidase and probable glutathione peroxidase in the leaves of Cachaco showed a decrease but this was not significant as that observed in Mbwazirume. The high level of superoxide dismutase ensures efficient conversion of ROS to oxygen, hydrogen peroxide and hydroxyl ions. Hydrogen peroxide produced is itself detrimental to cells therefore it has to be removed quickly as soon as it is produced by enzymes that convert it to water and oxygen (Manda *et al*., 2009). Increased expression of catalase suggests that Cachaco is more efficient in dealing with reactive hydrogen peroxide thus preventing cellular deaths as compared to Mbwazirume. Although there was down-regulation of L-ascorbate peroxidase and probable phospholipid hydroperoxide glutathione peroxidase in both genotypes, the levels in Cachaco were higher than those in Mbwazirume. These contribute to the efficiency of removing hydrogen peroxide in Cachaco under drought stress. The role of glutathione reductase is to increase the levels of reduced glutathione that has multiple functions in antioxidant defense. Reduced glutathione directly scavenges the ROS and it is a co-substrate for peroxide detoxification (Manda *et al.,* 2009). In the context of the results, Mn-SOD decreased in tissues of both genotypes under drought stress yet catalase increased in Cachaco tissues and decreased in Mbwazirume tissues.

5.2.3 Signal transduction during drought stress

Bananas like any other plants respond to environmental changes through a number of chemical signaling pathways. Drought stress is associated with increase in abscisic acid (ABA). At the transcriptome level, an increase in ABA can be examined by determining the expression levels of genes for enzymes that are involved in its biosynthesis such as 9-*cis* epoxycarotenoid dioxygenase (NCED) (Moumeni *et al*., 2011), zeaxanthin epoxidase (ZEP), and ABA 8'hydroxylase (Xiong and Zhu, 2003) as well as receptors that bind ABA and allow downstream

regulation of genes that are ABA sensitive. Increased expression of the above genes suggested an increased production of ABA, which has multiple effects. Abscisic acid is also implicated in stomatal closure during water deficit that occurs during drought stress. Abscisic acid increase in leaves is one of the factors that induce photorespiration, which in turn triggers increased production of ROS (Noctor *et al*., 2002).

Increased ABA concentration would enhance the activity of some transcription factors such as the abscisic acid responsive element binding proteins (AREB), MYC, MYB, AP2 and bZIP that bind to promoters of other genes and increase their expression. Increase in ABA receptors increases the efficiency of signal transduction into the nucleus thus increasing the expression of downstream genes. In this study abscisic acid receptor PYL8 was up-regulated compared to other isoforms in Cachaco tissues but was significantly down-regulated in Mbwazirume. A study in *Arabidopsis thaliana* by Saavedra *et al.,* (2010) showed that overexpression of PYL8/RCAR3 increased tolerance to water stress in vegetative tissues, suggesting an important role in signaling and ABA-regulated genes involved in stress response. Increased levels of ABA and the high expression of PYL8 receptors may increase the ability of Cachaco to tolerate drought stress. Reduced expression of abscisic acid biosynthetic enzymes and its receptors in Mbwazirume may delay or diminish the relay of signals that alert the plant to prepare for and counteract the effects of drought stress.

In addition to ABA, changes in calcium concentration in cell cytoplasm play a significant role in signaling pathways. These changes occur under conditions such as drought stress. Proteins that act as calcium sensors and initiate signal transduction include calcineurin B-like proteins (CBL), calmodulin and calcium-dependent protein kinase (CDPK). The expression of these proteins is induced by multiple stresses. The increased expression of CBL-1, CBL-3 and calmodulin in

Cachaco suggests that these proteins may be involved in signal transduction during drought stress and supports the observation made by Cheong *et al.,* (2003). Calcium-dependent protein kinases are serine/threonine protein kinases that act by phosphorylation of other proteins such as phytohormone. Calcium-dependent protein kinases have been implicated to play a role in mediating signaling elicited by the phytohormones such as auxins, ABA, cytokinin and gibberellin (Botella *et al*., 1996; Urao *et al*., 1994; Mitra and Johri, 2000). Calcium_dependent protein kinases may play a role is regulating plant growth under drought stress. Out of the 16 isoforms of CDPKs identified in banana genome, CDPK2, CDPK3, CDPK13 and CDPK28 were more responsive to stress, perhaps implying a significant role in signal transduction in both genotypes.

5.2.4 Channel proteins

Channel/transporter proteins are located within the cell membrane or tonoplast. These are very crucial in moving materials in and out of the cell. Homeostatic control of water balance in plant cells and the uptake/release of inorganic and organic molecules/ions depend on the regulatory activities of these channel proteins. Water is an essential requirement for proper plant growth and must be taken in by plant cells and this occurs in channels called aquaporins (Hamanishi and Campbell, 2011). The expression of various aquaporins was examined and generally there was down-regulation of aquaporins in stressed root tissues of both Cachaco and Mbwazirume, but the overall levels in Cachaco were higher than those in Mbwazirume. The major classes of aquaporins expressed in banana included Plasma Intrinsic Proteins (PIP), Tonoplast Intrinsic Proteins (TIP), Small and basic Intrinsic Proteins (SIP) and the Nodulin Intrinsic Proteins (NIP). Channel proteins are affected by drought stress through both transcriptional and post-transcriptional changes, including conformational change that limits their ability to transport

specific molecules. When the expression of aquaporins is reduced in roots coupled with conformational changes in their proteins, plant water balance is severely affected and can affect both growth and wilting of the entire plant. When this happens in the leaf cells, the plant shows increased leaf folding due to low turgor pressure and eventual rapid senescence of leaves. For tolerant plants, the effects manifest as reduced growth rate as was observed in Cachaco. A study in *Eucalyptus* showed that PIPs are essential for normal growth and any reduction in PIPs resulted into suppressed growth (Tsuchihira *et al.,* 2010). The results suggest that both PIP and TIP aqauporins are essential in normal banana plant growth because their expression in roots of drought tolerant Cachaco was not significantly reduced as compared to the drought sensitive Mbwazirume. Water as a biological solvent has a great impact on the physiology of the plant when its supply is reduced. Aquaporin activities display a rate-limiting step in biological processes such as photosynthesis, growth and development in plant species. Also cellular dehydration has a great impact on the activity of other channel/transporter through conformational changes. Significant reduction in the expression of aquaporins may be one of the contributory factors for poor adaptation to drought stress in Mbwazirume.

5.2.5 Cell cycle regulating proteins

Plant cells are like any other eukaryotic cells; they are born, they reproduce and die in a programmed manner. These events constitute what is termed the cell cycle. The cell cycle is divided into gap 1 (G_1), gap 2 (G_2), gap 0 (G_0), synthesis phase (S-phase) and mitotic phase (M-phase). The S-phase separates the G_1 and G_2 while G_2 is a transition stage to M-phase. Several proteins control the progression of the cell through the cycle and these include the cyclins and cyclin-dependent kinases, CDKs (Murray *et al.*, 2003). When conditions are not favorable for the cell to progress into G_1 after mitosis, it becomes quiescent (G_0) until favorable conditions are

available. Drought stress had an effect on cell cycles regulating gene expression. There was an increase in the expression of MAPKs in the root tissue of Cachaco and leaf tissue of Mbwazirume and a reverse was observed in Cachaco leaf and Mbwazirume root. Mitogen-activated protein kinases (MAPKs) regulate cellular programs such as proliferation, differentiation, movement and death (Manda *et al.,* 2009). Plant roots play a significant role in the survival of the plant (Yu *et al.,* 2008). It is the site where water and nutrients are absorbed from the soil to the plant. Increased expression of MAPKs in the roots of Cachaco suggested their ability to undergo active cell cycle. In association with this was the increased expression of CDKs and programmed cell death proteins genes in the tissues of Cachaco. The reverse was observed in Mbwazirume. The reduced root mass of Mbwazirume under conditions of drought stress and the inability of the root cells to undergo normal cellular cycles with accelerated cell deaths may point to some of the reasons why this genotype is drought sensitive.

