Genetic Diversity, Evolutionary History and Epigenetic Analysis of East African Highland Bananas

Volume I of I

Mercy Kitavi

A thesis submitted to National University of Ireland Galway

For the degree of Doctor of Philosophy

Academic Supervisor:

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

IITA Supervisors:

Dr. Jim Lorenzen, International Institute for Tropical Agriculture (IITA)

Dr. Morag Ferguson, International Institute for Tropical Agriculture (IITA)

September 2014

TABLE OF CONTENTS

DECLARATION

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

Signed: _____
Mercy Kitavi

ACKNOWLEDGEMENTS

My desire to pursue PhD training was finally fulfilled with funding from the Irish Aid through International Institute of Tropical Agriculture (IITA) to pursue a PhD in the National University of Ireland, Galway (NUIG).

First and foremost, to the Almighty and awesome God.

Special thanks to Drs` Deborah Karamura and Eldad Karamura (Bioversity Uganda), Dr Margaret Onyango (Deputy Director KARI-Kisii) because of their introduction to the *Musa* world, I fell in love with the banana and was able to do this study with a lot of dedication. I am grateful to Dr Leena Tripathi (head of IITA-Nairobi) and the Head of NBRP, Wilberforce Tushemereirwe, for always facilitating the SMTA.

My profound gratitude to my excellent team of supervisors Professor Charles Spillane, Dr Jim Lorenzen and Dr Morag Fergusson for making it possible for me to achieve this dream through their tremendous guidance and critic.

My advisor Dr. Tim Downing contributed enormously to the final shaping of this study. He provided constructive critiques and suggestions in data analysis and interpretation, and comments on the many drafts I gave him. He was always available to attend to my numerous questions despite a very busy schedule. Thanks for teaching me how to do research.

Thanks to the NUIG Botany lab members, IITA-ILRI staff especially Susan and Moses Nyine; Dr Kassa Semagn of CIMMYT; Dr Ateka of JKUAT; BecA-ILRI staff; Jagger, Timothy, Racheal, Ephy and Lucy. To my office mates, ILRI-lab 5 data room, we have come a long way; I would never have had a better career family than you, a big thank you for your assistance and understanding when I needed those quiet moments.

Finally, to my husband Stephen, your love and enormous sacrifices have blessed my life. Daughters; Sherlin, you are very special and Milan, your pretty face and the cartoons you drew for mummy kept her going even when the times were tough. Thanks to my parents for all your prayers and support always. The support from all my 9-siblings, you guys rock. Friends and relatives; Bridgit, Collins and Teddy your contributions are unforgettable.

DEDICATION

Dedicated

To

My Parents

My Husband
Stephen

And
Children
Sherlin and Milan

LIST OF CONFERENCES

ORAL PRESENTATIONS

15/01/2013: Oral presentation at **PAG XX1 2013 *Musa* genomics session** (Plant and Animal Genome *XXI*, 2013) San Diego, USA
https://**pag**.confex.com/**pag**/**xxi**/recordingredirect.cgi/id/535

LIST OF TABLES

LIST OF FIGURES

SUPPLEMENTARY MATERIALS

LIST OF SUPPLEMENTARY TABLES

LIST OF SUPPLEMENTARY FIGURES

LIST OF ABBREVIATIONS

ABI	Applied Biosystems
AFLP	Amplified fragment length polymorphism
AMOVA	Analysis of molecular variance
ANOVA	Analysis of varianve
BPCA	Between principal coordinate analyses
BSA	Bulked segregant analysis
BWA	Burrows wheeler alignment
CGIAR	Collective group of International Agricultural research
CIRAD	Centre de Coopération Internationale en Recherche Agronomique pour le Développement (French Agricultural Research Centre)
CTAB	Cetyl trimethylammonium bromide
DArT	Diversity arrays technology
DNA	Deoxyribonucleic acid
EAHB	East African Highland Bananas
EBSP	Extended Bayesian skyline plot
ESS	Expected sample size
FAO	Food and Agriculture organization
GBS	genotyping by sequencing
GLM	General linear model
GWAS	Genome wide association studies
GS	Genomic selection
HPD	Highest posterior density
HWE	Hardy Weinberg equilibrium
IAM	Infinite allele mutation
IBD	Isolation by distance
IBS	Isolation by state
HKA	Hudson Kreitman Aguade
IITA	International Institute of Tropical Agriculture

INIBAP	International Network for the Improvement of Banana and Plantain
ISSR	Inter simple sequence repeats
IRAP	Inter-retro transposon amplified polymorphism
IPGRI	International Plant Genetic Resources Institute
MAS	Marker assisted selection
MCC	Maximum clade credibility
MCMC	Markov Chain Monte Carlo
MLM	Mixed linear model
MIP	Methylation insensitive polymorphism
MsAFLP	Methylation sensitive AFLP
MSL	Methylation susceptible loci
MSP	Methylation sensitive polymorphism
NGS	Next generation sequencing
NML	No Methylated loci
PCR	Polymerase chain reaction
PCA	Principal coordinate analysis
PLP	Percent loci polymorphic
RAD	Restriction site associated DNA
RAPD	Random amplified polymorphic DNA
RFLP	Restriction fragment length polymorphism
SRAP	Sequence-related amplified polymorphism
STMS	Sequence tagged microsatellite sites
SNP	Single nucleotide polymorphism
SSR	Simple sequence repeat
SSI	State of Sustainability Initiatives
SM	Simple matching
SMM	Stepwise mutation model
TPM	Two phase model
TASSEL	Trait Analysis by *aSSociation*, Evolution and Linkage

tMRCA	the Most recent common ancestor
UPGMA	Unweighted pair group method of averages
VCF	Variant Call Format

SUMMARY OF THE CONTENTS

Genetic variation describes naturally occurring genetic differences among individuals of the same species and permits flexibility and survival of a population in the face of changing environmental circumstances, diseases and pests. Genomic variation develops from a combination of evolutionary influences, among them, mutation process and demographic history. Understanding in greater detail the basis of the tremendous phenotypic variability in East African Highland bananas subgroup that are apparently clonal variants of a single original seedling is essential for developing improved breeding strategies for this subgroup. While genetic diversity studies have included cultivars from this subgroup, intra- population structure and phylogenetic relationships *per se* are still unknown. In addition, none of these studies have attempted to study the evolutionary history and epigenetic polymorphism in this subgroup.

In this thesis, I have used EAHB cultivars to assess the genetic variation, population structure and evolutionary history. I focus on the role of DNA methylation as an epigenetic mark that contributes to phenotypic diversity and determine inheritance of DNA methylation patterns in sexual and vegetative propagation models. The results show that despite being phenotypically distinct, these cultivars are strikingly genetically similar with a narrow genetic base. While DNA methylation polymorphisms are common amongst EAHB cultivars, MSAP does not detect any obvious relationship between DNA methylation variation and phenotypic variation in East African Highland bananas.

This study demonstrates that the EAHB subgroup has low mutation rates, show past population expansion but may have suffered a genetic bottleneck that may have led to the low genetic diversity. Extensive linkage disequilibrium and balancing selection were observed. Finally, I discovered that EAHB cultivars and Zebrina (wild AA cultivar) underwent a speciation event 928 thousand years and their most recent common ancestor dates back 2980 thousand years ago.

The world at present stage demands a very sizeable toll of uniformity for finished product. This of course may not have been the case when the late Neolithic or Bronze Age man made his choice (Simmonds 1962).

CHAPTER 1

INTRODUCTION

1.1 GLOBAL BANANA CULTIVATION

Bananas (*Musa* spp) are grown in more than 120 tropical and subtropical countries, mainly by smallholder farmers. The State of Sustainability Initiatives; SSI-Review (2014) and Perrier *et al.*, 2011, ranked banana as the world's most popular fruit and one of the world's most important staple foods, along with rice, wheat and maize in terms of its importance as a food crop. In 2011, 107 million metric tons of bananas and plantains were produced in more than 130 countries on 0.1 per cent of the world's agricultural area; Africa was rated as the third biggest producer in the world (Figure 1; FAO, 2013). Uganda is a ranked as the world`s second largest banana producer after India. For example in 2012 alone Uganda`s banana production was 9.3 million tonnes and 1.4 million tonnes produced by Kenya.

Such numbers indicate the importance of bananas as a strong commodity, playing key economic and social roles worldwide, even though, more than 85% of bananas are grown for local consumption in tropical and subtropical regions.

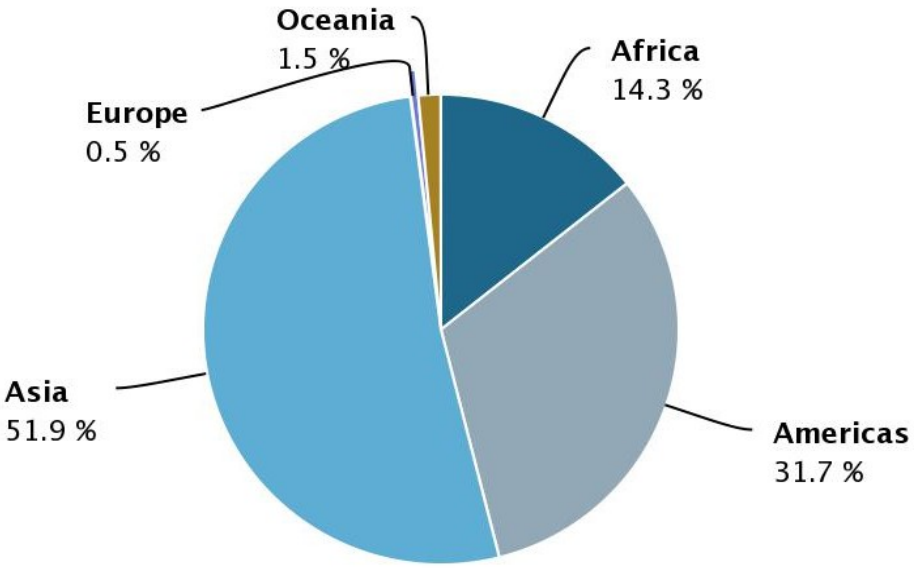


Figure 1: World`s banana production between 1993 to 2013. Source http://faostat3.fao.org/faostat.

Native to the old world tropics from eastern India to the Solomon Islands, bananas are monocotyledonous plants of the Musaceae family that includes the Asian and African genus Ensete, the genetically proximal Asian Musella genus, and the East Asian genus Musa. The genus Musa was formerly taxonomically divided into five sections, *Ingentimusa, Australimusa, Callimusa, Musa*, and *Rhodochlamys;* of which, *Australimusa* (20 chromosomes) and *Musa* (22 chromosomes), included domesticated bananas. Currently, the five sections have recently been reduced to three, *Ingentimusa, Callimusa* (incorporating *Australimusa*) and *Musa* (incorporating *Rhodochlamys*) (Wong *et al.*, 2002; Perrier *et al.*, 2011). Most edible bananas belong to the *Musa* section and are diploid or triploid hybrids from *M. acuminata* (Perrier *et al.*, 2011).

1.2 POLYPLOIDIZATION EVENTS AND THE ORIGINS OF EDIBLE BANANAS

Many crop species of agricultural importance are polyploids (Table 1). Polyploids are organisms having more than two complete sets of chromosomes in their cells. They are common in angiosperms, where at least 70% of the species experienced one or more events of genome doubling during their evolutionary history (Wendel, 2000; Aversano *et al.*, 2012). In plants, polyploidization is considered a major evolutionary force and also a definitive cause of sympatric speciation due to the immediate reproductive isolation between newly formed polyploids and their parents (Hendry *et al.*, 2009).

Table 1: Examples of polyploidy crops. Adapted from Aversano *et al.* (2012). The somatic chromosome number is reported in brackets

Crop	Species name	Chromosome number
Fruit trees	*Prunus domestica*	6× = 48
	Musa spp	3× = 33; 4× = 44
	Citrus aurantifolia	3× = 27
	Actinidia deliciosa	4× = 116
	P. cerasus	4× = 32
Tuber plants	*Solanum tuberosum*	4× = 48
	Ipomoea batatas	6× = 96
	Dioscorea sativa	6× = 60
Cereals	*Triticum aestivum*	6× = 42
	T. durum	4× = 28
	Avena sativa	6× = 42
Forage grasses	*Dactylis glomerata*	4× = 28
	Paspalum dilatatum	4× = 40
Legumes	*Medicago sativa*	4× = 32
	Arachis hypogaea	4× = 40
	Glycine max	4× = 40
Industrial plants	*Nicotiana tabacum*	4× = 48
	Gossypium hirsutum	4× = 52
	Saccharum officinalis	8× = 80
	Brassica napus	4× = 38

The precise origin of edible bananas is not known but the generally accepted theory is that Malesia, a biogeographical region including the Malay Peninsula, Indonesia, the Philippines and New Guinea, was the primary centre of origin and India was a secondary centre of origin (Simmonds & Shepherd 1955). It is likely that dispersal out of Asia was linked entirely to human movement (Daniells *et al*. 2001).

Recent molecular research, plus the outcomes of previous genetic studies, elucidates major stages of banana domestication, such as the generation of edible diploids and triploids (Wong *et al.*, 2002; Perrier *et al.*, 2011). Wild bananas occur from India to Oceania. There are about 50-plus wild species of the genus Musa that are colonizers of rainforest gaps and disturbances. However, their fruit are berries of characteristic banana shape that are full of gravelly hard seeds with little pulp and thus inedible. The *Musa* domestication process started some 7,000 years ago in Southeast Asia (D'Hont *et al.*, 2012). It involved hybridizations between diverse species and subspecies (A-genome alone or from hybridization with *Musa balbisiana* B-genome). Structural heterozygosity of these hybrid AAcvs, caused by chromosomal rearrangements between parental subspecies of *M. acuminata* contributed to gametic sterility. Further consequence of hybrid status was erratic meiosis in edible AAcvs occasionally producing diploid gametes which fused with haploid gametes generating sterile triploid genotypes. Spontaneous triploidizations involved almost all diploid cultivars leading to the formation of cultivated triploids, including pure *M. acuminata* varieties (AAA) and interspecific *M. acuminata* × *M. balbisiana* varieties (AAB, ABB) (Till *et al.*, 2010; Perrier *et al.*, 2011; Pachuau *et al.*, 2014). For instance, *M. acuminata* subspecies zebrina and Banksii derived AA cultivars are thought to have contributed to AAA Highland bananas of East Africa while subspecies banksii derived AAcvs with the BB genome contributed to AAB plantains of West Africa and the Pacific. Human migrations and selection have produced a complex genetic history

leading to the diversity of modern cultivated triploids of diploid and triploid seedless, parthenocarpic hybrids thereafter widely dispersed by vegetative propagation (Figure 2) (Barigozzi, 1986; Perrier *et al.*, 2011; D'Hont *et al.*, 2012).

The geographic distributions of genotypes involved in banana domestication point towards human translocations of plants, most likely under vegetative forms of cultivation, across vast regions (Figure 2). Thus, the hundreds of cultivated varieties are products of centuries, in some cases millennia of clonal (vegetative) propagation. While the introduction to Africa is undocumented, archeological evidence indicates it may have been several thousand years ago (Lejju *et al.*, 2006), probably over 2500 years ago (Perrier *et al.*, 2011).

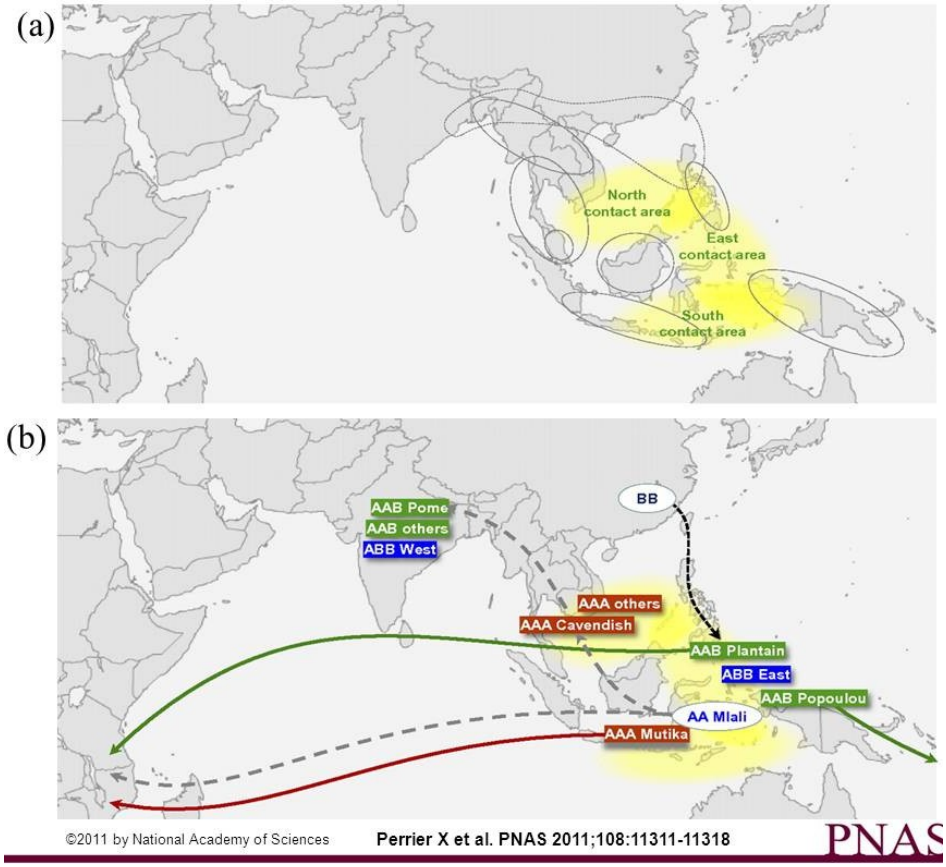


Figure 2: Origins and migrations of the main triploid subgroups (adopted from Perrier *et al.* (2011b). (a) Genetically derived contact areas between *M.*

acuminata subsp. at the origin of cultivated diploids. The three main contact areas: north among malaccensis, microcarpa, and errans; east between errans and banksii; and south among banksii, zebrina, and microcarpa (b) Origins and migrations of the main triploid subgroups. Plain arrows indicate long-term prehistoric migrations of triploid cvs to Africa and Pacific islands. Gray dotted arrows indicate (i) the migrations of Mlali AAcv subgroup, which is not found in Islands of Southeast Asia today, to mainland Southeast Asia, where it contributed to AAA Cavendish, then to India, where it hybridized with *M. balbisiana* to give AAB Pome; and (ii) migrations of the Mlali subgroup to the East African coast. Black dotted arrows indicate the route of *M. balbisiana* from south China to New Guinea over Taiwan and the Philippines, if Austronesian speakers were instrumental in the dispersal of this species.

1.3 IMPORTANCE OF CROP GENETIC DIVERSITY

Genetic diversity is the sum of the genetic characteristics within any species or genus (Rao & Hodgkin, 2002; Rauf *et al.*, 2010) and genetic variability explains the variation within these genetic characteristics. The loss of biodiversity is considered one of today's most serious environmental concerns by the Food and Agriculture Organization of the United Nations (FAO, 2002). Losses of genetically encoded characteristics in any population may limit its chances of survival (Trethowan & Mujeeb-Kazi, 2008). As the global population rises to nine billion by 2050, the loss of biodiversity will have a major impact on humankind's ability to feed itself (FAO, 2010). Generally, it is perceived that the overall genetic diversity in crop species has been reduced by a range of factors such as urbanization and replacement of traditional agriculture systems by modern farming methods (Rauf *et al.*, 2010).

There are several evolutionary processes that can impact the genetic diversity of natural populations. These are; spontaneously arising mutations; gene flow via migration; inbreeding; natural selection; the Wahlund effect; and random genetic drift (Porth & El-Kassaby, 2014). Genetic drift refers to any random activity (e.g. matings, environmental disasters) that introduces change in the allele frequency in a population by causing a certain allele (gene) to become

more popular or less popular in a population. Gene flow happens if a population interbreeds with another population (e.g. individuals start migrating between the two populations, or removal of some barrier, like a river, disappears between the two populations) causing change in the overall percentages (frequencies) of the alleles in one population, or both. Mutation does not cause change in allele frequencies (evolution) by itself, but only in combination with any of the other three forces; genetic drift, gene flow, or natural selection. Mutation introduces a new allele into an individual and therefore into the population but no change occurs until that mutation spreads into the population by genetic drift, gene flow, or natural selection. Natural selection is the best known and most important cause of evolution. If a given allele (gene) produces some advantage in survival or reproduction, then the individuals born with that allele will tend to have more offspring in their lifetime. So since alleles are inherited by offspring, this would cause the percentage of the population that has that allele to go up over time. Wahlund effect refers to reduction of heterozygosity in a population caused by subpopulation structure which in turn could be geographic barriers to gene flow followed by genetic drift in the subpopulations (Garnier-Géré & Chikhi, 2001). Genetic drift introduces random changes in allele frequencies over generations and becomes important for finite population samples and/or a large number of generations. These random allele frequency changes can, over time, lead to allele fixation or extinction. By all means, genetic drift represents a source of differences in genetic diversity among different populations. On the other hand, gene flow evens out among-population genetic differences, but increases genetic variation within populations, due to the introduction of new alleles.

1.3.1 Genetic diversity is important for human livelihood resilience

Due to ongoing climate change, some of the genetic diversity of the plants that we grow and eat could be lost forever, threatening future food supply (FAO, 2010a). According to McGrath (2012) (CCAFs policy report by Phil Thornton & Laura Cramer), for bananas and plantains, climate change may significantly alter yields, as well as vulnerability to diseases affecting the food security and incomes of millions of Africans and Latin Americans. They predicted that potato, which grows best in cooler climates, could also suffer as temperatures increase and weather becomes more volatile, and that banana and plantains may be a good substitute for potatoes in certain locations.

Climate change, by affecting temperature, can change the environmental conditions under which pests reproduce. The largest banana exporter, Costa Rica, recently declared a national emergency after the nation's crop was devastated by two separate outbreaks of mealy bugs and scale insects. The mealy bugs weaken the plants' overall health and disfigure the fruit causing exporters to reject up to 20 percent of it for shipment, a problem that is not restricted to Costa Rica or Central America (Morcroft, 2013). The wise use of crop genetic diversity in plant breeding can contribute significantly to protecting plant health, food production and the environment.

Crop varieties that are more resistant to pests and diseases can reduce the need for pesticide application, while more vigorous varieties can better compete with weeds (Kropff, 2001-2005) reducing the need for applying herbicides. Drought resistant plants can help save water through reducing the need for irrigation (Nautiyal & Kaechele, 2007); deeper rooting varieties can help stabilize soils; and varieties that are more efficient in their use of nutrients require less fertilizer (Smith, 2008). Most importantly, perhaps, productive agricultural systems reduce or eliminate the need to cut down biodiverse forest areas or clear fragile lands to create more farmland for food production.

1.3.2 Disease threats to crops with low levels of intra-specific genetic diversity

Species or population with a wide range of genetic variability are better positioned to survive in the presence of a stressor or disturbance. Monocultural agricultural approaches have a tendency to be based on low levels of crop varietal diversity (especially if the varieties were mass-produced or cloned). Monocultural stands of the same or very similar crop varieties makes the crop vulnerable to diseases (especially hypervariable fungi or viruses). It is possible that a single pest or disease strain could wipe out entire areas of a crop due to this genetic uniformity in the farmers fields (Martinez-Castillo, 2008). Historically known examples of harvests that severely suffered from low crop genetic diversity was the Irish Potato Famine of 1845-1847 caused by *Phytophthora infestans*; the US corn fungus disease of 1970 that caused a loss of over one billion dollars in production and the banana-killing fungus, *Fusarium oxysporum* f. sp.*cubense* that largely wiped out the cultivation of the previous Gros Michel variety of bananas in the 1950s. In the 1980s, dependence upon a single type of grapevine root forced California grape growers to replant approximately two million acres of vines when a new race of the pest insect, grape phylloxera (*Daktulosphaira vitifoliae*), attacked.

Of interest to this study, is the growing danger to present day agriculture caused by the *Fusarium oxysporum* f. sp.*cubense* that now threatens the Gros Michel successor, the Cavendish banana. The fungus, previously believed to be isolated to Australia and certain Asian venues has now been found in banana growing regions in Mozambique and Jordan and threatens to spread worldwide, including areas where it had not been reported before http://www.scientificamerican.com/article/banana-threatening-fungus/.
Genetically homogeneous crops also makes agriculture more vulnerable to major threats like drought, insect pests and diseases, which may become worse in many parts of the world as a result of climate change.

1.3.3 Crop diversity and the economy

Agriculture is the economic foundation of most countries, and for developing countries the most likely source of economic growth. Economic growth is most rapid where agricultural productivity has risen the most and the reverse is also true. Although beneficial for the wider economy, growth in agriculture benefits the most poor. Provision of affordable food extends these benefits beyond the 70% of the world's poorest people who live in rural areas and for whose livelihoods agriculture remains central (Yares, 2007). To ensure that agriculture plays this fundamental role in economic and social development, particularly for subsistence and very low income farming families, a range of improvements are required; among them is the growing of higher value crops. Fundamental to all of these efforts is crop (genetic) diversity. Crop genetic diversity enables farmers and plant breeders to develop higher yielding, more productive varieties having improved quality characteristics required by farmers and desired by consumers. They can produce varieties that resist pests and diseases and are drought tolerant, providing more protection against crop failure and better insulating poor farmers from risk (Smale, 2006). Agriculture's part in fighting poverty is complex, but without the genetic diversity found within crops, it cannot fulfill its potential.

1.4 ASSESSMENT OF GENETIC DIVERSITY IN *MUSA* SPECIES

Historically, three major areas have been important for molecular marker applications for plant improvement: (a) the determination of genetic diversity within and among populations; (b) verification and characterization of genotypes; and (c) marker-assisted selection (MAS). In the past, morphological and taxonomic systems have been used for differentiation of specific banana clones (Gibert *et al.*, 2010; Samarasinghe *et al.*, 2010). Currently, a range of molecular markers are being employed to investigate banana diversity. While

this is not a review paper on the topic, a range of molecular markers have been extensively applied in *Musa* studies and have sustained, and sometimes refined, the agro-morphological classification. These include studies involving random amplification of polymorphic DNA (RAPD) (Suttada *et al.*, 2007; Venkatachalam *et al.*, 2008), restriction fragment length polymorphism (RFLP) (Bhat *et al.*, 1994; Ning *et al.*, 2007), inter simple sequence repeats (ISSR) (Rout *et al.*, 2009; Lu, Y *et al.*, 2011), inter-retrotransposon amplified polymorphism (IRAP) (Häkkinen *et al.*, 2007), and sequence-related amplified polymorphism (SRAP) (Youssef *et al.*, 2011).

Simple Sequence Repeats (SSR also known as sequence tagged microsatellite sites - STMS) and Amplified Fragment length polymorphism (AFLP) markers have been popular molecular markers (Teixeira *et al.*, 2014), quite extensively used to study *Musa* diversity with some efficiency (Wong, 2001; Creste *et al.*, 2003; Ude *et al.*, 2003a; Amorim *et al.*, 2008; Resmi *et al.*, 2011; Hippolyte *et al.*, 2012; Mbanjo *et al.*, 2012; de Jesus *et al.*, 2013). Although AFLPs are anonymous markers, they have longer +1 and +3 selective primers and the presence of discriminatory nucleotides at 3′ end of each primer making the level of their reproducibility and sensitivity very high (Mammadov *et al.*, 2012). They can detect a large number of polymorphic bands in a single lane rather than high levels of polymorphism at each *locus* as in SSR methods. Disadvantages of this technique are that alleles are not easily recognized, has medium reproducibility, lengthy, not amenable to automation and has high operational and development (Abdel-Mawgood, 2012). Due to their binary nature, they possess inability to distinguish heterozygotes from homozygotes therefore have lower sensibility in detecting informative genotypic classes (Garcia *et al.*, 2004). However the high numbers of polymorphic loci revealed by AFLP methods counterbalance the loss of information resulting from dominance (Gerber *et al.*, 2000). SSRs are short stretches of DNA sequence occurring as tandem repeats of mono-, di-, tri-, tetra-, penta- and hexanucleotides. Due to mutation affecting the number of repeat units they are highly polymorphic and informative markers. In addition to their genetic co-

dominance, the value of SSRs is due to their abundance, dispersal throughout the genome, multiallelic variation, high reproducibility and require tiny amounts of tissue and can work on degraded or "ancient" DNA (Appleby *et al.*, 2009; Senan *et al.*, 2014). Despite their efficiency, SSR mutation rates are too high causing changes in conserved regions between species and homoplasy (alleles identical in state not identical by origin) (Barkley *et al.*, 2009).

Most recent types of molecular markers in *Musa* research have included DArT (diversity array technology) (Hippolyte *et al.*, 2010), discovery of Single Nucleotide Polymorphisms (SNPs) (Till *et al.*, 2010) and also markers at the chromosomal level e.g. molecular cytogenetics (Heslop-Harrison & Schwarzacher, 2007; Hribova *et al.*, 2011). A SNP represents a single nucleotide difference between two individuals at a defined location and is the ultimate form of molecular genetic marker, as a nucleotide base is the smallest unit of inheritance (Appleby *et al.*, 2009). The sequence information of SNPs provides the exact nature of the allelic variants and have a major impact on how the organism develops and responds to the environment (Appleby *et al.,* 2009). SNPs are abundant in plant systems (one SNP every 100–300 bp) and are evolutionarily stable, not changing significantly from generation to generation (low mutation rates). They are therefore excellent markers for studying complex genetic traits and as a tool for understanding genome evolution (Syvänen, 2001; Mammadov *et al.*, 2012). Even though SNP's have several advantages over other technology they have limitations to their discovery in the non-model organism (Abdel-Mawgood, 2012). SNP markers are very challenging in terms of bioinformatics and need huge data storage space.

1.5 THE EAST AFRICAN HIGHLAND BANANAS

The East African Highland banana (EAHB) is one of the highly valued staple food crop, providing starch for over 80 million people (CCAFS Report 2011)

in the East African Great Lakes region (Uganda, Tanzania, Burundi and eastern DR Congo (Figure 3) (Karamura, 1998; Bagamba *et al.*, 2010). They are a distinct triploid subgroup of the *Musa acuminata* species that grow at altitudes between 900 and 1900 m above sea level of the East African plateau (Simmonds, 1966). The importance of the EAHB is reflected in the diversity of uses and the number of cultivars recognized by the local people in the East African region. The vast majority of producers are small scale farmers growing the crop either for home consumption or for local markets. These EAHB bananas contribute significantly to food and income security of people engaged in its production and trade (Figure 4).

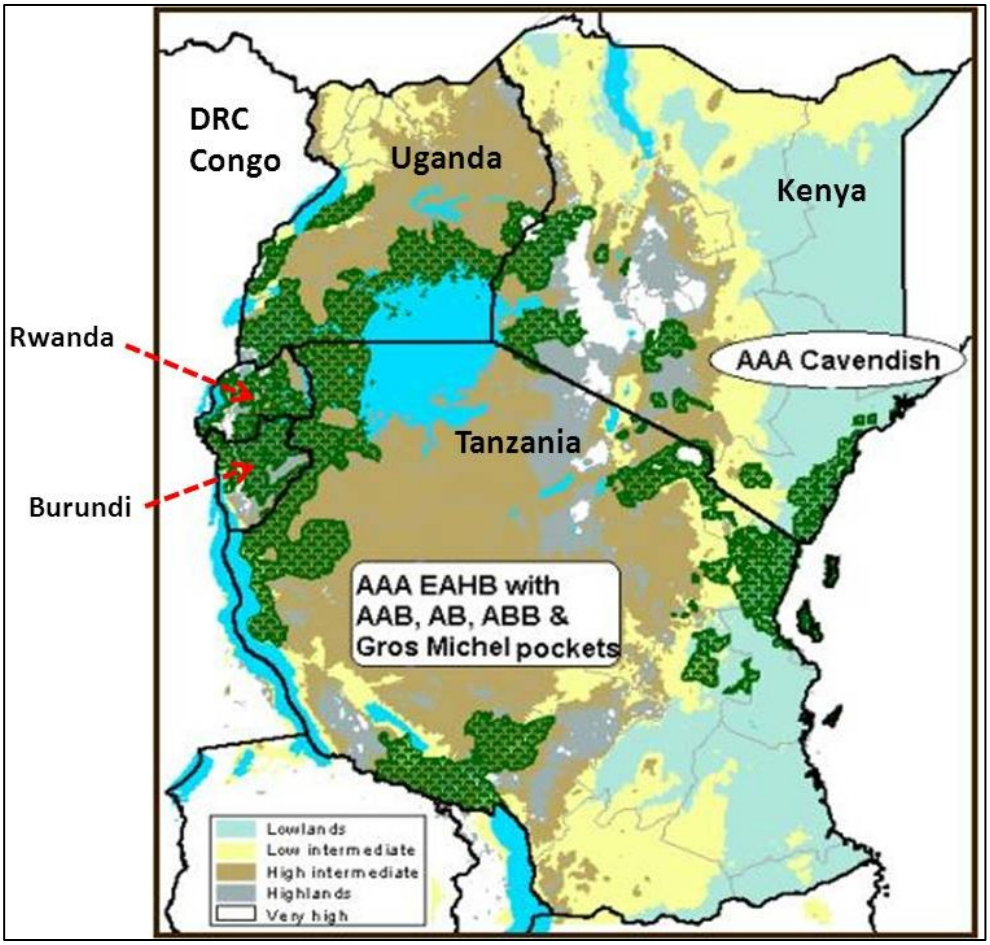


Figure 3: Principal banana growing areas of East Africa with *Musa* genome differentiation (Edmeades *et al.*, 2005).

Figure 4: East African Highland bananas play a role as an income source for small scale farmers, mostly women. Keumbu market in Kisii county (Kenya) is a well-known place that acts as banana collection point for middlemen traders and travellers going to the Kenya`s Capital -Nairobi.

The East African Highland bananas are said to be endemic (restricted) to the East Africa region with no clear analogue elsewhere in the world. The endemic clones, collectively termed as 'Lujugira-Mutika' subgroup (AAA-EA) consist of both cooking (locally known as '*matooke*') and beer ('*mbidde*') bananas (Karamura, 1998; Pillay *et al.*, 2001; Perrier *et al.*, 2011). Although the highland bananas have an AAA genome composition (Shepherd, 1957) they are distinctly different from the dessert or sweet bananas that have similar triploid genomes (Pillay *et al.*, 2001), and EAHBs are identified by their characteristic features (Figure 5). A remarkable morphological diversity of bananas (*Musa* spp) exists in the East Africa Great Lakes plateau, with at least

84 unique farmer selected, locally evolved clones (Karamura, 1998b; Bagamba *et al.*, 2010). Thus, East African region has been considered a secondary centre of *Musa* diversity (Tugume *et al.*, 2002).

Figure 5: Diagnostic characteristics of the Lujugira-Mutika subgroup.

1.5.1 Genetic diversity inferred within the East African Highland Banana subgroup

For decades, breeders have estimated genetic diversity on the basis of data generated by different molecular markers, providing a means of rapid analysis of germplasm and estimates of genetic diversity often found to corroborate phenotypic data. While this is not a review of the topic, we begin our study of the genetic diversity in the East African Highland bananas with a brief historical retrospect concerning marker types that have been employed for studying genetic variation in this subgroup.

The first type of markers, (and the most easily accessible type of plant marker), are morphological markers that can be monitored based on simple inheritance (Porth & El-Kassaby, 2014). The taxonomy of triploid cultivars was first studied by Simmonds and Shepherd (1955) using 15 morphological characters, and thereafter using chromosome counts (basic chromosome number) (Ude *et al.*, 2003a; Christelova *et al.*, 2011). Based on 73 morphological traits and fruit quality attributes (standardized in the *Musa* descriptors reference list (IPGRI-INIBAP-Bioversity, 2003) the EAHBs were for the first time extensively characterized and classified into five clone sets; Nfuuka, Musakala, Nakabululu, Nakitembe and Mbidde (beer) (Karamura, 1998).

Plant phylogenetic and diversity studies have successfully exploited relatively low marker densities or regional markers to determine relationship in plants at the interspecific and intraspecific levels (Nybom, 2004; Arif *et al.*, 2010; Deschamps *et al.*, 2012). A range of molecular techniques have facilitated the classification of new banana cultivars, and have played a big role in sustaining and sometimes refining the morphological classification of bananas over the past decades (Hippolyte *et al.*, 2012). Each molecular marker technique is based on different principles but their generic application is to harness genome-wide variability.

The first study to make use of molecular markers to examine genetic diversity in East African bananas was done by Pillay *et al.* (2001) using RAPDs to characterize 29 EAHBs. It was concluded that the EAHBs were closely related with a narrow genetic base (Cooper *et al.*, 2000). Using AFLP, Tugume *et al.* (2002) reported low levels of DNA diversity of 115 EAHB cultivars, however, his molecular classification somehow matched the morphological characterization of Karamura (1998). Indications from preliminary data generated by CIRAD (in context of the CGIAR Generation Challenge Program) seem to indicate that most of the triploid East African highland banana cultivars are nearly identical with regard to SSR markers. The CIRAD-generated dataset (of about 22 SSR loci for a dataset of 550 genotypes that

included ~20 EAHB varieties) indicated that the EAHB varieties clustered very tightly (i.e. were very closely related). Similarly, IITA Uganda found no differences for 10 SSR loci in a set of 16 EAHB genotypes that represented the 5 major clone sets identified by Karamura (1998). Other studies have included EAHB cultivars to represent out-group taxa or as a genomic group (AAA) and have corroborated the above findings (Ude *et al.*, 2003a; Christelova *et al.*, 2011; Changadeya *et al.*, 2012; Hippolyte *et al.*, 2012; de Jesus *et al.*, 2013). However, contradictory results regarding the diversity of the EAHB have also been reported based on 24 SSR markers (Buwa, 2009).

Early and recent reports have suggested the origins of all the East African Highland banana group cultivars from a very ancient single introduction (Pillay *et al.*, 2001; Perrier *et al.*, 2011). However, a large amount of variation in inflorescence characters, fruit shape, plant size and several other characteristics used for classifying germplasm (Tezenas du Montcel et al. 1983) are observed in this subgroup even between sister clones (Figure 6).

Figure 6: Phenotypic variations in inflorescence and bunch types observed within EAHB. a, b, c and d are major characteristic which classify the EAHB clonesets Musakala, Nfuuka, Nakabululu and Nakitembe with the fifth (Mbidde) having members from the other four clonesets but members have fruits with an astringent taste, therefore used in brewing. e, f, g, h, I and j are cultivars of the same cloneset (Nfuuka) while g and j are sister clones of the same mother plant.

The variation in the EAHB germplasm is speculated to be as a result of somatic mutations and subsequent preferential cultivation of the mutants (Simmonds, 1966; Karamura, 1998; Ude *et al.*, 2003a; Perrier *et al.*, 2011) and/or transposon activity as in the case of clonally propagated citrus plant species (Asíns *et al.*, 1999). Although transposable elements account for almost half of the *Musa* sequence (D'Hont *et al.*, 2012) their evolutionary role has not been fully investigated and their involvement in inducing genetic variability has not been demonstrated in *Musa* but cannot be discounted. The contrasting physical features and climates of East Africa and the social backgrounds of the East Africa region are also suspected to have played a role in the diversification of the different clones (Karamura, 1998).

Although a morphological and taxonomic systems allow for differentiation of specific banana clones (Samarasinghe *et al.*, 2010), insufficiencies of this approach start to emerge as the genetic basis of the plants under study gets narrow (Christelova *et al.*, 2011). Additionally, morphological diversity measures are rather complex, few and often underestimate or overestimate the actual amount of genetic diversity (Resmi *et al.*, 2011). Evaluation based on phenotypic data is also laborious and takes years to draw conclusion (Rauf *et al.*, 2010b). Morphological markers are also affected by the environment, so some have rather low heritability. Furthermore, a classification system that relies exclusively on the phenotypic manifestations of the genome obviously suffers from limited accuracy (Langhe *et al.*, 2005; Noyer *et al.*, 2005; Abdullah *et al.*, 2012) but can be made robust if supported by molecular-based characterization; either dominant (RAPD, AFLP and ISSR) or codominant (allozymes, RFLP and SSR, SNP) markers.

Until the 1970s, East African highland bananas were traditionally grown in Central Uganda. But EAHB production has since declined due to pests and diseases (Nyombi, 2013). This decline has led to the replacement of cooking bananas by exotic banana cultivars and annual food crops. Efforts to improve the EAHB bananas require good male parents and an assessment of their male fertility. This is because use and transfer of useful traits (genes), such as resistance to pests and pathogens, to banana cultivars from other cultivars or from wild relatives is greatly hampered by sterility/crossability problems with many banana lines. Banana breeding is exceedingly difficult. For example, cultivars of the important Cavendish group have remained female sterile even in cases involving pollination of hundreds of bunches (Shepherd, 1999).

Many crop improvement programs are based on an approach whereby useful genes/traits are identified in related cultivars or wild relatives, which can then be used in the conventional cross-breeding programmes to develop improved hybrids. Cultivated bananas are primarily triploid with 2n = 3x = 33

chromosomes, though accessions with diploid and higher ploidy levels (tetraploid) also exist.

Conventional crossing approaches have been successful in producing inter- and intra-specific banana hybrids by minimizing infertility barriers. For example, desired traits can be introduced from one triploid to another through genetic bridging. Improvement of triploid Musa species has been achieved through crossing 3x landraces with 2x (diploids) wild or improved lines, to produce 4x tetraploids that generally display greater male and female fertility. Selected tetraploids are then crossed with improved diploids to produce sterile secondary triploids (Pillay *et al.*, 2002). Ultimately, the success in banana breeding relies on the identification of female fertile landraces.

The *Musa* breeding scheme of IITA aims to produce seedless (parthenocarpic) hybrids, preferably in the triploid background. This usually involves crossing triploid cultivars with fertile diploids to produce tetraploids that generally display greater male and female fertility. Selected tetraploids are then crossed with improved diploids to produce sterile secondary triploids. Recurrent diploid breeding is done by intercrossing improved diploids (Tenkouano & Swennen, 2004). Inheritance studies in 3x-2x or 4x-2x cross-breeding suggested that traits of economic importance (e.g. yield) are more predictably inherited from the diploid parent than from parents with a higher ploidy status (Tenkouano *et al.*, 1998, 1999). The *Musa* breeding scheme of IITA is summarized in Figure 7.