In addition to MAPKs and CDKs, leaf senescence-related protein (LSRP) and senescence-inducible chloroplast stay green proteins (SIC-SGP) are involved in the regulation of leaf cell cycle. In Cachaco leaf tissues, there was a down-regulation of LSRP with an up-regulation of SIC-SGP and the reverse was observed in Mbwazirume. This would tend to delay senescence in Cachaco but accelerate it in Mbwazirume. Senescence is the final stage of plant development during which plants reclaim the valuable cellular building blocks that have been deposited in leaves and other parts of the plant during growth (Buchanan-Wollaston, 2007). The results suggest that drought stress increases leaf senescence in drought sensitive plants by inducing up-regulation of LSRP genes whereas in the drought tolerant banana it is the SIC-SGP genes, which are up-regulated. Senescence-inducible chloroplast stay green proteins are involved in the regulation of chlorophyll degradation by inducing light-harvesting chlorophyll binding protein II (LHCPII) disassembly through direct interaction, leading to the degradation of chlorophyll and

chlorophyll-free LHCPII by catabolic enzymes and proteases, respectively (Park *et al.,* 2007). The visual marker for leaf senescence is yellowing due to loss of chlorophyll and eventual drying of leaves. Chlorophyll is a key component in the photosynthetic apparatus. Drought stress destroys the photosynthetic apparatus in sensitive plants (Yordanov *et al*., 2003). Loss of green color in leaves is a sign of chlorophyll breakdown. Enzymes such as chlorophyllase-2 chloroplastic and chlorophyllide-a oxygenase chloroplastic are involved in chlorophyll catabolism. Chlorophyllases are involved in the first step of chlorophyll catabolism during leaf senescence (Park *et al.,* 2007). These enzymes genes were up-regulated in Mbwazirume and down-regulated in Cachaco leaf tissues. This possibly explains why Cachaco retained more green leaves than Mbwazirume under drought stress.

Cytochrome c is an evolutionary stable biological molecule that is at the heart of aerobic life. It shuttles electrons in the narrow spaces between the two-mitochondrial membranes so that the last step of aerobic respiration takes place. However, cytochrome c is also involved in apoptosis (Goodsell, 2004). When the expression of cytochrome c was examined, down-regulation was observed in tissues of both genotypes under stress. Apoptosis is referred to as programmed cell death which is regulated by proteins. Cytochrome c and mitochondria play an important role in initiating apoptosis in response to signals such as DNA damage that likely occurs during drought stress. Cytochrome c acts by associating with apoptotic protease activating factor 1 (Apaf-1) to form the seven-armed apoptosome a complex that finally activates the protein-cutting caspases that begin the heavy work of cell deaths (Goodsell, 2004). Reduced expression of cytochrome c may be associated with delayed apoptosis. In animals delayed apoptosis is associated with cancer cells but in plants it is a survival strategy during unfavorable conditions. The efficiency of this strategy is dependent on the effectiveness of other strategies.

5.3 Detection of SNPs

Single nucleotide polymorphisms (SNPs) are common variations found in plant genomes (Grattapaglia *et al.,* 2011). These occur by single base substitution (Jehan and Lakhanpaul, 2006) and these can be used as markers to differentiate even closely related organism. Based on the DH Pahang genome 'AA', there was a 2.6 fold difference in detected SNPs between Mbwazirume 'AAA' and Cachaco 'ABB' under highly stringent conditions. The high number of SNPs detected in Cachaco may be explained by the presence of two copies of B genome in this genotype whereas the lower number of SNPs detected in Mbwazirume may be explained by the high degree of sequence homology with the DH Pahang genome. The high number of heterozygous SNPs observed in both Cachaco and Mbwazirume could be attributed to the polyploid nature of the plant, and the fact that even Mbwazirume 'AAA' is derived from different subspecies of *M. acuminata*. Differences in gene expression, possibly affected by epigenetic differences, could lead to detection/not of heterozygosity. Single nucleotide polymorphisms are categorized as silent SNPs or functionally active SNPs. Silent SNPs do not affect amino acids when they occur within the coding region of genome, whereas functionally active SNPs result in amino acid change thus potentially affecting the functionality and activity of the gene. Although many SNPs in protein-coding regions are silent, some result into amino acid substitution, which manifest by phenotype change (Jehan and Lakhanpaul, 2006).

The number and validity of SNPs detected depends on the SNP calling tool used and these vary based on the algorithms used. There are many tools that can be used for SNP detection, among which are SNPServer, PolyBayes and PolyFreq (Souche *et al*., 2007). In this study, CLC genomic workbench 5.1 was used. It has user-friendly high-throughput sequence analysis tools that include SNP detection tool. The availability of large number of sequences generated by high-throughput sequencing technologies allows for *in silico* detection of candidate SNPs. Not all

in silico detected candidate SNPs are true SNPs. Therefore, there is need to validate these SNPs before using them as genetic markers for studying polymorphisms in organisms. To validate SNPs, primers are designed that target candidate SNPs. A PCR based assay is then carried out to generate products that are re-sequenced and remapped on the reference sequence for SNP detection or the amplicons are used for high resolution melting curve analysis. Han *et al.,* (2011) used these strategies to discover and validate SNPs in tetraploid alfalfa. The relatively large average distance between SNPs observed based on the DH Pahang genome could possibly be explained by the large introns in the genome since the reads used were from cDNA and the high stringency conditions used to call SNPs. The distance between SNPs ranged from1 bp to several thousand base pairs.

Mbwazirume belongs to a group of bananas commonly known as the East African highland bananas (EAHB). They are triploid 'AAA' and the members in this group are highly homogenous at genomic level but morphologically diverse. In Uganda, 84 distinct clones of cultivated EAHB were identified among which 238 accessions were grouped under five clone sets based on morphological characteristics (Karamura, 1999) but no gel-based molecular marker characterization has been able to fully confirm this classification (Onyango *et al.,* 2010,). This ambiguity could be resolved by the application of next generation sequencing technologies for detection of unique SNPs that are specific for identification of EAHB cultivars. Also genotyping by sequencing (Elshire *et al*., 2011) could be carried out, which is a far much easier and robust way to discern the long-term ambiguity that has always existed between morphological and molecular markers in characterization of this group of bananas.

5.4 Banana response to drought stress

Bananas like other plants respond to different environmental stresses such as drought stress, cold stress and high salinity. Under field conditions plants are subject to multiple stresses that require an integrated coordination of protective mechanisms within the cell and plant systems that aim at reducing the effects of the stress to enhance survival and performance. It is evident that no single pathway is sufficient enough to control the effects of environmental stresses but cross talk is the main strategy employed by the tolerant varieties. As observed from the different plant studies, drought stress leads to production of abscisic acid (ABA) in both roots and leaves. The first impact of drought stress is to reduce growth, a highly-sensitive process requiring turgor pressure. Abscisic acid in leaves induces stomatal closure by acting on the guard cells (Xiong and Zhu, 2003; Luo *et al*., 1992; Noctor *et al.,* 2002). The plant does this to limit water loss, but as a result photorespiration is increased and results in production of reactive oxygen species (ROS) that are detrimental to the plant. In response to increased ROS, plants need antioxidant enzymes and non-enzymatic antioxidants such as tocopherol, β-carotene and ascorbic acid (Abedi and Pakniyat, 2010; Liu *et al.,* 2011; Yordanov *et al,* 2003; Noctor *et al.,* 2002; Bartoli *et al.,* 1999 and Vidi *et al*., 2006) that help disable ROS that are capable of destroying cells. For transcription of genes coding for antioxidant enzymes and other genes to take place certain proteins known as transcription factors must be expressed and bind the promoter region so that RNA polymerase is recruited to carry out transcription (Hardy, 2010, Fujita *et al.,* 2004). Some known transcription factors are expressed in response to ABA. Transcription factors may not access the promoter region of the gene and even RNA polymerase may not be able to progress unless the DNA is somehow loosened from the histone proteins. Loosening of DNA from the histone protein requires the activity of an enzyme histone acetyltransferase (Murray *et al.,* 2003). This shows that even though histone acetyltransferase is not specifically linked to drought response its activity

has great impact on the plant response to environmental stress. In a broad perspective, genomes of organisms respond to changes in the environment via chemical reactions such as DNA methylation, histone acetylation, and RNA interference. These epigenetic mechanisms do not alter the DNA sequence but have effects in gene expression. Although similar strategies to respond to drought stress are available in different plant species, variation in the combination of strategies does exists. This may be attributed to variation in gene structure, gene copy number and stage of growth. For example in the leaves of cowpea plants (*Vagna unguiculata* L.) higher expression of multicystatin proteins, which act as inhibitors of cysteine proteinases was associated with drought tolerance (Diop *et al.,* 2004). However, in the present study on banana this did seem not to be the case; low expression of these genes was observed in drought tolerant Cachaco. A similar trend was observed with regard to expression of patatin-like genes. Increased expression of osmotin, and dehydrins as well as proline has been associated with drought stress, cold stress and disease infection as defensive mechanism (Zhou *et al*., 2010; Shinozaki and Yamaguchi-Shinozaki, 2007 and Moumeni *et al*., 2011). In drought tolerant Cachaco, increased expression of these genes was observed (Table 6). The good cooking qualities and organoleptic appeal (sensory perception) presented by the East African highland bananas such as Mbwazirume acts as the driving force for the need to improve these plants in terms of drought stress tolerance and disease resistance.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

Mapping of Cachaco and Mbwazirume 454 and Illumina reads to the reference genome revealed that both cultivars share many genes that were annotated in the DH Pahang genome but what differed was the expression levels of some genes when subjected to drought stress. Cachaco in many cases had an up-regulation of genes that are involved in diverse pathways associated with drought tolerance. There was an increased expression for transcription factors such as MYB44 and MYC4, NAC, bZIP, DREB1, DREB3 and AP2 that are known to be drought-induced via the ABA signaling pathway and that control downstream genes. The same trend was observed for the cell cycle regulating genes, genes for antioxidant enzymes, and biosynthetic enzymes for ABA and related signal transduction molecules. In Mbwazirume, these gene sets were down-regulated. Though the aquaporins were down-regulated in roots of both genotypes under drought stress, Cachaco maintained a significantly higher level of PIP and TIP compared to Mbwazirume, suggesting the reduced ability of Mbwazirume to respond to limited water supply. The increased expression of chlorophyll catabolic enzymes and leaf senescence-related protein genes in the leaves of Mbwazirume might explain its more rapid loss of green leaves under dry conditions. The concept of B genome conferring drought tolerance to some banana cultivar may be explained by difference in expression of genes associated with drought tolerance in Cachaco compared to Mbwazirume. Additionally, Cachaco shows greater divergence from the reference DH Pahang genome as compared to Mbwazirume.

It is no longer an excuse to have limited information about the dynamics of gene expression in the absence of reference genome for any given organism. Next generation sequencing

technologies allow *de novo* assembly of transcripts into a reference transcriptome which is sufficient enough to study the gene structure, behavior and variation in different organisms under any given set of conditions. In banana genomics research, more is still desired and this study just opens an opportunity for research in transcriptomics.

6.2 Recommendations

A fruitful area of follow-up research might be to investigate silent vs. functional SNPs in candidate genes. This may shed light as to why cultivars with B genome are drought tolerant. Since the reference genome is now available, genotyping by sequencing can be increased to gain more understanding on banana genetic diversity and allow breeders accelerate the banana and plantain improvement programs. As the cost of sequencing is reducing on the next generation sequencing platforms, it is recommended that similar studies be repeated with many cultivars representing different subgroups of bananas, using different tissues at different stages of growth and also increase biological replications to get a comprehensive understanding of banana response to drought stress. This will help contribute to sustainable food production through providing drought tolerant and disease resistant banana varieties.

REFERENCES

Abedi, T., and H. Pakniyat (2010). Antioxidant Enzyme Changes in Response to Drought Stress in Ten Cultivars of Oilseed Rape (Brassica *napus* L.). Czech J. Genet. *Plant Breeding*, 46(1): 27-34.

Aert, R., Sági L., and G. Volckaert (2004). Gene content and density in banana (*Musa acuminata*) as revealed by genomic sequencing of BAC clones. *Theoretical and Applied Genetics*, 109:129-139.

Akram, M., Malik, M.A. Ashraf M.Y., Saleem M.F. and M. Hussain (2007). Competitive seedling growth and K+/Na+ ratio in different maize (*Zea mays* L.) hybrids under salinity stress. *Pakistan Journal of Botany*, 39: 2553-2563.

Arango, R.E., Togawa R.C., Carpentier S.C., Roux N., Hekkert B.L. Kema G.H.J., and M.T. Souza Jr (2011). Genome-wide BAC-end sequencing of *Musa acuminata* DH Pahang reveals further insights into the genome organization of banana. *Tree Genetics & Genomes* DOI 10.1007/s11295-011-0385-3.

Banu, N.A., Hoque, A., Wantanbe-Sugimoto, M., Matsuoka, K., Nakamura, Y., Shimoishi, Y., and Y.Murata (2009). Proline and glycinebetaine induce antioxidant defense gene expression and suppress cell death in cultured tobacco cells under salt stress. *Journal of Plant Physiology*. 166 (2): 146-156.

Bao, H., Guo H., Wang J., Zhou R., Lu X., and S. Shi (2009). MapView: visualization of short reads alignment on a desktop computer. *Bioinformatics*, 12: 1554-1555.

Barnes, W.M.(1994). PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proceedings of the Natural Academy of Sciences* USA, 91:2216-2220.

Bartoli, C.G., Simontacchi M., Tambussi E., BeltranoJ., Montaldi E., and S. Puntarulo (1999). Drought and watering-dependent oxidative stress: Effect on antioxidant content in Triticum aestivum L. leaves. *Journal of Experimental Botany,* 50(332):375-383.

Biruma, M., Pillay M., Tripathi L., Blomme G., Abele S., Mwangi M., Bandyopadhyay R., Muchunguzi P., Kassim S., Nyine M., Turyagenda L., and S. Eden-Green (2007). Banana Xanthomonas wilt: a review of the disease management strategies and future research directions. *African Journal of Biotechnology*, 6 (8): 953-962.

Botella, J.R., Arteca J.M., Somondevilla M., and R.N. Arteca (1996). Calcium-dependent protein kinase gene expression in response to physical and chemical stimuli in mungbean (*Vagna radiate*); *Plant Molecular Biology,* 30:1129-1137.

Bouwmeester, H., Van Asten P., and E. Ouma (2009). Mapping key variables of banana based cropping systems in the Great Lakes region, partial outcomes of the base-line and diagnostic surveys. *CIALCA Technical Report*, 11. http:www.cialca.org/IITA, Ibadan.

Brem, R.B., Yvert, G., Clinton, R. and L. Kruglyak (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296: 752–755.

Buchanan-Wollaston, V. (2007). Senescence in plants. *Wiley online library*, DOI: 10-1002/9780470015902.a0020133.

CGIAR (2003). Applications of molecular biology and genomics to genetic enhancement of crop tolerance to abiotic stresses – a discussion document. Interim Science Council Secretariat, FAO.

Cline, W.R, (2007). World agriculture faces serious decline from global warming. *Center for global development*, 1-5.

Cheong, Y.H., Kim K., Pandey G.K., Gupta R., Grant J.J., and S. Luan (2003). CBL1, a Calcium Sensor That Differentially Regulates Salt, Drought, and Cold Responses in Arabidopsis. *The Plant Cell*, 15:1833–1845.

Cheung, F. and C. D. Town (2007). A BAC end view of the *Musa acuminata* genome. *BMC Plant Biology*, 7:29.

Choi, Y.B, Ko J.K, and J. Shin (2004). The transcriptional corepressor, PELP1, recruits HDAC2 and masks histones using two separate domains. *Journal of Biological Chemistry.* 3;279(49):50930-41.

Daniells, J.W. (1990). The Cavendish subgroup, distinct and less distinct cultivars. In: Jarret, R.L. (ed.) Identification of genetic diversity in the genus *Musa*, *INIBAP*, Montpellier, France, 29-35.