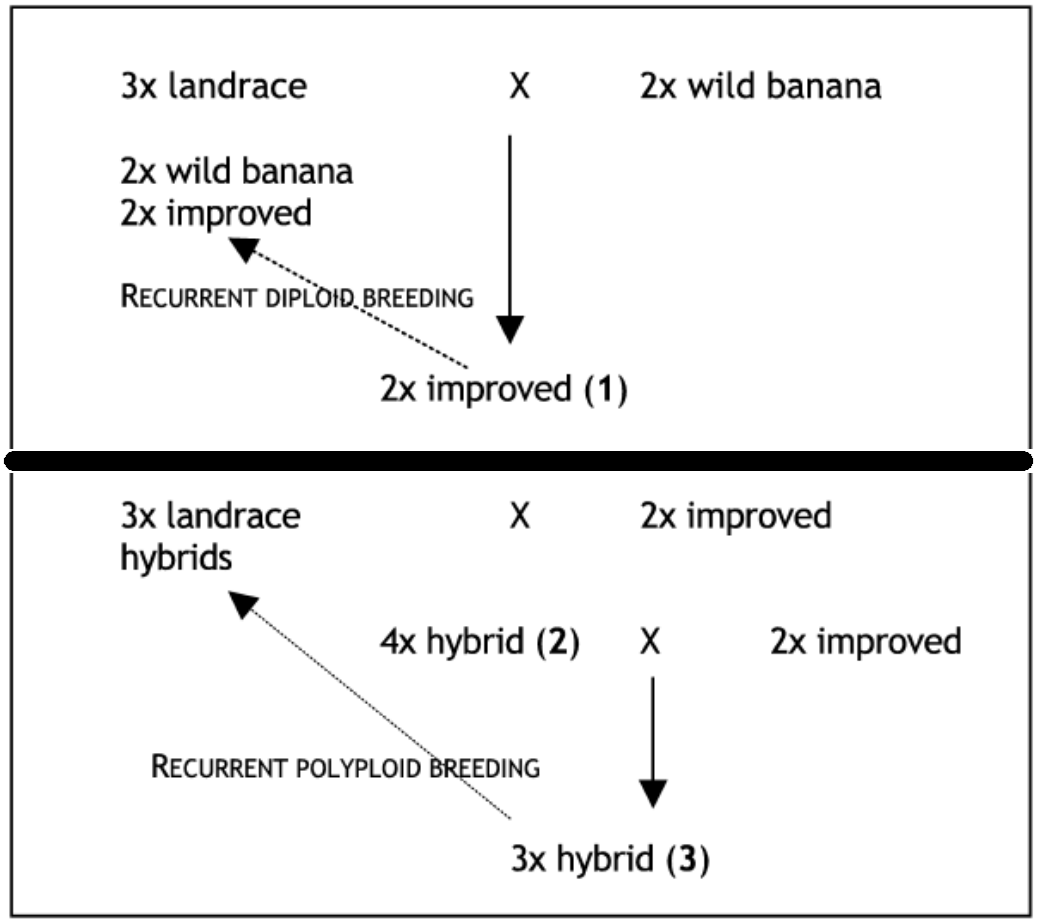


Figure 7: IITA's Musa breeding scheme. (1) Production of improved diploids; 2) production of resistant hybrids (preferably tetraploids) from East African highland bananas and other Musa cultivars; 3) production of secondary triploids from the tetraploids with improved diploids as male parents.

Because of the polyploid nature of banana, its complex cytogenetics (Shepherd, 1999), its physically large size, long life cycle (10 to 15 months of frost-free weather) and also issues relating to sterility, the improvement of banana cultivars through breeding is very challenging. In particular, among the factors hampering banana breeding are low male and female fertility and/or male and female sterility resulting in very low seed set per bunch (Vuylsteke & Ortiz, 1995; Ssebuliba *et al.*, 2006). For instance, the IITA Uganda program can currently generate about 1000 progeny plants per year from >20,000 F1 seeds that are obtained from a full-time crew of 6 field workers responsible for

pollination. The scarcity of suitable male (pollen) parents harbouring genes of interest is also a contributing factor to the slow pace of the genetic improvement of banana (Pillay personal observation). Banana breeding programs spend considerable time in either developing or identifying suitable male parents with characteristics of interest for breeding. Furthermore, the occurrence of somatic mutants for some cultivars coupled with many local names, synonyms and homonyms has limited the full knowledge of the available EAHB genetic resources (Pillay *et al.*, 2001).

Detailed information on the genetic diversity and phylogenetic relationships within the East African banana germplasm remains scarce. An understanding of the genetic relationships is essential to develop an efficient breeding program by providing basic information for breeders, such as the selection of suitable material for new cultivar development. Advances in *Musa* breeding over the past decades have established that crossing of divergent genotypes and subsequent selection of improved hybrids are important steps for the production of new banana cultivars (Ortiz & Vuylsteke, 1996). Further information and knowledge which can be generated to facilitate banana breeding/crop improvement will lead to a more diverse basket of choices of EAHB cultivars for smallholder farmers in East Africa to make their selections from.

1.6 PROBLEM STATEMENT AND JUSTIFICATION

The existing EAHB cultivars pose a challenge to the development of scientific banana breeding strategies. At present it is difficult to understand why there is tremendous phenotypic variability among old cultivars that are apparently clonal variants of a single original seedling. The current working hypothesis is that they are probably clonal derivatives. To develop improved breeding strategies for East African Highland Bananas, it will be necessary to understand in greater detail the basis of these large phenotypic differences in

presumed clonal variants. Such knowledge will allow the breeders to make more informed decisions about which parents to include in crossing programs. For example, EAHB varieties differ tremendously in female fertility (Ssebuliba *et al.*, 2006). If economically important differences between EAHB varieties are due to epigenetic factors (such as DNA methylation and histone modification) that are not stably transmitted to progeny, there is no reason to try to cross those varieties that have attractive bunches but low female fertility; rather the focus should be on the most female-fertile varieties. However, if clonal differences are stably transmitted to progeny, then significant effort should be made to generate seeds from the most desirable varieties, even though their female fertility is low.

Advances in *Musa* breeding have established that crossing of divergent genotypes and subsequent selection of improved hybrids are important steps for the production of new banana cultivars (Pillay *et al.*, 2001). Reliable identification and genetic information regarding the existing EAHB genetic resources will be useful for the effective breeding and conservation strategies. Noyer *et al.* (2005) and Ganapathy *et al.* (2011) have also highlighted the importance of a solid understanding of the genetic diversity available breeding material resources as prerequisite for setting up an efficient strategy for breeding improved banana varieties that support the choice of crossing parents. The IITA breeding scheme utilizes the current EAHB germplasm, cultivars with moderate sterility as female parents for breeding against diseases; black sigatoka, bacterial wilt and pests specifically nematodes. Different authors have shown the value of genetic diversity providing new options to combat different biotic and abiotic stresses. Therefore knowledge of the genetic diversity in cultivated EAHB is important for pest and disease management and provides important options for further improvement of the species.

There is no current report on the population genetic structure and demographic history of the EAHBs. Examining the intra-population variation and understanding its causes can provide insights regarding the particular selection pressures that drive phenotypic divergence among populations. This accentuates the need to collect, characterize, and document germplasm before its extinction from these areas.

A combination of diverse marker types is usually recommended to provide an accurate assessment of the extent of intra- and inter-population genetic diversity of naturally distributed plant species, on which proper conservation strategies for species that are at risk of decline can be developed (Porth & El-Kassaby, 2014). Due to their codominant nature, SSR markers are known to be superior to AFLP which is a dominant marker. However the development cost of SSR markers is high and time consuming (Fu *et al.*, 2014) limiting research on EAHB subgroup. We therefore took the advantage of newly developed SSR markers from *Musa acuminata* Malacensis subsp (AA) (Mbanjo *et al.*, 2012) to genetically characterize the EAHB. On the other hand, AFLP markers can detect a large number of polymorphic bands and are genome wide. Based on this reason and due to the unavailability of *Musa* species reference genome (for genome wide markers) in the beginning of this study, AFLP was chosen. In this PhD we (i) used SSRs to assess the genetic variation and relationships and determine the genetic basis of the EAHB morphological clonesets; (ii) distinguish between the EAHB morphotypes and assess the population structure using AFLP; (iii) evaluate nucleotide polymorphisms and selection signatures through genome-wide SNP analysis; (iv) use Methylation sensitive Amplified Fragment Length Polymorphism (MsAFLP) to determine if epigenetic diversity would mirror genetic or morphological diversity observed in the EAHB subgroup; and finally (v) investigate inheritance of methylation patterns/polymorphisms among EAHB breeding material both sexually and asexually generated offspring.

Even though this research does not link the morphological diversity to either genetics or epigenetics, to our knowledge, this is the first report on extensive genome wide nucleotide diversity analysis, linkage disequilibrium and signatures of selection in the EAHB subgroup using SNP markers. The report on the epigenetic polymorphism and their inheritance will also be the first for this subgroup of bananas. While other studies have used SSR and AFLP for diversity studies of this EAHB subgroup, we present the first report on SSR mutation rates and genetic ancestry of the East African Highland Bananas. Information generated from this PhD will be used by the breeder in improvement and efficient breeding of the EAHBs to create superior hybrids. This will in the long run translate to higher yields, disease and pest resistance and improved nutrition for the population of East Africa.

CHAPTER 2

SIMPLE SEQUENCE REPEAT (SSR) MARKER ANALYSIS REVEALS LOW GENETIC VARIATION OF THE EAST AFRICAN HIGHLAND BANANAS

Background

East African Highland bananas (EAHB, colloquially 'Lujugira-Mutika') are a distinct group of agriculturally important bananas that dominate the Great Lakes region. Little is known about their genetic variation, population structure and recent evolutionary history.

Methods and Results

Ninety triploid AAA-genome EAHBs were genotyped with 100 Simple Sequence Repeats markers to assess their population genetic diversity, the correlation of genetic variability with current morphological classes, and the number of origins since EAHB introduction into Africa. Population-level statistics for these 90 EAHB were compared to those for six Plantain (AAB) and Dessert (AAA) cultivars that reflected subspecies-level diversity. Little variability was observed in the 90 cultivars and no genetic differentiation was seen among the morphological clonesets. Although EAHBs sampled in Uganda showed marginally more genetic variation than ones from Kenya, there was little differentiation between these regions. Investigation of genetic diversity and population structure highlighted a single origin of these 90 cultivars in the context of much more diverse Plantains and Dessert samples, one of which has potential hybrid origins. The homogeneous modern EAHB has a small ancestral population size, which has only recently expanded.

Conclusion

Our study not only demonstrates the genetic uniformity of the East African highland cultivars, but also demonstrates that there is no differentiation among the morphological clonesets. Importantly, our results indicate that the recent introduction and expansion of the EAHB cultivars in Africa has caused a clonal pattern of genetic diversity. This could be addressed by exploiting extensive variation observed in other Musa subspecies.

Keywords: East African Highland bananas (EAHB), Simple sequence repeats, population structure, domestication, parthenocarpy.

2.1 INTRODUCTION

Bananas (genus *Musa*, family *Musaceae*) are monocotyledons from the *Zingiberales*, a sister group of the *Poales* and are of major economic importance (food security and income) in many tropical and subtropical countries (Martin *et al.*, 2013). Most cultivated bananas are triploids derived from spontaneous hybridization between diploid *M. acuminata* subspecies containing an "A" genome and other diploid *Musa* species containing "B" genome or other A versions (Perrier *et al.*, 2011) . Molecular approaches have provided insight into some of the genomic events coinciding with visible changes in phenotype in highly inbred plants (Kempinski *et al.*, 2013). However, neither the genetic origins of domesticated or wild triploid AAA and AAB *Musa*, nor their molecular mechanisms of phenotypic variation are sufficiently studied given their agricultural importance.

The East African Highland bananas (EAHB) ('Lujugira-Mutika' subgroup) dominate the Great Lakes region of East Africa (Karamura, 1998) with no clear analogue globally (Pillay *et al.*, 2001). This region is regarded as a secondary center of banana diversification (Tugume *et al.*, 2003). Edible AA *M. acuminata* subspecies zebrina and banksii have recent common ancestry with AAA EAHB, all of whom have a shared diploid AA ancestor (Li *et al.*, 2013). *M. acuminate* subspecies banksii also is related to the AAB Plantains, who have "B" genome ancestry stemming from *M. balbisiana* (Perrier *et al.*, 2011). We refer to bananas and plantains collectively as bananas here.

A lack of historical records and archaeological evidence has led to limited tracking of the historical movement of bananas from South East Papua New Guinea, but the geographical ranges of modern diploid parents and triploid hybrids support numerous ancient migrations both before (in diploids) and after hybridization (in triploids). Current evidence suggests a single introduction of AA or AAA varieties 'Lujugira-Mutika' into Africa, perhaps

about 2.5 kya, separate from the origins of AAB Plantains (Perrier *et al.*, 2011). It is not well established how wild bananas became domesticated, but it is possible that the accumulation of sterility and acquisition of parthenocarpy with the increase of pulp mass and the absence of seeds, followed by human selection, gave rise to the modern predominantly sterile cultivars (Perrier *et al.*, 2009; Perrier *et al.*, 2011; de Jesus *et al.*, 2013). Like most cultivated bananas, in the post-domestication period strictly vegetative propagation (i.e., cloning) of EAHB cultivars over long periods has likely led to somaclonal variants (Perrier *et al.*, 2011). In fact, the current entire EAHB across Uganda, Kenya, western Tanzania, Rwanda, Burundi and eastern Congo comprise of about 120 cultivar names (Karamura, 1998; Ssebuliba *et al.*, 2005). However about 200 cultivars are recognized in the region since a number of local names may exist for the same cultivar, and many cultivars cannot be easily distinguished on the basis of their morphology, especially if they are closely related (Pillay *et al.*, 2001).

Higher plant genetic diversity can reduce the impact of biotic and abiotic stresses, and does provide important raw material for breeding (Rauf *et al.*, 2010). The tremendous loss of genetic diversity among cultivated crops and wild relatives is alarming (Hopkins *et al.*, 2013). *Musa* has a limited potential for producing genetically diverse offspring because it propagates mainly vegetatively (through suckers), like many parthenocarpic crop plants. The lack of *Musa* varieties that are resistant to disease and the limitations of breeding present an extinction threat for the banana (Li *et al.*, 2013). Of major concern is the *Fusarium* wilt (or panama) disease caused by strains of a soil fungus *F. oxysporum cubense* (Foc) that previously eliminated the Gros Michel cultivar, which was the main exported banana variety from the nineteenth century until the 1950s. In response, Gros Michel was replaced with the Cavendish variety, which was resistant to that Foc strain. Worryingly, the Cavendish is susceptible to a new Foc Tropical Race 4 (Foc-TR4) strain, and could meet the same fate as the Gros Michel variety. Both Cavendish and Gros Michel are derived from the same source: a mating of a Mlali subgroup 2N (AA) gamete and Khai

subgroup N (A) gamete (Perrier *et al.*, 2011). The fungus was first detected in Asia in the 1990s, during which it also spread to a region of Australia – recently has been reported in Jordan (Garcia *et al.*, 2013) and Mozambique (Butler, 2013). The dispersal of the Foc-TR4 to East Africa would be disastrous for food and livelihood security in the region.

Genetic studies of EAHB have been limited despite the socioeconomic importance of the EAHBs. Inadequate molecular knowledge of the genetic diversity and population history of the EAHB germplasm hampers the breeding and improvement of EAHB. Members of the EAHB subgroup have been classified into five morphological groups known as clonesets to reflect their extensive morphological variation in fruit size, shape and color (Karamura, 1998). This morphological and taxonomic systems provide coherent classification for cultivated banana (Li *et al.*, 2013) and has been widely-used to differentiate specific banana clones (Hippolyte *et al.*, 2012; de Jesus *et al.*, 2013). However, this system has inadequate resolution for investigating populations within species (Christelova´ *et al.*, 2011). In addition, morphological diversity measures are often inaccurate (Abdullah *et al.*, 2012), or under- or over-estimate the actual amount of genetic variability (Resmi *et al.*, 2011). An improved understanding of molecular genetic diversity of present EAHB populations will enable better approaches to banana breeding (Ganapathy *et al.*, 2012).

To elucidate the systematic relationships and genetic diversity of *Musa* germplasm, several studies have evaluated the genetic diversity of cultivated banana and its wild relatives using RFLPs, AFLPs (Li *et al.*, 2013), RAPDs (Venkatachalam *et al.*, 2007), ISSRs (Lu *et al.*, 2011) SRAPs (Youssef *et al.*, 2011), DArTs (Hippolyte *et al.*, 2010) , SSRs and SNPs (de Jesus *et al.*, 2013). SSRs (Simple Sequence Repeats) are a type of microsatellite that are effective genetic markers for investigating population structure and history

(Spencer *et al.*, 2000; Amos & Hoffman, 2010; Mariette *et al.*, 2010; Galov *et al.*, 2013; Hoban *et al.*, 2013).

To our knowledge, we here present the first comprehensive report on the genetic diversity and population history of the EAHB group. Our aims were to (i) determine the power of SSRs to discriminate EAHB samples from different morphological groups in Uganda and Kenya; (ii) relate EAHB genetic variability to that in other *Musa* subspecies; and (iii) assess support for recent population size changes among cultivars since triploidization and subsequent domestication.

2.2 MATERIALS AND METHODS

2.2.1 Sample collection and morphological classification

Based on 73 morphological traits, the EAHB have been conventionally classified into five morphological clonesets; Musakala, Nakitembe, Nfuuka, Nakabululu and Mbidde (Karamura, 1998), 90 cultivars representing phenotypic diversity within and between the EAHB clonesets were collected from Ugandan Kenyan. Other six genetically distinct African samples (4 AAB Plantain and 2 AAA Desert) were used in this study (Appendix Table 1). DNA was extracted following modifications of two protocols; Dellaporta *et al.* (1993) and Mace *et al*. 2004 (Appendix 2). DNA was eluted in 100 µl low salt TE buffer and diluted to 20 ng/ml working stocks based on spectrophotometric measurements.

2.2.2 DNA amplification of SSR loci

Previously developed SSRs the transcribed regions (Mbanjo *et al.*, 2012b) and others from express sequence tags (ESTs) of the *Musa* genome (Crouch *et al.*,

1998; Hippolyte *et al.*, 2010) (S Table 3) were used in this study. We screened 250 SSR markers for polymorphism and multiple alleles and a final list of 100 SSRs separately were selected for the study.

PCR was used to amplify the SSRs: reactions contained 1x standard *Taq* buffer with $MgCl_2$; 0.2 mM dNTP mix; 0.5 units/µl *Taq* polymerase (New England Biolabs); 30ng/µl DNA template; and 0.3 µM fluorescently labeled primer. Amplification steps followed (i) initial denaturation at 95 °C for 3 minutes; (ii) 40 cycles at 95°C for 0.30 minutes, 1 minute of 52°C to 61°C annealing temperature (primer pair specific), and 72°C for 2 minutes; and (iii) final extension of 20 minutes at 72°C. All loci were individually amplified and the post-PCR primer products were multiplexed based on the dye and expected size of the fragment prior to capillary electrophoretic separation (ABI 3730xl DNA Analyzer), sizing (GeneScanTM-500 LIZ internal size standard) and manual verification of allele calling (Genemapper v4.1).

A standardized platform for molecular characterization developed for Musa germplasm by Christelová et al. (2011) was followed. One hundred SSR loci were scored after PCR with the fluorescently labelled primers (6-FAM, VIC, NED and PET) and capillary electrophoretic separation with internal standard (GeneScanTM-500 LIZ size standard, Applied Biosystems). The PCR products were multiplexed (based on the dye and expected size of the fragment) prior to the separation and loaded onto the automatic 96-capillary ABI 3730xl DNA Analyser. Electrophoretic separation and signal detection was carried out with default module settings. The resulting data was then analysed and called for alleles using Genemapper v4.0 software (Applied Biosystems Foster City, CA). Allele sizing and calling was done as described in the user's manual and alleles were scored manually as fragment sizes in base pairs.

The multiallelic information at each SSR locus was treated as binary data, so each SSR allele was thus treated as a separate marker (Christelova´ *et al.*, 2011; de Jesus *et al.*, 2013). We assessed genetic diversity of the six cultivars

of other genomic groups representing the out-groups separately from the set of 90 Kenyan and Ugandan EAHB.

2.2.3 Intra-specific EAHB population genetic variation

Levels of genetic diversity in the EAHB population were evaluated by calculating the average Polymorphic Information Content (PIC) for each SSR locus as $PIC_i = 2f_i(1-f_i)$ where i is the information of the ith marker; f_i is the frequency of the amplified allele (the presence of a band) and $(1-f_i)$ is the frequency of null alleles that have no band (Botstein *et al.*, 1980; de Jesus *et al.*, 2013). Variability was also assessed by the average number of alleles per locus, percentage of alleles identical by state, the population allele frequency, and the expected heterozygosity with Powermarker v3.25. Confidence intervals were estimated by non-parametric bootstrapping.

To assess genetic differences between cultivars, we calculated the average number of alleles in each cultivar and private alleles per cultivar (de Jesus *et al.*, 2013). Pairwise genetic distance between cultivars was calculated as the shared allele distance (D_{AS}) with Powermarker v3.25: D_{AS} values are linearly related to the time since common ancestry for a stepwise mutation model (SMM) (Goldstein *et al.*, 1995). It was calculated as $D_{AS} = \sum_{j=1}^{m}\left(\sum_{i=1}^{k}\min(p_{ij}, q_{ij})\right)$ where p_{ij} and q_{ij} are frequencies of ith allele at the jth locus, while k is the number of alleles at the jth locus, and m is the number of loci examined (Liu & Muse, 2005).

The Bayesian model used for population assignment was with prior (5 morphological groups) and without prior assumptions for groups (that all cultivars belonged to one genetic group), each of which was characterized by a set of allele frequencies at each locus. The correlation in allele frequencies in the 90 EAHB was examined using a Bayesian framework with Structure v2.3.3

(Pritchard *et al.*, 2010) that probabilistically assigned cultivars to genetically distinct clusters (K) and estimated admixture proportions for each cultivar in a population-free manner independent of a mutation model. The proportion of membership for each cluster permitted incomplete membership to minimize overfitting (Falush *et al.*, 2007). To determine the most likely number of clusters, a range of values were tested ($1 \leq K \leq 10$). Analyses assumed an admixture model with correlated allele frequencies for a burn-in period of 10^5 steps prior to a run length of 10^5 with three independent iterations per K to confirm chain convergence (Pritchard *et al.*, 2000). The second-order rate of change of the likelihood function (ΔK) was used to determine the most likely number of clusters (Evanno *et al.*, 2005) with Structure Harvester (Earl & vonHoldt, 2011).

2.2.4 Population and cloneset variability in the context of genetically distinct out-groups

To determine the genetic variability of cultivars within and among the two EAHB populations and five morphological cloneset groups, measures of genetic diversity were calculated: Shannon's Information Index (Lewontin, 1972), Nei's genetic diversity (Nei, 1973), the percentage of polymorphic bands, and the number of private alleles. Heterozygosity was defined as the mean number of polymorphic alleles. Genetic diversity was partitioned within (Hs) and among (Hb) the five morphological groups using AFLPSurv v1.1. The population metrics were compared with the out-group cultivars to provide a context for the level of population variation.

To examine the proportion of the total variance among and within clonesets, we performed an analysis of molecular variance (AMOVA) in GenAlex v6.5. Cloneset pairwise PhiPT values were calculated in order to examine the distribution of genetic differences among and within clonesets and to identify

deviations from expected heterozygosity (Peakall & Smouse, 2006, 2012). PhiPT was computed as Nei's genetic diversity within clonesets divided by that within and among clonesets (equivalent to an F_{ST} value). The number of migrants between clonesets in each generation (Nm) (Slatkin, 1985) was calculated as $Nm = (1 - PhiPT)/(2PhiPT)$ in GenAlex v6.5.

The population structure of the EAHB set of 90 and the six genetically distinct outgroup cultivars was investigated using principal coordinates analysis (PCA) of pairwise SRR D_{AS} values using R (www.r-project.org). The pairwise SSR D_{AS} values were also visualized using phylogenetic networks of Neighbor-Net uncorrected p-distances with SplitsTree v4 (Huson & Bryant, 2006) to relate the population-level variation to that of the subspecies.

2.2.5 Inference of historical population sizes

We assessed a neutral hypothesis of a constant population size compared to an alternative one of a population expansion during domestication. Genetic signatures of a bottleneck associated with clonal propagation from a small founding population may be diminished during the subsequent recovery phase during which the population expands. Consequently, distinguishing a population in post-bottleneck recovery from one that was expanding after a small ancestral population size may not be possible for inbred groups with a single recent origin.

To gain an insight into the evolutionary history of 90 samples, we investigated variation in the mutation rate across SSRs in order to subsequently estimate the ancestral and recent effective population size (N_e) using the estimated mutation rates with Beast v1.8 and Tracer v1.5 (Drummond, 2005; Drummond & Rambaut, 2007). We examined the parsimony informative SSRs (27) that had at least three taxa per SSR as diploid data within the set of 90 EAHB with repeat lengths ranging from 1 to 115 for a one-phase site model since two-

phase models may not be significantly superior (Sainudiin *et al.*, 2004). A single representative from genetically identical (and therefore uninformative) taxa was used, meaning nine samples were omitted. Mutation rates were estimated using data for the full set of 96 for these 27 SSRs assuming a constant population size for 9.5×10^7 MCMC iterations after a burn-in 9.5×10^6 steps to ensure the ESS (expected sample size) > 100 for each SSR. Somatic mutation rates were estimated for the parthenocarpic set of 90 assuming a constant population size for 8.6×10^7 MCMC iterations after burn-in 8.6×10^6 steps.

To examine population size changes, the operator weights for demographic.populationMeanDist, demographic.indicators and demographic.scaleActive were changed to 40, 100 and 60, respectively, and the demographic.populationMean prior was set to one. The mutation rate estimates for the 96 samples were used as the individual SSR clock rates for an extended Bayesian skyline plot to infer the ancestral effective population size of the population of 90 for 5.8×10^7 MCMC iterations after a burn-in of 5.8×10^6 (Heled & Drummond, 2008). This was repeated for the set of 96 (6.1×10^7 iterations with a burn-in of 6.1×10^6). Posterior density intervals were calculated to compute the relative changes in N_e. Time was scaled using generations and N_e was estimated using the mean mutation rate across the 27 SSRs. The set of 96 provided a more accurate mutation rate calibration than the 90 alone because of the higher sample of total mutations and the lower estimated proportion of somatic mutations: broader sampling of *Musa* subspecies is required to assess this.

An expanding population should have lower heterozygosity (Cornuet & Luikart, 1996; Piry *et al.*, 1999) caused by the higher incidence of rare alleles (Luikart *et al.*, 1998) at selectively neutral loci like SSRs. In contrast, a population bottleneck distorts the allele frequency distribution such that low-frequency alleles (<0.1) would be lost more rapidly than ones at higher

frequencies (Maruyama & Fuerst, 1985; Luikart *et al.*, 1998). Heterozygosity is robust to complex ploidy states or examining inbred samples (DeGiorgio *et al.*, 2011). We examined the heterozygosity of each SSR relative to the observed number of alleles and sample size with Bottleneck v1.2.02 using coalescent simulations under two mutation models, SMM (Di Rienzo *et al.*, 1994) and a two-phase model (TPM). For the TPM, the SSR alleles were geometrically distributed (Fu & Chakraborty, 1998) with 90% of mutations following a SMM and with a variance of 30% for non-stepwise mutations (Amos & Hoffman, 2010). Evidence for an expansion was examined using a standardized differences statistic (T_2) and standardized sign test (Cornuet & Luikart, 1996).

2.3 RESULTS

We amplified 100 SSRs in a population of 90 Kenyan and Uganda EAHB along with six genetically distinct cultivated Plantain and Desert bananas. We demonstrate a lack of variation within the set of 90 EAHB, the absence of a genetic correlation of these markers with standard morphological groups, and higher genetic diversity among banana subspecies.

2.3.1 The East African Highland banana population is genetically monomorphic

Little genetic variability was observed in the 90 EAHB cultivars compared to the six subspecies. The set of 90 had low heterozygosity as evidenced by the mean PIC (0.058), gene diversity (0.070). This is further highlighted by: a lack of genetic differentiation between pairs (D_{AS}=0.071, range 0.000-0.176); 81.3% of the 90 cultivars had D_{AS}<0.1; and the mean minor allele frequency was just 0.05. Moreover, 58% of alleles were identical by state, even though an average of 209.3 alleles per cultivar were sampled (from a minimum of 205 for

NyarLuo Ratong to a maximum of 214 for Nakitembe Red). A total of 267 alleles were discovered, of which 11.7% were private (ranging from one to three per cultivar). The 100 SSRs could be grouped into seven categories based on the number of alleles scored per locus (Table 2): 84 out of 100 had three alleles or less with genomic SSRs showing significantly higher levels of allelic diversity than EST-SSRs.

Table 2: Genetic variation in EAHB SSRs. Little genetic variation in the 90 EAHB was found for the majority of the 100 SRRs used by grouping the SRRs based on the number of alleles: 84 corresponding SSRs had only one to three alleles per locus.

Groups	Number of alleles per SSR	Number of corresponding SSRs	Total alleles
A	1	14	14
B	2	36	72
C	3	34	102
D	4	6	24
E	5	6	30
F	6	3	18
G	7	1	7
Total		100	267

Population-free clustering with Structure identified two geographically distinguishable groups (Evanno *et al.*, 2005) (Kenyan and Uganda subpopulations) (Figure 8; S Table 2), with population membership probabilities greater than 0.8 (Pritchard *et al.*, 2000). However, very little genetic differences and differentiation was observed between them. The sole cultivar without characteristic black or brown blotches on its pseudostem (Namwezi) was from Uganda but showed Kenyan ancestry.

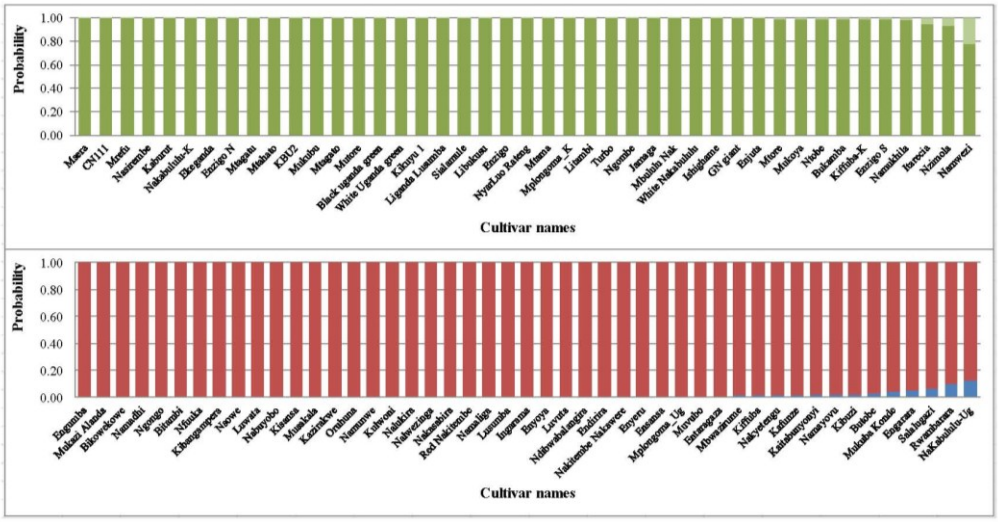


Figure 8: Inferred ancestry of the EAHB cultivars. Population membership for 90 East African Highland Banana cultivars from Kenya n=42, green), Uganda (n=47, red) at K= 2, the most likely number of groups based on Structure classification. Although the SSRs distinguished the Kenyan from Ugandan samples, their total diversity was low in the context of the six out-groups cultivars.

PCA of the population of 90 and six Plantain and Desert cultivars showed little differentiation between the two EAHB populations (Kenya and Uganda, PC2 with 1.9% of total variation). In contrast, the six subspecies were distributed across PC1 accounting for 94.2% of diversity (Figure 9A). There is a small level of differentiation between Kenyan and Ugandan groups, which could be due to recent genetic drift. Though much of the diversity was within subpopulations (Hs =0.056; $p<0.0001$) compared to between subpopulations (Hb = 0.052; $p<0.0001$), the Uganda subpopulation was more diverse than the Kenyan one. The Uganda group had a higher expected heterozygosity (0.065 vs 0.049) and proportion of polymorphic loci (13.5% vs 9.4%) suggesting the Kenyan EAHB sub-population may be derived from the Ugandan one. However, this may be complicated by genetic drift distinguishing the populations at few loci, and novel variants unique to certain plants (White Nakabululu and Mtore; Figure 9B).

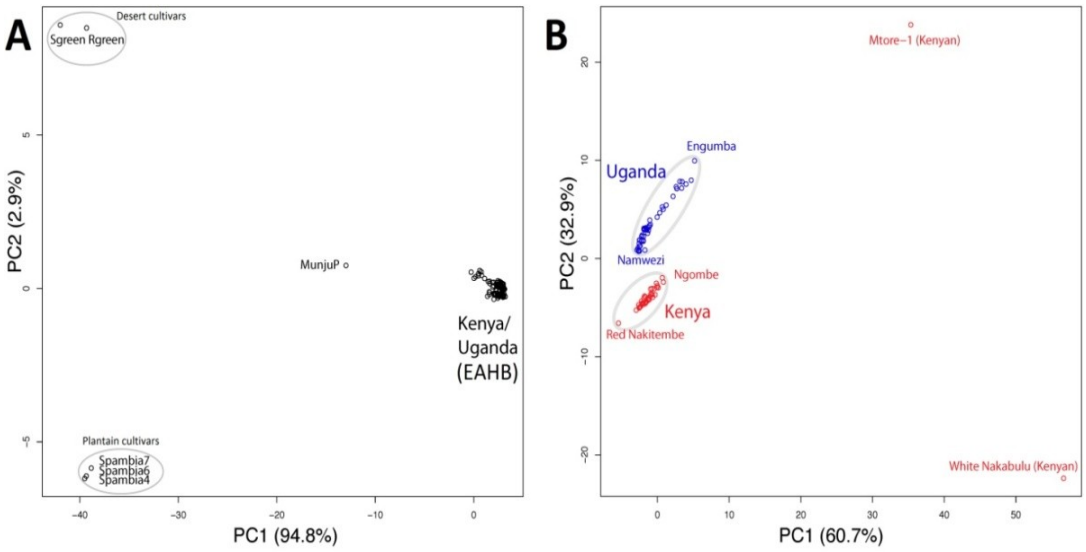


Figure 9: PCA (principal coordinate's analysis) shows that structure exists in the East African highland banana population. (A) The EAHB population and six out-group cultivars. EAHB-Uganda and Kenya are genetically close, whereas the Plantains (AAB) and Dessert (AAA) are from genetically different *Musa* groups, though MunjuP retains an intermediate classification. **(B) Clustering of the EAHB population showing little genetic differentiation of the Kenya and Uganda cultivars.** White Nakabululu and Mtore are substantially different to their main regional groups.

2.3.2 No genetic differentiation of morphological groups of EAHBs

There was much higher genetic variation within than between the five morphological groups known as clonesets (Hw = 0.0925 vs Hb=0.022): 96% of the variation was within cloneset groups and only 4% of the variation was between groups (Table 4). This was supported by a lack of differentiation between clonesets (mean PhiPT= 0.036, P = 0.01 with a range of 0.011 and 0.125, Table 3). Furthermore, the maximum D_{AS} value between cloneset pairs was just 0.004 (Table 4).

Table 3: Mean pairwise Nei's genetic diversity (Nei, 1973) of the EAHB clonesets. PhiPT and the number of migrants among clonesets per generation did not differentiate the five morphological clonesets (Mbidde, Musakala, Nakabululu, Nakitembe, Nfuuka). Phi_{PT} (equivalent to an F_{ST} value) was computed as Nei's genetic diversity within clonesets divided by within and among clonesets diversity. The number of migrants among clonesets per generation (Slatkin, 1985) was calculated as $\boldsymbol{Nm = (1 - PhiPT)/(2PhiPT)}$.

Nei's genetic diversity	Mbidde	Musakala	Nakabululu	Nakitembe
Musakala	0.995	****	****	****
Nakabululu	0.995	0.993	****	****
Nakitembe	0.992	0.996	0.996	****
Nfuuka	0.997	0.994	0.998	0.995

PhiPT	Mbidde	Musakala	Nakabululu	Nakitembe
Musakala	0.074	****	****	****
Nakabululu	0.080	0.107	****	****
Nakitembe	0.125	0.040	0.061	****
Nfuuka	0.024	0.090	0.011	0.071

Migrants per generation	Mbidde	Musakala	Nakabululu	Nakitembe
Musakala	6.296	****	****	****
Nakabululu	5.746	4.155	****	****
Nakitembe	3.508	11.965	7.648	****
Nfuuka	20.058	5.084	43.360	6.559

Table 4: Analysis of Molecular variance (AMOVA). Results for 90 EAHB clonesets illustrated much higher genetic variation within than between the groups (96% vs 4%).

Source	Degrees of freedom	Sum of squares	Mean of squares	Estimated variance component	%
Among clonesets	4	82.92	20.729	0.467	4%
Within clonesets	85	1054.91	12.411	12.411	96%
Total	89	1137.82		12.878	100%

2.3.3 Complex historical gene flow patterns in EAHB with Plantain samples

Diversity within the two EAHB subpopulations was much lower compared to the six out-group cultivars. This was illustrated above by PCA (Figure 2.3) and was highlighted in a phylogenetic network that partitioned the data in a similar manner: the set of 90 EAHBs form two closely related groups that are markedly homogeneous compared to the highly diverse six out-groups (Figure 10; S Table 1). These results highlighted the vegetative nature of propagation with the EAHB in contrast to the extensive divergence between subspecies. It also demonstrated the differentiation of the EAHB from the Plantains (AAB) and Dessert (AAA), although MunjuP retained an intermediate classification suggesting a potentially more complex origin.

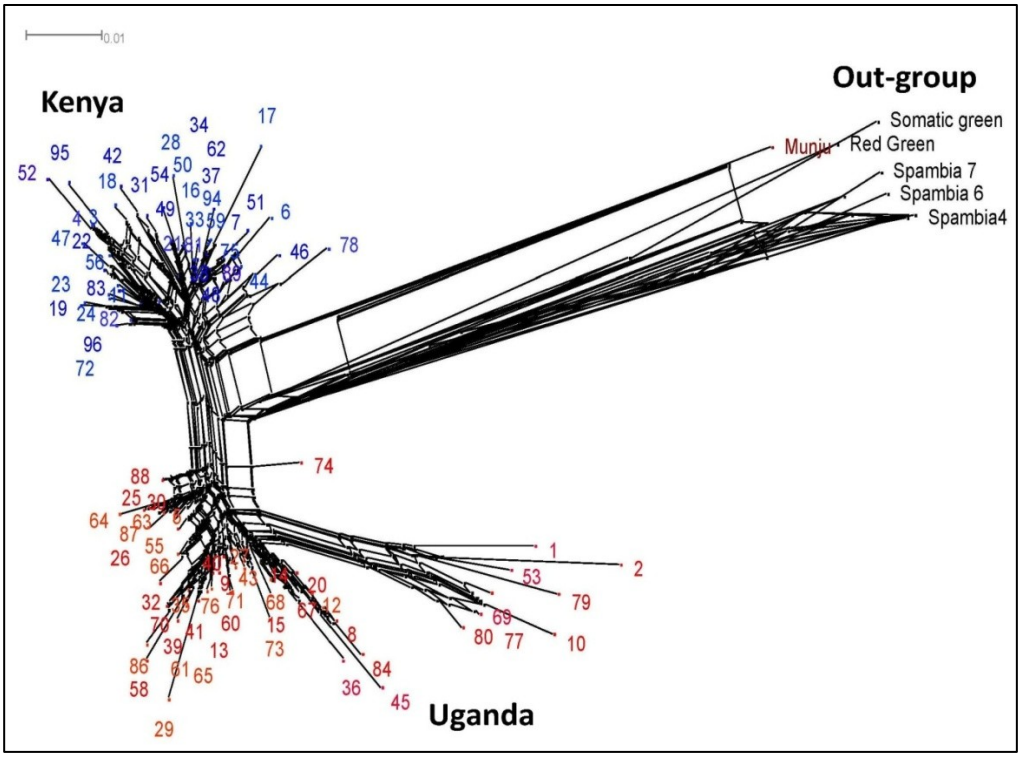


Figure 10: A Ne**ighbor-net network of Neighbor-net uncorrected p values (Bryant & Moulton, 2004) of 100 SSRs (Delta score 0.2812; Q residual score of 0.0095).** Phylogenetic Splits tree generated from the Shared allele distances (D_{AS}) for the set of 90 EAHB cultivars and six genetically distinctive out-groups: Plantains (Spambia 4, 6 and 7) and AAA-Desert (Somatic green and Red green). One sample (MunjuP) was genetically intermediate between the Plantains, Desert varieties, and the EAHB. Taxa numbers correspond to those in Appendix 1.

2.3.4 Evidence for a recent EAHB population expansion

Here, we performed the first estimation of SSR mutation rates in bananas. Considerable variation in the estimated rates across 27 parsimony-informative SSRs was observed, with a mean of 0.00166 substitutions per SSR per generation, ranging from 0.00027 (locus 44) to 0.0030 (locus 47). This 11-fold magnitude of variation illustrated the heterogeneity associated with STR mutation rates (Scarcelli *et al.*, 2013) (Figure 11). Moreover, somatic mutation rates estimated for the set of 90 differed: these had a mean value of 0.00804 substitutions per SSR per generation, ranging from 0.00065 (locus 44) to

0.07620 (locus 70). Compared to data for the 96, this produced an average excess of 4.8-fold, but this varied from 0.6- (locus 27) to 85.0-fold (locus 70), highlighting potential mutation rate differences between samples that were parthenocarpic with those partially sexually-reproducing.

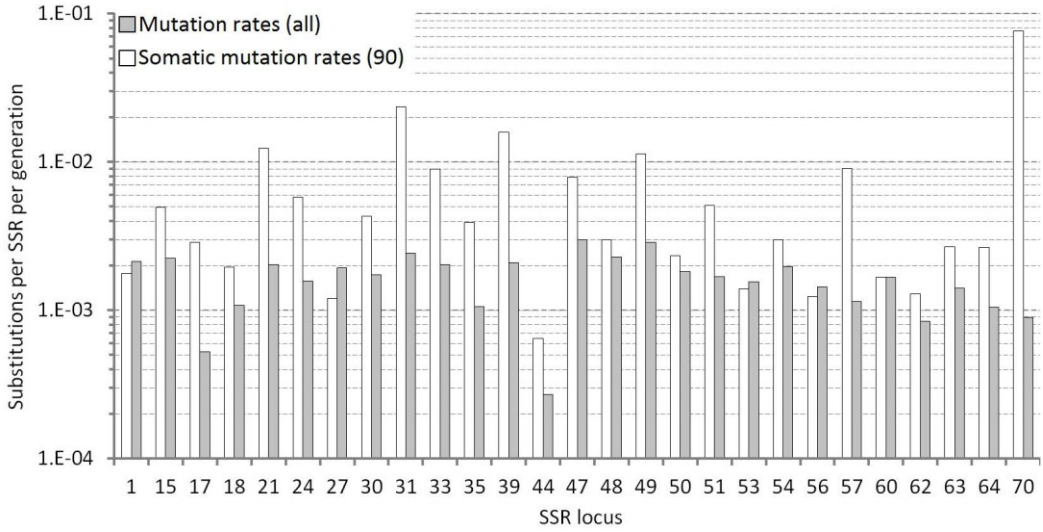
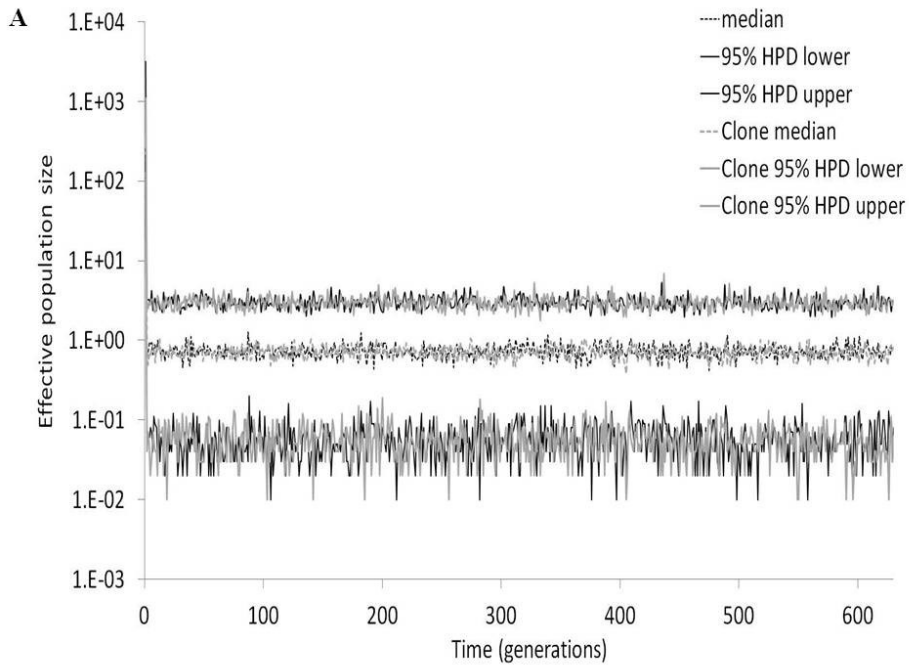


Figure 11: Estimated substitutions per SSR per generation (y-axis) for 27 parsimony-informative SSRs (x-axis) inferred for 96 samples using Beast v1.8 and Tracer v1.5. There was a mean value of 0.00166 substitutions per SSR per generation, though ranging 0.00027 (locus 44) to 0.0030 (locus 47). The unit of time approximates one generation of sexual reproduction in these partially clonal plants. Somatic mutation rates estimated for the set of 90 had a mean value of 0.00804 substitutions per SSR per generation, ranging from 0.00065 (locus 44) to 0.07620 (locus 70).