Daniells, J., Jenny C., Karamura D., and K. Tomekpe (2001). Musalogue: a catalogue of *Musa* germplasm. Diversity in the genus *Musa*. (E. Arnaud and S. Sharrock, Compil.). *INIBAP*, Montpellier, France 1-213.

De Langhe, E., Hřibová, Carpentier S., Doležel J., and R. Swennen (2010). Did backcrossing contribute to the origin of hybrid edible bananas? *Annals of Botany*, 106:849-857.

Delcher, A.L. Phillippy A., Carlton J., and S.L. Salberg (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*, 30**:** 2478-2483.

Diop, N.N., Kidri M., Repellin A, Gareil M., d'Arcy-Lameta A., Thi A.T.P., and Y. Zuily-Fodil (2004). A multicystatin is induced by drought-stress in cowpea (*Vigna unguiculata* (L.) Walp.) leaves. *FEBS Letters*, 577: 545-550.

Eaves, I.A., Wicker, L.S., Ghandour, G., Lyons, P.A., Peterson, L.B., Todd, J.A., and R.J. Glynne (2002). Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes. *Genome Research,* 12: 232–243.

Eisen, M.B., Spellman P.T., Brown P.O., and D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *Proceeding of the Natural Academy of Sciences,* USA, 95: 14863-14868.

Elshire, R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., and S.E. Mitchell (2011). A robust, simple genotypeing-by-sequencing (GBS) approach for high diversity species. *PLoS One*, 6(5):e19379.

FAO, (2010a). Acting together against banana diseases in Africa. News bulletin.

FAO, (2010b). The state of food insecurity in the world. http://www.fao.org/docrep/013/1683e/1683e.pdf.

FAOSTAT, (2009). World food production quantities, FAO. www.faostat.fao.org

Fleury, D., Jefferies S., Kuchel H., and P. Langridge (2010). Genetic and genomic tools to improve drought tolerance in wheat. *Journal of Experimental Botany*, 61(12): 3211-3222.

Fujita, M., Fujita Y., Maruyama K., Seki M., Hiratsu K., Ohme-Takagi M., Tran L.P., Yamaguchi-Shinozaki K., and k. Shinozaki (2004). A dehydration-induced NAC protein, RD26, is involved in a novel ABA-dependent stress-signaling pathway. *The Plant Journal* (2004) 39, 863–876.

Goodsell, D.S. (2004). The molecular perspective: Cytochrome c and apoptosis. *The Oncologists*, 9:226-227.

Götz, S., Garcı´a-Go´mez J.M., Terol J., Williams T.D., Nagaraj S.H., Nueda M.J., Robles M., Talo´n M., Dopazo J., and A. Conesa (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, 36(10): 3420-3435.

Grattapaglia, D., Silva-Junior O.B., Kirst M., Merco de Lima B., Faria D.A., and G.J. Ppappas Jr (2011). High-throughput SNP genotypying in the highly heterozygous genome of

Eucalyptus: assay success, polymorphism and transferability across species. *BMC Plant Biology*, 11:65.

Gunes, A., Pilbeam D., Inal A., and S. Coban (2008): Influ- ence of silicon on sunflower cultivars under drought stress, I: Growth, antioxidant mechanisms and lipid peroxidation. Commun. *Soil Science & Plant Nutrition*, 39**:** 1885-1903.

Hamanishi, E.T., and M. Campbell (2011). Genome-wide responses to drought in forest trees. *Forestry,* doi:10.1093/forestry/cpr012.

Han, Y., Kang Y., Torres-Jerez, Cheung F., Town C.D., Zhao P.X., Udvardi M.K., and M.J. Monteros (2011). Genome-wide SNP discovery in tetraploid alfalfa using 454 sequencing and high resolution melting analysis. *BMC Genomics*, 12:350.

Hardy, A., (2010). Candidate stress response genes for developing commercial drought tolerant crops. *MMG 445 Basic Biotechnology*, 6:54-58.

Hernandez, D., François P., Farinelli L., Osteras M., and J. Schrenzel (2008). *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research*, 18: 802-809.

Heuzé, V., and G. Tran (2011). Banana (GENERAL). Feedipedia.org and Tables Regions Chaudes. A project by INRA, CIRAD and AFZ with the support of FAO. http://www.trc.zootechnie.fr/node/4670 Last updated on December 8, 2011, 12:34.

Hřibová, E., Neumann P., Matsumoto T., Roux N., Macas J. and J. Doležel (2010). Repetitive part of the banana (Musa 77cuminate) genome investigated by low-depth 454 sequencing. *BMC Plant Biology*, 10:204.

Hu, H., Dai, M., Yao, J., Xiao, B., Li, X., Zhang, Q., and L.(Xiong 2006). Overexpressing a NAM, ATAF, and CUC (NAC) transcription factor enhances drought resistance and salt tolerance in rice. *Proceedings of the Natural Academy of Sciences*, USA 103:12987–12992.

Huang, W., and G. Marth (2008). EagleView: a genome assembly viewer for next-generation sequencing technologies. *Genome Research*, 9: 1538-1543.

Imelfort, M., and D. Edwards (2009). *De novo* sequencing of plant genomes using second-generation technologies. *Briefings in Bioinformatics*, 10 (6): 609- 618.

INIBAP, (1995). Banana and Plantain: the earliest crop? Annual report, 14-17.

INIBAP, (1999). Networking banana and plantain: INIBAP annual report 1998. *International Network for the Improvement of Banana and Plantain*, Montpellier, France. 1-64.

INIBAP, (2005). Unlocking the secrets of the banana genome. *Annual report*, 14-17.

JIRCAS, (2005). Improvement of drought stress tolerance by gene transfer of a transcription factor, AREB1, Involved in ABA-responsive gene expression. *Research Highlights*, 03.

Jeck, W.R., Reinhardt J.A., Baltrus D.A., Hickenbotham M.T., Magrini V., Mardis E.R., Dangl J.L., and C.D. Jones (2007). Extending assembly of short DNA sequences to handle error. *Bioinformatics* 23(21): 2942-2944.

Jehan, T., and S. Lakhanpaul (2006). Single nucleotide polymorphism (SNP)-Methods and application in plant genetics: A review. *Indian Journal of Biotechnology*, 5:435-459.

Jones, D.R. (2000). Diseases of Banana, Abacá and Enset. *CABI Publishing*, 1-495.

Karamura, D.A., (1999). Numerical taxonomic studies in the East African highland banana (*Musa* AAA-East Africa) in Uganda. *PhD thesis from the University of Reading*, 1-192.

Karp, C.L., Grupe, A., Schadt, E., Ewart, S.L., Keane-Moore, M., Cuomo, P.J., Kohl, J., Wahl, L., Kuperman, D., and S. Germer (2000. Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nature Immunology,* 1: 221-226.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research***,** 4:656-664.

Khan, M.A., Shirazi, M.U., Khan, A. M., Mujtaba, S.M., Islam, E., Mumtaz, S., Shereen, A., Ansari, R.U. and Y.M. Ashraf (2009). Role Of Proline, K/Na Ratio and chlorophyll Content

in salt tolerance of wheat (*Triticum aestivum* L.). *Pakistan Journal of Botany*, 41(2): 633-638.

Koorevaar, P., Menelik, G. and C. Dirksen (1983). Elements of soil physics. *Developments in soil science 13*, Lsevier-Amsterdam: 85.

Kumar, P.L. and R. Hanna (2010). BBTV in sub-Saharan Africa Established or Emerging Problem? http://www.iita.org/cms/articlefiles/1682-BBTV_poster.pdf, retrieved on 11-February 2010.

Kurtz, S., Phillippy A., Delcher A.L., Smoot M., Shumway M., Antonescu C., and S.L. Salzberg (2004). Versatile and open software for comparing large genomes. *Genome Biology,* 5: R12.

Langmead, B., Trapnell C., Pop M., and S.L. Salzberg (2009). Ultrafast and memory-efficient alignment of short DNA sequences to human genome. *Genome Biology*, 3: R25.