We investigated evidence for recent population size changes in the EAHB. Using these mutation rates estimated at 27 SSRs for the set of 96, extended Bayesian skyline plots (EBSP) indicated a historical low constant N_e with 95% highest posterior density (HPD) values < 4.9 individuals for both. Present generation N_e estimates for both datasets indicated a dramatic recent increase in N_e with values for the set of 90 (1,435.7 with a 95% HPD range of 1,175 to 12,234), which is supported by data for the 96 containing the six subspecies (2,868.6, 95% HPD range 2,352 to 35,301) (Figure 12). Despite this recent jump in N_e, no statistically significant evidence of population size change was

inferred using this multi-locus EBSP framework.



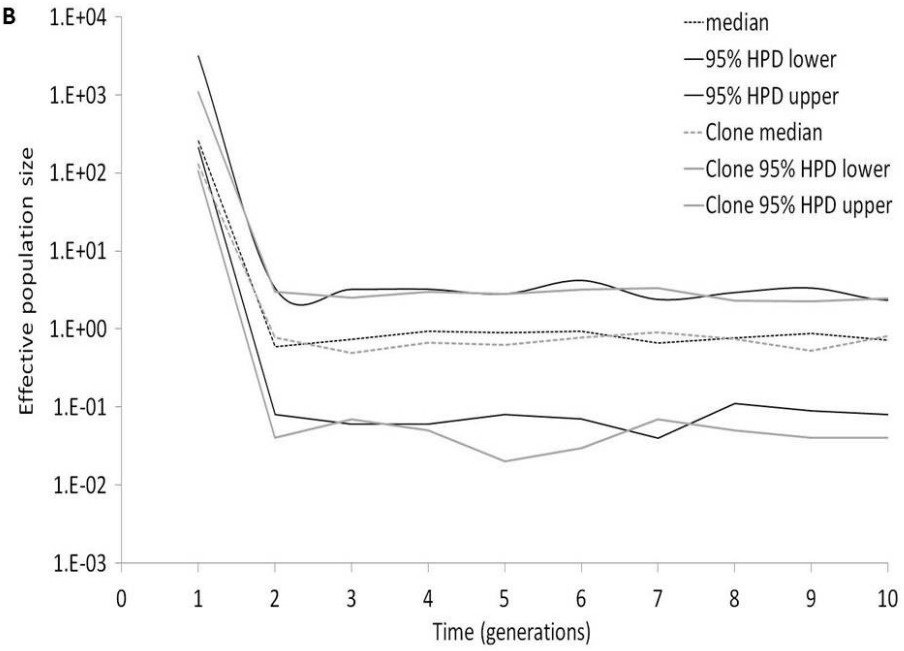


Figure 12: Extended Bayesian skyline plots (EBSP) of the EAHB showing (A) a low historical constant effective population of the EAHB and (B) a recent expansion of effective population size of the EAHB. The recent median effective population size ($log_{10}N_e$, y-axis) in the sets of 96 EAHB (black dashed line) and the subset of 90 quasi-clonal or parthenocarpic plants ("Clone", grey dashed line). Time (x-axis) is denoted in units of generations. The 95% HPD range are denoted by the flat lines. Historical N_e values were < 4.9 (A) but present N_e may be larger.

There was a significant heterozygote deficiency under both SMM and TPM mutation models, characteristic of a population expansion (Wilcoxon one-tailed p<0.0005, Table 3) (Cornuet & Luikart, 1996) or recovery after a decrease in size (McEachern *et al.*, 2011). In addition, a significantly negative standardized differences value (T_2) was observed (SMM: T_2=-3.774, P=0.00008; TPM: T_2=-2.668, P=0.00381). Allele frequency distributions of the 43 polymorphic SSRs (Table 5) showed a shifted mode of allele frequencies from the normal 'L-shaped' incompatible with a recent bottleneck (S Figure 1).

Table 5: The 90 EAHB showed a lower than expected heterozygosity consistent with a population expansion (Cornuet & Luikart, 1996; Piry *et al.*, 1999). Results were assessed with Bottleneck v1.2.02 using coalescent simulations under two mutation models: stepwise (SMM) and two-phase (TPM). The SSR alleles were geometrically distributed with 90% of mutations following a SMM and with a variance of 30% for the TPM. The significant of the heterozgosity deficit was supported by Wilcoxon standardized sign tests and negative standardized differences statistic values (T_2).

Mutation model	Expected SSRs with heterozygosity excess	Observed SSRs with heterozygosity excess	Observed SSRs with heterozygosity deficiency	Wilcoxon sign test
Stepwise mutation model (SMM)	50.95	32	78	0.00021
Two-phase mutation (TPM)	48.21	32	78	0.00048

2.4 DISCUSSION AND CONCLUSIONS

Knowledge of the extent of genetic diversity in crop germplasm is a prerequisite for developing a strategic breeding program, including in a crop like *Musa* which is recalcitrant to breeding owing to parthenocarpy, sterility and polyploidy. Somaclonal mutations combined with human selection has likely resulted in morphologically diverse current EAHB collections. The EAHB (AAA) subgroup may have unique morphotypes not found in any other *Musa* subgroup, but little is known about their genetic diversity and history. We used 100 SSR markers to assess the genetic diversity, population structure and evolutionary history of 90 EAHB cultivars growing in Uganda and Kenya, and six out-group cultivars representing plantains (AAB) and dessert (AAA)

and one unknown cultivar. Our study demonstrates that EAHB genetic variation had a single recent origin followed by vegetative reproduction generating limited variation through somaclonal mutations.

Although substantially reduced diversity within other *Musa* groups has been reported (Creste *et al.*, 2003; Hippolyte *et al.*, 2012), EAHB in this study exhibited significantly lower genetic variability than in other *Musa* sets (El-Khishin *et al.*, 2009; Opara *et al.*, 2010; Christelova´ *et al.*, 2011; Resmi *et al.*, 2011; Abdullah *et al.*, 2012; Shaibu, 2012; de Jesus *et al.*, 2013). Large variation in morphological characteristics did not reflect genetic variation also reported by Lu *et al.* (2011). Population classification and phylogenetic analysis showed little differentiation of Ugandan and Kenyan cultivars, suggesting both have a recent single origin, more likely to be ancestral to the Ugandan set on the basis of its comparatively higher genetic variability.

Simulations of the recent historical effective population size and expected heterozygosity were symptomatic of a recent population expansion (Cornuet & Luikart, 1996). Although low allelic diversity may be caused by a genetic bottleneck (Gebremedhin *et al.*, 2009) the sample set did not support this directly. Though the ancestral population size was extremely small, lack of evidence of a genetic bottleneck is consistent with previous work (Li *et al.*, 2013). The genetic pattern may be accentuated by a Meselson effect of heterozygous alleles persisting in an apomictic reproductive system (Butlin 2002). A recent expansion might be due to the human-mediated establishment of EAHB in new environments like East Africa (Schoebel *et al.*, 2014). Consequently, the hypothesis that cultivated bananas underwent a genetic bottleneck during domestication remains to be tested in bananas from other parts of East Africa (Li *et al.*, 2013). The small distinction of the 90 EAHB cultivars across geographical regions (Kenya and Uganda) was minute relative to the higher diversity in the Plantain and Desert cultivars suggesting lack of introgression of genetically distinct erasing genetic heterogeneity in the EAHB

(Miller *et al.*, 2012). No evidence for multiple origins nor admixture was observed. These results corroborate the hypothesis of a recent single seed origin (Hurtado *et al.*, 2012), consistent with hybridization followed by selection and clonal propagation (de Jesus *et al.*, 2013).

This close relatedness between EAHB cultivars has been observed in cultivars of the *Musa acuminata* group (Changadeya *et al.*, 2012). The Low level of genetic diversity and close relationship of the EAHB cultivars may be ascribed to founding effects due to a single origin from the same initial clone with subsequent vegetative propagation and somatic mutation. (Noyer *et al.*, 2005; Perrier *et al.*, 2011; Hippolyte *et al.*, 2012). This effect could have been accentuated by preferential breeding of specific clones as has been reported in other domesticated species (Hyten *et al.*, 2006; Rauf *et al.*, 2010; Bourguiba *et al.*, 2012). Banana domestication may have occurred up to 10 kya (Perrier *et al.*, 2011), but movement from centres of diversification like Papua New Guinea to Africa may have occurred much later (perhaps circa. 2-2.5 kya). This was most likely followed by further human manipulation and selection of clones, coupled with recent changes in agricultural practices may have reduced diversity by selecting for phenotypically productive varieties.

The intermediate phylogenetic placement of MunjuP between the Plantains, Dessert and EAHBs could suggest ancient interbreeding events (Bryant & Moulton, 2004), which is supported by previous work on African Plantains and EAHB, these may share at least one maternal A genome derived from *M. acuminate*- banksii (Kennedy, 2008). This observation of historical mixing events across subspecies indicates that deeper molecular investigation of these subspecies will be able to determine the hybrid origins of modern domestic breeds, for which the number and timing ancient migrations both before (in diploids) and after hybridization (in triploids) remains undetermined (Li *et al.*, 2013).

A finer elucidation of the events associated with the triploidization and domestication of EAHB can be garnered by comparing genetic variation in EAHB (AAA) and the Plantain (AAB) and Dessert (AAA) subspecies to that in *M. acuminata* wild and edible diploid (AA) and triploid (AAA) samples. This study supports the hypothesis that the EAHB arose from genetically monomorphic clones, possibly selected during domestication, and thus wild and cultivated diploids should harbor an array of genetically diverse clones.

High genetic diversity and scope for further breeding is important for resistance to infectious disease (Hajjar *et al.*, 2008; Rauf *et al.*, 2010). *Musa* variation is derived from four wild species: *M. acuminata* (A genome), *M. balbisiana* (B), *M. schizocarpa* (S) and *M. textilis* (T). The latter two show extensive genetic but not phenotypic variation (Li *et al.*, 2013) and only *M. acuminata* is parthenocarpic (Heslop-Harrison & Schwarzacher, 2007; Kennedy, 2008).

Our findings provide a genetic diversity context for the growing threat of migration of Foc-TR4 and other pathogens from Asia to Africa. Given that a 2013 outbreak occurred in northern Mozambique, the potential for devastation of EAHB cultivation in East Africa is clear. This outcome can be addressed with more extensive genetic characterization and broader sampling to discover genetically compatible wild and domestic subspecies that may be genetically resistant to such infections, and where such genetics can be introduced into farmer-preferred EAHB cultivars through breeding and/or gene transfer.

2.5 SUPPLEMETARY MATERIAL

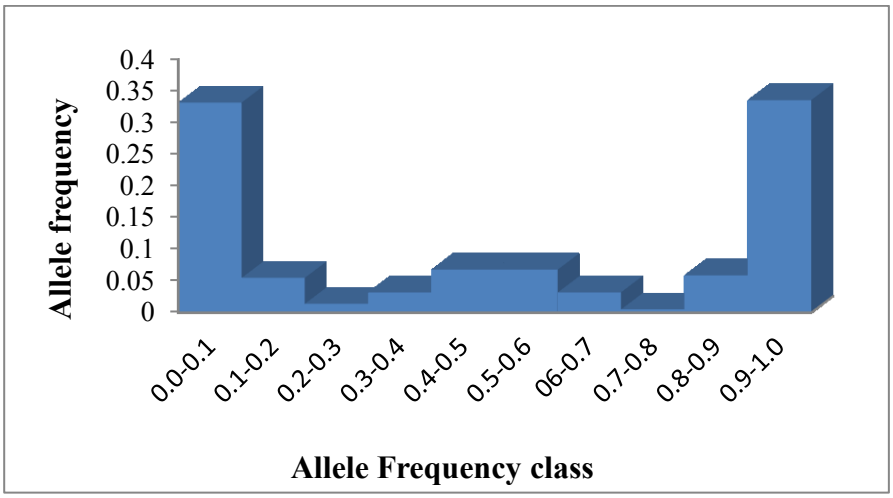
S Table 1: Allelic patterns of genetic variation at 100 SSR loci revealed low genetic diversity in Kenyan and Ugandan EAHB subpopulations compared to Plantain and Dessert subspecies.

Population	**Kenya**	**Uganda**	**Plantain**	**Dessert**
Mean Shannon's information index	0.035	0.046	0.280	0.025
No. Private Alleles	0.041	0.058	0.269	0.205
Expected heterozygosity	0.020	0.029	0.188	0.018
Observed heterozygosity	0.021	0.030	0.250	0.036

S Table 2: Allele frequencies correlation with Structure v2.3.3 (Pritchard *et al.*, 2010b). Two genetically distinct clusters (K=2) for the 90 EAHB based on the second-order rate of change of K (ΔK) (Evanno *et al.*, 2005) with Structure harvester (Earl & vonHoldt, 2011) were identified. K=8 had a high ΔK but differentiation among groups was low.

K	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	Ln''(K)	ΔK
1	-5569.87	16.67	—	—	—
2	-4662.23	27.69	907.63	824.77	29.79
3	-4579.37	25.57	82.87	431.43	16.88
4	-4927.93	344.356	-348.57	576.43	1.674
5	-4700.07	434.78	227.87	143.80	0.33
6	-4616.00	408.53	84.07	69.17	0.17
7	-4601.10	474.14	14.90	221.33	0.47
8	-4364.87	22.04	236.23	523.13	23.73
9	-4651.77	604.30	-286.90	366.20	0.61
10	-4572.47	383.72	79.30	—	—

S Figure 1: Distribution of allele frequencies in the EAHB population. This graphical method shows that a population has been recently bottlenecked if fewer alleles are found in the low-frequency class (0 to 0.1) than in 1 or more inter- mediate frequency classes (Luikart et al. 1998). Mode-shift distortions were not observed in the EAHB population. Error bars show the number of alleles found in each frequency class with 5% error rare for 1000 simulations.



S Table 3: The ID number, marker name, repeat motif, repeat lengths, forward and reverse primers and literature reference for the 100 microsatellite SRRs used in this study.

ID	Name	Repeat Motif	Repeat	chromosome	Forward primer (5'-3')	Reverse primer (5'-3')	Reference
1	Ta6025	(ATCT)9	4	2	gtggtgaagccgctcaagtg	cactggagttctggtgcagc	Mbanjo *et al.*
2	Ta6833	(AAG)7	3	2	gcaccactagttcttccaccacc	ggatccgggatgcagctc	Mbanjo *et al.*
3	Ma513034073	(CGC)6	3	10	ctccctgtactcgtccatgtgg	gcctcttcactgtgttaagtgcaca	Mbanjo *et al.*
4	Ma513052078	(CAG)8	3	6	ccatggaccaaaccgtgctg	ccctctttcatcaccaacccatct	Mbanjo *et al.*
5	Ma513032586	(CTACA)5	5	8	tggttggtggcttgcaaacg	ccgtcaccacccacaacac	Mbanjo *et al.*
6	Ta562	(AGC)7	3	10	cgcctcgtgtttcaacgagc	agaggcaggtcacggcac	Mbanjo *et al.*
7	Ta1553	(TTC)7	3	8	acgagacagatccctttcggtg	gctcatttcaccgacacgcac	Mbanjo *et al.*
8	Ta7514	(CTG)8	3	6,10	gctcagctgtccaggttgac	tgctgctgagtgaccgga	Mbanjo *et al.*
9	Ta3454	(TC)9	2	6	ggcgtcttggttactgtccttgg	gcaacaacaatcactgtcgtgttcc	Mbanjo *et al.*
10	Ta6942	(AG)13	2	4,7,8	ctgcaaggagctggaccc	cgagaggacgacacgacgtc	Mbanjo *et al.*
11	Ta3054	(TGTTT)6	5	6	tgccaacagcctataatacggca	gtatcaggggacgatcgacagtc	Mbanjo *et al.*
12	Ma513044953	(AGA)8	3	10	gttcgggtgatgatggcacc	ccaaacagcaccgtaggctg	Mbanjo *et al.*
13	Ta1885	(GAG)7	3	8	agcatatgcaaccacaacagttg	tgcgtcataatttgagacctgcca	Mbanjo *et al.*
14	Ta2979	(CT)10	2	4,	caggaaggtctgcagcgtg	acacagtccatcccatttggacg	Mbanjo *et al.*
15	Ma513047439	(ATC)7	3	4	gtaccaggcaacacccacc	tggcaacaccaacatcgctg	Mbanjo *et al.*
16	Ta1137	(CAG)CAT(CAG)2	3	9	ggttggcagagttgtcggtg	agctcccatcattcatctgcagg	Mbanjo *et al.*
17	Ta4184	(TGT)7	3	6	tgggtgaacacacacacacct	tggggagacatgaggccattt	Mbanjo *et al.*
18	Ta5540	(TCT)7	3	8	ccatgctgtgaatgcatcggag	cgcaggctgtagaagtaccacac	Mbanjo *et al.*
19	Ma513051880	(GA)9	2	10	cagctatttaggcgaagatcatc	tccaacaccagtcaaagctcca	Mbanjo *et al.*
20	Ta2203	(AG)9	2	8	ggtgcccagatgccatgc	agacatttatccaccaaaggcttcc	Mbanjo *et al.*
21	Ta6203	(CTG)6CAG(CTG)2	3	6	ggagaagacgagagacccgct	agccgccatcaaccaacg	Mbanjo *et al.*
22	DN239771	(TC)9	2	2	tcccctgtatcaccacagcag	tggaccatgcattactttgctgtgaa	Mbanjo *et al.*
23	Ma513037972	(CT)13	2	4	tcgacggagaactgcgacc	cctttgcccttcctattccggtg	Mbanjo *et al.*
24	Ma513045122	(TC)10	2	5	cgctctgtggcaggactg	gcaccgattggtcgaattagcg	Mbanjo *et al.*
25	Ta150	(GAA)6	3	2	agagcagcagaccgcacc	cacagtggcttccgacaagc	Mbanjo *et al.*
26	mMaCIR08	$(TC)_6N24(TC)_7$	2	1	acttattcccccgcactcaa	ctcctccatagcctgactg	Hippolyte *et*
27	mMaCIR21	(GA)8	2	3	tcccataagtgtaatcctcagtt	cgatgccacactggac	Hippolyte *et*
28	mMaCIR231	(TC)X10	2	3	gcaaatagtcaagggaatca	ctatttgacgttggtggtc	Hippolyte *et*
29	Ta5917	(GAC)7	3	8	accctgagggcaacggtg	ggtggctgaggaagctcctc	Mbanjo *et al.*
30	Ma513049034	(GA)11	2	8	aggccattcattccttaagggtgg	gctgcagctgacccaatcg	Mbanjo *et al.*
31	Ma-1-6	(GA)10	2	5	tttgcctggttgggctga	ccccccttcctcttttgc	Crouch et al
32	Ma-3-139	(GA)14	2	10	actgctgctctccacctcaac	gtcccccaagaaccatatgatt	Crouch et al
33	Ma-1-27	(GA)9	2	2	tgaatcccaagtttggtcaag	caaaacactgtccccatctc	Crouch et al
34	Ma513050081	(AG)16	2	5	gtgcgctccatcgttgttgag	gccactacccaatgcatcgag	Mbanjo *et al.*
35	mMaCIR12	$(GAGAA)_3GATGA(GA$	3	10	acagaatcgctaaccctaatcct	actctcgcccatcttcatc	Hippolyte *et*
36	Ta1069	(CTT)11	2	6	agagaagcgactttgtcatgcct	ggttcacaacaaagaggaataga	Mbanjo *et al.*
37	BJ1-10	$(TC)_{13}N3(CA)_9N2(GA)_{11}$	2				Hippolyte *et*
38	mMaCIR24	(TC)7	2	6	atcttttcttatccttctaacg	acccaggtctatcaggtca	Hippolyte *et*
39	mMaSTMS15		2	5	tgctcttccacatctcaagaac	gattgcacggagattcaaca	Kaemmer *et*
40	mMaCIR25	(GA)6	2	6	gtggtttggcagtggaatggaa	attagatcaccgaagaact	Hippolyte *et*
41	mMaCIR44	(GA)23	3	8	tggttgagtagatcttcttgtgt	tggagtgaagatgagacga	Hippolyte *et*
42	Ta2872	(CTG)6	3	4	acgccgtcgaccttctcg	cggttctccatcaccatgacca	Mbanjo *et al.*
43	Ta4757F-T	(GGC)6	3	4	ggaccccgaagagtcgtcc	tcacccagtagaagtaggccct	Mbanjo *et al.*
44	Ma513052491	(ACC)6	2	10	gggcttctttcgttagcggga	tcacggcgacgagctgct	Mbanjo *et al.*
45	Ma-1-24	(CT)12	3	6	gagcccattaagctgaaca	ggtcggtatgggaagcacc	Mbanjo *et al.*
46	Ta4187F-T	(CCT)6	2	10	tggatcaacctgtcctccaagg	caaggtgagcatgtcacagcg	Mbanjo *et al.*
47	Ta6377	(CG)9	2	7	cgacggagctcaaagtccct	tgaccagccggcaaatcc	Mbanjo *et al.*
48	Ma513052458	(CT)13	2	10	cgtgtttgtcgtcggagctc	ggaagcacgattcaccgactcg	Mbanjo *et al.*
49	mMaCIR03	(GA)10	3	1	tgacccacgagaaaagaagc	ctcctccatagcctgactg	Hippolyte *et*
50	Ta2139F	(CTC)7	3	7	ccgatggaagagctatccgagg	cgcctacctccatgcagagaag	Mbanjo *et al.*
51	Ma513037490	(TC)8	2	10	cttccgctccctttcaccc	cgaagcacggcgagtgttc	Mbanjo *et al.*
52	mMaCIR13	(GA)16N76(GA)8	3	3	tcccaacccctgcaaccact	ccctttgcgtgcccctaa	Hippolyte *et*

53	mMaCIR152	(CTT)18(CT)17(CA)6	2	4	ccacctttgagttctctcc	gaatgctgatacctctttgc	Hippolyte *et*
54	Ta376F-T	(GA)19	4	4	cgccattgcatttgctaatggct	tgttgatcgaaacagtagacagtac	Mbanjo *et al.*
55	Ma513026332	(GTAG)5	3	2	caactttctccaagatcag	acggagcagtaacacgggattg	Mbanjo *et al.*
56	Ta248F-T	(GGC)6	3	1	ctcccaccgcgaacaatgg	acggagctgctcaccagc	Mbanjo *et al.*
57	Ta578	(GCA)7	3	3	acttaccaggtcctggtgcag	actgaaccactacatcggccag	Mbanjo *et al.*
58	Ma513044920	(TC)13	3	6	tcgctttgtgatgcgtgcac	cgttggcattgattgatatgcgtgg	Mbanjo *et al.*
59	Ta3938	(CTA)6	5	2	tctggcccgcccactaaa	aaccatcacagagaactgtttggct	Mbanjo *et al.*
60	Ma513030283	(TCTGT)5	5	7	cgtgcagtgctgttgctgtg	tcccaacaagcagcccgt	Mbanjo *et al.*
61	Ma513048504	(GAA)6GAG(GAA)5	3	2	tgcacggagagatctgctcc	tgtcagcaagatcttaaccctgcag	Mbanjo *et al.*
62	Ma513036776	(TC)10	2	9	agataacgctcgagatcgccc	acgcagcacaagtcgtcca	Mbanjo *et al.*
63	Ta1401	(GA)10	2	7	acccgctattccgtttcgct	gagcatggaagaggcgttcc	Mbanjo *et al.*
64	Ma513043179	(TA)12	2	1	tgggcaaagaccggaagc	tgcatggagatacaaaaagaatcc	Mbanjo *et al.*
65	mMaCIR164	(AC)X14	3	6	aagacaagttccattgcttg	tttccctcttcgattctgt	Hippolyte *et*
66	Ma513047481	(CTC)7	3	2	ggtgatctctagttcatcggtgttg	agagcccaaagtcccaaggt	Mbanjo *et al.*
67	mMaCIR196	(TA)4,(TC)17,(TC)3	2	7	gctccaaacctcccttt	gttcgggctttcggt	Hippolyte *et*
68	Ma513051490	(CT)11	2	5	ccgcctcttccatagctgc	atcacaaggcgcctgctg	Mbanjo *et al.*
69	Ta6670	(GA)13	2	3	gcattccgcctatcaagtcgctg	tgttgccaacgtagatacctgctg	Mbanjo *et al.*
70	mMaCIR214	(AC)7	3	1,9,10	ccattgagagatcaaccc	ctccatcccccaagtcata	Hippolyte *et*
71	Ma513042326	(CAG)6	2	2	tcagaaggcagatcgaacagca	ccagaggagcatcccagagtg	Mbanjo *et al.*
72	Ma-2-10	$(CT)_{12}N_4(CT)_9$	2				Crouch *et al.*
73	Ma513046038	(AT)10	2	5	actgtccctcatgagttgcttacg	cgctacgactgggctcga	Mbanjo *et al.*
74	mMaCIR260	(TG)8	3	9	gatgtttgggctgtttctt	tgaccctccgacacctatt	Hippolyte *et*
75	Ta775	(GAT)7	2	2	catctgcacctgtggttgagg	tgcacgctcttcagctgc	Mbanjo *et al.*
76	Ta2955	(CT)15	3	4	cactacgctaacaggatagcaa	tgaagttgctagtgtgtttgctgact	Mbanjo *et al.*
77	Ma513019043	$(GAG)_3GNG(GAG)_2NA$	2	8	gttaacggccacctgcatgg	tcgaaggaacgatggccatctc	Mbanjo *et al.*
78	Ma513039300	(AG)18	2	1,2,8.10	tcgacggccaccgtgaac	cctggagattcagggttccgc	Mbanjo *et al.*
79	Ma513042336	(CT)9	2	10	gtgaagaacatctttggtggcctc	ggcatcacgcgactcgac	Mbanjo *et al.*
80	mMaCIR150	(CA)X10	2	11	atgctgtcattgccttgt	atgacctgtcgaacatcct	Hippolyte *et*
81	Ma513047251	(GT)11	2	7	accggtaaccaaatgcactgc	ggtcttcggtgttggcttgg	Mbanjo *et al.*
82	Ma513051273	(TCT)7	3	11	caagggaagtgaacagaaacct	agcttcctgtcgatgaggctg	Mbanjo *et al.*
83	mMaCIR307	(CA)X6	2	6	agacttgtatcgcttggtaaa	aagcaggtcagattgttcc	Hippolyte *et*
84	Ta1384	(GA)8CA(GA)3	2	1,2	aacttggcaacccacctgg	tgagtgcacggaaagcacatgtt	Mbanjo *et al.*
85	Ma513035997	(GA)10	5	10	gaggaccaatctgcgttcgc	acgcagcacaagtcgtcca	Mbanjo *et al.*
86	Ta253	(CT)16	2	11	ggacaaatcgacaaataaggga	acagtcatggtgggtgaggg	Mbanjo *et al.*
87	Ma-1-32	$(GA)_{17}AA(GA)_8AA(GA$	2	7	cacgtaaacaaggaggtgatc	caaaacactgtccccatctc	Crouch *et al*
88	mMaCIR42	(GA)16	3	6	cttttggagattattgcctaca	tgatggactcatgtgtacc	Hippolyte *et*
89	Ta4501	(AAG)7	3	8	cctccgcatttcgcaagcg	tggggggattcttggagtttcg	Mbanjo *et al.*
90	Ma513007351	$(GGA)_2N(GGA)_4$	2	6	ccctggagcaacagtgctactg	tcgaaggaacgatggccatctc	Mbanjo *et al.*
91	Ma513053096	(CT)10	2	4,7	cctccatcctttggccatcc	accctagtgacggcaacagag	Mbanjo *et al.*
92	mMaCIR38	(CT)12	2	6	gcaactttggcagcatttt	acgctgcaccagtcaa	Hippolyte *et*
93	Ma-1-18	(GA)11	2	7	gatgatggtgagaggctgatga	ccccctttcctcttttgc	Crouch *et al.*
94	Ta3183	(GCC)6	3	Not found	aaggccatccggctccag	tcgagcgctacgaggatgtc	Crouch *et al.*
95	Ma-1-2	(GA)10	3	7	gatgatggtgagaggctgatga	ggtcggtatgggaagcacc	Crouch *et al.*
96	Ta160	(TGC)6	3	9	ttgctaatatacgatgctgatgct	acccgtgttgatcgaacacca	Mbanjo *et al.*
97	Ta6591	(GAA)10TAA(GAA)2	3	7	cagcttcgtgatctcacccagaa	acaccgaggatgcgctgc	Mbanjo *et al.*
98	Ma513046494	(CTC)5CC(CTC)3	2	7	tggatgcggcgctccaag	gatgacgcagctgtggtcc	Mbanjo *et al.*
99	DN238509	(CACTG)4	2	7	tccgctgatgaactgtctgtctg	gctctgaggaaggccgtacc	Mbanjo *et al.*
100	Ta7568	(CTC)5CC(CTC)3	3	1	gaggggaagcttccagactacg	tgcgccgttgccgtagac	Mbanjo *et al.*

CHAPTER 3

MORPHOLOGICALLY DISTINCT EAST AFRICAN HIGHLAND BANANA CLONES ARE NOT GENETICALLY DIFFERENTIATED VIA AFLPs

Abstract

Background

The measurement and study of genetic differentiation among populations within a species' has been a fruitful approach to the detection of several evolutionary processes, including natural selection, gene flow, and genetic drift. When distinct populations or subpopulations are close to evolutionary equilibrium, differences in their genetic structure (i.e., the frequencies of different alleles and genotypes) reflect the potential role of natural selection in molding phenotypic and genetic variation. Assessing differentiation within populations (F_{ST}) is one of the current approaches to identify genome wide signatures of historic selective pressures on genome regions in the species.

Materials and Results

In this chapter, we used thirteen AFLP markers to genetically distinguish 90 EAHB morphologically distinct cultivars from Kenya and Uganda, determine their phylogenetic relationship and population structure and screen for genome wide "footprints" or signatures of selection. The markers demonstrated high polymorphism, 678/865 polymorphic bands were scored and primers pair polymorphic bands ranged from 37.5 % to 100%, but very low diversity indices (mean PIC and band diversity was 0.15 and 0.17 respectively). No fingerprints distinguished the various morphotypes and low diversity (Hp= 0.177) of AFLP bands among the cultivars was observed. Shared genetic distance (D_{AS}) and Dice dissimilarity between pairs of individuals belonging to the same or to different morphological groups was narrow (between 0.1000 and 0.3000) and largely overlapped suggesting lack of divergence. Diversity of cultivars was distributed within the morphological groups (cloneset) (Hw=0.1935) compared to among (Hb=0.0054) them, corroborating AMOVA results, 97% variation resided within cultivars in the groups compared to among the groups, 3%. The covariance component at both levels was found to be significant ($p<0.001$). Furthermore, the overall genetic differentiation values (ϕ-$_{PT}$ and F_{ST}) among the cultivars was significantly low (0.030 p=0.0001, 0.0271 respectively). Structure analysis, PCA, and neighbour joining tree grouped cultivars in two distinct clusters and all Kenya and Uganda cultivars separated in referred clusters, suggesting presence of geographic pattern and lack of genetic basis of their morphology. Most interestingly, Bayescan results indicated balancing selection of the EAHB

subgroup, highest $\log_{10}$ (PO) = -0.9159, and significant signatures of divergent natural selection appeared strongest between the EAHB and out-group cultivars, highest log10 (PO) = 2.6565. Evidence of recent population bottlenecks under the TPM and IAM models was significant (Wilcoxon test, P=0.0000) and both tests showed evidence for the loss of genetic diversity for the neutral loci.

Conclusion

In addition to the high genetic similarities between cultivars in the EAHB subgroup, this chapter demonstrates that the EAHB may have passed through a population bottleneck before a rapid expansion coinciding with migrations out of Papua New Guinea. A founder effect (founder populations bring only a subset of the genetic variation from their ancestral population) caused by the rapid expansion of a previously small population of the EAHB subgroup thus brought effects on the distribution of genetic variation. Smaller (founder) populations experience greater genetic drift because of increased fluctuations in neutral polymorphisms; this may serve as an explanation for the balancing selection observed in this population.

Keywords: Amplified fragment length polymorphism (AFLP), morphotypes, genetic diversity, balancing selection, bottleneck,

3.1 INTRODUCTION

Amplified Fragment Length Polymorphism (AFLP) (Vos *et al.*, 1995) is a versatile and firmly established molecular marker technique for genome-wide screening of genetic diversity (Blignaut *et al.*, 2013). AFLP has been an extensively used DNA fingerprinting method for many studies in plants (Meudt & Clarke, 2007; Garcı´a-Pereira *et al.*, 2010) animals and microorganisms (Nath *et al.*, 2013). The AFLP technique relies upon detecting genetic polymorphisms based on selective amplification of DNA fragments from digested total genomic DNA, through differential endonuclease restriction digestion total genomic DNA, generating reproducible fingerprints that are usually recorded as a 1/0 band presence–absence binary matrix (Vos *et al.*, 1995). AFLP is a useful marker system for resolving genetic relatedness among individual organisms, populations and species (Mueller & Wolfenbarger, 1999). The presence of phylogenetic signal in many AFLP data sets (Koopman, 2005) has stimulated its use as a source of genetic information for phylogenetic inference, particularly among closely related genera or species (Meudt & Clarke, 2007; Garcı´a-Pereira *et al.*, 2010).

High levels of polymorphism and high degree of discriminative capacity are the main advantages of AFLPs for the analysis of closely related genotypes (Ercisli *et al.*, 2011) and provide a rapid and inexpensive source of multilocus allele frequency data for making genomically robust inferences (Meudt & Clarke, 2007). AFLP markers are predominantly nuclear, and widely distributed throughout the genome, thus generating phylogenies based on multiple rather than single genomic regions (McKinnon *et al.*, 2008). AFLPs are particularly powerful for studying the phylogeny of organisms such as plants for which other nuclear and organellar markers are often lacking, insufficiently variable, or even inappropriate (Pellmyr *et al.*, 2007). Other strengths of AFLP markers include; efficiency because a pair of PCR reactions can be used to simultaneously amplify fragments from multiple chromosomal loci (Vos *et al.*, 1995; Zhang & Hare, 2012). The AFLP approach offers

repeatability and generates large numbers of potential markers across the genome that may counteract the low information content of its dominant markers (Blignaut *et al.*, 2013). In the previous chapter, low numbers of polymorphic SSR loci were observed. Given the comparative ease of gaining large numbers of data from few AFLP loci, compared to SSRs, the application of AFLPs for diversity studies would be feasible (Garcia *et al.*, 2004).

Molecular markers have played an important role in aiding the assessment of genetic diversity in a number of Musa species. Most *Musa* subspecies meet the conditions set by Meudt and Clarke (2007) in which the AFLP technique can be ideal for accurate phylogeny estimation. These include high genomic heterogeneity (i.e., when it is necessary to analyze many loci to ascertain an accurate measure of genomic diversity), low genetic variability (generally intra specific) and studies of polyploids where it is very difficult to use single-locus nuclear sequencing markers because of problems distinguishing the many alleles that may be present at each locus. In this regard, AFLP markers have been demonstrated to be a powerful tool capable of determining the genetic diversity wild and domesticated *Musa* subspecies and related species (Ude *et al.*, 2002a; Wong *et al.*, 2002; Ude *et al.*, 2002b; Tugume *et al.*, 2003; Ude *et al.*, 2003; Noyer *et al.*, 2005; Wang *et al.*, 2007; Opara *et al.*, 2010; Youssef *et al.*, 2011; Shaibu, 2012; Ahmad *et al.*, 2014). AFLPs are widely used to estimate phylogenies and population structure in a range of plant species (de Faria-Tavares *et al.*, 2013). AFLP has a number of broad applications, ranging from linkage mapping to analyses using population-based and phylogenetic methods. Of particular interest in this study is the use of AFLPs to generate data for genetic diversity and a phylogenetic study of the triploid East African Highland bananas.

Insufficient information exists for an adequate classification of the east African Highland bananas, and the system divisions that currently exist are wholly morphology based, and therefore not satisfactory due to their susceptibility to

changes in the environment. In this study, we employ an amplified fragment length polymorphisms (AFLP) approach to: (i) investigate genetic diversity of the triploid EAHB subgroup; (ii) explore population structure and relationships that exist within the subgroup and differentiate between cultivars grown in different regions of East Africa; and (iii) examine presence of outlier loci in the EAHB population. Results obtained are used as an additional tool to develop a more robust classification of the members of the EAHB subgroup.

3.2 MATERIALS AND METHODS

3.2.1 DNA samples

DNA samples and cultivars were same as used with SSRs (Chapter two). Since AFLP protocol uses higher amount of DNA compared to SSR, the DNA samples for this procedure were diluted to 250 ng/ml working stocks based on spectrophotometric measurements.

3.2.2 Amplification procedure

Amplification procedure AFLP fragments were generated according to a modified version of the procedure outlined by (Vos *et al.*, 1995). Thirteen primers (one fluorescently labeled), adaptors, sequences are provided in S Table 4. Enzymes and PCR components were all purchased from New England biolabs and restriction digests were simultaneously performed: the digestion included 1µg of genomic DNA in a 40 µl reaction volume with 5U MseI and EcoR1 enzymes each, 1X enzymatic buffer and 0.1mgml–1 BSA at 37°C for 3 hours. Pre-selective PCR included 0.1mgml–1 BSA , 0.5 µM of the EcoRI + 0/MseI + 0 primers 0.25 U Taq DNA polymerase and 5.0 µl of digestion ligation product) in 20 ul reactions. Selective PCR included 2.0 µl (1:10 dilution) of preamplified DNA, 0.6 nmol *Mse*I primer, 0.5 nmol *EcoR*1 primer,

0.2 mM dNTP mix, 0.24 mM $MgCl_2$, 1.0 standard Taq buffer, 0.625 U Taq polymerase and 0.0024mg ml–1 BSA in 10 µl reactions.

The PCR conditions differed depending on; the nature of the selective extensions of the AFLP primers used for amplification. AFLP reactions with primers having non selective nucleotide were performed for 30 cycles with the following cycle profile: 30s DNA denaturation step at 94°C, 1 min annealing step at 56°C and 2 min extension step at 72°C. Before and after the cycles initial denaturation at 95°C for 3 mins and final extension at 72 °C for 10 mins were done respectively. AFLP reactions with primers having three selective nucleotides were performed for: (i) 15 cycles with the following cycle profile: 30s DNA denaturation step at 94°C, 30s annealing step (see below), 1 min extension step at 72°C (ii) 23 cycles of 30s at 94°C, 1 min annealing step (see below), 2 min at 72 °C; followed by 10 min final extension at 72°C. The annealing temperature in the first cycle was 65 °C was subsequently reduced each cycle by 0.7 °C for the next 12 cycles and continued at 56 °C for the remaining 23 cycles. All amplifications were performed in 384-plate GeneAmp® PCR System 9700 Thermal Cycler (Applied Biosystems). For the reproducibility test, the AFLP reactions from DNA restriction to selective PCR amplification were repeated for 10 samples and one negative control to test the repeatability of the fragments. Accordingly, a total of 929 fragments with 100% reproducibility in the range of 50-500bp were considered for further analysis.

3.2.3 AFLP Scoring Details and Creation of primary binsets

The success of selective DNA amplifications was confirmed on a 2.0% w/v agarose gel. Due to the number of fragments generated using the AFLP technique and the effect of dye quenching, only two fluorescent dyes were used and post PCR co-loading of PCR product were done based on the dye of the primer, 1.5 µl for NED, and 1.0 µl for 6-FAM. AFLP fragments (1 µl of

PCR product cocktail) were run with an GeneScan 500 LIZ internal size standard (0.012 µl) and formamide (9 µl) on an ABI PRISM 3730 XL genetic analyzer (Applied Biosystems) after denaturation at 95° C for 5 mins and rapid cooling.

To guide the optimization in terms of minimizing genotyping error, while maximizing the number of loci retained a set of ten genotypes plus one outgroup was fragment separated (ABI 3730) and scored using GeneMapper 4.0 (Applied Biosystems). To ensure consistency among electrophoretic runs, a control comprising the same sample amplified with the same primer pair was included in every run, and the fragment profile from this control was compared across runs by eye. Several measures were taken; the advanced peak detection algorithm was used, with light smoothing turned on and all other settings left at defaults. Genotyping error, caused by co-migrating fragments from two or more loci (homoplasy), were avoidable by scoring only fragments of longer length, (Vekemans *et al.*, 2002) of between 50-500bp. Marker selection and fragment calling stages AFLP genotyping errors were minimized using locus elimination criteria and peak height (that is signal amplitude) fragment calling thresholds for each locus independently based on the peak height distribution across all samples to minimize false positives and negatives (Hornemanna *et al.*, 2012). Thus, only fragments with relative florescence units >100. The optimization of fragment size categories (referred to as bins), i.e. a correct assessment of statistical variability of electrophoretic mobility of fragments, which is necessary to avoid ‘‘oversplitting’’ of identical alleles into separate characters or merging of non-identical alleles into one character (technical homoplasy) was done by exploring scoring of fragments with various bin widths; 0.5, 1.0, 1.5 and 2.0. In preliminary analyses, bin widths of 2.0 produced topologies with the best resolution, so we used this bin width for our final analyses. Mean genotyping error rate per locus was 0.481% based on replicate analysis of 10 samples.

3.2.4 Data analysis

3.2.4.1 AFLP marker evaluation to detect polymorphisms in EAHB

Each AFLP fragment was considered as a putative locus and assumed a dominant marker with two alleles. Fragments were manually scored for their presence (1) and absence (0) in each sample generating a binary matrix that was then used for further analysis. Parameters for calculating the marker efficiency and genetic characteristics were done after removal of out-groups; rare bands were regarded as bands with frequency <5%; the percentage of polymorphic loci (P) and mean number of alleles per locus (A); were directly calculated from AFLP phenotypes. The polymorphic information content (PIC) for each primer combination was calculated as per Roldán-Ruiz *et al.* (2001): $PICi = 2fi(1 - fi)$,where PICi is the polymorphic information content of marker i, fi the frequency of the marker band which were present and 1−fi the frequency of marker bands which were absent. PIC is the relative discriminatory value of a locus which measures the information content as a function of a marker system`s ability to distinguish between genotypes (Weir 1990) and was averaged over the fragments for each primer combination using Powermarker v3.25. Allelic frequencies of AFLP marker were used separately to estimate number of average number of effective alleles (Ne) as; $\mathrm{Ne} = 1/(\mathrm{p\hat{}2} + \mathrm{q\hat{}2})$. Gene diversity, often referred to as expected heterozygosity and defined as the probability that two randomly chosen alleles from the population are different expected heterozygosity was calculated as; $He = 2 * p * q$ and Unbiased Expected Heterozygosity $uHe = 2N\ /\ (2N - 1)) * He$ (where, q = (1 - Band Freq.)^0.5 and p = 1 - q) with respect to Hardy–Weinberg using GenAlex ver 6.5 (Peakall & Smouse, 2006, 2012). Fis (Inbreeding coefficient) was calculated as; $1 - (\text{Observed He}) / (\text{Expected He})$. Shannon's Information Index was calculated to show the abundance of AFLP markers in the EAHB genome using GenAlex v6.5.

3.2.4.2 Genetic diversity within the EAHB population

To assess overall genetic variation within the EAHB subgroup, a population based comparison of allele frequencies of AFLP phenotypes was used to partition genetic diversity following the Bayesian approach with non-uniform prior distribution for allele frequency (Zhivotovsky, 1999) using AFLP-Surv 1.0 (Vekemans, 2002). For dominant markers, the presence of a band (or peak) can indicate either the homozygous condition or the heterozygous condition; therefore the frequency of the null allele must be estimated. Calculating allele frequencies (i.e. heterozygosity) from dominant markers is difficult but can be accomplished by using a Bayesian approach (Zhivotovsky, 1999) or the inbreeding coefficient and the square root of the frequency of the null homozygote (if Hardy–Weinberg Equilibrium; HWE) is assumed. For outcrossing species, their allele frequencies usually do not violate HWE, therefore, both of these approaches can yield good estimates of average heterozygosity (Krauss, 2000). For species such as self-fertilizing and parthenocarpic plants, the Bayesian approach does not assume HWE and is thus superior (Zhivotovsky, 1999; Meudt & Clarke, 2007; Zhang & Hare, 2012). These allelic frequencies were used as input for the computation of phenotypic gene diversity (Hp) following estimates of diversity from (Mariette *et al.*, 2002); $Hp = 1 - Pi2 - Qi2$, where Pi and Qi are the frequencies of band presence and absence, respectively. Estimates of Hp were calculated for each locus, and the mean over all loci was used as the overall estimate of diversity of the EAHB population. Frequencies of AFLP phenotypes were further used as input for the AFLPDiv 1.0 program (Coart *et al.*, 2005) to compute the percentage of polymorphic loci (PLP). The total number of AFLP bands present in the EAHB population here after referred to as AFLP band richness or Dγ were calculated in excel using the countif function, then partitioned into a within- and a between-individual component using an additive model (Lande, 1996; Puşcaş *et al.*, 2008). Within-individual AFLP band diversity (Dα) was calculated as the mean number of AFLP bands per individual. Between-individual AFLP band diversity (Dβ) was calculated as

the mean number of AFLP bands that were absent in an individual AFLP phenotype (Puşcaş *et al.*, 2008). An inbreeding coefficient of 0.0 was assumed due to self-incompatibility (Hornemanna *et al.*, 2012; Kolb & Durka, 2013). Unbiased estimates of average diversity within cultivars were given (Hw) with its variance components (total variance, Var(Hw); variance due to sampling of individuals, VarI(Hw); variance due to sampling of loci, VarL(Hw); and variance due to sampling of populations, VarP(Hw). The total gene diversity (Ht), i.e. expected heterozygosity or band diversity in the overall sample was calculated as the sum of the average diversity within cultivars (Hw) and the average diversity among morphological groups in excess of that observed within cultivars (Hb). To assess the level of differentiation between cultivars (Vekemans, 2002) Wright's F_{ST} value was estimated using AFLP-Surv 1.0.

3.2.4.3 Genetic similarity and relatedness of the cultivars

To get an overall measure of how similar (or different) the cultivars are, two approaches were used; (i) genetic dissimilarity matrix was computed based on AFLP phenotypic data using Dice's coefficient (Dice, 1945). Dice dissimilarity matrix was produced using the presence/absence dissimilarity index ($dij = (b + c)/(2a + (b + c)$ of DARwin version 5.0 (where dij=dissimilarity between units i and j; number of variables where a=Xi=presence and Xj=presence; b=number of variables where Xi=presence and Xj=absence; c=number of variables where Xi=absence and Xj=presence); (ii) Genetic distance (D_{AS}) between cultivars was computed using shared allele genetic distance calculated as; $1 - proportion\ of\ shared\ alleles$ (Bowcock *et al.*, 1994). Shared allele genetic distance matrix and Dice dissimilarity were represented as percent frequency counts.

3.2.4.4 PCA and population structure

To mimic the approaches used in population genetic variation analyses, Principal coordinate analysis (PCA) was conducted with NTSYS-pc software package v2.3.3 (Rohlf, 2001a), based on the simple-matching (SM) coefficient of Sokal and Michener (1958). This multivariate approach was chosen to complement the cluster analysis information, because cluster analysis is more sensitive to closely related individuals, whereas PCA is more informative regarding distances among major groups (Zhang *et al.*, 2007).

Genetic distance matrices were generated using the Nei and Li (1979) and used to run cluster analysis based on Neighbour-joining (NJ) using the un-weighted pair-group method with arithmetic averages (UPGMA) of Powermarker v3.25. Bootstrap analysis, which is a method for determining confidence limits in clusters produced by UPGMA-based dendrograms was performed. In order to obtain statistically accurate bootstrap *p* values at 99% level, all dendrograms had 10,000 replications. Thereafter, the cophenetic correlation value (*r-value*) coefficient was used to test for association between the clusters in the dendrograms and the dissimilarity matrices from which they were produced.

To view overall genetic structure between the cultivars, Bayesian population assignment tests were STRUCTURE 2.3.3 (Pritchard *et al.*, 2000; Falush *et al.*, 2007). Estimation of the number of populations (K) in Structure was conducted using three replicate exploratory runs at each level of K from K=1 to K=10 using an admixture model with correlated allele frequencies with 200,000 iterations for the length of burn-in period and subsequent number of MCMC (Markov-Chain-Monte-Carlo) repeats with lambda set at the program default of 1.0 for exploratory analyses. Chain convergence was assessed by examining the output graphs of alpha vs program run number provided by STRUCTURE. Individuals were grouped into genetic clusters representing homogeneous gene pools with and without a priori information about individual origin. The optimal level of K was calculated as the as the second order of likelihood

change (Δk) using CLUMP software (Evanno *et al.*, 2005; Earl & vonHoldt, 2011). This was adopted due to its sensitivity compared to the LnP(K) method to detect the number of subpopulations and in circumstances where the K value does not reach a clear plateau (Evanno *et al.*, 2005; Jesus *et al.*, 2013). We assigned each individual to a cluster whenever STRUCTURE estimated that at least 80% of its genome originated from that cluster.

3.2.4.5 Variation and differentiation among EAHB morphological groups (clonesets)

To measure the variation within and among the clonesets, the following parameters were computed using GenALEx v6.5 (Peakall & Smouse, 2012); observed number of alleles (Na), mean number of observed alleles (A) (Kimura & Crow, 1964); number of effective alleles (Ne), % of polymorphic loci; number of private bands (alleles/bands unique to a cloneset); Nei`s genetic diversity (h) (Nei, 1987) (equivalent to the average expected heterozygozity *He* in the population, (Bonin *et al.*, 2007) and Shannon`s information index (I) (Lewontin, 1974). The significance of these analyses was determined by the formula Mean±SE. Means were compared at 95% level of significance using ANOVA with Welch's correction performed by GraphPad Prism version 3.0 for windows (http://www.graphpad.com).

Genetic differentiation, which measures among population component of genetic variance, was calculated to determine the proportion of total variation that was due to differences between population allele frequencies. The coefficient of gene differentiation, Phi_{PT}, (the analogue of F_{ST} fixation index and Gsts), variance components and their significance levels was obtained using AMOVA (GenAlEX v6.5) following the methods of Excoffier *et al.* (1992), Huff *et al.* (1993) and Michalakis and Excoffier (1996). Levels of significance were based on 1000 permutations. To estimate band richness of each cloneset and a rarefaction measure of genetic variation independent of

sample size was standardized to the smallest sample size (n=14) (Zhang & Hare, 2012). Allele frequencies computed using AFLPSurv was used as input for AFLPDiv v1.1 package. Genetic distances between clonesets were calculated based on Nei (1983) genetic distance of Powermarker v3.25 (Liu & Muse, 2005).

Genetic differentiation among clonesets was quantified with F-statistics following Lynch and Milligan (1994), calculating overall and pairwise F_{ST} values between groups (Kolb & Durka, 2013). Monomorphic loci were excluded from analyses (187 loci) and mean allele frequency was calculated as the arithmetic average of the band-absent frequency in the five clonesets (referred to as groups) for the remaining loci. For each locus, genetic differentiation between populations was measured in terms of F_{ST}, calculated as; $1 - (\mathrm{HS\ \ HT})$ using AFLPSurv, where H_S is the mean locus-specific heterozygosity within clonesets and H_T is the locus-specific total heterozygosity (Nei, 1973). Negative values of F_{ST} were converted to zero. Then global F_{ST}, a measure of central tendency for the distribution of F_{ST} across loci, was calculated as; $1 - (mean\ HS\ \ mean\ HT$, where mean HS and mean H_T are the arithmetic averages across loci (Nei & Chesser, 1983). Standard error of F_{ST} for individual loci was estimated by calculating the F_{ST} and resampling statistics based on 1000 random permutations of individuals among populations.

Hierarchical analysis of molecular variance (AMOVA) was performed in GenAlEx 6.5 (Peakall & Smouse, 2006, 2012). This analysis enables partitioning of the total AFLP variation into within and among the populations variation components, and provides a measure of inter-population genetic distances as the proportion of the total AFLP variation residing between the EAHB cultivars of any five subpopulations (called Phi statistics). AMOVA calculates $\Phi_{PT,}$ an analogue of F_{ST} using the squared Euclidean distance matrix between allele phenotypes and allows hierarchical analysis of genetic

structures. Φ_{PT} is a band-based approach and does not depend so critically on specific assumptions that could underestimate genetic variability and is specifically recommended for band based data (Hufford *et al.*, 2013).

3.2.4.6 Phylogenetic relationships

To assess the relationship between pairs of cultivars and estimation of the putative amount of time since the two cultivars diverged from a hypothetical common ancestor, a genetic distance matrix was constructed using Nei and Li (1979) similarity coefficient. The genetic distance measure reflects the assumption that mutations occur independently throughout genome and that the events are exponentially distributed over time. Therefore, the time of the next mutation is assumed to be independent of the times of past mutations. Some number of such changes, beginning from a hypothetical shared ancestor, characterizes the relation between each pair of cultivars. The lowest of these distances is termed the "nearest genetic distance," and it indicated the degree of homology with the compared cultivar. Genetic relatedness of the cultivars and outgroups, was displayed using neighbor-joining tree using the unweighted pair group method with arithmetic mean (UPGMA) (Powermarker v3.25 (Liu & Muse, 2005). The tree was visualized in MEGA software package version 5.0 (Tamura *et al.*, 2011) and the relative support for the different groups and stability of the tree was assessed by bootstrap analysis (2000 replicates). The cophenetic correlation coefficient was calculated to provide statistical support for goodness-of-fit of the tree and cluster analysis of the matrix on which it was based.

3.2.4.7 Footprints of selection

To identify candidate loci under natural selection or strongly differentiated loci in the EAHB subgroup outlier analysis was conducted using BayeScan 2.01. The analysis aimed to detect loci under selection by comparing allele

frequencies, assuming they followed a multinomial Dirichlet distribution, which takes into account complex demographic models with varying gene flow between loci and between populations. One of the scenarios covered consists of an island model in which allele frequencies are correlated through a common migrant gene pool from which they differ in varying degrees. The difference in allele frequency between this common gene pool and each locus is by a specific F_{ST} coefficient. The posterior probabilities of two models are compared: one including selection via a locus specific F_{ST} component to explain observed allele frequency differences, and a 'neutral' model with only population-specific F_{ST} parameters which are shared across all loci. If the model including a locus-specific F_{ST} component is necessary to describe the observed allele frequencies, then a departure from neutrality is assumed for that locus. BayeScan runs were implemented using a uniform distribution of F_{IS} between 0–1, prior odds for the neutral model of 1, and default values for all other parameters, including 20 pilot runs at 100,000 iterations in total, 50,000 of which consisted of a burn-in period (Foll & Gaggiotti, 2008). According to Robert *et al.* (2009), a log posterior odds (log PO) equals (Bayes factor) x (prior model odds) > 1 is considered as strong evidence for positive selection and balancing selection is invoked if < 0 on Jeffrey's scale. For each locus, the probability of being under selection is then inferred using the Bayes factor (BF). Based on Jeffreys' (1961) scale of evidence, a $\log_{10}$ BF of 1.5–2.0 is interpreted as "strong evidence" of selection (Soto-Cerda & Cloutier, 2013). The posterior odds is the ratio of posterior probabilities of the selection and neutral models and also allows the control of the False Discovery Rate (FDR), the proportion of false positives among loci classified as under selection. We checked that all loci classified as significantly differentiated at $\log_{10}$ (PO) remained significant after applying an FDR P value 0.05 using the method provided in the user's manual (Foll, 2012).

To distinguish between selection and bottleneck effects, bottleneck analyses were conducted separately for neutral and outlier loci. Deviations from

expected heterozygosity using the program BOTTLENECK 1.2.02 were computed with 5,000 coalescent simulations assuming a two-phase mutation model (TPM) and infinite allele model (IAM (Cornuet & Luikartt, 1996). Significance of deviations was determined by sign test and standardized differences test in addition to a one tail for heterozygosity (H) excess or deficiency and a two tail for He excess and deficiency using the Wilcoxon test. The population was also tested for loss of rare alleles. Rare alleles were defined as those that occurred they occurred at a frequency of less than 0.05 in the examined populations (Wang *et al.*, 2012). Bottlenecks are known to cause a characteristic mode-shift distortion in the distribution of allele frequencies at selectively neutral loci. Moreover, low-frequency alleles (<0.1) would be more lost rapidly during bottleneck than ones at higher frequencies (Maruyama & Fuerst, 1985; Luikart *et al.*, 1998).

3.2.4.8 Comparisons of AFLP and SSR results

A Mantel test (Mantel, 1967) with 1,000 permutations was used to estimate the correlation (association) significance between the distance matrices resulting from SSR, AFLP and combined analyses, the test was done using the NTSYS pc 2.1 software (Rohlf, 2001).

3.3 RESULTS

3.3.1 AFLP efficiency

A total of 929 AFLP bands (loci) were reliably scorable for polymorphism, of these 865 bands were scored in both EAHB cultivars and 64 bands were scored in out-groups only. For calculation of genetic diversity of cultivars in the EAHB subgroup, only the 865 AFLP bands were considered and showed a

high percent of polymorphic loci, 78.3% (678 bands, both absent and present band considered). The number of loci (or fragments) scored per primer combination ranged from 44 (E-ATG x M-CGT) to 82 (E-AGT x M-CTG and E-AGA x M-CCA) with an average of 66.5 bands per combination. Primer combinations used were polymorphic ranging between 37.5% (E-AAG × M-CTA) to 100% (E-AGA × M-CTC, E-AGG ×M-CTA and E-AGT x M-CTT) however the number of bands with frequency <5% (rare bands) differed among them and their mean was low (6.4). Mean PIC value and gene diversity were low, 0.15 and 0.17, respectively (Table 6). Only 4 primers combination had PIC values > 0.2, almost half the highest PIC value (0.5) for dominant markers. Overall, measures of genetic diversity for the EAHB cultivars over all loci were relatively low; gene diversity or expected heterozygosity (He; Mean±SE, 0.150±0.003), unbiased expected Heterozygosity (uHe; 0.162±0.003), number of different alleles (Na; 1.441±0.010) and number of effective alleles (Ne; 1.266±0.005). However, Shannon's Information Index (I) was slightly higher at 0.250±0.004 and F_{IS} value was 0.058 (data not shown).

Table 6: Attributes of AFLP primers used in this study; number of fragments, polymorphic bands, band diversity and Polymorphism information content (PIC) of the 13 *EcoR*1 (denoted as E) and *Mse*1 (denoted as M) primers

Primer pair	No. of bands	Polymorphic bands (%)	Bands with frequency < 5%	Band diversity (He)	PIC
E_AAG x M_CTA	56	37.5	1	0.10	0.08
E_AGA x M_CCA	82	63.4	8	0.15	0.12
E_AGA x M_CTC	52	100.0	15	0.15	0.13
E_AGA x M_CTG	61	70.5	-	0.14	0.12
E_AGC x M_CTT	81	38.2	8	0.10	0.08
E_AGG x M_CCT	72	55.6	4	0.11	0.10
E_AGG x M_CTA	73	100.0	14	0.25	0.21
E_AGT x M_CGA	67	80.6	20	0.15	0.13
E_AGT x M_CTG	82	97.6	4	0.28	0.22
E_AGT x M_CTT	72	100.0	-	0.23	0.20
E_ATC x M_CTA	72	91.7	8	0.29	0.23
E_ATC x M_CTC	51	74.5	-	0.19	0.16
E_ATG x M_CGT	44	47.7	2	0.12	0.10
Mean	**66.5**	**73.6**	**6.4**	**0.17**	**0.15**

3.3.2 Low genetic polymorphisms detected among the EAHB-AAA cultivars

Diversity of AFLP bands among the cultivars was low (Hp= 0.177), the average frequency of present (p^2) and absent (q^2) bands was 0.5443 and 0.2784, respectively. The total number of AFLP bands present in the EAHB population Dγ (band richness) was 865 and the within (Dα) and between (Dβ) individual AFLP band diversity was 547.43 and 317.48 respectively. The total diversity (Ht), average diversity within cultivars (Hw) and diversity in the morphological groups (Hb) and their variances as calculated following Lynch & Milligan method are reported in (Table 7 and 8). Shared genetic distance (DA) and Dice dissimilarity calculated between pairs of individuals belonging to the same or to different morphological groups largely overlapped. The overall average genetic distance between cultivars was 0.1771 (SE=0.02)

within the range of 0.0439 (Red Nakitembe and Nabuyobo) and 0.3191 (Rwambarara and Ekeganda). The mean Dice dissimilarity between the cultivars was 0.1409 ranging between 0.0342 and 0.2747 minimum and maximum, respectively. A summary of genetic distance and their frequencies for all EAHB cultivars is presented in Figure 13, shows a high frequency of accessions (66.6% in AFLP) with genetic distance within the range of 0.1001 and 0.2000. It is not surprising that, genome-wide genetic differentiation of the cultivars was significantly very low (F_{ST}=0.0271, p=0.0000 Table 8).

Table 7: Gene diversity within the EAHB population (Lynch & Milligan method)

Hw	S.E.(Hw)	Var(Hw)	VarI(Hw)	VarL(Hw)	VarP(Hw)
0.935	0.00342	0.000012	0.000001	0.000006	0.000004
		Percent of Var(Hw)	11.42	52.92	35.66

Var(Hw); total variance,
VarI(Hw) variance due to sampling of individuals
VarL(Hw); variance due to sampling of loci,
VarP(Hw)variance due to sampling of morphological groups

Table 8: Population genetic structure (Lynch & Milligan method)

Ht	Hw	Hb	Fst
0.1989	0.1935	0.0054	0.0271
S.E.	0.003419	0.001462	0.268851
Var	0.000012	0.000002	0.072281

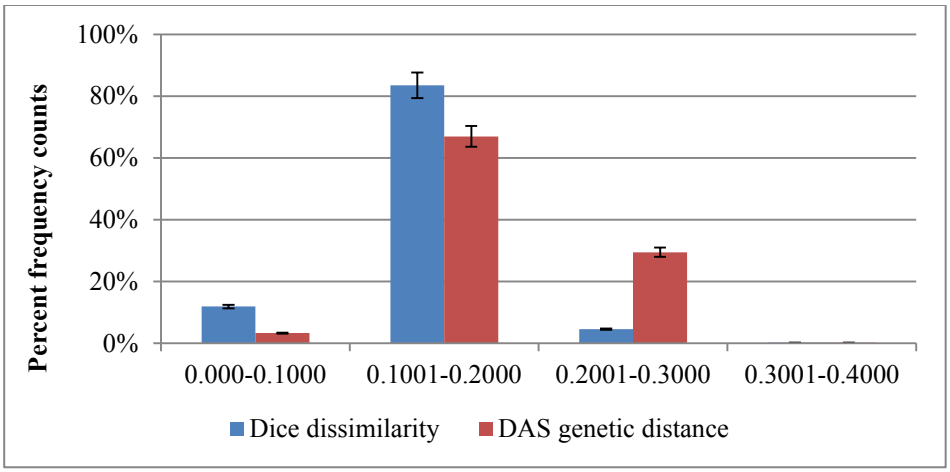


Figure 13: Column graph of the % of frequency counts of shared allele genetic distance (D_{As}) and Dice similarity between cultivars of the five EAHB morphological groups.

3.3.3 EAHB diversity is geographically structured

The PCAs provided visual representations of genetic proximities between all cultivars of (Figure 14). The PCAs revealed that 63.9% and 10.9% of the overall variation was accounted for by R1 and R2, respectively. Coordinates did not separate the cultivars into distinguishable EAHB populations and cultivars displayed a closer affinity with each other despite the differences in their morphology. However, there was a tendency for clustering based on the population origin of the 90 cultivars and a clear separation from the out-group cultivars though MunjuP was grouped closer with the EAHB but slightly different than for SSRs. For the whole data, set including all EAHB and out-group cultivars, ΔK approach peaked at K=2. The EAHB Uganda cultivars showed no admixture, however, the Kenyan group showed an admixture of the two groups (Kenya and Uganda). Surprisingly, all out-group cultivars clustered with the EAHB from Kenya but the proportion of admix differed. The level of assignation to one or the other cluster showed a continuous gradient across individuals and all samples were assigned to one or the other cluster with a probability of at least 70% (Figure 15A; S Table 5). For K=2, cluster

assignments were perfectly consistent with the two EAHB geographical regions and not morphological (Figure 15B).

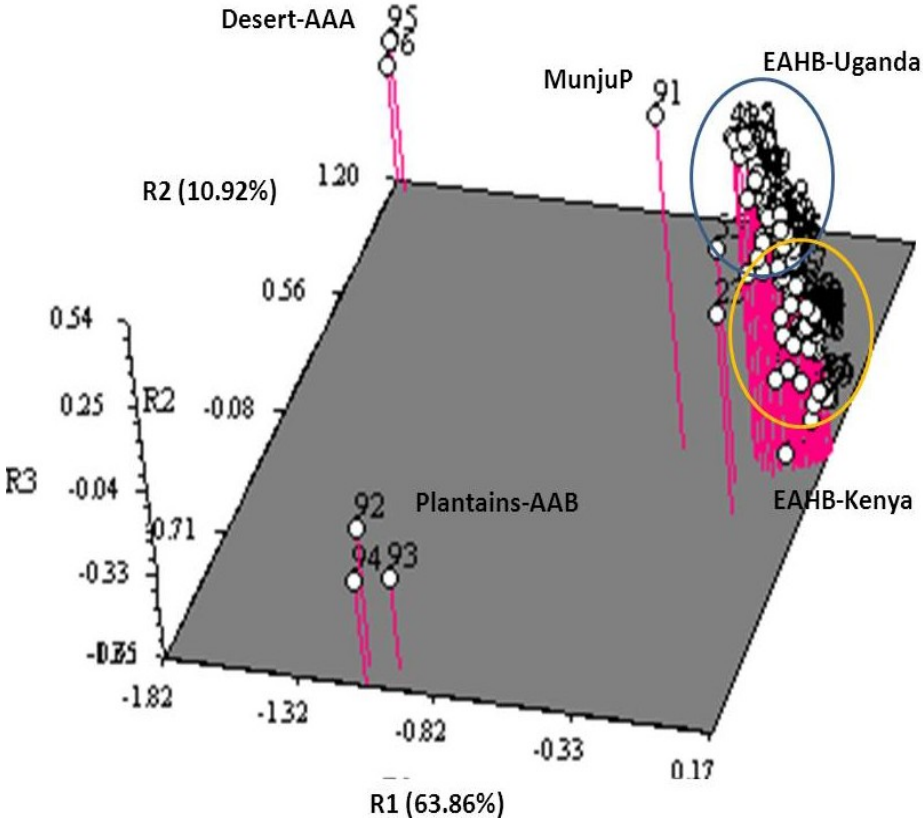


Figure 14: Principal coordinate analysis plot of EAHB. The PCA partitions the population into two groups (Kenya and Uganda) while the out-groups clusters separately. R1 and R2 explain 63.9% and 10.9% of the total variance, respectively**.**

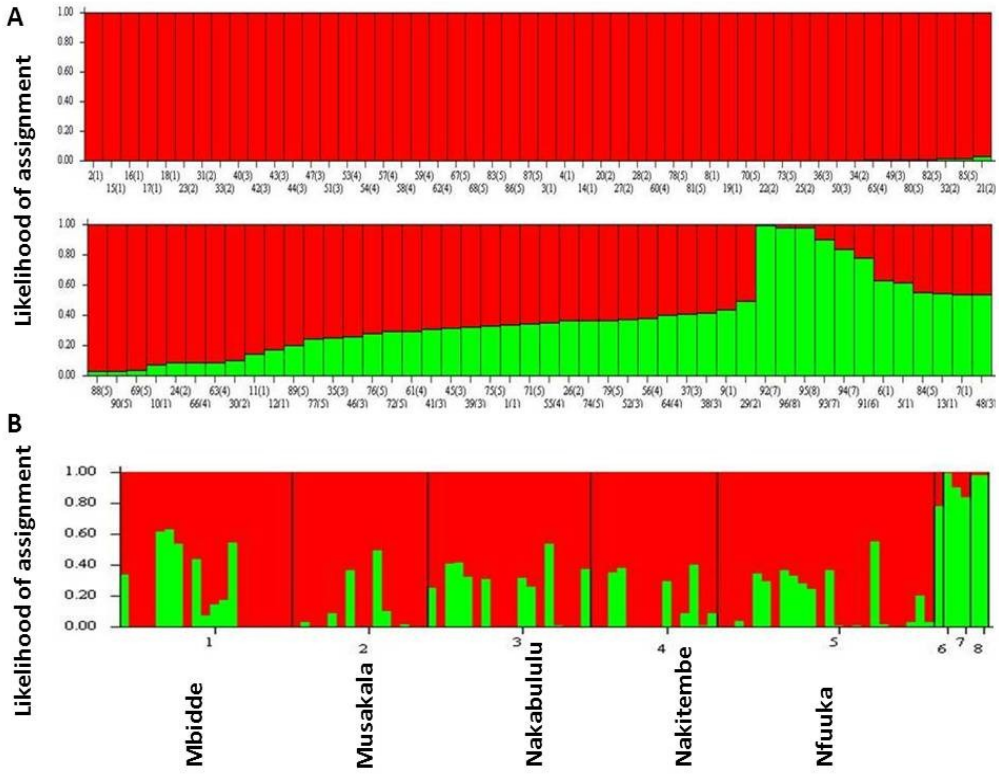


Figure 15: Summary plot from the STRUCTURE analysis. Presenting the proportional assignment of (A) each the 96 cultivars without prior population assignment and (B) numbers 1-5 represents the EAHB cloneset; 6 is Munju; 7 is AAA desert for the K inferred clusters for K=2. In the assignment level of 0.7 is indicated for each cluster with a hatched line. Each individual/cloneset is represented by a thin vertical line, which is partitioned into K coloured segments that represent the individual's estimated membership fractions.

3.3.4 Partitioning of Variation and Genetic Divergence among EAHB clonesets

Genetic diversity values varied amongst morphological groups but was relatively low, overall. Among groups, gene diversity (He), the percentage of polymorphic loci (PLP) and band richness (Br) varied between 0.157 and 0.194 (mean ± SD, 0.170 ± 0.0135), 50.1% and 54.7% (52.16 ± 2.06) and 1.061 and 1.097 (1.078 ± 0.01), respectively (Table 9). Mean expected heterozygosity values were generally similar across the morphological groups and no correlation was found between mean heterozygosity and sample size

(n). Significance tests are indicated for comparisons between the EAHB groups only.

Table 9: Comparison of the variation within the EAHB morphological groups (clonesets) and EAHB groups versus the out-group

Pop	n	Na	Ne	I	[a]He	UHe	[b]PLP	[c]Pb	[d]Br
Mbidde	19	1.36	1.274	0.247	0.163	0.167	54.7	2	1.070
Musakala	15	1.35	1.277	0.251	0.165	0.171	50.9	2	1.097
Nakabululu	18	1.31	1.246	0.221	0.146	0.150	50.1	1	1.081
Nakitembe	14	1.29	1.264	0.233	0.155	0.160	50.4	4	1.082
Nfuuka	24	1.45	1.306	0.274	0.181	0.184	54.7	5	1.061
Outgroup	5	1.39	1.358	0.308	0.208	0.231	55.44	54	1.120
P value		0.08	0.007	0.007	0.0934	0.0934	0.056	0.005*	0.0934

n : average number of cultivars scored
[a]*He* : expected heterozygosity under Hardy-Weinberg genotypic proportions, also called Nei's gene
[b]*PLP*: proportion of polymorphic loci at the 5% level, expressed as a percentage diversity
[c]Pb: private bands, bands unique to a single population
[d]Br: band richness
*P<0.05, statistical significance was tested by ANOVA

Pairwise clonesets F_{ST} ranged from 0.0102 (Nfuuka vs Nakabululu) to 0.0609 (Nakabululu and Musakala). Lack of genetic differentiation of the clonesets corroborates the narrow genetic distances between them. The overall averages of Nei's (1983) genetic distance between the clonesets was 0.0181 ranging between 0.013 (minimum; Nfuuka and Nakabululu) and 0.027 (maximum:Musakala and Nakabululu) (Table 10). The out-group cultivars seem to have differentiated from the EAHB. Partitioning of hierarchical genetic variation using AMOVA revealed that high variation, 97%, resided within cultivars in the groups compared to among the groups, 3%. The covariance component at both levels was found to be significant (P<0.001). Furthermore, the overall genetic differentiation ϕ-$_{PT}$ value) among the populations was significantly low (0.030 P=0.0001) (Table 11).

3.3.5 EAHB cultivars from the same geographic region are closely related

The unrooted Neighbour-joining tree revealed that cultivars of different morphological groups were intermingled (Figure 16) within the region (Kenya or Uganda) suggesting presence of geographic pattern and lack of genetic basis of their morphology. However, a low level of genetic diversity was displayed among them. The cophenetic correlation coefficient between the tree and the original similarity matrix was significant (r = 0.980) markers supporting a good degree of confidence in the association obtained for the 90 EAHB cultivars and out-groups. These results corroborated those obtained using PCA analysis and Structure. The out-groups did not form an independent branch from the EAHB cultivars and clustered with the Kenya subpopulation unlike the SSRs.

Table 10: Pairwise F_{ST} and Nei`s genetic distance between the EAHB populations

Pairwise Fst						
		Mbidde	**Musakala**	**Nakabululu**	**Nakitembe**	**Nfuuka**
	Musakala	0.0462	****	****	****	****
	Nakabululu	0.0335	0.0609	****	****	****
	Nakitembe	0.0162	0.0250	0.0153	****	****
	Nfuuka	0.0193	0.0289	0.0102	0.0152	****
	Out-group	0.6413	0.6855	0.6702	0.6597	0.6746
Pairwise Nei`s GD						
		Mbidde	**Musakala**	**Nakabululu**	**Nakitembe**	
	Musakala	0.024	****	****	****	
	Nakabululu	0.020	0.027	****	****	
	Nakitembe	0.017	0.019	0.017	****	
	Nfuuka	0.016	0.018	0.013	0.016	

Table 11: Analysis of Molecular variance (AMOVA) indicates higher diversity within the morphological groups vs among the groups and EAHB population differentiation; PhiPT

Source	df	SS	MS	Est. Var.	%	PhiPT
Among groups	4	468.50	117.12	2.32	3%	0.030
Within groups	85	6436.83	75.73	75.73	97%	
Total	89	6905.32		78.05	100%	

PhiPT= AP / (WP + AP) = AP / TOT (AP = Est. Var. Among Pops, WP = Est. Var. Within Pops).

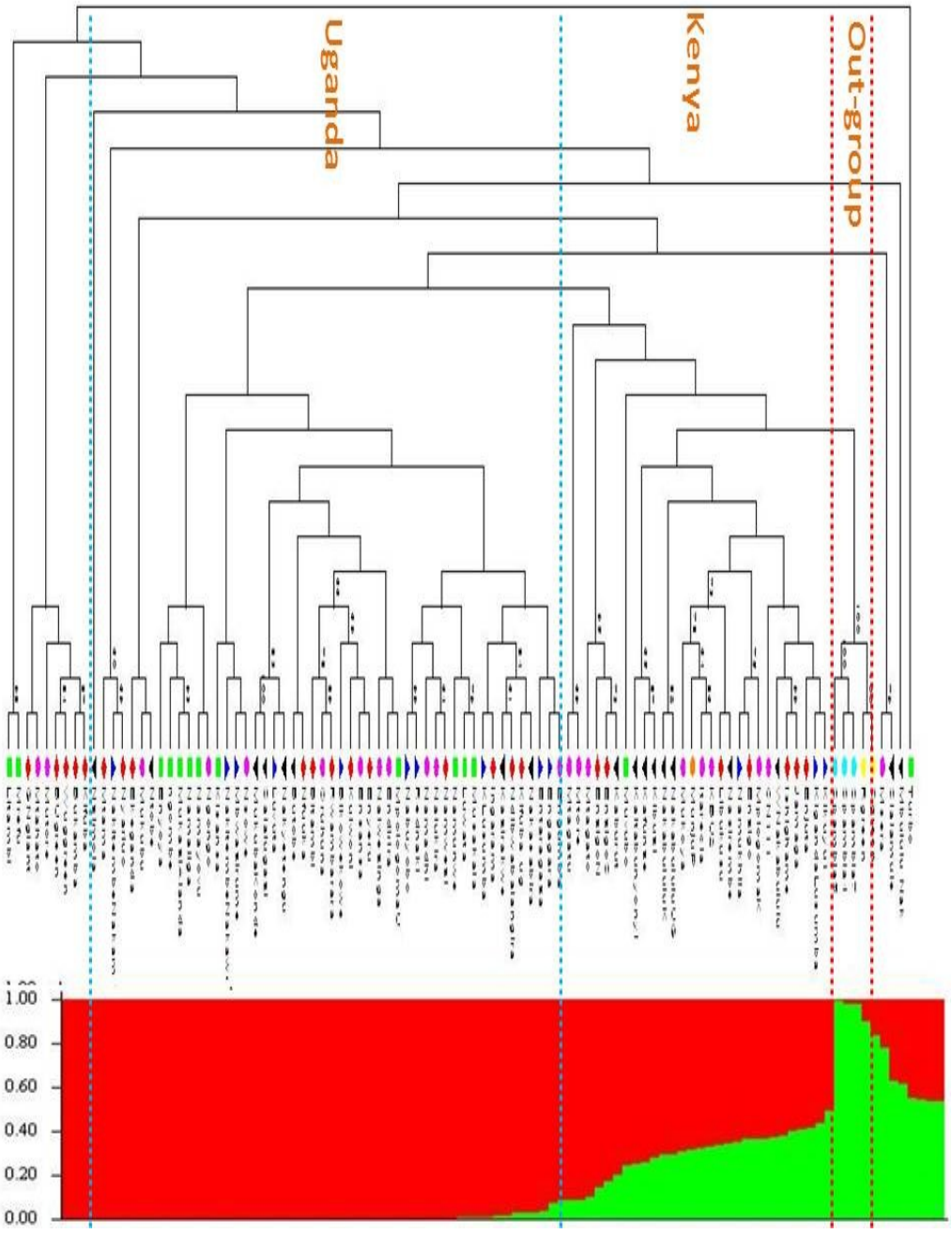


Figure 16: Neighbor-Joining tree drawn from the Nei and Li (1979) genetic distance of AFLP data and genetic population structure of the EAHB. The cultivars were clustered based on their country of collection colour codes of the tree represents the five morphological groups and population structure has two colour codes representing the most likely groups two regions (K=2) composed of cultivars from the two regions. The number on the nodes represents bootstrap values, only values >80 were shown.

3.3.6 No outlier loci detected in the EAHB population

For the Bayesian analysis, after 20 independent iterations, no outliers at $\log_{10}$ (PO) of >1 were identified (Figure 17A) for the EAHB population and formed a tight cluster in the posterior probability F_{ST} plot, but when the out-groups were included (Figure 17B) a log10 (PO) >2 was observed. The highest $\log_{10}$ (PO) was -0.9159 for the EAHB analysis and 2.6565 analysis inclusive of out-groups which, based on Jeffreys' (1961) scale, corresponds to "strong against" and "barely worth mentioning for" selection. The loci remained non-significant after applying a false discovery rate (FDR) P.0.05, and represented 6.9% of the total number of loci which were polymorphic between these populations. No significant differentiated cluster of loci was observed for the EAHB. In total, 51 rare alleles were identified at the cultivar level and 12 rare allele at the morphological groups' level. Analysis of bottleneck signatures in the EAHB subgroup showed highly significant heterozygosity excess both for stepwise mutation model (SMM) and the infinite allele model (IAM) (Table 12).

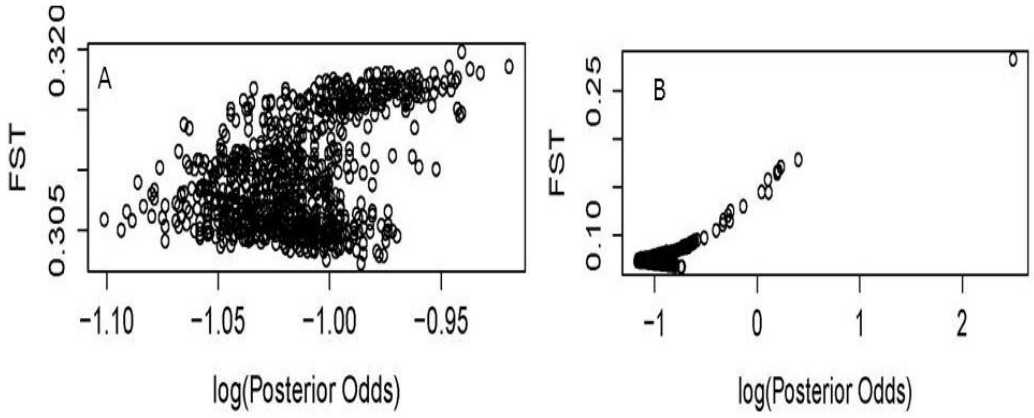


Figure 17: F_{ST} outlier locus identification. Locus-specific F_{ST} plotted against the posterior odds of the model including locus-specific selection effects versus the model excluding locus-specific selection effects, for the EAHB (A) and (B) including out-groups.

Table 12: Tests for evidence of genetic bottleneck. Highly significant heterozygote excess observed in a population that's has suffered a severe bottleneck event

Mutation Model	Expected Loci with He excess	Loci with He excess	Loci with He deficiency	Wilcoxon test P value
IAM	80.38	143	56	0.00000
SMM	92.13	143	56	0.00000

3.3.7 Comparison of SSR and AFLP markers for genetic diversity analyses of EAHBs

Some authors have compared the data produced by AFLP and SSR and showed that both markers have comparable efficiency in other crops. In this study, Kruskall-Wallis test indicated a significant difference (P=0.0001) in diversity indices obtained from SSR and AFLP. However, ranking for genetic diversity among the morphological groups were not significantly different (P=0.42) among populations, neither with AFLP nor with microsatellite data. Furthermore; genetic distances, population structure, differentiation and genetic relationships; results were similar for both markers. Mantel tests revealed that pairwise between-individual genetic distances calculated from AFLPs were significantly correlated with those calculated from microsatellite data (r^2 = 0.80573; P=0.00; Figure 18).

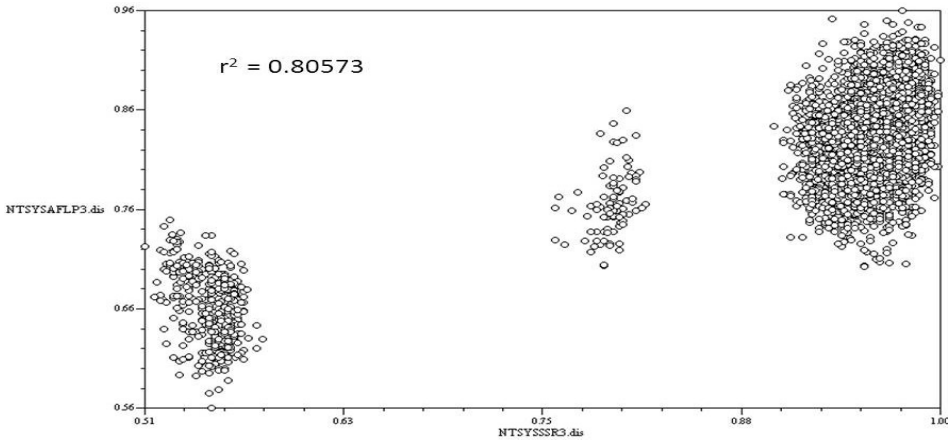


Figure 18: High significant correlation between SSR and AFLP dissimilarity matrices. Mantel correlation between SSR and AFLP dissimilarity matrices genetic distance matrices performed with one tailed probability at 1000 permutations.

3.4 DISCUSSION

This study has investigated the overall genetic diversity and population structure of the East African highland bananas. Such an analysis is necessary to explore possible mechanisms governing their phenotypic variations. Here the AFLP technique has been employed to assess the variability of EAHB obtained from two geographical origins. Our results demonstrate the utility of AFLP markers to assess genetic diversity among the cultivars of EAHB. The differences in the numbers of AFLP loci produced by each of the different primer pairs, likely reflect differences in sequence composition in the genome (Nath *et al.*, 2013). AFLPs are dominant markers (presence/absence of amplified fragments) with lower information content per locus than codominant markers (Zhang & Hare, 2012; Ley & Hardy, 2013). Using AFLP, Ude *et al.* (2002) reported mean PIC value of 0.24. This difference in PIC values obtained in other Musa studies and the present study could be linked with selection of different markers and a more diverse set of varieties, and also few numbers of loci with large discriminatory power. PIC and gene diversity is a reflection of allele diversity and frequency among the cultivars (Liu & Muse,

2005; Resmi *et al.*, 2011). The PIC and gene diversity values obtained in this study indicates lack of allelic diversity amongst the EAHB.