Lescot, M., Piffanelli P., Ciampi A.Y., Ruiz M., Blanc G., Leebens-Mack J., da Silva F.R., Santos C.M.R., D'Hont A., Garsmeur O., Vilarinhos A.D., Kanamori H., Matsumoto T., Ronning C.M., Cheung F., Haas B.J., Althoff R., Arbogast T., Hine E., Pappas G. J., Sasaki T., Souza M.T., Miller R.N.G., Glaszmann J.C. and C. D. Town (2008). Insights into the Musa genome: Syntenic relationships to rice and between Musa species. *BMC Genomics*, 10.1186/1471-2164-9-58.

Li, R., Li Y., Kristiansen K., and J. Wang (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics,* 5: 713-714.

Li, H., and R. Durbin (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 5: 589-595.

Li, H., Ruan J., and R. Durbin (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 11: 1851-1858.

Li, R., Yu C., Li Y., Lam T., Yiu S., Kritiansen K., and J. Wang (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 15:1966-1967.

Liu, C., Liu Y., Guo K., Fan D., Li G., Zheng Y., Yu L., and R. Yang (2011). Effect of drought on pigments, osmotic adjustment and antioxidant enzymes in six woody plant species in Karst habitats of southwestern China. *Environmental and Experimental Botany,* 71:174-183.

Luo, M., Liu J., Mohapatra S., Hill R.D., and S.S.Mohapatra (1992). Characterizationof a Gene Family Encoding Abscisic Acid- and Environmental Stress-induciblProteins of Alfalfa. *The journal of biological chemistry*, 267(22):15367-15374.

Magi, A., Benelli M., Gozzini A., Girolami F., Torricelli F., and M. L. Brandi (2010). Bioinformatics for Next Generation Sequencing Data. Genes 1: 294-307.

Malhis N., Butterfield Y.S.N., Ester M., and S.J.M. Jones (2009). Slider—maximum use of probability information for alignment of short sequence reads and SNP detection. *Bioinformatics*, 25(1): 6-13.

Manivannan, P., Abdul Jaleel C., Kishorekumar A., Sankar B., Somasundaram R., Sridharan R., and R. Panneerselvam (2007): Changes in antioxidant metabolism of *Vigna unguiculata* (L.) Walp. By propi- conazole under water deficit stress. *Colloids and Surfaces Biointerfaces*, 57**:**69–74.

Manda, M., Nechifor M.T., and T.M. Neagu (2009). Reactive oxygen species, Cancer and Anti-Cancer Therapies. *Current Chemical Biology,* 3:342-366.

Margulies, M., Egholm M., Altman W.E., Attiya S., Bader J.S., *et al.,* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437: 376-380.

Milne, I., Bayer M. Cardle L., Shaw P., Stephen G., Wright F., and D. Marshall (2010) Tablet—next generation sequence assembly visualization. *Bioinformatics*, 3: 401-402.

Mitra, D., and M.M. Johri (2000). Enhanced expression of a calcium-dependent protein kinase from the moss *Funaria hygrometrica* under nutritional starvation. *Journal of Biosciences*, 25(4):331-338.

Mortazavi, A., Williams B.A., McCue K., and B. Wold (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621-628.

Moumeni, A., Satoh K., Kondoh H., Asano T., Hosaka A., Venuprasad R., Serraj R., Kumar A., Leung H., and S. Kikuchi (2011). Comaparative analysis of root transcriptome profiles of two pairs of drought-tolerant and susceptible rice near-isogenic lines under different drought stress. *BMC Plant Biology,* 11:74.

Murray, R.K., Granner D.K., Mayes P.A., and V.W. Rodwell (2003). Harper's illustrated biochemistry; DNA organization, replication and repair. *Lange medical books*, New York. 26th edition, Chapter 36:314-340.

Nakashima, K., Ito Y., and K Yamaguchi-Shinozaki (200). Transcriptional regulatory networks in response to abiotic stresses in Arabidopsis and grasses. *Plant Physiology, 149:88-95.*

Nelson, S.C., Ploetz R.C., and A.K. Kepler (2006). Musa species (banana and plantain), *Species Profiles for Pacific Island Agroforestry* 2.2:1-33.

Ning, Z., Cox A.J., and J.C. Mullikin (2001). SSAHA: a fast search method for large DNA databases. *Genome Research*, 11: 1725-1729.

Noctor, G., Veljovic-Jovanovic S., Driscoll S., Novitskaya L., and C.H. Foyer (2002). Drought and oxitative load in leaves of C_3 plants: a predominant role for photorespiration. *Annals of Botany*, 89:841-850.

Nyren, P., and A. Lundin (1985). Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Analls of Biochemistry*, 151: 504-509.

Onyango, M., Haymer D., Keeley S., and R. Manshardt (2010). Analysis of genetic diversity and relationship in East African 'Apple bananas' (AAB genome) and 'Muraru' (AA genome) dessert bananas using microsatellite markers. *Proceedings of International Conference on Banana and Plantains,* 623-636.

Öpik, H., Rolfe, S.A. and A.J. Willis (2005). The physiology of flowering plants 4th edt, chapt. 13: 344-370.

Ortiz, R., and D.R. Vuylsteke (1994). Future strategy of Musa improvement. In: Banana and plantain breeding: Priorities and strategies. INIBAP, Montpellier, France, 40-42.

Pardo, J.M (2010). Biotechnology of water and salinity stress tolerance. *Current Opinions in Biotechnology*. 21:185-196.

Park, SY., Yu JW., Park JS., Li J., Yoo SC., Lee NY., Lee SK., Jeong SW., Seo HS., Koh HJ., Jeon JS., Park YI., and NC. Paek (2007). The senescence-induced staygreen protein regulates chlorophyll degradation. *The Plant Cell*, 19:1649-1664.

Pro*Musa* (2002). Third meeting of the PRO*MUSA* Sigatoka working group, 11: 1-24

Rismani-Yazdi, H., Haznedaroglu B. Z., Bibby K., and J. Peccia (2011). Transcriptome sequencing and annotation of the microalgae Dunaliella tertiolecta: Pathway description and gene discovery for production of next-generation biofuels. *BMC Genomics*, 471-2164/12/148.

Robinson, J. C., and V. G. Saúco (2010). Banana and Plantain, 2nd Edition. *Crop Production Science in Hotculture,* 19: 85.

Rowe, a., and P.E. Rosales (1996). Bananas and Plantains. Fruit Breeding vol.1: Tree and Tropical Fruit edited by Jules Janick and James N. Moore. ISBN 0-471-32014-X.

Saavedra, X., Medrego A., Rodriguez D., Gonzalez-Garcia MP., Sanz L., Niclas G., and O. Lorenzo (2010). The Nuclear interactor PYL8/RCAR3 of *Fagus sylvatica* $FsPP_2C_1$ is a

positive regulator of abscisic acid signalling in seeds and stress. *Plant physiology*, 152:133-150.

Sanger, F. and A.R. Coulson. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94 (3): 441–448.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R. and G. Cavet (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422: 297–302.

Schatz, M.C., Trapnell C., Delcher A.L., and A. Varshney (2007). High-throughput sequence alignment using Graphics Processing Units. *BMC Bioinformatics,* 8: 474.

Schena, M., Shalon D., Davis R.W., P.O. Brown (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270: 467 – 470.

Seki, M., Ishida J., and M. Narusaka (2002b). Monitoring the expression pattern of ca. 7000 Arabidopsis genes under ABA treatments using a full-length cDNA microarray. *Functional and Integrative Genomics,* (2) 282–291.

Seki, M., Narusaka, M., and J. Ishida (2002a) Monitoring the expression profiles of ca. 7000 *Arabidopsis* genes under drought, cold, and high-salinity stresses using a full-length cDNA microarray. *The Plant Journal,* (31**)** 279-292. http://jxb.oxfordjournals.org/cgi/external_ref?access_num=10.1046%2Fj.1365-313X.2002.01359.x&link_type=DOI

Setter, T.L. and M.A. Fregene (2007). Recent advances in molecular breeding of cassava for improved drought stress tolerance. M.A. Jenks et al. (eds.). Advances in Molecular Breeding Toward Drought and Salt Tolerant Crops, *Chapter*, 28: 701–711.