Overall AFLP genetic diversity indices for the EAHB-AAA subgroup was low, as also reported by as reported Tugume *et al.* (2003). However, diversity for this subgroup was extraordinarily lower compared to other *Musa* studies of cultivated; (El-Khishin *et al.*, 2009; Opara *et al.*, 2010; Shaibu, 2012) and wild bananas; (Ude *et al.*, 2002b; Wang *et al.*, 2007). Other studies have used cultivars of mixed genomic groups. Therefore such low genetic diversity observed in this study can be related to the occurrence of recombination between two closely related or genetically similar parents.

The genetic composition of a population is usually described in terms of number of allele, frequencies and heterozygosity. Polymorphism in a given population is often due to the existence of genetic variants represented by the number of alleles at a locus and their frequency of distribution in the population, the results obtained indicate lack of genetic composition in the EAHB subgroup. High genotypic diversity among parthenocarpic lineages is often interpreted as evidence for multiple lineage origin with each reflection of the genetic variation found in their sexually producing progenitors (Pongratz *et al.*, 1988). Multiple lineages of the edible bananas have been suggested by Kennedy (2008) and Perrier *et al.* (2011), however there is no evidence of multiple lineage of the Lujugira-mutika subgroup. Noyer *et al.* (2005) hypothesized a single origin of all accessions in this subgroup and postulated that they were derived from the same initial clone from whence the have evolved by somatic mutations fixed through vegetative propagation. However, large variations in morphological characteristics do not necessarily have to reflect the same degree of genetic variation (Lu *et al.*, 2011).

Other non-mutually exclusive factors may have also contributed to the low levels of genetic variation. For instance, one likely explanation is that these populations experienced a genetic bottleneck at some point in time.

Alternatively, the population could have been founded by only a few founder vegetative suckers (Kolb & Durka, 2013). Life history traits such as mating system together with ecological factors are expected to have a discernable effect on genetic diversity (Charlesworth & Wright, 2001). Therefore, the low genetic diversity seen in the EAHB subgroup may be due to their self-fertilization and parthenocarpic life. Lower genetic diversity has been observed in selfers compared to obligate out-crossers in 12 species of flowering plants (Leffler *et al.*, 2012) and in the genus Capsella (Foxe *et al.*, 2009). This is because, self–fertilization causes inbreeding which in turn reduces Ne under complete inbreeding to half its value thus affect neutral diversity (Nordborg, 2000). Furthermore, if selection that reduces variation at linked sites is widespread, the lower effective recombination in self-fertilizing species could also reduce neutral diversity by accentuating the effects of selection on linked sites (Charlesworth, 2009). For instance, low genetic diversity had been reported in flax collections, likely as a consequence of the mating system, limited gene flow, and breeding methods commonly applied in a rather narrow breeding gene pool (Soto-Cerda & Cloutier, 2013). There is a need to broaden the genetic diversity to conduct successful breeding in the EAHB subgroup.

Loss of diversity in the EAHB study group may have been due to human selection and domestication of a reduced number of elite clones, originally conducted to improve for pulp enhancement leading to parthenocarpic fruits and edibility (Perrier *et al.,* 2011). Selection of the EAHB banana clones began about 11,7000yrs ago from AAA progenitors after hybridization causing gametic sterility due to chromosomal rearrangements between parental subspecies (Sherperd 1999). It is possible that the large number of cultivated morphotypes arose from only a few imported introductions and suggests a long period of somaclonal mutations within these regions (Perrier *et al.,* 2011). Today EAHB banana clones continue to face selection pressures, most importantly, for traits such as taste and colour of Matooke as well as other agronomic traits; yield, disease and pest resistance and adapatability to the

changing climate. Loss of genetic diversity due to selection has been witnessed in other clonal plants; e.g. grapevines (Pelsy, 2010). Domestication has been reported to cause decrease of genetic diversity in soyabean, (Hyten *et al*. 2006), apricot (Bourguiba *et al*. 2012), wheat (Haudry *et al*. 2007), maize, barley, sunflower and sorghum (Rauf *et al*. 2010).

PCA and Structure results suggest that the EAHB Kenya subpopulation may have been established from the EAHB Ugandan subpopulation through transfer of planting material across the two countries, hence the admixture detected. In addition, structure reveals a more homogenous Uganda subgroup compared to the Kenyan subgroup. From this we can deduce two things; firstly, the observed admixture of the latter group could mostly be due to the presence of shared alleles with high frequencies, which were also present in the Uganda subpopulation. This suggests that these alleles either represent shared ancestral polymorphism through a common ancestral parent during hybridization or are homoplasious (alleles identical by state not by descent) (Duputie *et al.*, 2007; Perrier *et al.*, 2011). Secondly, if the two subpopulations are product of same parents during hybridization and were singly introduced into Africa (Kennedy, 2008), after successive generations, substantial differences accumulated, and the sets of genes in isolated populations began to diverge through the action of evolutionary factors expressed in each group resulting into some differentiation between the two geographical populations (de Faria-Tavares *et al.*, 2013).

The fixation index is a measure of how populations differ genetically and its value theoretically ranges from 0.0 (no differentiation) to 1.0 (complete differentiation, in which subpopulations are fixed for different alleles). As F_{ST} measures the amount of the excess of homozygotes a low F_{ST} value can be interpreted to mean that individuals from the population tend to share alleles. In this study, the genome-wide mean F_{ST} value across all AFLP loci was found to be 0.0271, interpreted as a low level of differentiation and lack of population

structure, or only marginally structured (F_{ST} values below 0.05 indicating low differentiation and values above 0.65 as indicating extreme differentiation).

The dendrogram reconstituted based on the genetic similarity coefficient summarizes the interrelationship among the EAHB cultivars. The majority of the cultivars irrespective of their morphological groups were clustered together based on geographical origin, meaning the genetic distance is not correlated with morphological groups. Confidence limits obtained through bootstrap analysis were high providing strong evidence for the reliability of the clustering of AFLP dataset. The genetic similarity estimates obtained through AFLP analysis displayed minimal differences between cultivars. Clustering of crop species is based normally on common origins of cultivars or shared mutations (Changadeya *et al.*, 2012; Hippolyte *et al.*, 2012). Among vegetative propagated crops like bananas, variations within each cluster is mainly dependent on genotype and genome differences arising from mutations whose frequency is dependent on how often a clone has been multiplied and planted (Changadeya *et al.*, 2012). The genetic relationships observed in this study do not resolve the morphological differences observed between cultivars or reveal any morphological clustering. Similar observations were reported by Baneh *et al.* (2009) in a genetic diversity study of grapevine clones.

In other studies, global genome comparisons (Nielsen *et al.*, 2009) have identified lower percentages of outlier loci, which could be construed as conservative, but also help to identify either neutral or outlier loci with applications to wider demographic scenarios rather than very specific environmental conditions. Two tests were applied based on different algorithms and assumptions to minimize the possibility of selecting false positives (Foll & Gaggiotti, 2008; Soto-Cerda & Cloutier, 2013). In his study, Pérez-Figueroa *et al.* (2010) compared three alternative F_{ST}-based outlier program to detect to loci under positive selection and observed that the most favorable situation for detecting loci under positive selection is that of a low estimated neutral FST distribution (<0.20) as selective loci would tend to show

high F_{ST} values. In this study, however, the neutral F_{ST} distribution was 0.203 implying that this factor could affect the efficiency of Bayescan in detecting positive selection. On the other hand, under balancing selection, a high neutral F_{ST} distribution would be more favorable for detecting selective loci. The estimated mean alpha coefficient which indicates the strength and direction of selection was 0.00014. A positive value of alpha suggests diversifying selection, whereas negative values suggest balancing or purifying selection (Foll 2012). Diversifying selection, also known as disruptive selection removes individuals from their center of phenotypic distribution and thus caused the distribution to become bimodal. It occurs when natural selection favors both extremes of continuous variation. Over time, the two extreme variations will become more common and the intermediate states will be less common or lost. Disruptive selection can lead to two new species.

BOTTLENECK analyses provided evidence of recent population bottleneck under the TPM and IAM models, and both tests showed evidence of loss of genetic diversity for the neutral loci (Table 7). Lack of rare alleles and decreased heterozygosity support the results of the bottleneck test. The occurrence of a population bottleneck causes a significant reduction in the effective population size and represents a major reason for the loss in allelic diversity, first by the loss of rare alleles, then by the successive loss of heterozygosity in the population (Porth & El-Kassaby, 2014).

The genetic diversity indices from SSR and AFLP significantly differed; therefore, direct comparisons were impossible. Gaudeul *et al.* (2004) has argued that due to different mutation levels of different markers (e.g SSR and AFLP mutation rates are 10^{-3}-10^{-4} and 10^{-6} respectively), comparing absolute diversity and differentiation values estimated with different types of markers, should not done but global qualitative patterns (ranking of populations for genetic diversity or differentiation, or the agreement (or not) of the data with a given biological model (e.g isolation by distance).

Congruent patterns of genetic similarities between SSR and AFLP were observed, contrary to results obtained from most comparisons in studies undertaken to identify and classify distinct cultivars/varieties of crop species involved in breeding programs. Poor correlation between estimates of genetic similarities derived from RAPDs, AFLPs and microsatellites in *Musa* were reported by Crouch *et al.* (1999). They attributed this to the different techniques that selectively screened complementary, and not overlapping, regions of the genome (Wang *et al.*, 2007). In contrast, Roa *et al.* (2000) reported significant Mantel tests between AFLP and microsatellite genetic similarities calculated across seven species of the Manihot genus. Congruent patterns of genetic distances in SSR and AFLP have been observed (Gaudeul *et al.*, 2004) have also been observed in other studies.

The Mantel test has been used in analysis of genetic diversity in crop plants, particularly in ascertaining the correspondence of matrices derived by means of different marker systems over the same set of genotype (Semagn *et al.*, 2012). The mantel test product-moment correlation value ($r = 0.91$, $p = 0.001$) showed a strong relationship between AFLP and SSR similarity matrices. A method yielding a high co-phenetic correlation coefficient can be considered as an appropriate method for a particular analysis (Mohammadi & Prasanna, 2003). The degree of fit can be interpreted subjectively as: $0.9 \leq r$, very good fit; $0.8 \leq r < 0.9$, good fit; $0.7 \leq r < 0.8$, poor fit; $r < 0.7$, very poor fit (Semagn *et al.*, 2012). The co-phenetic correlation values showed that the genetic clusters accurately represented the estimates of genetic similarity.

3.5 CONCLUSION

To more fully understand the genetic diversity of the East African Highland bananas, we need to reassess the diagnostic morphological characters. For

example, the Nakabululu and Nfuuka cultivars seem to be genetically closer than any other clonesets, which is also morphologically reflected. However, our results clearly demonstrate that cultivars of the EAHB have a narrow genetic base and are genetically uniform. The morphological differences (bunch shape/size, pulp astringency etc) that exist among them are not readily explainable on the basis of AFLP analysis that screens a reduced representation of the possible polymorphisms across the genome. The morphological differences between the EAHB bananas could be due to rare somatic polymorphisms arising from divergence between cultivars due to vegetative propagation (and possibly farmer selection) and/or heritable epigenetic polymorphisms arising from vegetative propagation and agri-environmental selection pressures. Next generation sequencing approaches that can generate more comprehensive marker densities are necessary to determine how rare any potential genetic polymorphisms are that could potentially differentiate EAHB cultivars.

The existence of a correlation between changes in the methylation state of particular gene sequences and the presence of a mutant phenotype has been shown clearly shown in other species (Jaenisch & Bird, 2003; Fujimoto *et al.*, 2012; Wang *et al.*, 2014). Therefore phenotypic variation could also be explained by differential expression of certain structural genes regulated by epigenetic changes, or by the occurrence of DNA/chromosomal mutations.

The sequencing of the Musa genome, recently completed (D'Hont *et al.*, 2012) opens the exciting possibility to apply improved DNA technologies among them, genotyping by sequencing and to also consider to analyze specific candidate genes related to different morphological and agronomical traits. The new next-generation sequencing based technologies will be necessary for clarifying the genetic bases of clonal differences, and in excluding any homonymous or wrong attributions based only on observations of morphological characters.

3.6 SUPPLEMENTARY MATERIALS

S Table 4: Sequences of EcoR1 and Mse adaptors, pre-selective primers (E+0, M+0) and 13 selective primer combinations used in this study (E+3, M+3, E stands for EcoR1 and M is Mse1 primers)

Name	Function	AFLP stage	Sequences
EcoR1 F	Adaptor	Digestion-Ligation	cgtagactgcgtacc
EcoR1 R	Adaptor	Digestion-Ligation	gactgcgtacatgcag
Mse1 F	Adaptor	Digestion-Ligation	gacgatgagtcctgag
Mse1 R	Adaptor	Digestion-Ligation	tactcaggactcatc
EcoR1+0	primer	Pre-selective PCR	gactgcgtaccaattca
Mse1+1	Primer	Pre-selective PCR	gatgagtcctgagtaac
E-AAG	Primer	Selective PCR	Fam-gactgcgtaccaattcaag
M-CTA	Primer	Selective PCR	gatgagtcctgagtaacta
E-AGA	Primer	Selective PCR	Ned-gactgcgtaccaattcaga
M-CCA	Primer	Selective PCR	gatgagtcctgagtaacca
E-AGA	Primer	Selective PCR	Ned-gactgcgtaccaattcaga
M-CTC	Primer	Selective PCR	gatgagtcctgagtaactc
E-AGA	Primer	Selective PCR	Ned-gactgcgtaccaattcaga
M-CTG	Primer	Selective PCR	gatgagtcctgagtaactg
E-AGC	Primer	Selective PCR	Ned-gactgcgtaccaattcagc
M-CTT	Primer	Selective PCR	gatgagtcctgagtaactt
E-AGG	Primer	Selective PCR	6Fam-gactgcgtaccaattcagg
M-CCT	Primer	Selective PCR	gatgagtcctgagtaacct
E-AGG	Primer	Selective PCR	6Fam-gactgcgtaccaattcagg
M-CTA	Primer	Selective PCR	gatgagtcctgagtaacta
E-AGT	Primer	Selective PCR	Fam-gactgcgtaccaattcagt
M-CGA	Primer	Selective PCR	gatgagtcctgagtaacga
E-AGT	Primer	Selective PCR	Fam-gactgcgtaccaattcagt
M-CTG	Primer	Selective PCR	gatgagtcctgagtaactg
E-AGT	Primer	Selective PCR	Fam-gactgcgtaccaattcagt
M-CTT	Primer	Selective PCR	gatgagtcctgagtaactt
E-ATC	Primer	Selective PCR	Ned-gactgcgtaccaattcatc
M-CTA	Primer	Selective PCR	gatgagtcctgagtaacta
E-ATC	Primer	Selective PCR	Ned-gactgcgtaccaattcatc
M-CTC	Primer	Selective PCR	gatgagtcctgagtaactc
E-ATG	Primer	Selective PCR	Ned-gactgcgtaccaattcatg
M-CGT	Primer	Selective PCR	gatgagtcctgagtaacgt

S Table 5: Optimal value of K, the highest Delta K value was K=2 obtained from AFLP data of the EAHB cultivars using admixture model with and without priori population assignment.

K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
1	3	-42015.17	37.391	—	—	—
2	3	-38162.70	30.767	3852.47	1501.27	48.794
3	3	-35811.50	302.787	2351.20	1362.07	4.498
4	3	-34822.37	236.790	989.13	161.27	0.681
5	3	-33994.50	287.533	827.87	229.03	0.797
6	3	-32937.60	25.569	1056.90	135.00	5.28
7	3	-32015.70	302.813	921.90	375.50	1.24
8	3	-31469.30	527.775	546.40	20914.13	1.46
9	3	-51837.03	22573.216	-20367.73	40863.00	1.81
10	3	-31341.77	1048.577	20495.27	—	—

CHAPTER 4

LOW DIVERSITY LEVELS AND SIGNATURES OF BALANCING SELECTION FOR SINGLE NUCLEOTIDE POLYMORPISMS (SNPs) IN EAST AFRICAN HIGHLAND BANANAS

Abstract

Background

Determining the level of genetic variation within and between species or populations is necessary to study the effects of mutation, natural selection and genetic drift. Genotyping by sequencing (GBS) has recently emerged as a promising next generation sequencing based approach for assessing genetic diversity on a genome-wide scale. We report an analysis of more than 14K SNPs genotyped using Illumina GBS (SNP) in 89 East African Highland banana (EAHB) cultivars from two regions in Africa (Uganda, Kenya).

Materials and results

To provide a more extensive and complete sampling of genetic variation, we included samples with unique phenotypes. Consistent with observations made by SSR and AFLPs, our results highlight significant shared variation amongst the EAHB cultivars (Nucleotide diversity (π); 0.03, and demonstrate that much of the genetic variation is within the population (AMOVA; within population variation 95% vs between regions 5%) and that population structure is geographically continuous. A genome-wide pattern of patchy heterozygosity strongly suggests that these cultivars originated from a cross between two genetically distinct Musa complex species. Indeed, principal components analyses reveal no discernible genetic differentiation between the EAHB from the two regions and cultivars of other genomic groups in our sample, and in most cases, individuals cannot be clearly assigned to defined morphological groups on the basis of SNP genotypes. All individuals are accurately classified into geographical groups using a model-based clustering algorithm with no conflicts. While we could not identify any loci under positive selection, 3937 and 2766 loci were detected to be under balancing and neutral selection, respectively. Extensive significant ($p<0.05$) linkage disequilibrium ($r^2=0$), 51.66%, and moderate linkage equilibrium ($r^2>0$), 39.22%, was observed in 3314 SNP pairs. Low (2.51%), but significant ($P<0.05$) recombination ($D`=0$) was detected in SNP pairs of the total pairs while the bulk of pairs showed no evidence of recombinations. All neutrality tests, Tajimas D`(4.9512 $P<0.001$), Fu and Li`s ($D^* = -4.5059$, $^{**}P < 0.02$), Fu's, *Fs* (*Fs* = -6.75, $^{**}P < 0.02$), raggedness statistics (r =0.0006, $^{*}p<0.004$) and Ramos-Onsins and Rozas's *R*2 statistics (R^2; 0.0478, $^{**}p<0.001$) were congruent with balancing selection and

population growth. Beast v1.8 results estimated the speciation time of the cultivars to 928 years in the past and time to the most recent common ancestor to 2590yrs before present.

Conclusion

We suggest that patterns of low genetic diversity could be interpreted as reflecting a recent reduction in diversity as a result of selection of farmer preferred banana clones, or alternatively, a historical lack of diversity caused by the hybridization of genetically similar sexual parents. This chapter provides evidence of balancing/purifying selection associated with EAHB domestication and crop improvement and extensive linkage disequilibrium caused by selective sweeps in this subgroup. Significant GWAS indicate the potential use of genome wide SNP markers in genomic selection, QTL mapping and breeding of important agronomic traits in EAHB population.

Key words: Single nucleotide polymorphism (SNP), Genotyping by sequencing (GBS), Nucleotide diversity, balancing and neutral selection, Linkage disequilibrium

4.1 INTRODUCTION

Genomic variation analysis is an essential component of plant genetics and crop improvement programs (Deschamps *et al.*, 2012) and can be associated with phenotype differences, be genetically linked to its causative factor, or indicate relationships between individuals in populations. Use of genotyping has enabled the characterization and mapping of genes in plants as well as the study of species diversity and evolution, marker-assisted selection (MAS), germplasm characterization and seed purity, over the last 30 years. Genetic markers are heritable polymorphisms that can be measured in one or more populations of individuals. Molecular markers lie at the heart of modern day genetics and enable the study of important questions in population genetics, ecological genetics and evolution (Davey *et al.*, 2011). The ideal molecular approach for population genomics should uncover hundreds of polymorphic markers that cover the entire genome in a single, simple and reliable experiment (Luikart *et al.*, 2003). Now, with the advent of next-generation sequencing (NGS) technologies, there are several such approaches, which are capable of discovering, sequencing and genotyping not only hundreds but thousands of markers across almost any genome of interest in a single step, even in populations in which little or no genetic information is available. These technological advances have facilitated the characterization of genes and genomes and started to provide a more comprehensive view of diversity and gene function in plants (Deschamps *et al.*, 2012).

One of the recently emerged techniques is, genotyping-by-sequencing (GBS), where the detection of sequence differences (namely SNPs) in a large segregating or mutant population is combined with scoring, thus allowing a rapid and direct study of its diversity targeted towards the mapping of a trait or a mutation of interest (Deschamps *et al.*, 2012). The GBS approach is based on genome reduction with restriction enzymes (Altshuler *et al.*, 2000; Elshire *et al.*, 2011), does not requires a reference genome for single nucleotide polymorphism (SNP) discovery, is a one combined step process of marker

discovery and genotyping and provides a rapid, high-throughput, and cost-effective tool for a genome-wide analysis of genetic diversity for a range of non-model species and germplasm sets (Poland & Rife, 2012; Fu, 2014). Genome complexity reduction combined with multiplex sequencing was first demonstrated through restriction site associated DNA (RAD seq) tagging (Baird *et al.*, 2008); and NGS of the RAD tags to genetically map mutations (Miller *et al.*, 2007). Genotyping-by-sequencing (GBS) was developed as a simple but robust approach for complexity reduction in large complex genomes (Elshire *et al.*, 2011). Both RAD sequencing and GBS target the genomic sequence flanking restriction enzyme sites to produce a reduced representation of the genome. However, GBS library development is greatly simplified compared to that of RAD and requires less DNA, avoids random shearing and size selection and is completed in only two steps on plates followed by PCR amplification of the pooled library (Elshire *et al.*, 2011). Nevertheless, both techniques now have a wide application in plant breeding.

Because of the advances in next-generation sequencing technologies (Metzker, 2010) Genotyping by sequencing (GBS) has recently emerged as a promising genomic approach for exploring genetic diversity and association mapping on a genome-wide scale (Poland & Rife, 2012; Fu, 2014). SNP-based marker technologies has increased marker density (abundance in the genome) and reduced genotyping costs and time by orders of magnitude in relation to earlier approaches, and are, therefore currently, the most widely used genotyping markers (Elshire *et al.*, 2011; Deschamps *et al.*, 2012). SNP discovery and use to explore genetic diversity in non-model species has greatly increased (Baird *et al.*, 2008; Fu & Peterson, 2011; Poland *et al.*, 2012; Poland & Rife, 2012; Fu *et al.*, 2013; Lu *et al.*, 2013; Sonah *et al.*, 2013; Fu, 2014; Schilling *et al.*, 2014) and in approximately, 7.4 million of ex situ plant germplasm samples conserved in world genebanks (FAO, 2010; Fu & Peterson, 2011). More importantly, the increase in information about potentially millions of genome-wide SNPs or small insertion-deletions and their surrounding sequence context

has set the foundation of high-throughput genotyping (Deschamps *et al.*, 2012).

The East African highland banana (triploid *Musa acuminata*) is an annual crop that produces starchy fruits and is widely cultivated in tropical countries, especially in East Africa (Karamura, 1998). Like many edible plants, the EAHB is clonally propagated by farmers using suckers produced by the mother plants after bunch production. This farming system is advantageous since plants with a high fitness rating can be maintained identically over the years (Scarcelli *et al.*, 2013). The domestication of banana dates back 10 000 years to South East Asia, most probably New Papua Guinea (Perrier *et al.*, 2011), but the scarcity of archeological remains and of in-depth genetic data does not allow the precise date or movement and introduction into Africa. Analyses of the diversity of the EAHB revealed marked variability at morphological level (Karamura, 1998) diversity at molecular level remains questionable. Clearly, farmers' management of banana cultivars strongly selects against off-types when they choose suckers to be used for the next generation (Karamura *et al.*, 2012). Farmers may end up selecting suckers of the same mother plant and discard the rest, potentially lowering any extant genetic diversity. It is therefore possible to consider that the EAHB subgroup is a single genotype that has evolved by accumulating somatic mutations (Pillay *et al.*, 2001; Perrier *et al.*, 2011). Hence, knowing the mutation rate and the demographic evolution of the varieties and estimating the possible ages of the clones will be important for rational breeding of EAHB.

Somatic mutations in bananas (*Musa* spp.) have been exploited for selection of favorable traits, both for consumption and commercial purposes (Karamura *et al.*, 2010). However, the implications of somatic mutations are not usually obvious, depending on which traits have been affected. In their work Karamura *et al.* (2010) showed how 13 traits that have been affected by mutations are selected by farmers and subsequently conserved on-farm in EAHB banana

based farming systems, but the genetic basis of these traits is still not yet known.

NGS approaches have facilitated genome-wide scans of positive selection and outlier loci in moderate and low coverage sequenced samples in model and non-model organisms such as Arabidopsis (Stapley *et al.*, 2010; Deschamps *et al.*, 2012). Genome-wide association studies (GWAS) have the potential to pinpoint genetic polymorphisms underlying human diseases and agriculturally important traits (Zhang *et al.*, 2010) and has been found to be more successful in crop plants than in humans (Brachi *et al.*, 2011). GWAS studies have been applied in several plant species; e.g lettuce, maize, rice, sorghum and peaches among others (Huang *et al.*, 2010; Dhanapal & Crisosto, 2013; Kwon *et al.*, 2013; Kannan *et al.*, 2014) using the general linear model (GLM) and mixed linear model (MLM) approaches (Zhang *et al.*, 2010). Genome-wide association studies (GWAS) utilize the natural diversity present in a multi-generational population and require hundreds of thousands to millions of markers to generate sufficient information and coverage (Edwards & Batley, 2010). Genome-wide association analysis (GWAS) is a powerful approach to identify the causal genetic polymorphisms underlying complex traits (Zhao, K *et al.*, 2011). It has also been applied in Bulked Segregant Analysis (BSA), Genomic Selection (GS), genotype correlation to appropriate phenotypic values, map various traits of interest in specific environments (Gore *et al.*, 2009; Deschamps *et al.*, 2012), construction of linkage maps and detection of QTLs associated with disease resistance (Pfender *et al.*, 2011).

The recent genomic sequencing and release of the Musa genome has been a significant advance. The double haploid banana-Pahang CIRAD 930 ITC 1511 (DH–Pahang), genome size is 523Mb which in a 91% assemblage revealed 36,542 protein-coding genes anchored to the 11 *Musa* chromosomes (D'Hont *et al.*, 2012). This provides a unique genomic platform for genetic improvement of the under researched EAHB crop. DH Pahang was derived

from the Pahang wild diploid ($2n$ = 22) *Musa acuminate* Colla. subspecies *Malaccensis* accession which shares its genetic lineage with dessert and cooking bananas (Maldonado-Borges *et al.*, 2013).

Genome-wide patterns of diversity and selection are critical measures for understanding how evolution has shaped the genome. Yet, these population genomic estimates are available for only a limited number of model and non-model organisms. Here, we focus on the population genomics of the EAHB, we: (i) assess if the recent population of EAHB is characterized by impoverished genetic diversity caused by founder events; (ii) evaluate if the EAHB population structure concur with either the 'propagule pool' or the 'migrant pool' model of Slatkin; (iii) determine the amount of background linkage disequilibrium in the entire EAHB population; (iii) use SNP data to detect recent signatures of selection, demographic history and "footprints" of natural selection; (iv) estimate mutation substitution rates and probable age of the current clones of the EAHB subgroup and (v) evaluate genome-wide trait-marker association of SNPs with thirteen phenotypic traits (known to be vulnerable to somatic mutations) in this subgroup.

4.2 MATERIALS AND METHODS

4.2.1 DNA preparation, quantification and Quality

High quality DNA (S Figure 2) was extracted from 95 individuals (91 EAHB and 4 out-group cultivars) using a combination of Mace *et al*. (2004) and Dellaporta *et al.* (1983) CTAB protocols with minor modifications (Appendix 2). Extraction of high quality genomic DNA samples of the 95 cultivars was essential to avoid differentially complete digestion across the samples or differing amounts of sample DNA that may cause varying ratios of adapters to sticky ends and therefore affect variation in the number of reads from any given sample (www.igd.cornell.edu/). Quantification of DNA employed the

pico green method for genomic DNA quantification (www.promega.com/). Spectroscopy and restriction digests with ApeK1 enzyme was used to assess DNA quality by assessing protein (OD 260/280 ratio >1.7) and polysaccharide contamination (OD 260/230 ratio (>1.7) (Sedlackova *et al.*, 2013).

4.2.2 Preparation of sequencing libraries and complexity reduction

Library preparation followed the protocol of Elshire *et al.* (2011) and was optimized for *Musa* species such that no adapter dimers were created. Optimization of sequencing libraries was done with EcoT221, ApeK1 and Pst1 restriction enzymes (S Figure 3). Using barcodes unique for each of the 95 genotypes and one blank, the sequencing libraries were prepared in a 96-plex each a with unique barcode (for adapters sequence information and library preparation PCR protocols see; www.igd.cornell.edu/.../**buckler**_lab_genotyping_by_sequencing)

4.2.3 Genome data and alignment of sequences

Illumina HiSeq 2000 System was used to sequence the samples. Since Musa species have a reference genome sequence (D'Hont *et al.*, 2012), the GBS pipeline in Trait Analysis by aSSociation, Evolution and Linkage (TASSEL Version 3.0.160) was used to analyze DNA sequences. GBS pipeline is a Java program implemented in Tassel (Bradbury *et al.*, 2007) and specifically tailored to the GBS protocols of Elshire *et al.* (2011) and Poland *et al.* (2012). Multiple sequenced GBS libraries and identical samples were merged prior SNP calling. To obtain a list of variants, or differences between EAHB, the individuals genome sequence reads were aligned to the reference genome sequence. To achieve both speed and memory efficiency, indexes of the reference genome (Langmead, 2010) were built with Burrows-Wheeler

Alignment (BWA) index build tool (bwa index my.fasta) and reads were aligned with the aid of the indexes using BWAVersion 0.7.5a-r405, MEM algorithm. This method is faster, more accurate and produces high quality querries, compared to other BWA methods, therefore achieving better performance for 70-100bp Illumina reads (Li & Durbin, 2009). Unique keyfile(s) were used to associate barcodes with sample IDs while running the GBS pipeline and SNPs were called using parameters in S Table 6. Genotypes were filtered to those with genotype quality 98 or higher (high confidence SNP calls) with VCFtools v0.1.10. Failed samples (non-blank) were defined as those with less than 10% of the mean reads per sample coming from the lane on which they were sequenced. The remaining individuals/sites were filtered on missingness and allele frequency (with the same parameters as those used for the GBSHapMapFiltersPlugin (S Table 6) generating VCF file with a total 45895 SNP loci.

4.2.4 Data analysis

4.2.4.1 SNP variation

Estimation of genome wide alleles (SNPs and Indels) in the EAHB was done using the VCF file. However for diversity and other subsequent analysis the VCF snps were merged using Tassel Version: 4.3.0. VCFtools version [v0.1.10] was used to calculate Depth and Missingness from the unfiltered file all.mergedSNPs.vcf.gz. Further filtration and removal of sites and taxa with more than 10% missing data was done generating a final filtered data file with 14121 SNP loci and 92 genotypes (89 EAHB, 3 out-group cultivars).

4.2.4.2 Population Polymorphism and diversity of the EAHB population

Nucleotide diversity (π) was calculated as the average number of nucleotide differences per site between two sequences and haplotype diversity (*Hd*) as the

probability that two randomly chosen haplotypes from a given population were different. Alignment gaps may lead to underestimated diversity values hence, to avoid potential bias, Insertion or deletions (indels) were excluded from all estimates (Li *et al.*, 2011). Average nucleotide diversity (π) and θ (4Neμ) over all chromosomes was calculated using concatenated sequences in software TASSEL v5.0 and haplotype diversity was calculated in GenAlex v6.5 (Peakall & Smouse, 2012). Diversity within Synonymous and non-synonymous substitutions regions were calculated in DNAsp v5 using the method of Nei and Gojobori (1986) and a Jukes–Cantor correction was applied to correct for multiple hits.

Population diversity was estimated by calculating heterozygosity (proportion of heterozygous individuals in the population); unbiased gene diversity (probability that two randomly chosen alleles from the population are different); Polymorphism information content (PIC) (Botstein *et al.*, 1980) were estimated. Within-population inbreeding coefficient (f) was estimated using the method-of-moments FST estimator proposed by Reynolds *et al.* (1983) in Powermarker v3.25. The overall estimates were calculated as the average across all loci, whereas variances and confidence intervals are estimated by nonparametric bootstrapping across different loci.

To examine whether the pattern of polymorphism observed over the whole region was in agreement with the neutral mutation hypothesis (Kimura, 1983). Tajima's D` test (Tajima, 1989) was applied. Kimura`s theory states that, the vast majority of molecular differences that arise through spontaneous mutation does not influence the fitness of the individual. Tajima`s statistic compares the difference between two estimates of the amount of nucleotide variation in the number of segregating sites (Watterson, 1975) and the average number of pairwise differences (Nei & Li, 1979; Tajima, 1989). The extent of DNA divergence between the EAHB and out-group cultivars was computed using; nucleotide diversity of each population, average number of nucleotide

substitutions per site between populations, Dxy and the number of net nucleotide substitutions per site between populations, Da (Nei, 1987).

4.2.4.3 Population structure, ancestry and relationships of the EAHB cultivars

Genetic differentiation between individual pairs was calculated with FST (Fumagalli *et al.*, 2013) and between groups calculated in STRUCTURE. For F_{ST}, we employed Hudson *et al.* (1992) $FST = 1 - Hw/Hb$ (Hw is mean number of differences between sequences from the same subpopulation, and Hb is mean number of differences between sequences from the different populations) using AFLPSURV v1.1. The hierarchical analysis of population differentiation was conducted using AMOVA implemented in GenALEX v6.5 (Peakall & Smouse, 2012). Population differentiation coefficient (PhiPT) was calculated $PhiPT = AP / (WP + AP) = AP / TOT$ and Nm (Haploid) calculated as $Nm = [(1 / PhiPT) - 1] / 2$.

To determine genetic relationships between the EAHB cultivars, we performed PCA, hierarchical clustering and Structure analysis. Principal component analysis (PCA) was performed on genotype data (converted into numerical data set) without missing values using the correlation method eigen value ≥ 0 (the minimum eigen value associated with each axis) in TASSEL v5.0. Hierarchical clustering was done in R using the FactoMiner package (Husson *et al.*, 2014). The model-based (Bayesian) cluster software STRUCTURE v2.3.4 (Pritchard *et al.*, 2000) was chosen to estimate the population structure of the EAHB cultivars and assign accessions to groups or subgroups with the SNP molecular markers distributed across all *Musa* chromosomes. For structure analysis, each individual was codedusing a two-row format: (xji, 1, xji, 2), which represents the genotype of individual i at locus j as described by Pritchard et al. (2000). We ran STRUCTURE under the 'admixture model' with a burn-in period of 100 000 followed by 100 000 replications of Markov

Chain Monte Carlo. Three independent runs each were performed with the number of clusters (K) varying from 1 to 10. An ad hoc measure delta K based on the relative rate of change in the likelihood of the data between successive K values was used to determine the optimal number of clusters using the Structure Harvester (Earl & vonHoldt, 2011) implementing the Evanno *et al.* (2005) method. Inferred ancestry estimates of individuals (Q-matrix) were derived for the cultivars (Pritchard *et al.*, 2000) and those with less than 0.60 membership probabilities were retained in the admixed group. Classification of the accessions was based on the STRUCTURE results with no priori population information, due to lack of cultivar pedigree information.

To compare degree of relatedness and coancestry of the EAHB cultivars, we calculated Kinship (the identity by descent; IBD/state; IBS). Kinship was computed using scaled IBD method implemented in TASSEL v5.0. This method was preferred over the pairwise IBS method because the latter method may result in an inflated estimate of genetic variance.

4.2.4.4 Screening for adaptation signatures

To detect outlier loci that has been or are still being under selection for local adaptation of the EAHB population, Bayescan method of Foll and Gaggiotti (2008) (http://www-leca.ujf-grenoble.fr/logiciels.htm) was applied. Bayescan software v2.1 directly estimates the posterior probability of a given locus to be under selection assuming that allele frequencies within the population follow a Dirichlet distribution. The analysis is based on a logistic regression to decompose F_{ST} into a β component (shared by all loci) and a locus specific α component (shared by all the populations) (Soto-Cerda & Cloutier, 2013). Departure from neutrality at a given locus is assumed when the locus-specific component is necessary to explain the observed pattern of diversity, either as an indication of positive (diversifying) selection, if $\alpha>0$ or balancing (purifying selection, if $\alpha<0$). The probability of being under selection is then inferred

using the Bayes factor (BF) for each locus (Jeffreys 1961). For our analysis, the estimation of model parameters was set as 20 pilot runs of 5000 iterations and a burn-in of 50000 MCMC was employed. The sample size was set to 5000 and the thinning interval to 10, resulting in a total chain length of 150 000 iterations (Perez-Figueroa *et al.*, 2010). The loci were ranked according to their estimated posterior probability and all loci with a value over 0.993 were retained as outliers. This corresponds to $\log_{10}$ PO.2.0, which provides decisive support for acceptation of the model. In our genome scan, the $\log_{10}$ PO 2.0 was considered a threshold value for determining loci under selection according to Jeffreys' interpretation (Foll, 2012), which is a logarithmic scale for model choice as follows: Log_{10} PO.0.5 (substantial); $\log_{10}$ PO.1.0 (strong); $\log_{10}$ PO.1.5 (very strong); and $\log_{10}$ PO.2.0 (decisive support for accepting a model) (Jeffry 1961).

4.2.4.5 Detection of the evolutionary forces determining pattern of genetic variation

To characterize the coding-sequence divergence of the closely related EAHB genomes, we compared DNA sequence divergence between sequences from all EAHB cultivars. *K*a (the number of non-synonymous differences divided by the number of non-synonymous sites) and *K*s (the number of synonymous differences divided by the number of synonymous sites) (Rozas & Rozas, 1999). We identified the relative strengths of the evolutionary forces acting on the population by estimating the ratio of substitutions at replacements sites (*K*a) to substitutions at synonymous sites (*K*s) (Hughes, 1999; Yang & Bielawski, 2000)

We used several parameters to test whether the studied populations have experienced size changes in their history. Fu and Li's D* (Fu' & Li, 1993) neutrality test, also equivalent to Tajima's *D* (Tajima, 1989; Innan & Stephan, 2000) and R^2 statistics (Ramos-Onsins & Rozas, 2002) are based on mutation frequency distribution, the *Fs* statistic (Fu, 1997) is computed from the

haplotype distribution, and the raggedness statistic r (Harpending, 1994) is derived from the pairwise differences between sequences (i.e., the mismatch distribution). These tests were shown to be the most robust for detecting population size (Ng & Stephen, 2013). In our simulations, neutral genealogies without recombination were used. A graphic representation of the observed and expected values for expanding and stationary populations was plotted (Tajima, 1989; Slatkin & Hudson, 1991; Harpending, 1994).

Separately, recombination rate (R) was computed as $R = 4Nr$ (N = population size and r = recombination rate per sequence or between adjacent sites (Hudson, 1987). To estimate the minimum number of recombination (RM), events in the history of the sample we used the algorithm (the four-gametic test) described in Hudson and Kaplan (1985). Using the complete sequence alignments between the all EAHB cultivars and zebrina (wild diploid, thought to be a progenitor of the EAHB cultivars), we classified fixed differences as well as polymorphic sites within and between EAHB cultivars and Zebrina synonymous or non-synonymous. We did the Hudson-Kreitman-Aguadé (HKA) test (Hudson *et al.*, 1997) to evaluate the effect of selection by comparing the levels of polymorphism and divergence of the EAHB from the out-group and estimated the divergence time, of the two categories. The test is based on the neutral theory of molecular evolution (Kimura, 1983) which predicts that for a particular region of the genome, its rate of evolution is correlated with the levels of polymorphism within species. All parameters and tests were calculated in DnaSp v5.0.

4.2.4.6 Estimation of ancestral population sizes and speciation times

Estimating branch lengths in proportion to time is confounded by the fact that the rate of evolution and time are intrinsically linked when inferring genetic differences between species. We used the uncorrelated lognormal relaxed

clock model (Drummond, 2005; Lepage *et al.*, 2007) with the prior probability for the substitution rate uniformly distributed. This model assumes that the rate associated with each branch is independently drawn from a single underlying parametric distribution. A Yule tree prior was used for estimation of the divergence time at the species level. A Yule tree prior, assumes a constant lineage birth rate for each branch in the tree and the most suitable for species-level phylogenies (Heled & Drummond, 2012) was used, whereas priors at population level analyses were based on the coalescent model. Population genetic analysis involved estimation of demographic parameters population sizes, growth/decline and migration using Bayesian skyline plots (Heled & Drummond, 2010). We used 50,000 iterations as the burn-in and then take100 million samples, sampling every 10,000 iterations. Analysis with Tracer v1.7.5 (Drummond & Rambaut, 2007) confirmed convergence of analyses and adequate sample sizes, with ESS values above 200. LogCombiner v.1.6.2 was used to combine trees in a single file after discarding the first 10% generations of each run as burn-in. TreeAnnotator v1.7 (Drummond & Rambaut, 2007) was used to select the maximum clade credibility (MCC) tree that has the maximum sum of posterior probabilities on its internal nodes and summarizes the node height statistics in the posterior sample. Clade support was represented by posterior probability (PP) values, with PP values of more than 0.95 indicating strong support. The Bayesian MCC tree was used to define calibration points and other key nodes in the dating analysis.

4.2.4.7 Intralocus and interlocus Linkage Disequilibrium

To determine whether recombination or homoplasy has occurred between alleles and assess the correlation between alleles at two loci, LD was estimated as D' and r^2. Where, D' is the standardized disequilibrium coefficient and determines whether recombination or homoplasy has occurred between a pair of alleles while r^2 represents the correlation between alleles at two loci. D' and

r^2 can only be calculated when two alleles are present, therefore, only biallelic SNPs with at least 10% frequency and polymorphic loci were considered. LD (D' and r^2) D was estimated for each polymorphic SNP pair using TASSEL v 5.0.9 and represented in scatter plots of r^2 and D' values versus genetic/physical distances between all pairs of alleles along the 12 chromosomes. The frequency of LD with distance in base pairs (bp) was also evaluated and plotted. P-values were determined by a two sided Fisher`s exact test because only two alleles are present at both loci. Haplotype blocks of all 12 chromosomes and for the entire genome was plotted in a heatmap using TASSEL v5.0.9.

4.2.4.8 Trait-marker association analysis

A phenotypic data set of 13 traits (S Table 7) found to be associated with mutations in East African highland banana (Karamura *et al.*, 2010) and previously identified as among the characters that differentiated the EAHB morphological clonesets (Karamura, 1998; Daniells *et al.*, 2001) were collected from 90 EAHB cultivars in Mbarara germplasm collection. For the 13 characters 5 plants belonging to each cultivar were scored. If there was variation within an accession for qualitative characters, the overall score considered was for the majority of the five plants under study i.e three or four out of five plants scored. The non-ordered characters; bunch shape, fruit shape and male bud shape; were coded as series of discrete states because ratios were not giving a true picture of their shapes. Characters related to colour were mainly examined indoors. The standard Royal Horticultural Society Colour Charts Edition Version 2 (measured with spectrophotometer) was used in colour scoring. Pieces of leaves, fruit skins and pulp were cut and examined under the hole in the colour patch of the RHS chart so that natural colours could easily be matched with the colour of the chart. The colour numbers of the chart displayed the standard expression of the described colour state of that character. Quantitative characters were measured to the nearest centimetre or

nearest millimetre using calibrated tapes and were entered directly as raw data. Mean values for continuous quantitative characters for the five randomly selected healthy plants per accession were calculated and most quantitative data were converted to ratios to reduce environmental effects.

Marker–trait association was done to elucidate the genetic basis of 13 phenotypic traits (S Table 7). GWAS analysis was done using two statistical models: (i) general linear model (GLM) with K - matrix (excluding Q - matrix) and (ii) mixed linear model (MLM) model with Q -matrix and K -matrix (MLM $Q + K$) used to correct for population structure following the compressed approach using EMMA (Kang *et al.*, 2010) and P3D method (Zhang *et al.*, 2010) algorithms implemented in TASSEL (Bradbury *et al.*, 2007). Genome-wide association analyses based on these models were conducted with the software TASSEL v5.0.9. Markers were defined as being significantly associated with traits on the basis of their significant association threshold at $P \leq 0.01$ ($-\text{Log}_{10}P \geq 10.00$, $P \leq 7.0\text{x}10^{-7}$ after Bonferroni multiple test correction (0.01/14121). Permutation tests were done at 1000 number of permutations after removal of monomorphic sites (Anderson and Ter-Braak (2003).

4.3 RESULTS

4.3.1 Genome alignment results

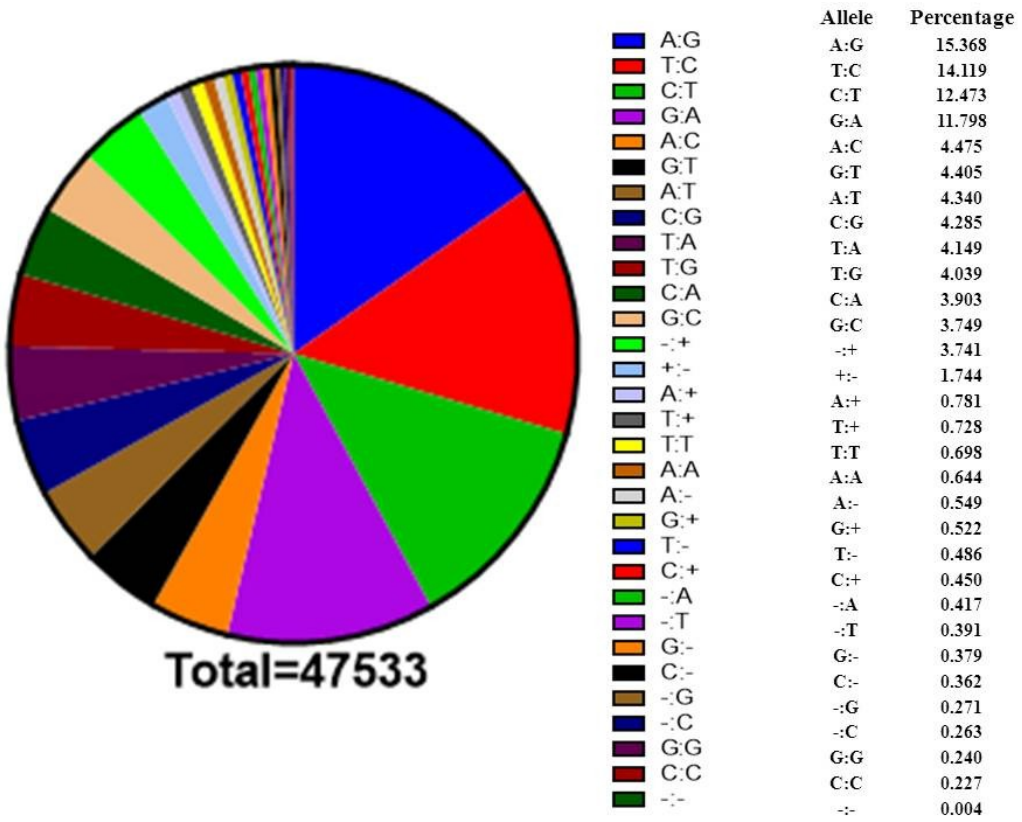
The total number of reads generated by GBS using Pst1 restriction enzyme were 1021793780 and 69.21% (707239032) were of high quality. Our genome-wide assessment of SNP variation, revealed relatively low levels of sequences. We produced a total of 858293 tags from the high QC sequences; 588479 (68.6%), 91291 (10.6%) aligned to unique positions and multiple positions in respective on the reference genome, and 178523 (20.8%) could not be aligned. Samples that failed (with more than 10% missing data) were 4.17% of total

reads in three individuals (Nakitembe_Red, Namadhi and Ornata). The mean sequencing individual and site depth was 60.06 and 53.75 (sdv. 17.89, 12.36) respectively. Individual and site missingness was 0.184 and 0.184 (sdv 0.152 and 0.267) respectively. A total of 45958 SNPs and indels were discovered in the aligned bases, representing, 0.118 SNP per kilobase (Table 13; Figure 19). However, it is likely that the number of SNPs per base was overestimated (at a genome-wide level) and true nucleotide diversity across the genome is much lower. Nonetheless, these data constitute substantially more genome coverage than achieved with previous analyses based on AFLPs and SSRs.

Table 13: Mapping of Single Nucleotide polymorphisms (SNPs) on the EAHB-AAA chromosomes

Chr	ChromLength	BasesCovered	% covered	Mean	SD	Candidate SNPs*	SNPs per kb
1	27,573,629	2,151,983	7.80%	126.1	814	8920	0.32
2	22,054,697	2,283,381	10.40%	110	742.7	2316	0.11
3	30,470,407	1,979,904	6.50%	122.3	807	3500	0.11
4	30,051,516	1,701,158	5.70%	121.4	803.7	3120	0.10
5	29,377,369	2,444,918	8.30%	115.1	774	2887	0.10
6	34,899,179	2,485,099	7.10%	116.1	789.6	3986	0.11
7	28,617,404	2,322,484	8.10%	103.3	729.7	3018	0.11
8	35,439,739	2,790,660	7.90%	116.4	781.9	3458	0.10
9	34,148,863	2,166,515	6.30%	101.7	712.9	3347	0.10
10	33,665,772	2,604,331	7.70%	108.9	760.1	3251	0.10
11	25,514,024	2,351,510	9.20%	108.7	757.5	2685	0.11
Un	141,154,048	4,952,271	3.50%	112.4	742.6	5407	0.04
Total	472,966,647	30,234,214	88.50%	1362.4	9215.7	45895	1.41

*False positives are likely



Allele	Percentage
A:G	15.368
T:C	14.119
C:T	12.473
G:A	11.798
A:C	4.475
G:T	4.405
A:T	4.340
C:G	4.285
T:A	4.149
T:G	4.039
C:A	3.903
G:C	3.749
-:+	3.741
+:-	1.744
A:+	0.781
T:+	0.728
T:T	0.698
A:A	0.644
A:-	0.549
G:+	0.522
T:-	0.486
C:+	0.450
-:A	0.417
-:T	0.391
G:-	0.379
C:-	0.362
-:G	0.271
-:C	0.263
G:G	0.240
C:C	0.227
-:-	0.004

Figure 19: Representation of SNPs and indels (insertion and deletions) scored in 89 EAHB and 3 out-group cultivars. Proportions of alleles in the genome are represented in percentage.

Of the total number of sites scored (14836) in 92 taxa (89 EAHB and 3 out-group cultivars) the number of invariable (monomorphic) and variable (polymorphic) sites were 5984 and 8852 respectively. Singleton variable sites, 2617, were much lower than Parsimony informative sites, 6235. Average observed heterozygosity, PIC and gene diversity across all SNPs was 0.4103, 0.1987 and 0.2549, respectively. The average frequency of the minor allele was 0.4388 (S Figure 4) and the inbreeding coefficient was -0.6062. Pairwise sequence diversity averages were; π = 0.2506 (range from 0.1860 to 0.2642); theta value (θ) ($4N_e\mu$), 0.1978 (range from 0.1953 to 0.2001), Tajima's D, 4.9512 ($0<t = -6.43$, d.f. = 14120, $P<0.001$) (Figure 20 A&B; Figure 21) and Haplotype diversity (Hd) was 0.147 (SE, 0.002). However lower values were

observed when the out-groups were excluded; average π (per bp), θ (per bp) and Tajimas D value was 0.2495, 0.1875 and 1.1407 respectively (S Table 8).

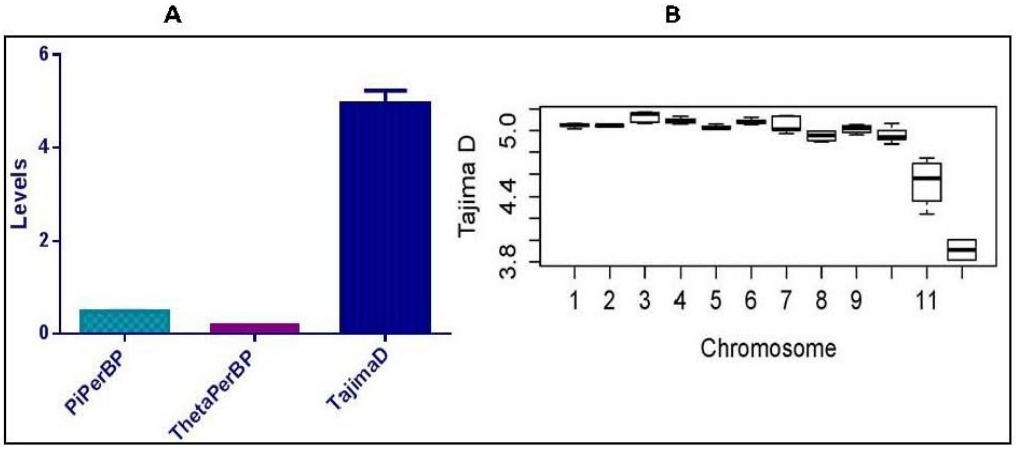


Figure 20: Diversity analysis (A) Average Pi, θ and Tajima D obtained in 89 EAHB and 3 out-group cultivars for 14121 SNP loci. B; average Tajima D calculated in the 11 and one unmapped *Musa* chromosomes.

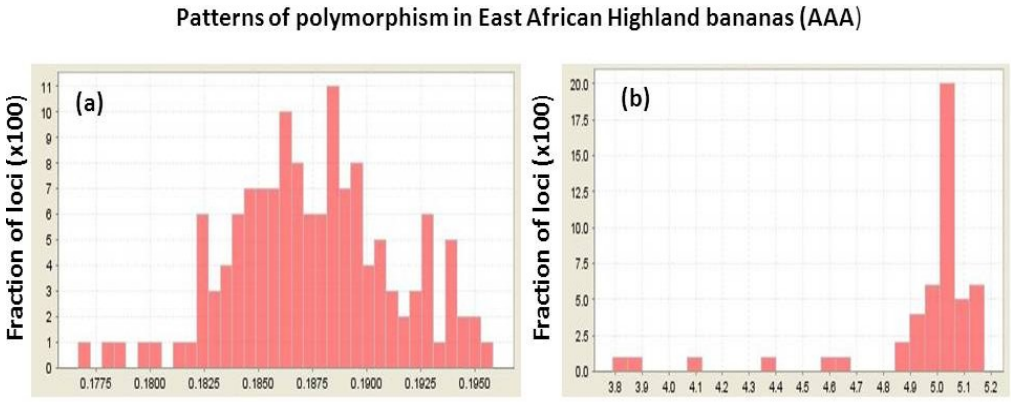


Figure 21: The distribution of nucleotide diversity and Tajima's D among loci. (a) The frequency distribution of diversity. (b) The frequency distribution of Tajima's D.

4.3.2 Genome-level polymorphism of the cultivars

Diversity analysis was done with the data set of 14121 SNPs excluding the out-group cultivars revealing 8675 and 6919 were monomorphic and variable sites respectively. Variable site with singleton mutations were 3721 and 3198 parsimony informative sites were observed. Nucleotide diversity, π, was 0.03532 (Sampling variance of π 0.0000719), Theta (per site) from S, θ_W = 0.08768. In the 14121 SNP loci 3858 codons were reported with higher nucleotide diversity in 2856.12 synonymous sites; π(s) =0.05616 (Jukes & Cantor; 0.06272) compared to the 8717.88 non-synonymous sites, π(a) = 0.04497 (Jukes & Cantor: 0.04924).

4.3.3 Diversity within the EAHB and between the EAHB vs out-group cultivars

Nucleotide diversity was significantly higher within the out-group cultivars (π=0.2830) vs the EAHB group (π=0.0368) (Table 14). However between population differences were low, we observed 36 number of population fixed nucleotide differences (EAHB vs out-group). Mutations that were EAHB polymorphic but out-group monomorphic and out-group-polymorphic but EAHB-monomorphic were 2518 and 2250, respectively and shared a total of 4048 mutations. Average number of nucleotide differences and nucleotide substitutions per site (Dxy) between the two groups was 3424 and 0.23079 (Dxy (JC 0.2795, SD 0.0482) respectively and a number of net substitution per site between populations at 0.0716.

Table 14: Diversity within the EAHB and between EAHB vs out-group cultivars

Parameter	EAHB	Outgroup
No of sequences	89	3
No of polymorphic sites	6566	6298
Total number of mutations	6566	6298
Average number of nucleotide differences, k	520.11	4198.67
Nucleotide diversity (π)	0.03506	0.28301
Nucleotide diversity with Jukes and Cantor, π(JC)	0.03679	0.35684
Standard deviation of π (JC)	0.00557	0.1005

4.3.4 Population structure of the EAHB

The PCA analysis of 89 EAHB cultivars and 3 out-group individuals showed similar patterns of genetic differentiation (Figure 22), as did the hierarchical classification (Figure 23). Despite their very different assumptions and modeling approaches, STRUCTURE, PCA, and tree-based analyses all provided very similar results, giving confidence in the robustness of these inferred population groups. On the PCA, the first two axes accounted for 34.61% of the total variation and the first five axes accounted for 64.49%.

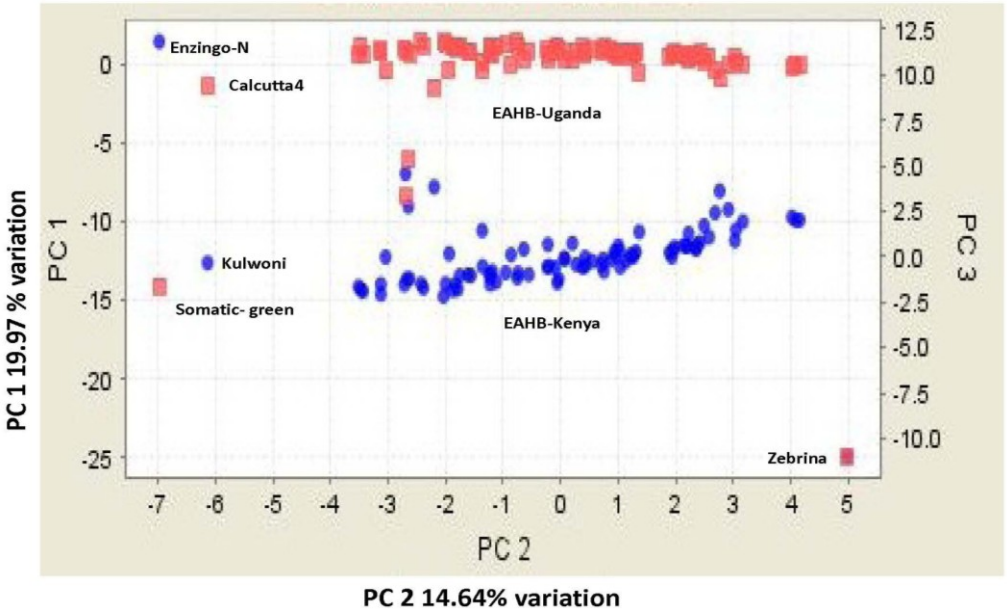


Figure 22: Principal component analysis plot of EAHB and out-group cultivars. PC1, PC2 and PC3 explained 19.97%, 14.64% and 9.98% of the total variance, respectively. The first 5 axes accounted for 63.49% of the total variation. The eigen analysis show a tendency to cluster based on their geographic origin structure in the EAHB population, but separation between the two regions with either of the eigenvectors is not apparent.

Applying the Evanno *et al*. (2005) method suggests K=3 as the optimal partition (Figure 23;), which is the uppermost relevant hierarchy reflecting the EAHB Kenya, EAHB Uganda and out-group cultivars Zebrina, Calcutta, somatic_green split. However, not much difference in terms of ancestry was observed for K=3 to K=5 (S (Figure 5). Above K=5, the increase in goodness of fit with larger K values are only incremental, suggesting that they do not reveal significant phylogenetic structure (Falush *et al.*, 2003; Pritchard *et al.*, 2010). STRUCTURE results of K=3 are in agreement with prior information about the main genetic architecture of the EAHB cultivars using microsatellite loci and AFLP loci.

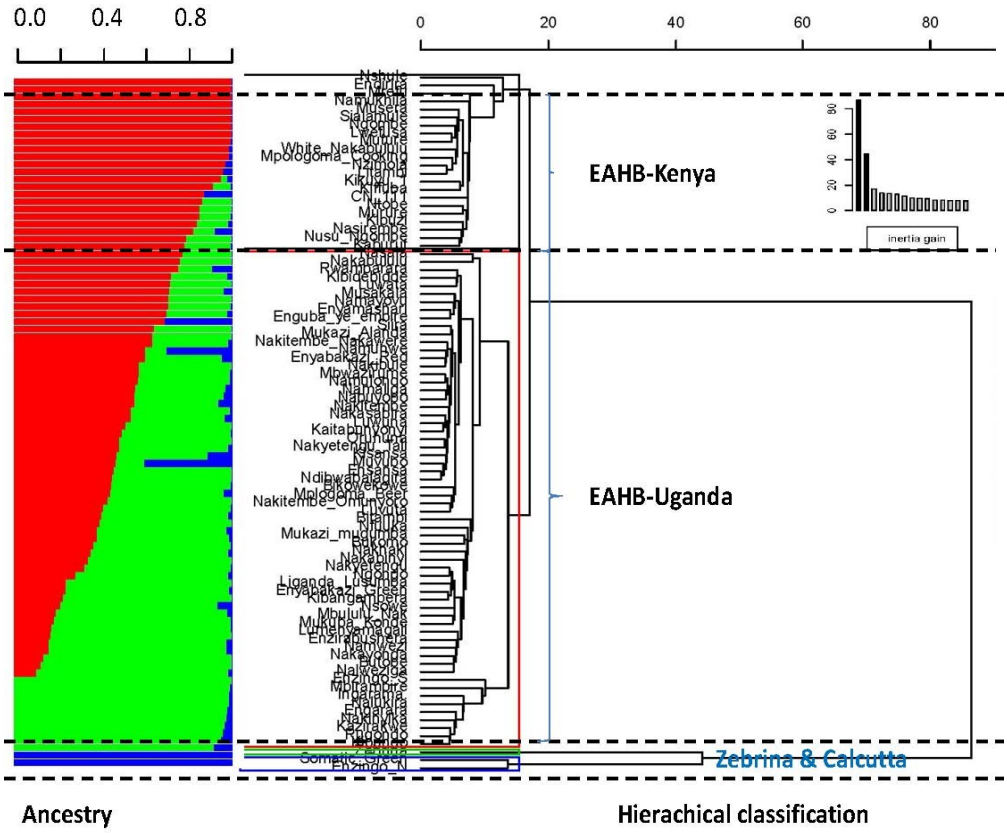


Figure 23: Hierachical classification and population structure of the EAHB. This figure depicts genetic relationships and ancestry of the EAHB cultivars, confirming that cultivars from the same region share a common ancestry and are more related compared to cultivars from different regions. However the out-group cultivars come from a different ancestral group. STRUCTURE bar plots of genetic membership proportions (K=3). Each cultivar is represented by a vertical line divided into K colors.

To corroborate the EAHB PCA, Structure and hierarchical clustering, we examined allele sharing across the 92 individuals by calculating identity by state (IBS) coefficients (i.e., the proportion of times a given pair of individuals have the almost same genotype across SNPs) among all pairs of individuals for autosomal genomic regions. For the entire population the estimated level of allele sharing was consistent with the PCA, structure and hierachical analysis (Figures 22&23) which suggests that the majority of genetic variation is found within (and not among) morphological groups and regions.

4.3.5 Genetic differentiation within population and between geographical regions

The average distances (heterozygosity) of cultivars in the three STRUCTURE groups; out-groups, Uganda and Kenya; were 0.3839, 0.1276 and 0.1689, respectively. The allele frequency divergence among the three groups was low, pairwise F_{ST} estimates between the three STRUCTURE groups averaged 0.0723 (out-group vs Kenya), 0.0618 (out-group vs Uganda) and 0.0242 (Kenya vs Uganda) on the basis of sequence data. Contrasting results for the F_{ST} values for the entire cultivars in groups (groups identified in the structure analyses) showed a significantly higher level of differentiation at cultivar-wide sampling (paired t-test, all P, 0.001). Mean F_{ST} values between cultivars within the groups was high 0.5393 and 0.4738 in Kenya and Uganda, respectively but low, 0.0942, between the out-group cultivars. Population genetic differentiation ϕ_{ST} measured by AMOVA analysis was 0.049; 95% of the variation was partitioned within groups, and only 5% was attributed to differences between groups (Kenya and Uganda) (the two hierarchical levels were significant with P = 0.001) (Table 15). Furthermore, allele frequencies of the Kenya and Uganda groups were highly homogeneous showing lack of differentiation (Figure 24).

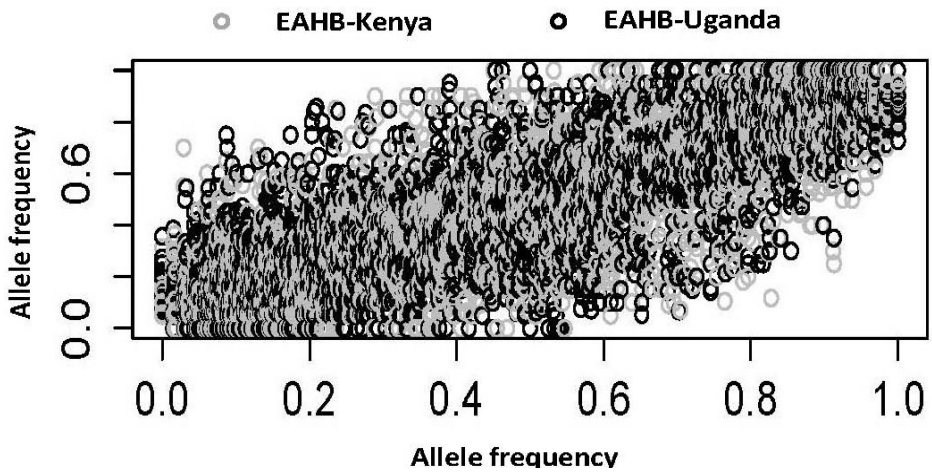


Figure 24: Geographical allele frequencies of the EAHB. Allele frequencies in samples of Kenya plotted against allele frequencies in Uganda.

Table 15: Summary AMOVA Table showing partitioning of variation within and between the EAHB morphological and geographical groups

Source	df	SS	MS	Est. Var.	%	PhiPT
Among groups	1	3274.50	3274.50	65.07	5%	0.049
Within groups	87	109329.80	1256.66	1256.66	95%	P=0.001
Total	88	112604.30		1321.73	100%	

AP = Est. Var. Among Pops,
WP = Est. Var. Within Pops
Probability, P (rand >= data), for PhiPT is based on standard permutation across the full data set.

4.3.6 Adaptation signatures

Bayescan v2.1 analysis produced no differentiation loci at a threshold of $\log_{10}$ PO = 2.0 (posterior probabilities higher than 0.99) corresponding and no outlier locus under selection was detected (Figure 25B & C). Of the total scanned SNP loci, 3937 SNP loci potentially affected by balancing selection and 2766 loci affected by neutral selection were identified (Figure 25A). Surprisingly, outlier loci affected by positive selection were not consistently identified by the outlier test (Figure 25A). In our study, however, the neutral F_{ST} distribution was high at 0.40 (Figure 25C).

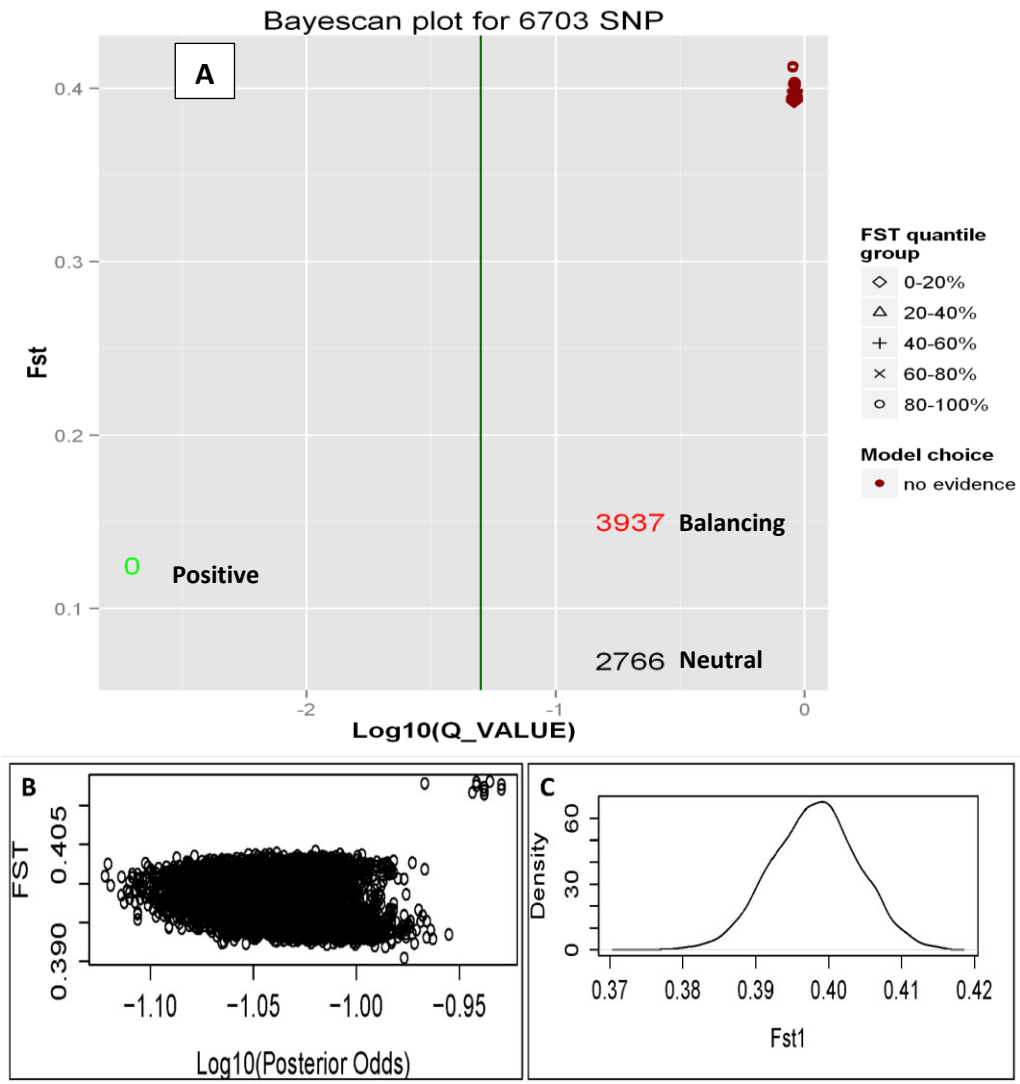


Figure 25: Genomic scan to identify outlier loci subject to selection by Bayescan approach. (A) Bayescan plot identifying number of loci under selection. (B) The results shows loci under balancing selection; Log_{10}<-0.95 but lacks evidence of outlier loci corresponding to Log_{10}> 2.0 (posterior odds) Each point corresponds to a SNP locus and F_{ST} is plotted against the Log_{10} of the posterior odds (PO), which provides evidence whether the locus is subject to selection or not (C) posterior distribution of F_{ST} show high F_{ST} values. The threshold value used for identifying outlier loci is (Log_{10}= 2.0). No outlier locus under selection is detected.

4.3.7 Evolutionary forces shaping genetic variation of the EAHB group

4.3.7.1 Polymorphism and divergence

In total, 11574 sites were analyzed, 2856.12 and 8717.89 synonymous and non-synonymous respectively. The number of codons detected in both sites was 9414 codons but 5556 of these were found to have alignment gaps or missing data, therefore only 3858 codons were analyzed. Both, synonymous and non-synonymous substitution stop codons were found in the coding region and were considered that they could be coding for a rare amino acid (the 21st amino acid: for example for Selenocysteine, Secys). Synonymous nucleotide diversity was higher; Pi(s); 0.0418 (Pi(s) Jukes &Cantor; 0.04308) compared to non-synonymous nucleotide diversity, Pi (a): 0.03279 (Pi (a) Jukes &Cantor; 0.03353) and the Pi (a)/Pi(s) ratio was 0.778. Synonymous nucleotide divergence, ks was 0.26303; ks (JC):0.3239 versus non-synonymous nucleotide divergence, ka: 0.2210; Ka (JC): 0.2618; the ka / Ks ratio was 0.808 indicating negative selection. A significant difference in ks and ka was observed when out-group values were included and were therefore excluded. The between cultivar sequence pairs had mean values of *K*a, *K*s, and *K*a:*K*s of 0.0231, 0.0351, and 0.0273 respectively. The coefficient of variation (CV), a measure of the variability of a sample relative to the sample mean, of *K*a and *K*s was 1.49, and 2.16 respectively. A *Z* test (Zar 1996) indicated that the variability in *K*a was significantly lower than in *K*s (*Z*_5.6, $p < 0.001$; to meet assumptions of normality samples were square root transformed prior to conducting the *Z* test).

4.3.7.2 Neutrality tests

Pairwise differences in EAHB were significantly similar with the distribution expected under a model of population growth and decline. All computed parameters, including (Fu and Li`s D^{*} = -4.5059, $^{**}P < 0.02$), Fu's, *Fs (Fs* = -6.75, $^{**}P < 0.02$), raggedness statistics *r* (0.0006, $^{*}p<0.004$) and Ramos-Onsins and Rozas's *R*2 statistics (R^2; 0.0478, $^{**}p<0.001$) were statistically significant and indicated population growth of the EAHB population.

4.3.7.3 Linkage disequilibrium

Of the total comparisons (704775), LD was reported for 3314 SNP pairs. When no recombination has occurred between two markers, D' will equal 1.0 (in the absence of mutation or genotyping error), while r^2 will be dependent on both markers' allele frequencies. Moderate but non-significant recombination (D' =0) were observed in 24.93% (2.51% significant), however, 31.32 % pairs significantly (Fishers exact test, $p<0.05$) lacked evidence of recombination (D'>0). Linkage equilibrium ($r^2=0$) was observed in 14.27% ($P>0.05$) of the total pairs. A large majority of pairs, 85.69%, (36.97% significant Fishers exact test, $p<0.05$) pairs were in linkage disequilibrium ($r^2>0$).

A dramatic reduction in number as well as in strength of LD was observed (in both r^2 and D') as the distance between marker pairs increased, inclusive of pairs of SNPs which can be defined as belonging to the same haplotype block (S Figure 6); r^2 seems to decay <10kb while D' decays much fast, <5kb, (Figure 26(b) and (c) respectively). None of the SNP pairs separated by more than 10 kb were in such absolute LD (Figure 26(a). There was almost no LD between two loci from different chromosomes. Whole genome LD showed a remarkably few significant number of haplotype blocks (S Figure 6) with chromosome 8 showing the highest number of significant haplotype blocks (Figure 27), this signifies that these loci are both statistically and physically linked.

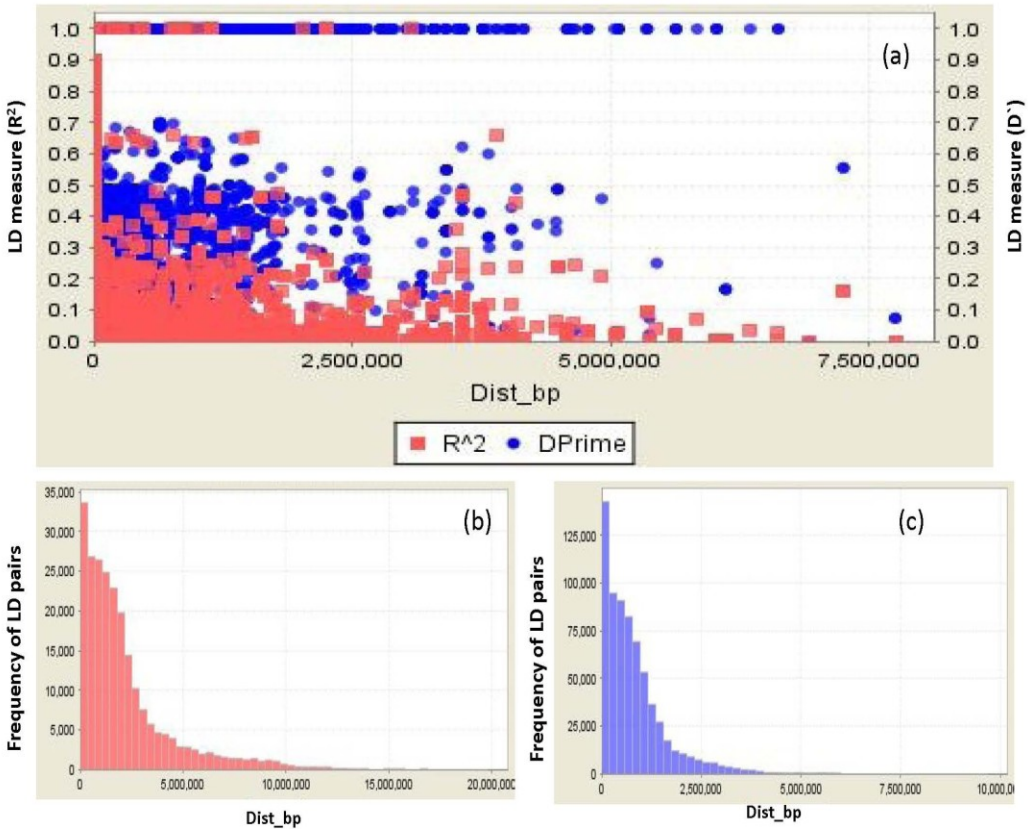


Figure 26: Extent of LD in SNP pairs of the EAHB population. (a) LD distribution presented by r^2 and D' as a function of distance (in bp). Each *spot* represents distance (bp) between the two polymorphic sites (*x*-axis) and LD of them as measured by r^2 and D'(*y*axis). (b) and (c) frequency of locus pair in LD (r^2 and D' respectively) plotted against marker intervals in bp. LD decays with increase of distance.

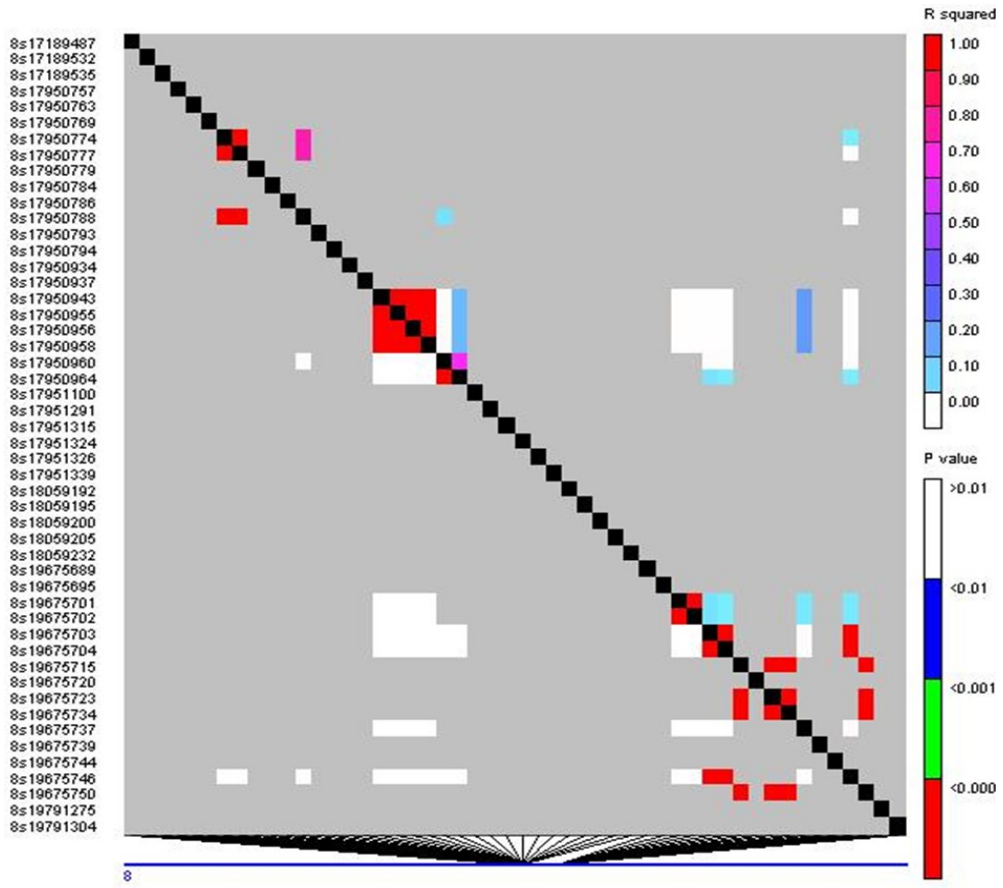


Figure 27: **LD heat plot of loci in chromosome 8**. Black triangles represent polymorphic sites. Each grid represents the strength of LD estimated by r^2 for each pairwise comparison between polymorphic sites with a minor allele frequency (MAF) > 0.1. The colour legend for r^2 values is given on the right side.

4.3.8 Demographic history of the EAHB population

4.3.8.1 Ancestral population size

The pairwise differences (mismatch sites) yielded signatures of waves of population growth and decline (Figure 28) which is taken as further strong evidence for population expansion of the studied species. No recombination was detected between adjacent sites (recombination parameter, R =0.000).

Minimum number of recombination, Rm, events observed over the entire autosomal region was 2843. Based on the Hudson-Kreitman-Aguadé (HKA) test we report a significant (X-square value: 6.086, P-value: 0.0136*) estimation of divergence time (T=6.7349) equivalent of 1198.81 generations (Tx2N) since the split of EAHB and Zebrina. The Expected number of segregating sites among the 89 sequences, Sn(t) was 2.713 and the value of Sn(t)/al was 0.536. Expected Number of Segregating sites among 2 sequences, S2(t), was estimated to 0.259 (i.e. the average number of pairwise differences) and the value of S2(t)/a1 was 0.25. The number of segregating sites in the 89 cultivars was higher than expected in a population in equilibrium upto the 9th generation (Figure 29).

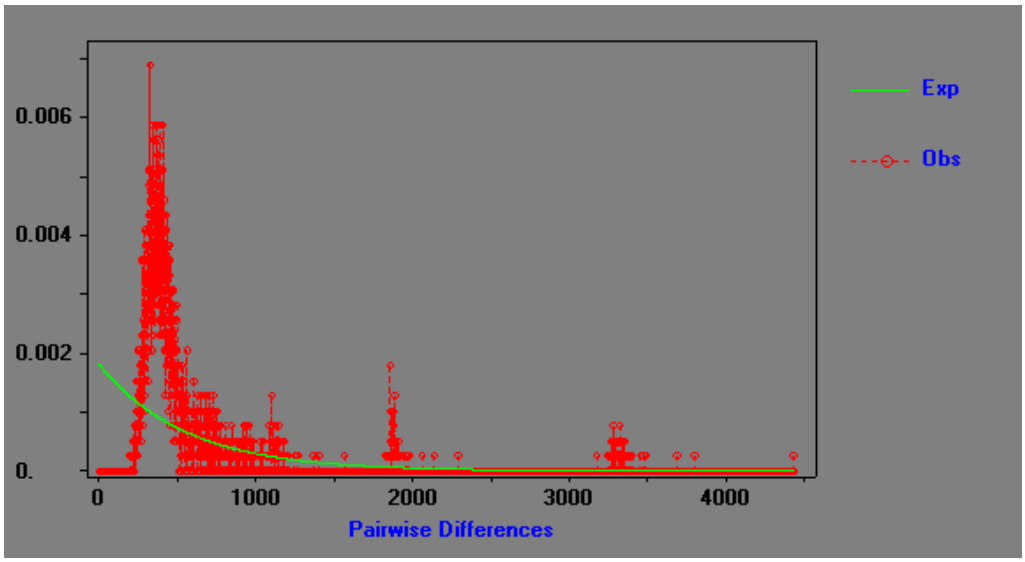


Figure 28: Mismatch distribution for EAHB population. Showing observed distribution of pairwise differences (open circles) and the expected distribution under a model of population growth and decline as calculated by DnaSP with initial theta = 506.424, final theta = 1000, and final tau = 44.324.

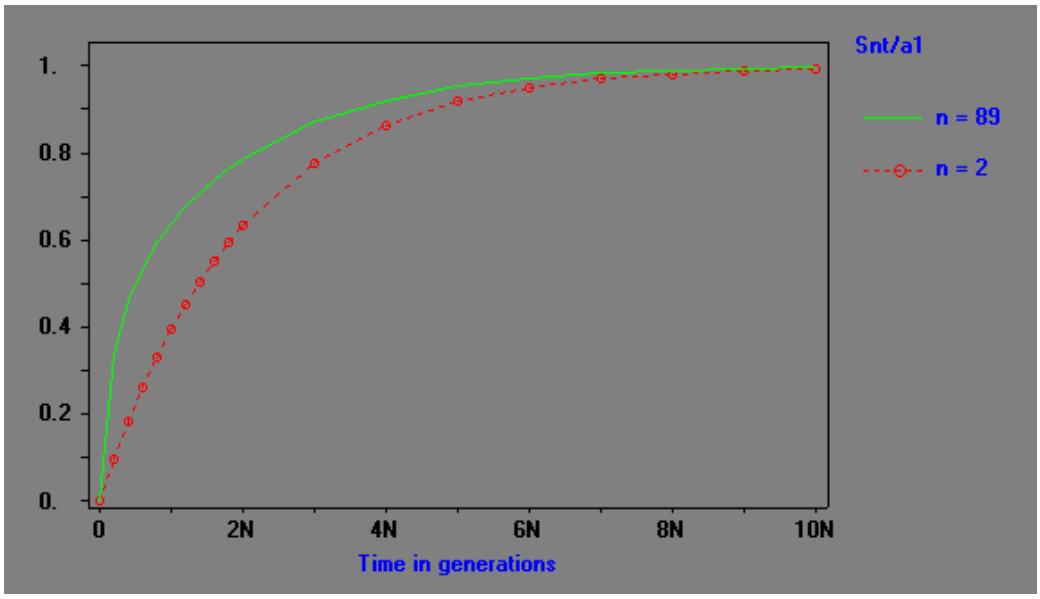


Figure 29: Frequency spectrum. The expected number of segregating sites among 89 sequences (Sn(t): 2.713) compared to the expected number of segregating sites among 2 sequences, S2(t): 0.259 (i.e. the average number of pairwise differences) computed with initial theta = 0.000, final theta=1.000 and Time=0.600N generations. Expected value of Sn(t)/a1 : 0.536, expected value of S2(t)/a1: 0.259 and a1 is the sum of (1/i) from i=1 to n-1.