Shagin D.A., Rebrikov D.V., Kozhemyako V.B., Altshuler I.M., Shcheglov A.S., Zhulidov P.A.,

Bogdanova E.A., Staroverov D.B., Rasskazov V.A., and S. Lukyanov (2002). A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. *Genome Research*, 12(12):1935-1942.

Sheperd, K. (1957). Banana cultivars in East Africa. *Tropical Agriculture*, 34:277-286.

Shendure, J., and H. Ji (2008). Next generation DNA sequencing. *Nature Biotechnology*, 26(10): 1135-1145.

Shinozaki, K., Yamaguchi-Shinozaki, K., and M. Seki (2003). Regulatory network of gene expression in the drought and cold stress responses. *Current Opinion in Plant Biology,* 6:410–417.

Shinozaki, K. and K. Yamaguchi-Shinozaki (2007). Gene networks involved in drought stress response and tolerance. *Journal of Experimental Botany*, 58(2):221-227.

Simmonds, N.W. (1962). The evolution of bananas. Longman, Londan.

Slater, G.S.C., and E. Birney, (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6**,** 31.

Souche, E.L., Hellemans B., Van Houdt J.K.J., Canario A., Klages S., Reinhardt R., and F.A.M. Volckaert (2007). Mining for single nucleotide polymorphisms in expressed sequence tags of the European Sea Bass. Journal of Integrative Bioinformatics, 4(3)73.

Stover, D.H., and N.W. Simmonds (1987). Bananas. Third edition. *Longman Scientific/John Wiley*, New York, USA. 468.

Tsuchihira, A., Hanba Y.T., Kato N., Doi T., Kawazu T., and M. Maeshima (2010). Effect of overexpression of radish plasma membrane aquaporins on water-use efficiency, photosynthesis and growth of *Eucalyptus* trees. *Tree physiology*, 30:417-430.

Turcatti, G., Romieu A., Fedurco M., and A.P. Tairi (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Research,* 36: e25.

Urao, T., Katagiri T., Mizoguchi T., Yamaguchi-Shinozaki K., Hayashida N., and K. Shinozaki (1994). Two genes that encode Ca(2+)-dependent protein kinases are induced by drought and high salt stress in *Arabidopsis thaliana*. *Molecular and General Genetics*, 244:331-340.

Van Asten, P.J.A., Fermont A.M. and G. Taulya (2011). Drought is a major yield loss factor for rainfed East African highland banana. *Agricultural Water Management*, 98(4): 541-552.

Vidi, P., Kanwischer M., Baginsky S., Austin J.R., Csucs G., Dormann P., Kessler F., and C. Brehelin (2006). Tocopherol Cyclase (VTE1) Localization and Vitamin E Accumulation in Chloroplast Plastoglobule Lipoprotein Particles. *The Journal of Biological Chemistry*, 281(16): 11225-11234.

Wall, P.K., Leebens-Mack J., S Chanderbali A., Barakat A., Wolcot E., Liang H., Landherr L., Tomsho L.P., Hu Y., Carlson J.E., Ma H., Schuster S. C., Soltis D.E., Soltis P. S., Altman N. and C.W. dePamphilis (2009). Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics*, 10**:**347.

Warren, R.L., Sutton G.G., Jones S.J.M., and R.A. Holt (2007). Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 23(4): 500-501.

Wayne, M.L., and L.M. McIntyre (2002). Combining mapping and arraying: an approach to candidate gene identification. *Proceedings of the Natural Academy of Sciences,* USA 99: 14903-14906.

Wullschleger, S.D., and S.P. Difazio (2003). Emerging use of gene expression microarrays in plant physiology. *Comparative and Functional Genomics,* 4: 216–224.

Xiao, X., Xu X., and F. Yang (2008): Adaptive responses to progressive drought stress in two *Populus cathayana* populations. *Silva Fennica*, 42: 705–719.

Xiong, L., and JK. Zhu (2003). Regulation of abscisic acid biosynthesis. *Plant Physiology*, 133:29-36.

Xu ,P.L. Guo Y.K., Bai J.G., Shang L., and X.J. Wang (2008): Effects of long-term chilling on ultrastructure and antioxidant activity in leaves of two cucumber cultivars under low light. *Physiologia Plantarum*, 132**:** 467–478.

Yordanov, I., Velikova V., and T. Tsonev (2003). Plant response to drought stress and stress tolerance. *Bulgarian Journal of Plant Physiology,* special issue 2003, 187-206.

Yu, H., Chen X., Hong YY., Wang Y., Xu P., Ke SD., Liu HY., Zhu JK., Oliver D.J., and CB. Xiang (2008). Activated expression of an Arabidopsis HD-START protein confer drought tolerance with improved root system and reduced stomatal density. *The plant cell*, 20:1134-1151.

Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R. and L. Kruglyak (2003). Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics,* 35: 57–64.

Zerbino, D.R., and E. Birney (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome. Research,* 5: 821-829.

Zhao W., Wang J., He X., Huang X., Jiao Y., Dai M., Wei S., Fu J., Chen Y., Ren X., Zhang Y., Ni P., Zhang J., Li S., Wang J., Wong G.K., Zhao H., Yu J., Yang H., and J. Wang: (2004). BGI-RIS, An integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Research*, 32:377-82.

Zhou, M., Ma J., Pang J., Zhang Z., Tang Y., and Y. Wu (2010). Regulation of plant stress response by dehydration responsive element binding (DREB) transcription factors. *African Journal of Biotechnology*, 9 (54):9255-9279.

Zhu, Y.Y., Machleder E.M., Chenchik A., Li R., Siebert P.D (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques*, 30: 892-897.

APPENDICES

Appendix I: Small scale RNA isolation from plant tissue

A protocol adopted from Invitrogen, revised on 14th December 2005

Materials needed

- Liquid nitrogen
- Mortar and pestle
- RNase-free microcentrifuge tube
- 5M NaCl
- Chloroform
- Isopropyl alcohol
- 75% ethanol
- RNase-free water

Sample preparation

i. Cool RNase-free micro-centrifuge tubes in dry ice before transferring frozen tissue into the tube

ii. Grind fresh or frozen plant tissue in liquid nitrogen to a powder using mortar and pestle. Grind dry seed samples at room temperature.

iii. Store all ground plant material at -70^{o}C until further use. Frozen tissue must remain frozen at -70^{o}C prior to extraction with plant RNA reagent. Accidental thawing may result in RNA degradation.

Procedure

i. Add 0.5 ml cold (+4^{o}C) PureLink Plant RNA Reagent to up to 0.1 gram of frozen, ground plant tissue. Mix by brief vortexing or by flicking the bottom of tube until the sample is thoroughly re-suspended.

ii. Incubate the tube for 5 minutes at room temperature.

   a. **Note: Lay the tube down horizontally during incubation to maximize surface area.**

iii. Clarify the solution by centrifuging at 12000 x g in a micro-centrifuge tube for 2 minutes at room temperature. Transfer the supernatant to a clean RNase-free tube.

iv. Add 0.1 ml of 5 M NaCl to the clarified extract. Mix by tapping the tube.

v. Add 0.3 ml chloroform to the sample. Mix thoroughly by inverting the tube.

vi. Centrifuge the sample at 12000 x g for 10 minutes at +4^{o}C to separate the phases. Transfer the upper, aqueous phase to a clean RNase-free 1.5 ml eppendorf tube.

vii. Add to the aqueous phase an equal volume of isopropyl alcohol. Mix and let stand at room temperature for 10 minutes.

viii. Centrifuge sample at 12000 x g for 10 minutes at +4^{o}C.

ix. Decant the supernatant, taking care not to lose the pellet, and add 1 ml of 75% ethanol to the pellet.

Note: Pellet may be difficult to see.

x. Centrifuge at 12000 x g for 1 minute at room temperature. Decant the supernatant carefully, taking care not to lose the pellet. Briefly centrifuge to collect the residual liquid and remove it with pipet.

xi. Add 10-30 µl RNase-free water to the RNA pellet. Pipet up and down over the pellet to re-suspend RNA. If any cloudiness is observed, centrifuge the solution at room temperature for 1 minute at 12000 x g and transfer the supernatant containing RNA into a

clean 1.5 ml eppendorf tube and store at -70^{o}C. OR add 0.1 Volume of 3 M Sodium acetate and 2.5 volumes of 96% Ethanol, mix thoroughly and store at -20^{o}C for years.

Appendix II

Names and sequences of oligonucleotides used during the study

SMART-Sfi1A oligonucleotide	5'-AAGCAGTGGTATCAACGCAGAGTGGCCATTACGGCCrGrGrG-3'
CDS-SfiIB-GC T23 primer	5'-AAGCAGTGGTATCAACGCAGAGTGGCCGAGGCGGCCTTTTGTTTTTTCTTTTTTTTTTTTTVN-3
SMART PCR primer	5'-AAGCAGTGGTATCAACGCAGAGT-3'
CDS-SfiBI T19-454 primer	5'-ACGCAGAGTGGCCGAGGCGGCCTTTTGTCTTTTCTTCTGTTTCTTTTVN
SfiIA PCR primer	5'-CAACGCAGAGTGGCCATTAC
SfiIB PCR primer	5'-ACGCAGAGTGGCCGAGGCG

Appendix III

Tables showing absolute RPKM values of different genes and their relative expression under drought stress

A: Transcription factors

Gene name	Tissue	CD	CW	MD	MW	CDRE	MDRE
DREB	Leaf	25.905	8.215	6.492	68.329	**3.153**	**0.095**
	Root	31.967	15.034	6.36	13.177	**2.126**	**0.483**
MYB transcription factor, putative	Leaf	178.479	87.425	89.696	75.071	**2.042**	**1.195**
	Root	255.724	151.451	112.409	116.485	**1.688**	**0.965**
MYB transcrption factor	Leaf	197.247	199.624	199.341	135.96	**0.988**	**1.466**
	Root	151.624	105.394	214.419	99.314	**1.439**	**2.159**
MYC transcription factor	Leaf	301.852	128.221	181.34	376.332	**2.354**	**0.482**
	Root	160.253	126.978	198.718	151.21	**1.262**	**1.314**
B3 domain containing protein, putative	Leaf	44.736	34.198	22.158	14.629	**1.308**	**1.515**
	Root	60.985	56.084	24.973	50.457	**1.087**	**0.495**
Ethylene responsive transription factor, putative	Leaf	1104.583	862.769	904.362	951.111	**1.280**	**0.951**
	Root	1843.823	1174.422	1005.997	1147.462	**1.570**	**0.877**
Early responsive to dehydration stress-related protein	Leaf	84.65	111.407	73.223	103.254	**0.760**	**0.709**
	Root	365.75	237.344	78.442	162.354	**1.541**	**0.483**
bZIPtranscription factor domain containing protein	Leaf	517.817	413.645	323.233	488.437	**1.252**	**0.662**
	Root	618.367	540.201	360.638	506.578	**1.145**	**0.712**
AP2-Like ethylene responsive transcription factor	Leaf	68.304	67.166	45.894	65.131	**1.017**	**0.705**
	Root	170.23	124.089	40.677	98.56	**1.372**	**0.413**
NAC domain containing protein	Leaf	287.844	175.745	145.668	363.894	**1.638**	**0.400**
	Root	773.153	272.725	133.592	272.493	**2.835**	**0.490**

Gene name	Tissue	CD	CW	MD	MW	CDRE	MDRE
Transcription factor MYB44, putative	Leaf	152.599	40.719	67.012	42.742	**3.748**	**1.568**
	Root	189.424	83.646	85.586	70.316	**2.265**	**1.217**
Transcription factor MYB44	Leaf	72.382	66.155	68.897	13.474	**1.094**	**5.113**
	Root	86.835	38.879	86.7	35.394	**2.233**	**2.450**
Transcription factor MYB2	Leaf	9.949	24.839	18.958	3.27	**0.401**	**5.798**
	Root	4.042	12.126	22.611	8.781	**0.333**	**2.575**
Transcription factor MYC4	Leaf	277.197	99.707	157.366	344.286	**2.780**	**0.457**
	Root	103.107	64.396	173.944	101.727	**1.601**	**1.710**

B: Antioxidant enzymes

Gene name	Tissue	CD	CW	MD	MW	CDRE	MDRE
Catalase	Leaf	2740.597	2555.323	2322.559	5075.117	**1.073**	**0.458**
	Root	1866.684	1681.276	2339.093	2448.82	**1.110**	**0.955**
L-ascorbate peroxidase	Leaf	839.729	1001.473	598.901	828.972	**0.838**	**0.722**
	Root	1328.517	1458.684	633.376	1316.823	**0.911**	**0.481**
Superoxide dismutase	Leaf	1154.361	1619.273	1558.624	2610.625	**0.713**	**0.597**
	Root	1485.615	1383.871	1804.544	1531.804	**1.074**	**1.178**
Glutathione reductase	Leaf	89.694	99.664	84.244	158.09	**0.900**	**0.533**
	Root	242.857	365.343	82.492	277.947	**0.665**	**0.297**
Probable phospholipid hydroperoxide glutathione peroxidase	Leaf	588.803	608.726	468.155	511.783	**0.967**	**0.915**
	Root	759.76	654.112	474.027	665.915	**1.162**	**0.712**
Polyphenol oxidase, chloroplastic	Leaf	1.195	0.83	0.483	5.627	**1.440**	**0.086**
	Root	106.383	69.574	0.697	26.755	**1.529**	**0.026**

Gene name	Tissue	CD	CW	MD	MW	CDRE	MDRE
Superoxide dismutase [Cu-Zn]	Leaf	940.056	1373.732	1339.922	2265.153	0.684	0.592
	Root	1105.332	982.351	1544.837	1105.6	1.125	1.397
Superoxide dismutase [Fe]	Leaf	68.986	79.501	67.439	41.382	0.868	1.630
	Root	24.113	19.159	79.578	26.261	1.259	3.030
Superoxide dismutase [Mn]	Leaf	145.319	166.04	151.263	304.09	0.875	0.497
	Root	356.17	382.361	180.129	399.943	0.932	0.450

C: Signal transduction molecules

Gene name	Tissue	CD	CW	MD	MW	CDRE	MDRE
9-cis epoxycarotineod dioxygenase	Leaf	22.06	15.342	13.085	34.643	**1.438**	**0.378**
	Root	40.677	12.64	14.581	22.332	**3.218**	**0.653**
Zeaxanthin epoxidase	Leaf	95.566	64.832	28.033	31.629	**1.474**	**0.886**
	Root	23.205	14.497	20.707	9.453	**1.601**	**2.191**
Calcineurin B-Like	Leaf	247.363	229.288	237.5	268.977	**1.079**	**0.883**
	Root	256.644	231.549	225.436	262.648	**1.108**	**0.858**
ABA receptors	Leaf	282.723	176.477	193.25	345.817	**1.602**	**0.559**
	Root	236.863	214.92	194.74	218.533	**1.102**	**0.891**
Calmodulin	Leaf	1019.733	1284.103	869.378	1613.633	**0.794**	**0.539**
	Root	3859.214	3214.494	966.707	2266.122	**1.201**	**0.427**
Calcium-dependent protein kinases	Leaf	496.508	452.897	377.956	705.604	**1.096**	**0.536**
	Root	1078.704	789.897	333.32	692.831	**1.366**	**0.481**