4.3.9 EAHB coalescent time and speciation

Ancestral population size and speciation was calculated with; gamma parameter, α, chosen to be >1 so that the distribution peaks at the positive value instead at zero. The mean rate of evolution over the whole speciation tree was 0.998 and 95% of the density in the interval 0.859 and 1.1259. Speciation time of the EAHB cultivars in the prior was estimated to 928.37 years before present (between 670.361 and 1208.478 95% credibility interval) and population coalescent time estimated (tMRCA) to 2590.17 years (95% CI 2588.65 and 2590.62) (Figure 30 a & b; S Figure 7). The average rate of substitutions was 0.002 per site per generation (Figure 31) and the rate of evolution differed substantially amongst different lineages in the tree, coefficient of variation, estimating the variation of evolution from lineage to

lineage (expressed as a proportion of the mean rate) was 0.378 (95% CI between 2.0222E-6 to 1.1742).

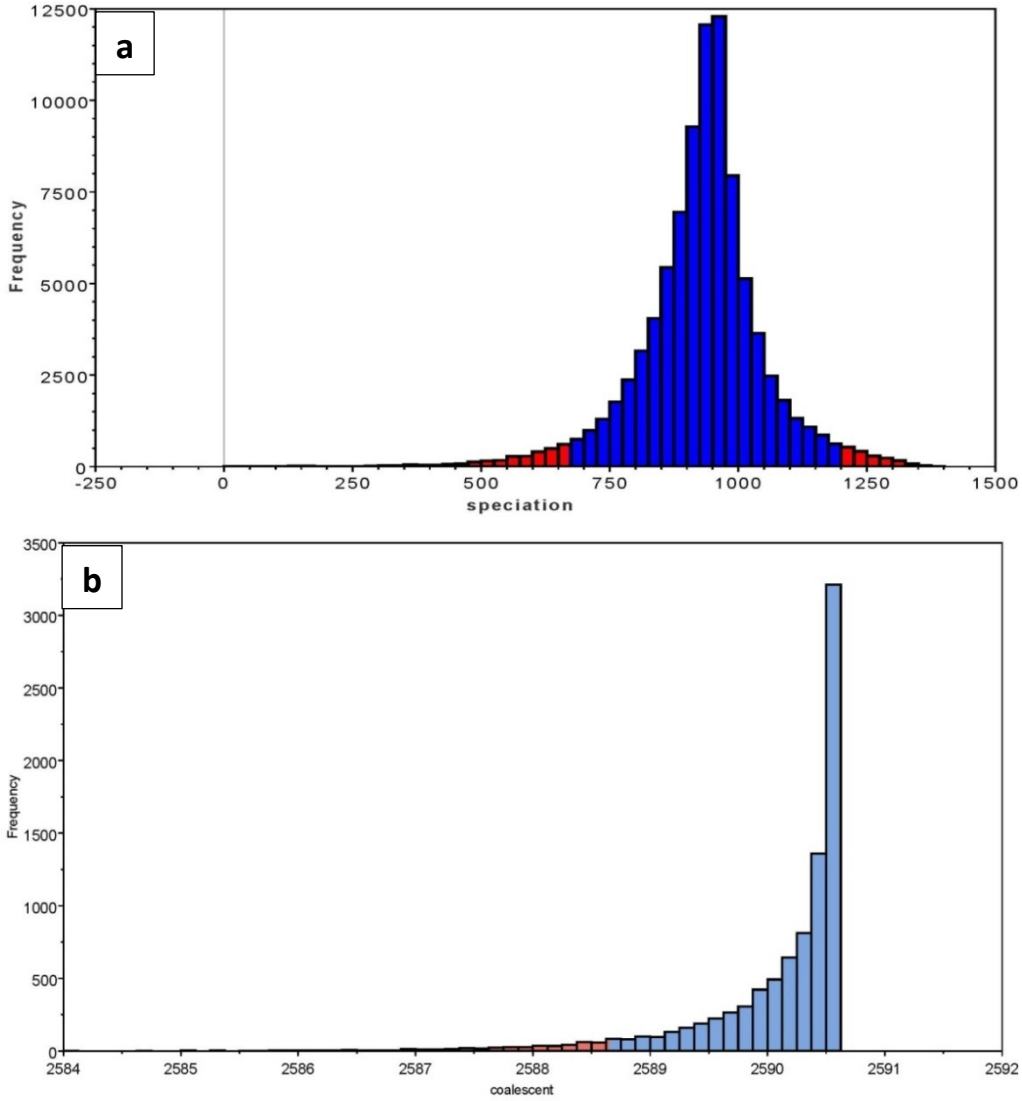


Figure 30: Demographic history of the EAHB (a) speciation time between the EAHB and Zebrina and (b) coalescent time of the EAHB population indicating the time (yrs) of most recent shared ancestor.

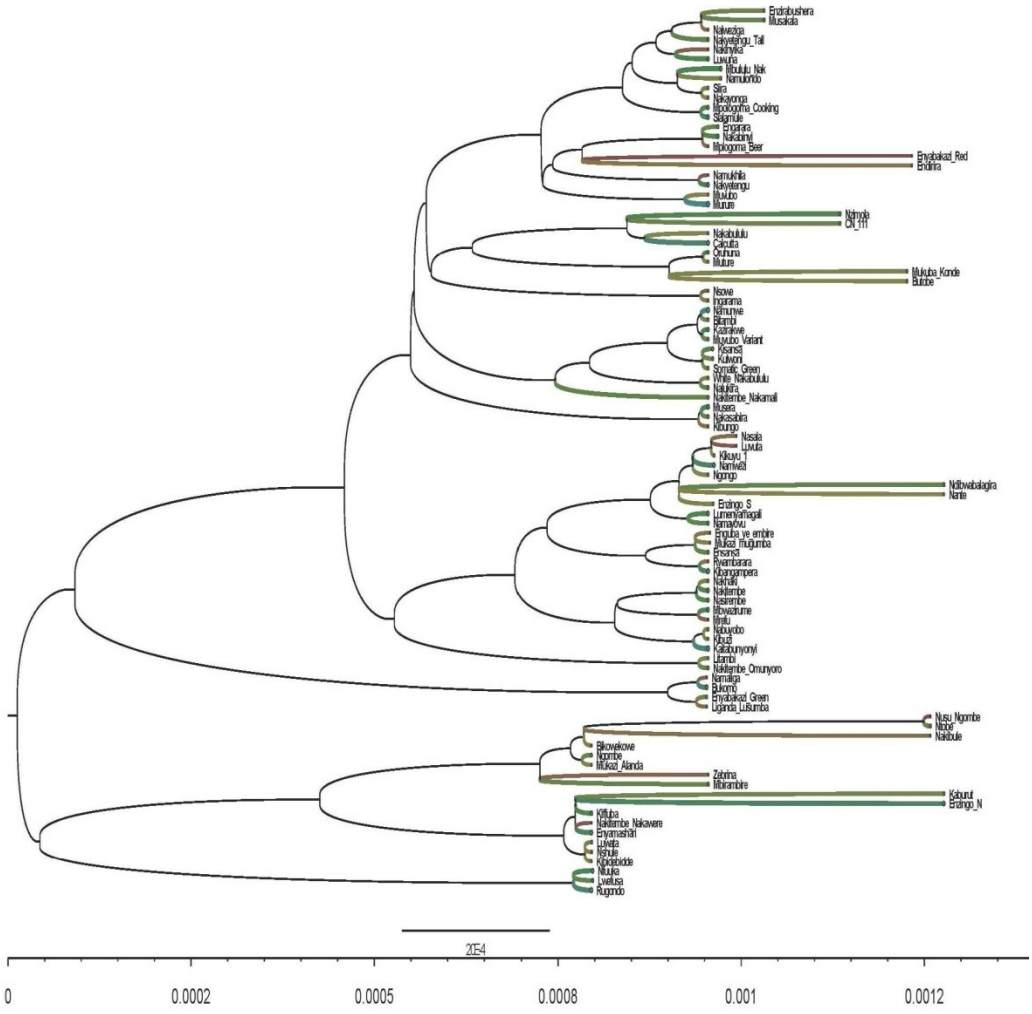


Figure 31: Species tree of the EAHB population based on variation in nucleotide substitution rates. The branch lengths estimated under uncorrelated relaxed Lognormal with Yule speciation prior. The species trees indicate variation in nucleotide substitution rates. The hue colour of the branches indicate the substitution rates

4.3.10 Trait-marker association

GLM analysis of marker-trait association among the cultivars revealed putative association of two markers (S9_12284986, S11_10140849) with bunch compactness per cultivar and four markers associated with fruit apex and shape and degree of fruit astringency (S9_12284986,

S11_10140849,S12_74648201 and S12_74648249). Plant stature had one marker (S9_12284986) associated with it (Table 16). All the markers were significantly associated with the 5-traits at 0.05% threshold level. Marker trait associations for degree of astringency in fruit pulp and fruit apex/shape were highly significant with markers located on chromosomes 9, 11 and 12. Markers S11_10140849 explained 39.03% explained maximum phenotypic variation for degree of astringency in fruit pulp followed by S9_12284986 (38.746%), S12_74648201 and S12_74648249 (31.596%). On the contrary, MLM analysis identified 621 SNPs that were significantly associated with all the 13 phenotypic traits (Figure 32 (a); S Table 2). The degree of fruit astringency had the highest number of associating SNPs (Figure 32(b).

Table 16: Significant association of SNP marker loci with five phenotypic traits identified by GLM analysis

Trait	Chromosome	Locus_position	Marker_F	R^2	Marker_p
Bunch compactness	9	12284986	23.66	0.22084	5.38E-06*
	11	10140849	13.98	0.25306	5.98E-06*
Fruit apex/shape	9	12284986	32.57	0.27669	1.73E-07***
	11	10140849	20.51	0.32739	5.97E-08***
	12	74648201	22.90	0.21227	7.34E-06*
	12	74648249	22.90	0.21227	7.34E-06*
Plant stature	9	12284986	26.36	0.23408	1.84E-06*
degree of astringency in fruit	9	12284986	53.00	0.38746	1.73E-10***
	11	10140849	26.50	0.39037	1.32E-09***
	12	74648201	38.66	0.31596	1.94E-08***
	12	74648249	38.66	0.31596	1.94E-08***

*** highly significant
* Significant

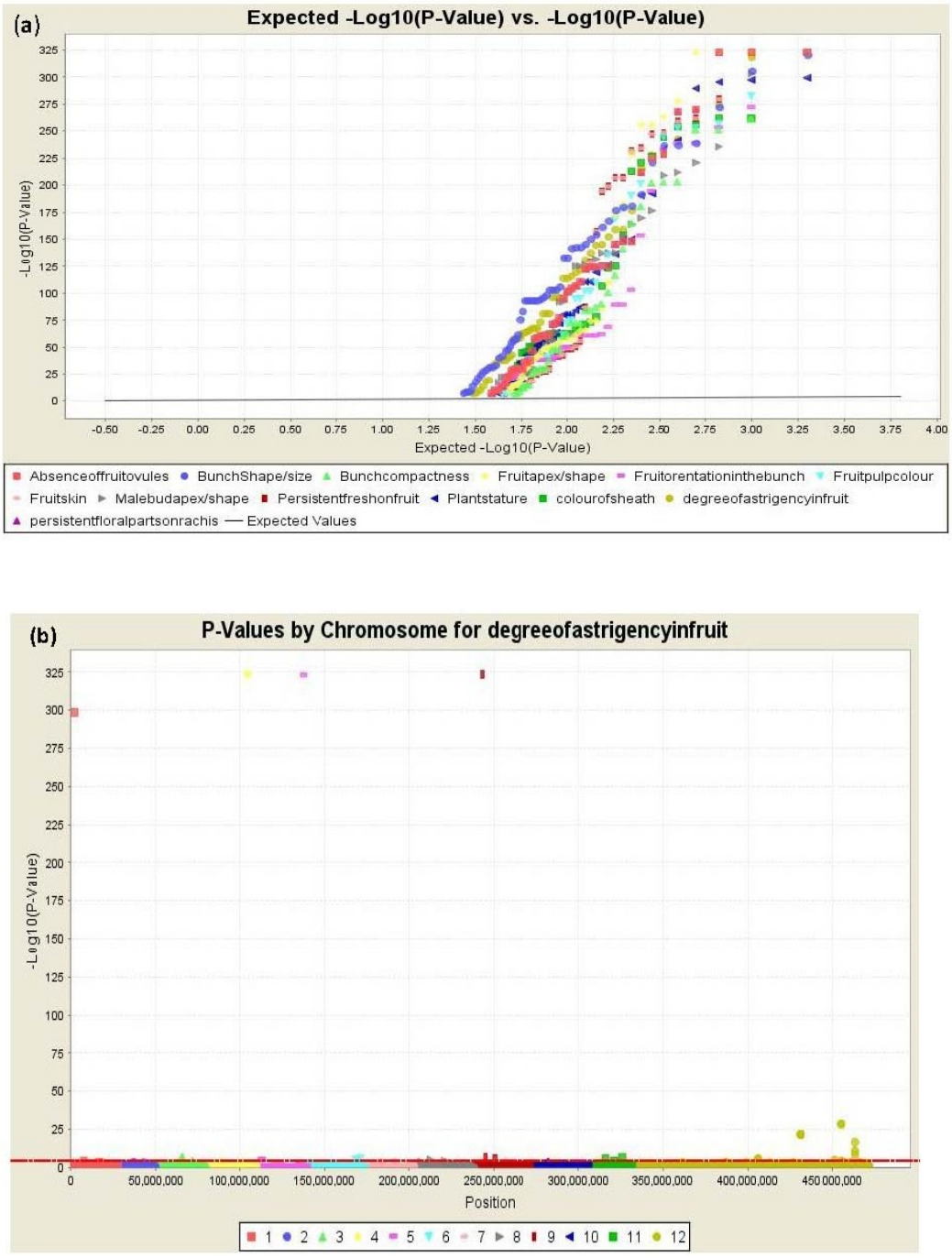


Figure 32: Genome-wide associations of SNPs with 13 traits found in EAHB subgroup and vulnerable to somatic mutations. (a) shows highly significant association between SNP markers and 13 phenotypic traits suggested to be vulnerable to somatic mutations was observed. (b) Manhattan plots of the MLM model for degree of fruit astringency in chromosomes 1-11 of Musa genome and in unmapped chromosome 12.Significance was evaluated at 7.0e-7 Bonferroni and cut-off point is shown by the cross-section line.

Negative log10-transformed *P* values from a genome-wide scan are plotted against position on each of 12 chromosomes.

4.4 DISCUSSIONS

Alignment probability depends on the length of alignment, on the number of mismatches and gaps and on the uniqueness of the aligned region on the genome and it should reflect the probability of the reads being of origin from the aligned region on the reference. It is possible that low number and variation in SNPs scored in this study may be because the reference genome used (double haploid Pahang-CIRAD 930 ITC 511) is derived from the Pahang wild diploid (*Musa acuminata* Colla ssp. Malaccensis) accession which shares its genetic lineage with dessert and cooking bananas (Josefina Ines *et al.*, 2013).

It is noteworthy that nucleotide diversity in EAHB was higher than observed for the Arabidopsis genome (Nordborg *et al.*, 2005) and potato (Simko *et al.*, 2006), but not remarkably elevated. Though not comparable, elevated genetic diversity has been observed in mitochondrial genes in asexual populations of Bark lice and was explained using three hypotheses that may have caused the elevated diversity; (i) larger effective population size, (ii) greater mutation rate or (iii) possible recent origin of sexuals (Shreve *et al.*, 2011). Results obtained in this study indicate that our samples studied here have a short coalescence time, and originate from a small number of founders that presumably were genetically close (Li *et al.*, 2011). Furthermore, highly selfing populations exhibit moderate reductions in diversity compared to outcrossing and mixed mating system (Ness *et al.*, 2010).

Tajima`s D (Tajima, 1989) is a measure of nucleotide diversity used to compare an observed nucleotide diversity against the expected diversity under the assumption that all polymorphisms are selectively neutral and constant population size. Demographic parameters would be expected to affect the genome evenly than selective pressures. Therefore, using the empiric distribution of Tajima`s D from a collection of regions across the genome

provides advantages in assessing whether selection or demography might explain an observed deviation from expectation. The positive Tajima's D for the SNP loci observed in this study indicates an excess of intermediate (both low and high frequency polymorphisms) frequency (polymorphic) alleles and signifies either a decrease in population size and/or balancing selection (Schmidt & Pool, 2002).

Even though the geographical regions seem to have a role to play in the genetic structure of the EAHB, low genetic diversity is observed within and between the EAHB subgroup. High genetic diversity can be created by multiple introductions, which bring together large amounts of genetic variation and novel genetic combinations (Wang *et al.*, 2012). Thus, single introductions may be inferred among the extant cultivars of the EAHB population. Moreover, we observed weak genetic differentiation between the cultivars, an indication that balancing selection might have occurred in this subgroup rather than population subdivision (He *et al.*, 2008). We did not observe divergence or isolation by distance between cultivars of this subgroup and neither between the two geographic regions. A weaker pattern of isolation by distance in most analyses of *A. thaliana* have been attributed to homogenization of allele frequencies resulting from thousands of years of human disturbance (Sharbel *et al.*, 2000; Nordborg *et al.*, 2005). Furthermore cultivar allele frequencies and region allele frequencies observed in this study were very homogeneous.

A number of evolutionary processes can impact the genetic diversity of natural populations among them; spontaneously arising mutations; inbreeding; natural selection and the Wahlund effect (Porth & El-Kassaby, 2014). Selection influences within-population diversity, but the effects are dependent on the nature of these selection processes, specifically balancing selection. Furthermore, the effects of natural selection are interwoven with stochastic effects, such as genetic drift. Mutations can counterbalance the loss of allelic diversity; however, natural mutations are rare, furthermore, harmful allelic variants are removed by purifying selection.

BAYESCAN results of this study detected a higher percentage of loci 27.88% under balancing selection versus loci under 19.59% under neutral selection and no loci under positive selection were detected. Balancing selection is known to increase the level of polymorphism because multiple alleles are likely maintained for a long time (Hudson & Kaplan, 1988), however; the level of polymorphism is reduced shortly after a fixation of adaptive mutation. This event is called a selective sweep because the fixation of a beneficial allele could sweep out the variation in the surrounding region of the selection target site by the hitchhiking effect (Kaplan *et al.*, 1989). BAYESCAN is known to be more efficient and detects a higher percentage of outlier loci than the other methods (Perez-Figueroa *et al.*, 2010) but could not detect loci under positive selection in this study. This is possibly due to weak signal of the selection (Karlsson *et al.*, 2014) . Similar results have been reported in Flax (Soto-Cerda & Cloutier, 2013). Perez-Figueroa *et al.* (2010) observed that the most favorable situation for detecting loci under positive selection is that of a low estimated neutral F_{ST} distribution (<0.20) as selective loci would tend to show high F_{ST} values. In our study, however, the neutral F_{ST} distribution was 0.40, implying that this factor could affect the efficiency of Bayescan in detecting positive selection. Conversely under balancing selection, a high neutral F_{ST} distribution would be more favorable for detecting selective loci (Perez-Figueroa *et al.*, 2010). Bayescan successfully identified loci affected by balancing selection. Comparisons of F_{ST} outlier tests indicated that Bayescan has the lowest type I error (Perez-Figueroa *et al.*, 2010).

The rates at which non-synonymous (ka) mutations are retained in a population indicate the presence and strength of selection in a coding region. The rate of mutation is expressed as the number of substitutions per non-synonymous site (Ka) or the number of substitutions per synonymous site (Ks). In neutrally evolving sequences, no difference should be observed between the two measures, or Ka = Ks. Positive selection in a region results in an increase in

the number of non-synonymous mutations, such as Ka >Ks. Conversely, if functional mutations are constantly removed from a population by purifying selection, the opposite trend can be expected (or Ka<Ks). The ratio Ka/Ks is evaluated among different coding regions. The ka/ks ratio obtained in this study was <1, and is consistent with a history of negative selection, purifying selection although it does not rule out positive selection (Yang & Bielawski, 2000; Roth & Liberles, 2006). There were no sequences with *K*a:*K*s >1, and thus we found no strong evidence that positive selection has contributed to the interspecific sequence divergence of any of these cultivars. Although the CV at synonymous sites was significantly less than the CV at non-synonymous sites, there was still a wide range in the rates at which synonymous substitutions accumulate, perhaps reflecting intergene differences in mutation rates or the relative strength of selection acting on codon. Although Ka/KS is a good indicator of selective pressure at the sequence level, it only calculates selective pressure within protein coding regions and therefore cannot detect evolutionary change that does not cause differences at an amino acid level; for instance, balancing selection (Yang & Bielawski, 2000).

Dissimilar values of pairwise differences and segregating sites reported in our study suggest that some form of selection could be acting on the sequences Genome wide changes in the shape of the frequency distribution (spectrum) of genetic variation is a signatures of the relative increase in the proportion of either low or high frequency mutations in the selected region (Tajima, 1989; Fu' & Li, 1993) within a population and suggests a positive or balancing selection of less than 200 000 years (Oleksyk *et al.*, 2010). Generations after the selective sweep, new (derived) mutations are slowly introduced back into the recently selected region, and most appear at low frequencies than expected under mutation/drift equilibrium, resulting in a skewed frequency distribution.

A large majority of combinations showed significant LD ($P < 0.05$), thus limited recombination. The degree of linkage disequilibrium is high; vegetative

propagation mimics complete physical linkage over the entire genome (Vandepitte *et al.*, 2010). Significant LD has been reported in other clonal populations (Lott *et al.*, 2010; Vandepitte *et al.*, 2010). A number of factors are known to contribute to the emergence and maintenance of LD; mutation drift, population bottlenecks, population substructure, population admixture, levels of inbreeding and selection (Mather *et al.*, 2007). In this study, moderate but non-significant recombination was observed. Since LD between two loci is degraded by crossover between genes, the extent of LD is dependent on the effective recombination rate; this explains the observed high number of SNP pairs in LD.

A high level of inbreeding is evident in this study and may have led to the decline in SNP pairs showing evidence of recombination. This is because high levels of inbreeding cause populations to become composed of homozygous, inbred lines that in turn limit the effectiveness of recombination (Morrell *et al.*, 2005). A strong correlation is expected between interlocus distance and LD in a population of constant size if recombination rates do not vary across the genome, in this study no correlation between interlocus distance was observed and LD.

LD decay within 10kb was observed with increasing distance. Patterns of linkage disequilibrium have been characterized in several crop species and their relatives; with rapid LD decay observed in outcrossing species (Remington *et al.*, 2001; Tenaillon *et al.*, 2001; Garris *et al.*, 2003; Hamblin *et al.*, 2005; Rakshit *et al.*, 2007). However, high levels of marker association has been observed to persist in selfing species e.g. barley, soybean (Caldwell *et al.*, 2006; Hyten *et al.*, 2007). Low levels of linkage disequilibrium and intragenic LD decay with a range of only a few kilobases were reported in wild barley (*H. vulgare* ssp. spontaneum), highly selfing species (Morrell *et al.*, 2005). SNPs separated by long distance are for the most part in linkage equilibrium (r^2=0) and therefore the LD seems to decreases with increasing genetic distance

(Shifman, 2003). LD decay can also vary considerably from locus to locus due to different recombination rates and selection pressures at different regions of the genome. In addition, higher levels of LD are observed in self-pollinating species compared to outcrossing species, indicating that mating systems play a role (Flint-Garcia *et al.*, 2003).

Three types of populations were suggested to have an increased LD: (i) admixture of two populations with different allele frequencies (Smith *et al.*, 2001) (ii) small stable populations (Terwilliger *et al.*, 1998) and (iii) isolated populations with a few founder individuals, facing rapid expansion (Sheffield *et al.*, 1998; Shifman & Darvasi, 2001). In our study a good proportion of SNP pairs shows no evidence of recombination (D'=1.0). SNPs common in all samples are suggested to have emerged from old mutations however, in regions with very low recombination rates, some of the haplotype combinations for these SNPs are relatively rare or absent in a particular population due to genetic drift and recent historical events. Our results are consistent with history of the EAHB that suggest the population may have descended from a small number of founders and then rapidly expanded 2500 years ago (Perrier *et al.*, 2011). Interesting, no recombination events were detected between adjacent sites. Increased homozygosity observed in clonal plants reduces opportunities for recombination to break down associations among alleles by crossing over between heterozygous loci (Nordborg 2000). This further increases linkage disequilibrium and decreases the effective rate of recombination (Wright *et al.*, 2008). Signatures of genetic recombination have been reported in several studies in clonal lineages of asexual species (Stewart *et al.*, 2013).

In many plant populations, especially those with annual life histories and small structured populations, demographic processes may play a more prominent role in causing reduced diversity than increased genetic hitchhiking (draft) associated with selfing (Ness *et al.*, 2010). The effects of draft are strongest in

asexual organisms where the entire chromosome stays linked forever. Based on the observation of Harpending (1994), the pairwise differences among all DNA sequenced in our study revealed a constant population size for a long time followed by expansion (population growth) (Figure 28). The distribution of pairwise differences in a sample from populations that has been stationary for a long time are ragged and erratic, whereas a population that has been growing generates mismatch distributions that are smooth and have a peak (Harpending, 1994). The position of the peak reflects the time of the population growth; our data shows a signature of a recent population expansion (Figure 28).

Episodes of population growth and decline leave characteristic signatures in the distribution of nucleotide (or restriction) site differences between pairs of individuals. The implications of continued exponential growth are indistinguishable from those of a sudden burst of population growth. For instance bottlenecks in population size also generate waves similar to those produced by a sudden expansion, but with elevated upper tail probabilities. Initially reductions in population size generate L-shaped distributions with high probability of identity, but these converge rapidly to a new equilibrium. In equilibrium populations the theoretical curves are free of waves. The most serious one could be that demography also affects the amount and the pattern of polymorphism. In other words, the effects of selection and demography are confounded (Innan, 2006).

The GLM analysis does not account for kinship as a potential cause of the genotype-phenotype relationship. The structured association mapping revealed four quite interesting marker-trait associations for four agronomic traits. This approach could identify markers with pleiotropic effects (S_12284986 in chromosome 9 was found associated with all four traits; bunch compactness, plant stature, degree of astringency in fruit and fruit apex/shape), as well as markers that were identified to be epistatic indicating that population wide

analysis served as an effective tool in deciphering marker–trait associations in EAHB population. Though not at the same degree, significant associations of markers with agronomic traits has been reported in rice (Huang *et al.*, 2010), sorghum (Kannan *et al.*, 2014), lettuce (Kwon *et al.*, 2013) and peaches (Dhanapal & Crisosto, 2013). Even though GLM and MLM analysis are basically similar, highly significant associations were obtained using MLM because it has a statistical power equivalent to that of the full optimization approach only for traits with low heritability (Zhang *et al.*, 2010). Genome wide association studies (GWAS) identified SNPs in the genomic DNA of the EAHB cultivars that were highly associated to economically important traits of interest (bunch compactness, plant stature, degree of astringency in fruit and fruit apex/shape). SNPs are the indirect markers associated with the quantitative trait loci (QTL); therefore, they cannot specify the causal genes on their own. Further work should be done to associate these SNPs with the positions of genes and QTLs or even identification of novel QTLs. These presumptive QTL loci would then provide opportunities for improvement of EAHB based on a marker approach. The results suggest that GWAS has potential for use in future breeding programs in *Musa* spp.

4.5 CONCLUSIONS AND RECOMMENNDATIONS

Analysis of DNA sequence diversity of >14k SNPs across 89 EAHB sample provides evidence of balancing/purifying selection associated with EAHB domestication and crop improvement. Although it is possible that some of these cases of low genetic diversity are due to recent demographic crashes, it is unlikely they can all be explained by recent demographic events. We suggest that patterns of low genetic diversity could be interpreted as reflecting a recent reduction in diversity as a result of selection of farmer preferred banana clones, or alternatively, a historical lack of diversity caused by the hybridization of genetically similar sexual parents. Like SSR and AFLP data, the analysis supports the hypothesis of a genetic bottleneck caused by the introduction of a

single clone, or a limited number of clones, followed by rapid population expansion. To extrapolate the exact cause of the low genetic diversity in the East African highland banana subgroup, sampling from other potential regions would be beneficial to understand the movement of the EAHB from its centre of origin, and patterns of past hybridization with wild relatives.

4.6 SUPPLEMENTARY MATERIALS

S Table 6 GBS reference pipeline. Bwa parameters and options used for read alignment and SNP calling used for EAHB GBS analysis

	Option	Value	Description
FastqToTagCountPlugin	c	1	Minimum number of times a tag must be present to be output. Default: 1
FastqToTagCountPlugin	s	300000000	Max good reads per lane. (Optional. Default is 300000000).
MergeMultipleTagCountPlugin	c	5	Minimum number of times a tag must be present to be output. Default: 1
TagCountToFastqPlugin	c	1	Minimum count of reads for a tag to be output (default: 1)
FastqToTBTPlugin	y	-y	output to tagsByTaxaByte (tag counts per taxon from 0 to 127) instead of tagsByTaxaBit (0 or 1)
FastqToTBTPlugin	c	1	Minimum taxa count within a qseq file for a tag to be output. Default: 1
MergeTagsByTaxaFilesPlugin	s	200000000	Maximum number of tags the TBT can hold while merging (default: 200000000)
TagsToSNPByAlignmentPlugin	y	-y	Use byte-formatted TBT file (*.tbt.byte)
TagsToSNPByAlignmentPlugin	errRate	0.01	Average sequencing error rate per base (used to decide between heterozygous and homozygous calls) (default: 0.01)
TagsToSNPByAlignmentPlugin	mnLCov	0.1	Minimum locus coverage i.e. the proportion of taxa with at least one tag at the locus. Default: 0.1
TagsToSNPByAlignmentPlugin	mxSites	2000000	The maximum number of SNPs per chromosome for hapmap files (default = 2000000)
TagsToSNPByAlignmentPlugin	mnMAC	999	Minimum minor allele count. Defaults to 10. SNPs that pass either the specificed minimum minor allele count (mnMAC) or frequency (mnMAF) will be output.
TagsToSNPByAlignmentPlugin	mnMAF	0.01	Minimum minor allele frequency. Defaults to 0.01. SNPs that pass either the specificedminimum minor allele frequency (mnMAF) or count (mnMAC) will be output.

MergeDuplicateSNPs Plugin	misMat	0.05	Threshold mismatch rate above which the duplicate SNPs won't be merged. Default: 0.05.
MergeDuplicateSNPs Plugin	callHets	-callHets	When two genotypes at a replicate SNP disagree for a taxon call it a heterozygote. Defaults to false (=set to missing)
FastqToTBTPlugin	y	-y	output to tagsByTaxaByte (tag counts per taxon from 0 to 127) instead of tagsByTaxaBit (0 or 1)

S Table 7: Morphological traits used for GWAS analysis of East African Highland banana. List of 13 traits known to be vulnerable to somatic mutation within the EAHB subgroup (passport data taken from (Karamura *et al.*, 2010).

Morphology/Traits
Bunch Shape/size
Bunch compactness
Fruit apex/shape
Fruit orientation in the bunch
Fruit pulp colour
degree of astringency in fruit
Fruit skin
Absence of fruit ovules
Persistent fresh on fruit
persistent floral parts on rachis
Male bud apex/shape
Plant stature
colour of sheath

S Table 8: Nucleotide variation in 139 blocks of 14121 SNP loci in 89 East African Highland bananas

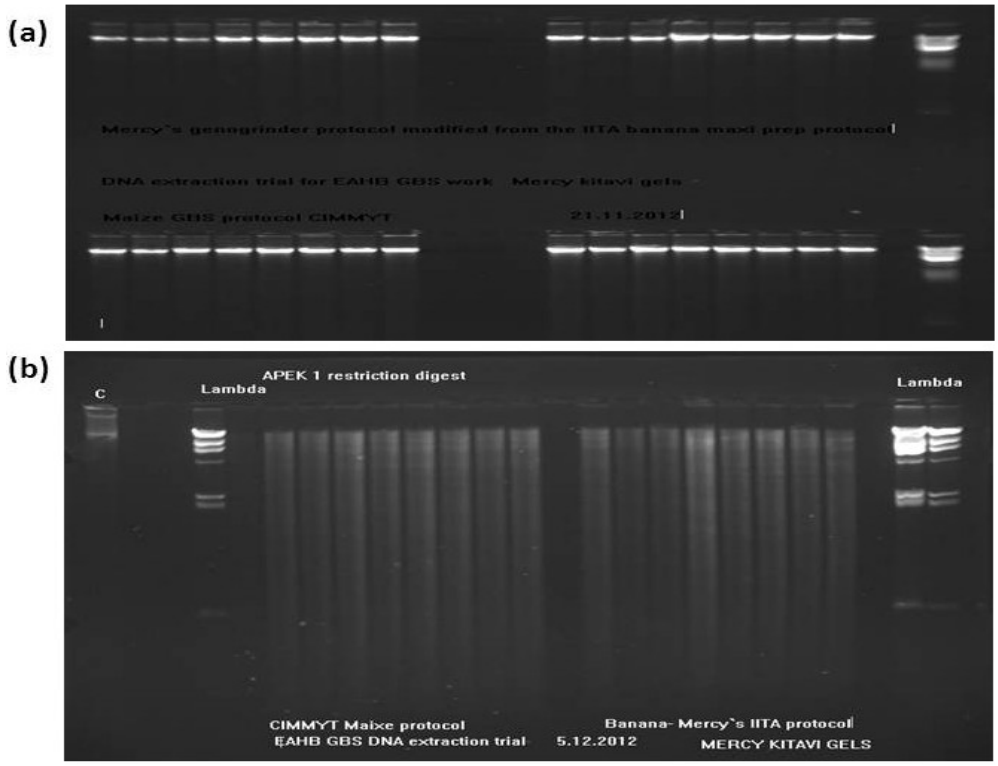
Chromo-some	Start Chr Position	End Chr Position	Pi Per BP (π)	Theta Per BP (θ)	TajimaD
1	93,199	11,227,099	0.28469	0.19226	1.65535
1	2,797,527	14,300,641	0.28775	0.1942	1.65696
1	5,418,322	16,310,403	0.2736	0.19426	1.4065
1	6,566,402	17,330,955	0.26026	0.1938	1.17941
1	8,979,938	19,145,938	0.25465	0.19261	1.10795
1	11,324,779	23,121,117	0.2378	0.19181	0.82462
1	14,300,691	3,941,769	0.23618	0.19067	0.82189
1	16,310,417	7,504,976	0.24577	0.18952	1.0229
1	17,340,514	10,657,730	0.24949	0.18872	1.10985
1	19,145,989	13,374,806	0.23635	0.18907	0.86091
1	23,637,445	15,284,083	0.23252	0.18708	0.8362
2	3,941,781	16,982,786	0.24009	0.18783	0.9566
2	7,539,992	19,013,412	0.24926	0.18899	1.09561
2	10,695,766	20,954,392	0.24859	0.18899	1.08345
2	13,388,916	742,252	0.24425	0.18541	1.08986
2	15,284,084	2,988,618	0.26252	0.18859	1.3466
2	17,049,488	6,432,849	0.26391	0.18942	1.35228
2	19,013,419	8,349,927	0.25969	0.18903	1.2856
2	20,954,408	10,132,635	0.2609	0.18982	1.28769
3	742,274	13,128,372	0.26945	0.19352	1.35254
3	3,016,834	19,426,962	0.26983	0.19232	1.38933
3	6,461,754	23,348,046	0.28928	0.18992	1.8032
3	8,568,427	25,558,919	0.29474	0.18958	1.91424
3	10,132,636	28,283,644	0.30604	0.18952	2.11916
3	13,128,541	30,449,874	0.30958	0.18823	2.21693
3	19,536,590	3,705,884	0.28784	0.18704	1.85332
3	23,348,088	6,898,752	0.25162	0.1878	1.16749
3	25,579,561	8,845,525	0.24031	0.18819	0.95134
3	28,283,766	11,745,238	0.23347	0.18621	0.87176
3	30,449,876	17,929,725	0.23209	0.18823	0.80125
4	3,705,888	22,049,623	0.24304	0.18863	0.99197
4	6,898,764	24,411,611	0.26776	0.18668	1.4952
4	9,089,578	26,404,017	0.26548	0.18863	1.40105
4	11,745,239	29,068,914	0.228	0.18783	0.73536
4	17,929,729	1,668,011	0.22352	0.18462	0.72372
4	22,057,353	4,354,177	0.21015	0.18226	0.52612
4	24,447,353	7,113,684	0.1869	0.17946	0.1423
4	26,468,022	10,118,345	0.19724	0.17669	0.39971

4	29,068,937	15,371,281	0.2382	0.17868	1.14494
5	1,671,562	20,621,113	0.23785	0.17788	1.15881
5	4,354,203	24,791,834	0.24412	0.18186	1.17686
5	7,113,719	27,359,405	0.2647	0.18783	1.40724
5	10,118,381	227,346	0.25385	0.19062	1.14091
5	15,371,312	3,136,240	0.2495	0.18903	1.1002
5	20,621,278	5,049,516	0.26003	0.19141	1.23284
5	24,791,987	6,264,327	0.26436	0.19022	1.34042
5	27,425,775	8,286,461	0.2718	0.18823	1.52664
6	227,349	10,181,146	0.2782	0.18468	1.743
6	3,136,243	12,628,926	0.28698	0.18545	1.8826
6	5,049,545	17,497,334	0.2816	0.18508	1.79511
6	6,264,333	23,513,252	0.2661	0.18428	1.52821
6	8,342,180	26,849,228	0.26299	0.18428	1.47002
6	10,181,150	28,862,635	0.26215	0.18748	1.37132
6	12,638,685	30,894,280	0.2512	0.18827	1.15062
6	17,497,335	32,844,823	0.2463	0.18624	1.10889
6	23,513,256	812,484	0.24976	0.18465	1.21238
6	26,849,326	2,851,692	0.25377	0.18545	1.2669
6	28,862,664	4,389,362	0.2566	0.18462	1.33898
6	30,937,296	7,033,482	0.25339	0.18385	1.30036
6	32,846,192	8,560,591	0.26023	0.18624	1.36602
7	812,489	11,662,287	0.27942	0.18664	1.70949
7	2,851,782	19,554,430	0.25892	0.18668	1.33214
7	4,404,767	23,098,657	0.25876	0.18468	1.38071
7	7,033,484	25,398,706	0.27579	0.18428	1.70914
7	8,560,600	27,548,262	0.29254	0.18306	2.05624
7	11,662,294	1,698,469	0.30994	0.18584	2.29605
7	19,554,473	3,476,494	0.32235	0.18346	2.60308
7	23,153,098	5,231,921	0.31187	0.18465	2.36907
7	25,398,810	8,439,257	0.29568	0.18226	2.1395
7	27,562,149	12,110,335	0.25562	0.18269	1.37406
8	1,764,764	17,068,983	0.23058	0.18349	0.88339
8	3,476,536	21,771,492	0.20919	0.18637	0.42251
8	5,250,149	26,103,014	0.19761	0.18717	0.19239
8	8,439,427	27,839,872	0.20862	0.18958	0.34661
8	12,110,441	30,367,982	0.23688	0.18552	0.95404
8	17,069,076	32,370,489	0.24154	0.18392	1.07943
8	22,069,579	405,867	0.2441	0.18226	1.16638
8	26,103,020	1,982,588	0.26088	0.18306	1.4617
8	27,839,874	5,137,875	0.22008	0.18226	0.71333
8	30,367,999	7,216,634	0.21922	0.18624	0.60882
8	32,370,616	9,392,825	0.22711	0.18584	0.76363

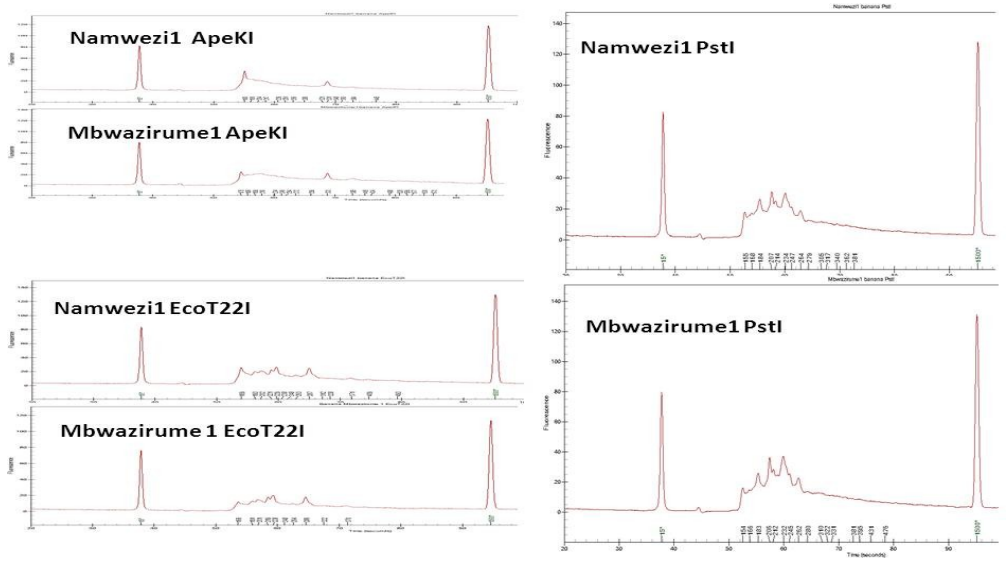
9	405,891	14,274,275	0.2548	0.18783	1.22606
9	1,982,595	25,785,937	0.2751	0.18907	1.56647
9	5,150,101	28,185,872	0.32453	0.19067	2.41723
9	7,228,216	30,886,242	0.33686	0.19027	2.65265
9	9,567,860	33,138,697	0.32254	0.19107	2.36933
9	14,274,425	7,644,844	0.29887	0.18508	2.11622
9	25,785,940	11,900,992	0.27349	0.18508	1.64415
9	28,375,062	16,534,752	0.26373	0.18512	1.46332
9	30,886,263	19,425,157	0.26172	0.18352	1.46823
9	33,138,711	22,323,476	0.26963	0.18352	1.61669
10	7,644,850	24,413,470	0.28217	0.18872	1.70674
10	11,902,714	26,114,284	0.28913	0.18624	1.89956
10	16,534,925	28,045,963	0.28472	0.18744	1.78491
10	19,663,221	30,130,210	0.28179	0.18823	1.70932
10	22,383,966	32,308,353	0.28048	0.18624	1.74007
10	24,413,588	1,032,282	0.27375	0.18581	1.62553
10	26,114,287	3,939,496	0.27211	0.18462	1.62768
10	28,045,992	5,968,107	0.26855	0.18385	1.5839
10	30,130,233	8,766,859	0.25851	0.18303	1.41619
10	32,308,356	11,557,403	0.2466	0.18704	1.0951
11	1,037,338	15,393,266	0.24961	0.187	1.14996
11	3,939,505	19,709,123	0.26038	0.18624	1.36886
11	5,968,112	22,049,115	0.26254	0.18541	1.42869
11	8,766,868	23,616,089	0.24354	0.18226	1.156
11	11,557,492	25,382,090	0.25567	0.18025	1.43683
11	15,722,862	5,849,873	0.25638	0.18147	1.41916
11	19,709,139	8,363,085	0.24221	0.18552	1.05289
11	22,049,158	15,391,442	0.23242	0.18757	0.82495
11	23,616,092	24,942,785	0.25808	0.19366	1.14942
11	25,382,094	29,468,264	0.23821	0.19286	0.81249
12	5,862,280	36,095,969	0.23154	0.19044	0.74548
12	8,363,087	48,149,238	0.22838	0.18958	0.70622
12	15,391,652	54,266,510	0.22778	0.18757	0.73956
12	24,942,798	64,555,436	0.2137	0.18522	0.53114
12	29,893,252	66,371,231	0.21572	0.18763	0.51701
12	36,241,731	70,796,098	0.21651	0.18924	0.49788
12	48,155,451	76,933,735	0.21754	0.19004	0.49979
12	54,266,522	81,688,245	0.21419	0.19125	0.41448
12	64,555,454	84,723,973	0.20943	0.19406	0.27358
12	66,371,234	90,412,654	0.22481	0.19293	0.57163
12	70,796,102	93,106,244	0.2092	0.19213	0.30728
12	76,934,414	96,531,157	0.2033	0.18931	0.25579
12	81,688,281	99,785,888	0.1975	0.18777	0.18004

	12	84,724,096	107,211,037	0.2003	0.18608	0.2638
	12	90,412,682	108,319,478	0.1809	0.18689	-0.11113
	12	93,123,348	113,033,806	0.1886	0.18649	0.03929
	12	96,531,169	117,082,247	0.1930	0.19052	0.0457
	12	100,025,223	122,947,034	0.1972	0.19165	0.09935
	12	107,211,043	124,072,433	0.1846	0.19293	-0.14915
	12	108,319,490	125,555,265	0.1843	0.19293	-0.15472
	12	113,033,813	129,371,212	0.1930	0.19414	-0.02034
	12	117,082,249	131,067,955	0.1924	0.19302	-0.01182
	12	122,947,035	133,063,627	0.2022	0.19503	0.12747
	12	124,072,479	135,686,359	0.2144	0.19584	0.3277
	12	125,555,267	138,184,051	0.2258	0.19495	0.547
	12	129,371,215	139,931,680	0.2137	0.19414	0.34888
Average				**0.2495**	**0.1875**	**1.1407**

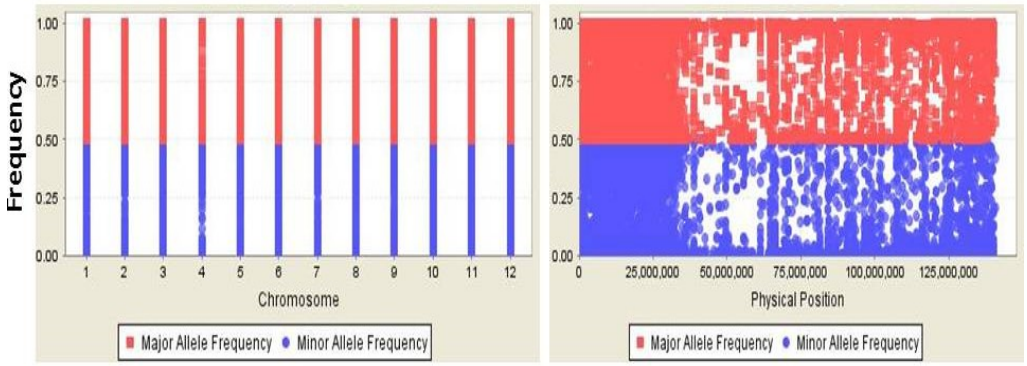
S Figure 2: Optimization of DNA extraction for quality and purity checks (a) Banana DNA extracted using two CTAB protocols (b) quality check using ApeK1 restriction digests.