Gene name	Tissue	CD	CW	MD	MW	CDRE	MDRE
ABA receptor PYL8	Leaf	238.087	153.399	155.312	306.767	**1.552**	**0.506**
	Root	188.863	155.394	159.359	168.265	**1.215**	**0.947**
CBL-1	Leaf	74.386	80.068	87.694	94.945	**0.929**	**0.924**
	Root	108.575	77.602	87.194	113.661	**1.399**	**0.767**
CBL-3	Leaf	96.407	96.215	80.626	115.897	**1.002**	**0.696**
	Root	94.503	86.229	77.615	82.336	**1.096**	**0.943**
CDPK-2	Leaf	62.622	44.344	53.691	177.619	**1.412**	**0.302**
	Root	84.248	70.845	50.133	83.948	**1.189**	**0.597**
CDPK-3	Leaf	56.429	55.939	33.28	68.554	**1.009**	**0.485**
	Root	362.933	231.373	25.471	161.964	**1.569**	**0.157**
CDPK-13	Leaf	81.843	69.436	57.173	96.862	**1.179**	**0.590**
	Root	97.609	104.288	57.852	84.332	**0.936**	**0.686**
CDPK-28	Leaf	92.824	72.679	63.635	149.976	**1.277**	**0.424**
	Root	146.472	79.884	48.923	69.6	**1.834**	**0.703**

D: Channel proteins (aquaporins)

Gene name	Tissue	CD	CW	MD	MW	CDRE	MDRE
Aquaporin-NIP	Leaf	107.584	211.6	56.017	34.962	**0.508**	**1.602**
	Root	53.272	89.253	75.412	80.478	**0.597**	**0.937**
Aquaporin-PIP	Leaf	882.153	806.071	585.671	1821.157	**1.094**	**0.322**
	Root	1979.857	2526.952	613.874	3955.305	**0.783**	**0.155**
Aquaporin-TIP	Leaf	184.199	358.671	143.198	91.276	**0.514**	**1.569**
	Root	1351.415	1407.36	169.163	1523.777	**0.960**	**0.111**
Aquaporin-SIP	Leaf	40.557	46.881	38.834	33.575	**0.865**	**1.157**
	Root	80.832	80.248	43.822	52.291	**1.007**	**0.838**

E: Cell cycle regulating proteins

Gene name	Tissue	CD	CW	MD	MW	CDRE	MDRE
MAPK	Leaf	505.205	569.409	507.919	434.623	**0.887**	**1.169**
	Root	867.368	641.22	464.629	613.226	**1.353**	**0.758**
CDK	Leaf	677.584	489.32	596.023	826.704	**1.385**	**0.721**
	Root	744.383	670.216	539.726	656.698	**1.111**	**0.822**
Chlorophyllase-2 (chloroplastic)	Leaf	45.238	47.157	40.261	1.207	**0.959**	**33.356**
	Root	3.08	2.324	36.777	2.493	**1.325**	**14.752**
Chlorophyllide_a_oxygenase, chloroplastic	Leaf	48.804	46.878	29.983	6.428	**1.041**	**4.664**
	Root	4.028	3.388	29.623	4.538	**1.189**	**6.528**
LSRP	Leaf	119.498	253.083	223.56	94.389	**0.472**	**2.368**
	Root	196.025	226.879	255.839	188.656	**0.864**	**1.356**
SIC-SGP	Leaf	14.164	3.164	1.521	6.698	**4.477**	**0.227**
	Root	15.259	6.543	1.718	4.578	**2.332**	**0.375**
Cytochrome c	Leaf	63.635	118.56	60.21	88.238	**0.537**	**0.682**
	Root	326.834	438.327	66.585	286.206	**0.746**	**0.233**

F: Other genes

Gene name	Tissue	CD	CW	MD	MW	CDRE	MDRE
Chalcone synthase	Leaf	705.695	2085.629	140.055	814.397	**0.338**	**0.172**
	Root	606.344	349.982	157.709	330.313	**1.733**	**0.477**
Chalcone flavonone isomerase, putative	Leaf	495.135	937.201	148.194	70.452	**0.528**	**2.103**
	Root	160.46	168.733	154.248	166.665	**0.951**	**0.925**
Tocopherol cyclase, chloroplastic, putative	Leaf	39.298	18.523	8.671	14.389	**2.122**	**0.603**
	Root	12.335	5.929	9.739	3.851	**2.080**	**2.529**
Multicystatin, putative	Leaf	3.172	4.413	1758.974	16.606	**0.719**	**105.924**
	Root	15.022	17.56	1917.627	11.106	**0.855**	**172.666**
Cysteine protease	Leaf	1285.685	1199.723	838.078	926.539	**1.072**	**0.905**
	Root	655.502	659.479	815.193	944.683	**0.994**	**0.863**
Cysteine protease inhibitor	Leaf	583.34	292.939	299.901	581.676	**1.991**	**0.516**
	Root	863.609	392.484	302.151	331.491	**2.200**	**0.911**
Patatin	Leaf	323.944	466.884	257.222	1968.68	**0.694**	**0.131**
	Root	650.96	629.838	222.938	1242.667	**1.034**	**0.179**
Epoxide hydrolase	Leaf	160.007	129.493	144.086	148.308	**1.236**	**0.972**
	Root	199.563	145.526	134.061	137.874	**1.371**	**0.972**
Sucrose synthase	Leaf	221.563	667.782	457.412	1439.103	**0.332**	**0.318**
	Root	1219.608	2344.02	403.49	2511.369	**0.520**	**0.161**

Glyceraldhyde-3-phosphate dehydrogenase	Leaf	2276.764	2911.942	2282.316	1371.491	**0.782**	**1.664**
	Root	2484.921	2653.66	2473.713	2464.899	**0.936**	**1.004**
Phosphoenolpyruvate carboxylase	Leaf	29.747	43.096	10.948	19.132	**0.690**	**0.572**
	Root	62.449	121.201	12.793	106.348	**0.515**	**0.120**
Acidic endochitinase	Leaf	20604.53	6228.288	10113.543	1863.957	**3.308**	**5.426**
	Root	1270.774	203.108	14099.764	197.746	**6.257**	**71.302**
Fructose bisphosphate aldolase, cytoplasmic	Leaf	2718.513	1946.628	2347.292	797.245	**1.397**	**2.944**
	Root	1138.849	1410.507	2616.066	2269.491	**0.807**	**1.153**
Betain aldehyde dehydrogenase, chloroplastic	Leaf	64.889	66.805	80.672	336.3	**0.971**	**0.240**
	Root	208.6	194.552	83.23	216.867	**1.072**	**0.384**
Lipoxygenase	Leaf	320.331	297.642	1034.904	295.722	**1.076**	**3.500**
	Root	1079.631	1570.689	980.472	2264.547	**0.687**	**0.433**
Dehydrin	Leaf	20.27	2.821	1.63	424.147	**7.185**	**0.004**
	Root	662.498	529.83	3.887	378.316	**1.250**	**0.010**
Osmotin-like protein	Leaf	23.455	4.621	3.245	112.254	**5.076**	**0.029**
	Root	303.402	288.085	4.864	199.43	**1.053**	**0.024**
ZF-HD homeobox protein	Leaf	42.698	27.468	31.37	2.43	**1.554**	**12.909**
	Root	2.302	2.462	30.726	5.229	**0.935**	**5.876**
Putative Histone H1	Leaf	1161.041	542.071	704.439	1089.513	**2.142**	**0.647**
	Root	327.099	185.191	683.531	170.561	**1.766**	**4.008**
Putative Transducin beta-like protein	Leaf	50.615	34.506	31.934	49.16	**1.467**	**0.650**
	Root	57.296	55.379	39.148	42.481	**1.035**	**0.922**

Relative expression value of 1.0 = no difference in expression between well-water and drought stressed tissue, >1.0 = up-regulation in stressed tissue and <1.0 = down-regulation in stressed tissue. CDRE = Cachaco dry relative expression, MDRE = Mbwazirume dry relative expression, DREB = dehydration responsive element binding protein, AP2 = apetala 2, NAC = **N**AM, **A**TAF1,2 and **C**UC2, ABA = abscisic acid, CBL = calcineurin B-like, CDPK = calcium-dependent protein kinase, NIP = nodulin intrinsic protein, PIP = plasma intrinsic protein, TIP = tonoplast intrinsic protein, SIP = small and basic intrinsic protein, MAPK = mitogen-activated protein kinase, CDK = cyclin-dependent kinase, LSRP = leaf senescence related protein, SIC-SGP = senescence-induced chloroplast stay green protein and ZF-HD = zinc figure homeodomain.