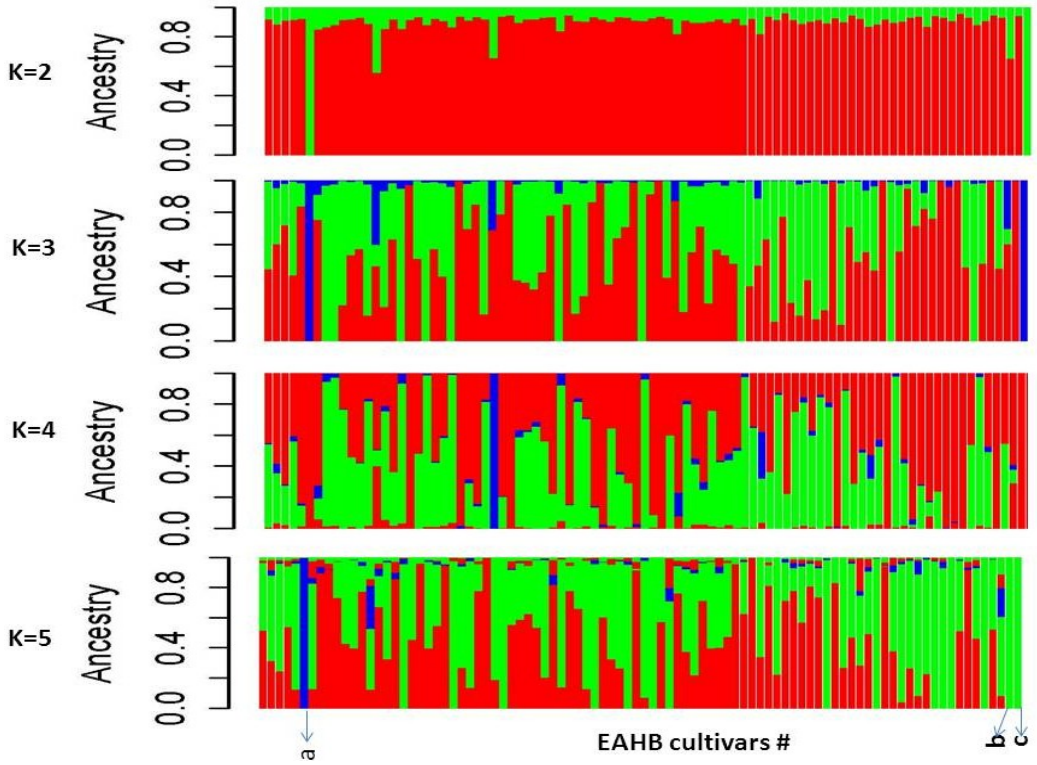
S Figure 3: EAHB GBS sequencing library run on the BioRadExperion. Optimization for complexity reduction using Pst1, ApeK1 and EcoT221



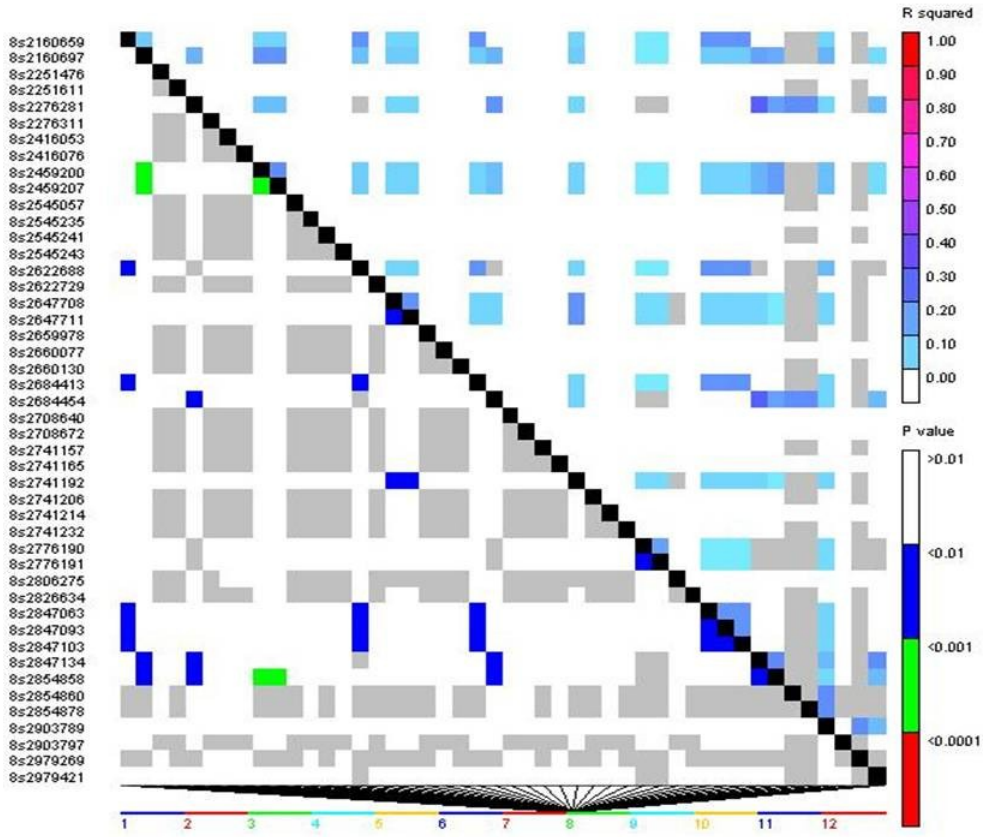
S Figure 4: Major and minor allele frequency in 89 EAHB. Distributions of SNP allele frequencies computed on the basis of chromosome from which they were scored (A) and (B) based on their physical positions on the genomes



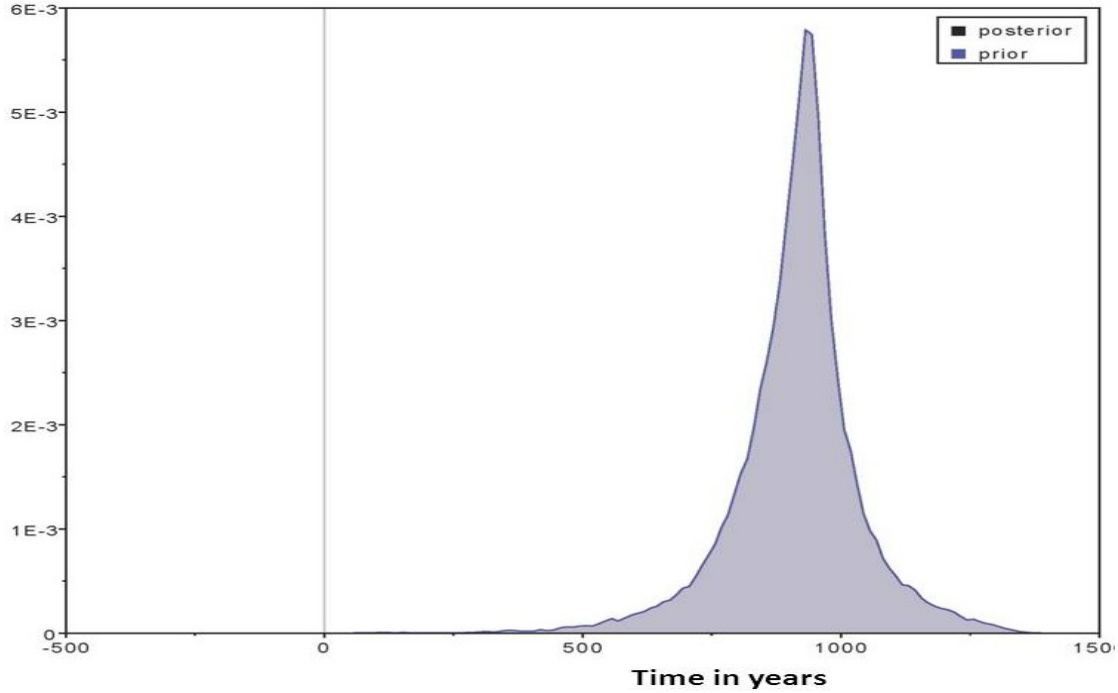
S Figure 5: STRUCTURE bar plots of genetic membership proportions (K=2 to K=5). Each cultivar is represented by a vertical line divided into K colors. Letters a, and c at the indicated at the bottom of the bar plots reperesent out-group cultivars, Calcutta-4 (AA), somatic-green (AAA desert) and Zebrina (AA) respectively.



S Figure 6: Linkage disequilibrium analysis. Showing lack of inter loci linkage in 5135 SNP pairs



S Figure 7: Marginal prior distributions (gray) versus marginal posterior distributions (dark gray) for the calibrated nodes from EAHB population



The philosophy of chapter 5 and 6 is reflected in the following statements made by Jablonka and Lamb (1995) and echoed by Tsaftaris and Polidoros (2000):

"Not all changes are the result of Darwinian selection of random variations created by the shuffling of genes and rare chance mutations. The nature of different types of heritable variation is now beginning to receive closer attention, and there is growing realization …that there are non-DNA sequence heritable variations that play a crucial part…".

CHAPTER 5

DNA METHYLATION ANALYSIS AMONGST GENETICALLY SIMILAR EAST AFRICAN HIGHLAND BANANA CLONES

Abstract

Background

We focus on the role of DNA methylation as an epigenetic mark that contributes to epiallelic diversity. Here, we estimate epigenetic diversity and differentiation among EAHB cultivars that are genetically similar and test the association of morphological groupings with genome-wide and locus-specific methylation states.

Materials and Results

Methylation sensitive amplified fragment length polymorphism (msAFLP) technique was used by simultaneous restricting the DNA of 90 EAHB and 6 outgroup cultivars with Hpa2 and Msp1 restriction enzymes. Following methylation-sensitive AFLP (MSAP), 670 bands, representing 315 (47.02%) methylation-susceptible and 355 non methylated epi-loci, were scored across 90 individuals. The number of polymorphic MSL and NML was 191 (61% of the total MSL) and 164 (46% of total NML) respectively showing extensive genomic methylation level polymorphism. The overall CG and CHG methylation levels were 24.17% and 17% respectively with extensive diversity, Shannons diversity index =0.49. However the methylated loci showed significantly high diversity (Shannons diversity index, 0.49) compared to the genetic loci (Shannon diversity index, 0.18). Genome-wide CG and CHG methylation levels were significant (P=0.001) within the morphological groups, however a non-significant among group difference was found in the CG and CHG methylation (P>0.99. The cultivars were were also significantly epigenetically (PhiST = 0.0114; P < 0.001) and genetically (PhiST = 0.0026; P < 0.001) differentiated. Highly significant (p<0.002) epigenetic variation that is structured into distinct within - (98%) and among (2%) population was observed. Epigenetic and genetic profiles displayed similar distributions in the co-inertia subspace and were significantly correlated based on Dice coefficients, r =0.89, P=0.0001 (permutations 10^4); ßST coefficients, r =0.90 p=0.0001; RV coefficient = 0.89, P=0.001 (Monte Carlo test. Our results indicate that while DNA methylation polymorphisms are common amongst EAHB cultivars, MSAP does not detect any obvious relationship between

DNA methylation variation and phenotypic variation in East African Highland bananas.

Conclusion

Results clearly indicate that the EAHB genome is highly methylated with extensive variation. Comparing the genetic loci vs the methylated loci, the latter is shows more polymorphism and differentiation which could be a way of epigenetics compensating for the lack of genetic diversity in EAHB. However the two profiles contribute equally to the co-interia indicating a genetic control of DNA methylation in EAHB. In Arabidopsis it has been found that epialleles that are genetically controlled do not result in phenotype change, this may explain why cultivars did not cluster on PCA and Neighbor joining tree based on their morphological groups. More experiments need to be done to ascertain how much DNA methylation is dependent on the underlying genome sequence before we can fully appreciate the extent to which natural epigenetic variation contributes to phenotypic variation.

Keywords: Methylation-sensitive amplified polymorphism, DNA methylation, East African highland banana (EAHB), phenotypic variation, Epigenetic

5.1 INTRODUCTION

The potential for, and rate of adaptive response to natural selection is determined by type and structure of the genetic variation underlying phenotypic traits. Abundant phenotypic diversity within a species is exhibited by many organisms (Candaele *et al.*, 2014). Much of the heritable variation within a species is a consequence of differences in the primary DNA sequence of different individuals. Genetic causes of phenotypic variance are attributable to many sources such as; mutations that create allelic variation and recombination that alters the genetic background in which alleles are expressed. However, there is a large part of this diversity that cannot be explained by genetic polymorphisms alone (Manolio *et al.*, 2009). More subtle sources of genetic variations that alter phenotypic variations also exist.

There is growing evidence that natural variation exists not only at the DNA sequence level, but also that heritable variation occurs in the absence of DNA sequence polymorphisms, termed epigenetic variation (Herrera & Bazaga, 2010; Becker *et al.*, 2011; Paszkowski, J. & Grossniklaus, U., 2011; Richards *et al.*, 2012) and has been proposed to be one component of this missing heritability, particularly common in plants. Several studies suggest that epigenetic variation alone can cause significant heritable variation in phenotypic traits (e.g., (Cubas *et al.*, 1999; Scoville *et al.*, 2011), play a crucial role in in mediating environmentally induced phenotypic variations and may be stably inherited by future generations (Guerrero-Bosagna & Skinner, 2009; Bossdorf *et al.*, 2010; Lira-Medeiros *et al.*, 2010; Paszkowski, Jerzy & Grossniklaus, Ueli, 2011; Liu & Feng, 2012). Currently, there is a growing appreciation that epigenetic variation, resulting from a multitude of diverse chemical modifications to the DNA and chromatin, can have profound effects on phenotype.

Therefore, heritable epigenetic variation as well as genetic variation has the potential to underpin natural variation (Fujimoto *et al.*, 2012). Epigenetic variation includes all those mechanisms often associated with a variety of chromatin marks that give rise to differential gene expression in specialized cells; these may include methylation of cytosine in genomic DNA, or chromatin configurations, or combinations of the two (Jablonka & Raz, 2009). Epigenetic marks can contribute to altered gene expression states which could underlie the phenotypic differences seen in genotypes considered to be largely isogenic (Bird, 2007).

DNA methylation is the addition of a methyl group to the 5-carbon of cytosine and perhaps the best studied of epigenetic phenomena. The DNA methylation process in eukaryotic cells is carried out by a family of DNA methyltransferase enzymes, which transfers methyl groups from the methyl donor S-Adenosyl methionine (SAM) to the cytosine. This results in a 5-methyl cytosine (^{5m}C) which is often repressive and can be associated with gene silencing (Cedar & Bergman, 2012) by inhibiting the binding of DNA by transcription factors (Watt & Molloy, 1988) or by recruiting additional chromatin proteins to form heterochromatic state that is inaccessible for transcription (Cedar & Bergman, 2012; Zhang & Hsieh, 2013).

DNA methylation in mammals, occurs almost exclusively in the symmetric CG context and is estimated to occur at ~70–80% of CG dinucleotides throughout the genome (Ehrlich *et al.*, 1982) although, non-CG methylation is observed in embryonic stem (ES) cells in small amount (Lister *et al.*, 2009). In plants, methylation of DNA occurs in the symmetric CG, CHG and asymmetric CHH contexts (H = A, C, or T) (Cokus *et al.*, 2008; Osabe *et al.*, 2014), predominantly on transposons and other repetitive DNA elements (He *et al.*, 2011). In Arabidopsis, genome wide DNA methylation levels of approximately 24%, 6.7% and 1.7% are observed for CG, CHG, and CHH contexts, respectively (Cokus *et al.*, 2008; Law & Jacobsen, 2010). DNA methylation in the CG context is carried out by METHYLTRANSFERASE1 (MET1) which

is a homologue of the mammalian DNA METHYLTRANSFERASE1 (DNMT1) while CHG methylation in maintained by the plant-specific DNA methyltransferase CMT3 whose chromodomain recognizes and binds to H3K9me2 marks (Du *et al.*, 2012). RNA-directed DNA methylation (RdDM) is responsible for the maintenance of CHH methylation as well as de novo methylation the other sequence context.

The functional role of DNA methylation is poorly understood. However, genetic changes such as transposon insertions can also lead to changes in DNA methylation (Eichten *et al.*, 2013). DNA methylation in animals and plants may be involved in regulation of diverse biological processes including cell differentiation, X-chromosome inactivation, transposon and gene silencing, and genomic imprinting (Law & Jacobsen, 2010). It has also been known to strongly influence chromatin structure and gene expression and is frequently associated with epigenetic regulation in plants and mammals (Saze *et al.*, 2012). There are far-reaching implications of epigenetic research for agriculture and for plant breeding, particularly as some epigenetic marks in plants are clearly stably heritable over many generations (centuries) (Cubas *et al.*, 1999).

5.1.1 Effects of epigenetic variation on plant phenotypes

The levels and patterns of DNA methylation are highly variable in animals, ranging from no detectable 5mC in the nematode *Caenorhabditiselegans* and limited, developmentally restricted methylation in *Drosophila melanogaster* to widespread genomic methylation in vertebrates (Rabinowicz *et al.*, 2005). Conversely, DNA methylation seems to be ubiquitous among plants. CG and non-CG methylation in plants, can silence transposons and pseudogenes, and regulate plant development and tissue specific gene expression (Schob & Grossniklaus, 2006). The first important step to understanding the significance of epigenetic effects is characterizing the phenotypic response to epigenetic

variation. However, there are few known simple and obvious phenotypic effects that result from changes in epigenetic marks at single genes. Researchers have made substantial progress in demonstrating important phenotypic effects that result from changes at only the epigenetic level, occurring naturally (Table 17) (Richards *et al.*, 2012) and through manipulation of methylation levels and isolation of methylation mutants. Other epigenetic related phenotype plasticity are environmentally related e.g phenotypic plasticity leaves of *Ilex aquifolium* (Aquifoliaceae) trees (Herrera & Bazaga, 2013).

Table 17: Examples of naturally occurring epigenetic modifications causing phenotypic changes in plants (Zhang & Hsieh, 2013).

Target	Epigenetic effect	Species	Phenotype	References
Lcyc	DNA methylation and silencing of Lcyc	*L. vulgaris*	Change in floral symmetry	Cubas *et al.* (1999)
CNR	DNA methylation and silencing of CNR	*S. lycopersicum*	Fruit ripening defect	Manning *et al.* (2006)
OsSPL14	OsSPL14 Promoter hypomethylation	*O. sativa*	Panicle branching and higher grain yield	Miura *et al.* (2010)
DWARF1 (D1)	DNA methylation and silencing of D1	*O sativa*	Dwarf	Miura *et al.* (2009)
OsFIE1	OsFIE1 Promoter hypomethylation and ectopic OsFIE1expression	*O. sativa*	Dwarf	Zhang *et al*. 2012

5.1.2 Cytosine methylation levels in plant

Plants, show a progressive DNA methylation trend from cotyledons to vegetative organs to reproductive organs (Ruiz-Garcia *et al.*, 2005). The level of cytosine methylation is variable in plants, from 6% of cytosines in *Arabidopsis* to 25% in maize (Rabinowicz *et al.*, 2005). Numerous epigenetic related studies have been done on the model species, *Arabidopsis thaliana* (Richards, 2011). Less is known about the levels and importance of epigenetic processes in natural populations. Studies on the differential cytosine methylation levels and polymorphism in agricultural crops are beginning and several have been reported; cotton (Osabe *et al.,* 2014), rice (Xiong *et al.*, 1999), maize (Lu *et al.*, 2008; Candaele *et al.*, 2014), sorghum (Zhang *et al.*,

2011), plantains (Noyer *et al.*, 2005) and tobacco (Zhao *et al.*, 2011); *Acacia mangium* (Baurrens *et al.*, 2004) and grapes (*Vitis vinifera L.*) (Schellenbaum *et al.*, 2008).

The possibility of non-genetic inherited effects on phenotype has excited great interest among both evolutionary biologists and plant breeders. Because of these observations, there is currently increasing interest in understanding the role of epigenetic processes in plant breeding. Recently, researchers demonstrated that, in addition to genetic processes, epigenetic processes may play an important role in causing inbreeding effects (Vergeer *et al.*, 2012; Cheptou & Donohue, 2013). It was shown that epigenetic variation is affected by inbreeding and that epigenetic variation can be modified in such a way that negative effects of inbreeding largely disappear. These results are of great interest to plant breeders because of the high economic costs involved in dealing with inbreeding depression in breeding programmes.

To date, analyses of natural epigenetic variation have either used high-resolution genetic information to understand variation in specific traits (usually on model species without explicit links to populations or environments), or using low-resolution genetic information such as MS-AFLPs to address population-level questions. In plants, analysis of cytosine methylation has been approached by studying either global levels of methylated cytosines (Cervera *et al.*, 2002) or by examining specific gene sequences (Soppe *et al.*, 2000; Riddle & Richards, 2002) using either bisulfite treatment (Frommer *et al.*, 1992; Sadri & Hornsby, 1996; Xiong & Laird, 1997) or restriction enzyme isochizomers that differ in their sensitivity to methylation of their recognition sequences (Vongs *et al.*, 1993).

*Hpa*2 and *Msp*1 are isoschizomers that that show differential sensitivity to cytosine methylation and are frequently used to detect cytosine methylation and both recognize the tetranucleotide sequence 5`-CCGG-3`. *Hpa*2 is inactive if one or both cytosines are fully methylated (both strands methylated) but cleave the hemimethylated sequence (only one DNA strand methylated)

whereas *Msp*1 cleaves $C^{5m}CGG$ but not ^{5m} CCGG sequences (McClelland *et al.*, 1994). The AFLP (Amplified Fragment Length Polymorphism) technique has been widely adapted for the analysis of cytosine methylation in plants (Xiong *et al.*, 1999; Cervera *et al.*, 2002; Lu *et al.*, 2008; Schellenbaum *et al.*, 2008; Gao *et al.*, 2010; Zhao *et al.*, 2011; Osabe *et al.*, 2014) and animals, e.g great ground leaf bat (Liu *et al.*, 2012), birds (Schrey *et al.*, 2012; Liebl *et al.*, 2013) and fish (Blouin *et al.*, 2010; Moran & Perez-Figueroa, 2011). This technique is suited for non-model species with little genomic information and provides rapid epigenetic fingerprints for a large number of samples (Liu *et al.*, 2012).

While in sexually reproducing organisms each individual possesses a different genotype, asexually reproducing individuals from the same clonal lineage are presumed to be genetically identical, like the case of the East African Highland bananas. Even though the full extent to which epigenetic variation contributes to phenotypic variation remains to be determined (Richards, 2011). There is a potential for epigenetics to play a role in crop improvement, including regulation of transgene expression and creation of novel epialleles. Here, we; (i) assess the level of naturally occurring DNA methylation (epigenetic diversity) and provide an estimation of methylation status of EAHB population; (ii) determine DNA methylation variation between cultivars and among EAHB morphological groups; (iii) investigate epigenetic structure and relationships of the EAHB, (iv) compare if epigenetic diversity and population structure mirror genetic diversity, and; (v) assess if epigenetic variation is caused by neutral drift.

5.2 MATERIALS AND METHODS

5.2.1 MSAP technique and DNA amplification

MSAP analysis was performed for each genotype (Vos, P *et al.*, 1995), with modifications. Modifications made were: 500 ng of template DNA, double-digest using combinations of *EcoR*I with *Hpa*II (New England Biolabs, Arundel, Australia) or *Msp*I (New England Biolabs, Gold Coast, Australia), fluorescently labeled reactions (FAM and NED) were mixed, and peaks were separated using an ABI3730XL capillary sequencer (Applied Biosystems, Melbourne, Australia). Adaptors and oligonucleotides used are listed in Table 1. The pre-amplification and selective amplification cycling conditions were performed as manufacturer's instruction with 40 cycles for the selective amplification. A subset of 12 cultivars, two each of the EAHB morphological groups and two out-group cultivars in three replicates were used for optimization of scoring procedure. Scoring of each CCGG site was automated to assess the presence (''1'') or absence (''0'') of peaks using GeneMapper 4.1. Apanel for each oligonucleotide pair was constructed based on present peaks in the 12 genotypes and was manually refined by selecting the peaks that were strong and consistent in at least two of the three replicates. This panel was applied to the subset genotype samples to produce binary data for epigenetic analysis. Fragment analysis and scoring of peaks for the entire set of 96 DNA samples was done as optimized in the subset. To reduce the potential impact of size homoplasy, only unambiguous and intense bands, ranging in size from 150 to 500 bp, were scored (Caballero & Quesada, 2010; Liu *et al.*, 2012). A total of 30 oligonucleotide pairs were screened for selective amplification specificity and multiple alleles, but only 6 primer pairs (Table 19) produced clear and unambiguous bands that could be scored reliably across the 96 genotypes and were considered for further analysis.

5.2.2 Data analysis

Methylation Sensitive Polymorphism (MSP) binary matrix was scored as 1 (?/- or -/?), 0 (?/?), and missing (-/-). The other non-methylated loci, which were identical in their presence/absence in the *EcoR*I/ *Hpa*II and *EcoR*I/*Msp*I patterns, were considered as CCGG-genetic markers; a site was considered as ‘‘methylation- insensitive polymorphism’’ (MIP) band if the fragments were absent in both enzymes in at least one individual. The MIP binary matrix was scored as 1 (?/?) and 0 (-/-) (Table 18). To evaluate methylation status of the EAHB genome, each loci was classified as either ’methylation-susceptible locus’ (MSL) or ’non-methylated locus’ (NML) at an error rate of 0.05 per primer combinations using R (msap) package (Pérez-Figueroa, 2013).

The polymorphism ratio within each *EcoR*I/ *Hpa*2/*Msp*1 was determined by calculating the total number of polymorphic sites within the genotypes divided by the total number of sites analyzed. The percentage polymorphism of the methylation sensitive enzyme was calculated by total number of DNA methylation polymorphic sites identified, divided by the total number of sites analyzed. The calculated DNA methylation polymorphic sites do not exclude sites that are polymorphic both genetically and by DNA methylation. The DNA methylation level was quantified for each genotype using the MSAP binary data by using the sum of CG and CHG methylated bands divided by the total number of scored bands.

The percentage polymorphism of the methylation sensitive enzyme was calculated by total number of DNA methylation polymorphic sites identified, divided by the total number of sites analyzed. The calculated DNA methylation polymorphic sites do not exclude sites that are polymorphic both genetically and by DNA methylation. GenALEx v6.5 (Peakall & Smouse, 2012) was used to calculate mean proportion of polymorphic loci, mean number of observed alleles (A) (Kimura & Crow, 1964) (Na), number of effective alleles (Ne), Nei`s biased and unbiased genetic diversity (h) (Nei, 1983), number of private bands (alleles/bands unique to a single population). To estimate overall

epigenetic diversity (Ht) was calculated. Significance of these analyses was determined by the Mean±SE and significance test at 95% confidence level of significance using unpaired *t* - test with Welch's correction performed by GraphPad Prism version 3.0. We calculated the Shannon diversity index based on the frequency of each band among the 90 individuals to estimate the overall population epigenetic diversity.

5.2.2.1 Methylation variation within and between groups

The four patterns of amplification products from *EcoR*I/*Hpa*II and *EcoR*I/*Msp*I were produced and compared among the morphological groups (termed as populations in this study) (Table 18): (a) non methylation (Hpa^{+}/Msp^{+}); (b) Hemimethylated or CHG methylation (Hpa^{+}/Msp^{-}); (c) internal cytosine methylation ($Hpa11^{-}/Msp1^{+}$) or CG methylation; (d) full methylation or absence of target (Hpa^{-}/Msp^{-}) (uninformative bands, as this could be caused by genetic mutation or hypermethylation). MSAP fragments that differ in their presence/ absence in the EcoRI/HpaII and EcoRI/MspI patterns in at least one individual were considered as methylated bands. The Kruskal–Wallis H test to estimate the significance of the difference in the CG and CHG methylation patterns among populations.

Table 18: *Hpa*II and *Msp*I sensitivities to 5`-CCGG-3` methylation status from REBASE specifications

HpaII	MspI	Type	Notes
+	+	Non methylated	
+	-	CHG Methylated	Hemimethylated
-	+	CG Methylated	Internal cytosine methylation
-	-	Uninformative	Fragment absence or hypermethylated

+ and - represent the presence or absence of a fragment respectively

To estimate the within-population epigenetic diversity, Shannon diversity index (H_{pop}) was calculated based on the frequency of each band among the 90 individuals. Significant differences of Shannon index among populations were assessed by the Kruskal–Wallis H test, and the significance of the test was adjusted by the sequential Bonferroni correction (Rice, 1989). Pairwise coefficient of epigenetic differentiation among the groups was computed as Phi-ST.

We calculated population differentiation based on MSL and NML and Kruskal-wallis h test used to estimate the significance of the difference in the CH and CHG methylation patterns among populations. We used multivariate analyses to explore between population epigenetic structure. Principal component analysis (PCA) on inter-profile covariance matrix based on MSP and MIP binary profiles were performed to provide a genome-wide variability point of view, summarized in a few synthetic variables (Liu *et al.*, 2012). The between Eigen analysis (BPCA-PCA among groups based on PCA among individuals, (Parisod & Christin, 2008) (Parisod & Bonvin, 2008) was processed to group PCA profiles into populations maximizing the between group variance. Statistical significance was assessed by the Romesburg randomization test (10^4 permutations). Multivariate analyses were performed by ADE-4 software (Thioulouse *et al.*, 1997). Multivariate analyses were performed by ade4's dudi.pco and s.class to obtain a PCA.

Between-group Eigen analysis divides the variance into within- and between population components and it is based on Euclidean distances and can be considered as analogous to F statistics (called βST) (Lira-Medeiros *et al.*, 2010; Liu *et al.*, 2012). Phi-ST value is equal to the ratio of inertia of the between PCA to the total inertia. However, βST is not equivalent to F statistics and may be overestimated because BPCA maximises the between-group variance (Parisod & Christin, 2008; Liu *et al.*, 2012).

Symmetrical co-inertia analysis was used to further explore the contribution of both epigenetic and genetic profiles to the EAHB population structure. Symetrical co-inertia maximizes shared structures among multiple datasets drawn from the same samples (Lira-Medeiros *et al.*, 2010) and can be safely used for multivariable that are related because it does not rely on linear regressions (Thioulouse *et al.*, 1997). Statistical significance was assessed by 10^4 Monte Carlo permutations in the ADE-4 software (Thioulouse *et al.*, 1997). We used Mantel's tests (Mantel, 1967) to assess relationship of epigenetic (MSP) and genetic (MIP) profiles by two main indexes (10^4 permutations) (i) we compared Dice coefficients (Sneath & RR, 1973) which were calculated independently from both MSP and MIP profiles in NTSYS 2.0 Software according to (Cervera *et al.*, 2002) (ii) we compared pairwise βST (BPCA) values of both profiles (Parisod & Christin, 2008; Lira-Medeiros *et al.*, 2010), calculated by the ADE-4 software (Thioulouse *et al.*, 1997).

5.2.2.2 Epi-genetic genotype similarities, relationships and population structure

To assess and visualize the similarities or dissimilarities of the EAHB based on DNA Methylation based genetic distance (GD) and PCA analysis were done. Pairwise (epi)-genetic distance matrix was generated using share allele genetic distance implemented in Powermarker v3.25. Principal component analysis (PCA) was performed in Darwin v5.0. Dissimilarity matrix was generated

using modalities dissimilarity coefficient (Sokal & Michener, 1958) which is equivalent to the Simple matching coefficient. Simple matching method considers the double-absence of peaks as additional information in a pair-wise comparison for closely related species (Osabe *et al.*, 2014). This coefficient is therefore appropriate for assessing the EAHB genotypes as these genotypes are expected to have low heterozygosity and the presence/absence of the bands are likely due to homology rather than homoplasy (comigrating DNA fragments from different ancestral origin). For closely related species, use of Jaccard is recommended when it is not known whether the double-absence of peaks in pair-wise comparison are due to DNA sequence polymorphism or homoplasy (Laurentin, 2009). However, in the previous chapters, we have shown that the methylation insensitive (*EcoR*I/*Mse*1) data had bands present across most genotypes and showed that the absence of peaks were not due to sequence polymorphism, and the potential contribution of homoplasy is expected to be very small. Therefore, absence of bands in pair-wise comparison of genotypes in both *EcoR*I/*Hpa*II and *EcoR*I/*Msp*I data indicates that this region is likely to share the same methylation state.

Relationships of the cultivars were visualized using a Neighbor joining tree (unweighted pair group method with arithmetic mean UPGMA) constructed using the simple matching dissimilarity coefficient and significance evaluated using 10^3 bootstraps. STRUCTURE v2.3.4 was used to assess population structure was and ancestry of the cultivars based on DNA methylation.

Structure program approach was used to assign individuals to (epi)-populations and identify migrants and admixed individuals using MSAP multilocus genotype data independent of prior population information. The inherent (epi)-genetic structure of EAHB population was assessed directly using a method developed by Pritchard *et al.* (2000) and implemented in the program STRUCTURE, that implements a model-based clustering method to infer population structure. The approach assumes a model in which there are *K* populations (where *K* may be unknown), each of which is characterized by a

set of allele frequencies at each locus (Pritchard *et al.*, 2000). Individuals in the sample are assigned probabilistically to populations or jointly to two or more populations if their genotypes indicate them to be admixed. Structure analysis was run with K=10, each with three replicates assuming the admixture model with correlated allele frequencies with 200,000 iterations for the length of burn-in period and subsequent number of MCMC (Markov-Chain-Monte-Carlo) repeats with lambda set at the program default of 1.0 for exploratory analyses without a priori information about individual origin. The optimal level of K was determined using an ad hoc statistic K based on the rate of change in the log probability of data between successive K values using CLUMP software (Evanno *et al.*, 2005; He *et al.*, 2011). This was adopted due to its sensitivity compared to the LnP(K) method to detect the number of subpopulations and in circumstances where the K value does not reach a clear plateau (Evanno *et al.*, 2005). PCA, cluster analysis and Structure were used to show whether patterns of methylation polymorphisms mirrored phenotypic groups' population structure and relationships

The overall coefficient of epigenetic differentiation (PhiPT) was computed in GenAlex v6.5 as $PhiPT = (AP/(WP + AP) = AP/TOT$, here AP= estimated variation Among populations, WP =estimated Variation Within populations.Number of migrants (Nm) for haploid data was estimated as $Nm = [(1/PHIPT) - 1]/2$. To test significance of the structure of genetic diversity, the AMOVA was carried out. This is a population genetics statistical tool (Excoffier *et al.*, 2009) based on the analysis of variance principle. AMOVA analysis was applied to the pairwise distance matrix to partition the sources of the observed variation by component parts.

5.3 RESULTS

5.3.1 Relative Genomic Methylation Levels in EAHBs

MSAP assays investigated the context of DNA methylation in EAHB and out-group cultivars using the methylation sensitive isoschizomers *Hpa*II and *Msp*I, allowing discrimination between CG and CHG methylation. Following methylation-sensitive AFLP (MSAP), 724 bands, representing 251(34.70%) methylation-susceptible and 473(65.3%) non methylated epiloci, were scored across 96 individuals (90 EAHB and 6 out-group cultivars). Of the total fragments scored, a high percent, 75.96%, of polymorphism was observed. Methylation was divided into two categories, scored as CG, CHG methylation according to MSAP data (Table S3). All primers pairs showed a high level of polymorphism and demonstrated significant differences in % of CH and CHG methylated fragments within- and between the primers pairs (Table 19).

To characterize variation in methylation profiles in the EAHB cultivars (without out-group cultivars), we analyzed genomic DNA methylation patterns of 90 East African Highland Banana cultivars from five morphological populations using MSAP. We scored 670 loci using six selective primer combinations with an average of 111.7 loci per primer combination. By comparing the presence/absence of restricted fragments in Hpa11 and Msp1 assays for each individual, 355 loci (52.98% of the total loci scored) showed similar digestibility, but varied among morphological groups (referred to as populations), these fragments were termed as No Methylated Loci (NML). The other 315 loci (47.02% of total loci scored) showed differential digestibility suggesting methylation and was termed as Methylation-Susceptible Loci (MSL). Number of polymorphic MSL and NML was 191 (61% of the total MSL) and 164 (46% of total NML) respectively.

Table 19: Selective primer pair polymorphism. Number of fragments, % of polymorphic fragments, % of CHG and CG methylated fragments observed in Oligo pairs used in Epi-diversity study of the EAHB

Oligo pair	No. of fragments	% polymorphic	% CHG methylated	% CG methylated
EAAC_Hpa2AAT	153	96.73	7.54	22.29
EAGA_Hpa2AAC	76	98.68	26.44	16.91
EACT_Hpa2ATG	142	89.44	8.51	36.12
EATC_Hpa2AGA	110	97.27	23.73	14.30
EACA_Hpa2AGC	93	84.95	6.19	18.45
EAGT_Hpa2ATC	150	92.00	4.26	25.42

5.3.2 Diversity of Genome Methylation in 90 EAHB cultivars

The cultivars showed a low percent of CHG methylated fragments compared to CG, on average 83 (11.46%, range 58 to 122 loci in Ekeganda and Oruhuna) of the total loci scored in cultivars was CHG methylated. Conversely, CG methylation was modest 170 (23.48%) ranging between 138 to 198 loci in Liganda-Lusumba and Bitambi, respectively (Figure 33). The mean percentage of polymorphic loci at the cultivar level was 57.68%.

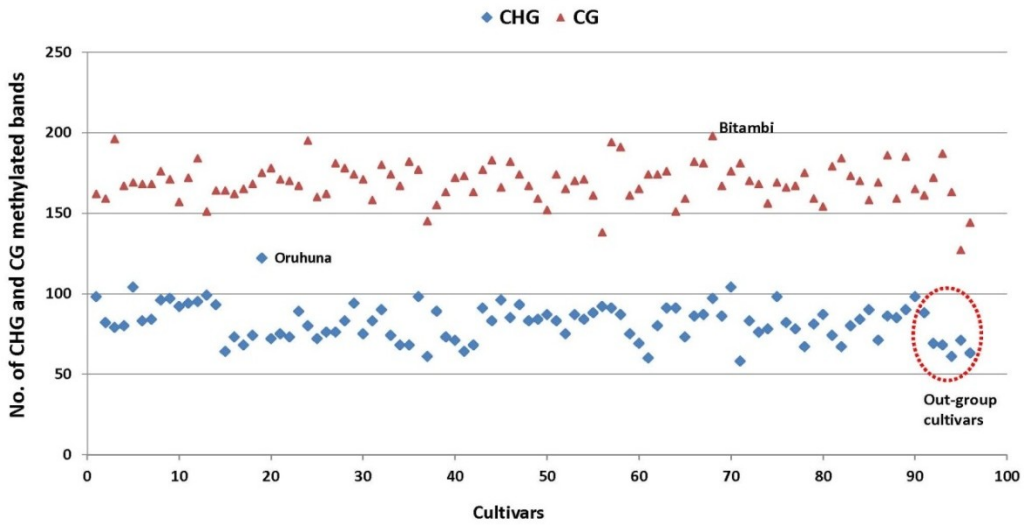


Figure 33: Genome wide differential cytosine methylation levels (CG and CHG) of CCGG sites in 90 EAHB-triploid cultivars and six outgroup cultivars. Methylation level (number of fragment) was calculated by counting MSAP bands representing methylated 5`-CCGG sites (differential presence/absence of restricted fragments in HpaII and MspI assays)

High epigenetic diversity was observed among the cultivars, Shannon's diversity index was 0.437; Number of different alleles (Na), 1.89±0.02, Number of effective alleles (Ne), 1.38±0.009 and observed Nei`s genetic diversity (h), 0.21±0.04. The pairwise GD values revealed moderate level of epigenetic divergence among the EAHB cultivars. Average of Nei's Genetic distances (GD) between cultivars was 0.3391, ranging between 0.1554 (Red Nakitembe and Luvuta) and 0.5649 (Mbululu NAK and Mbwazirume) (Figure 34).

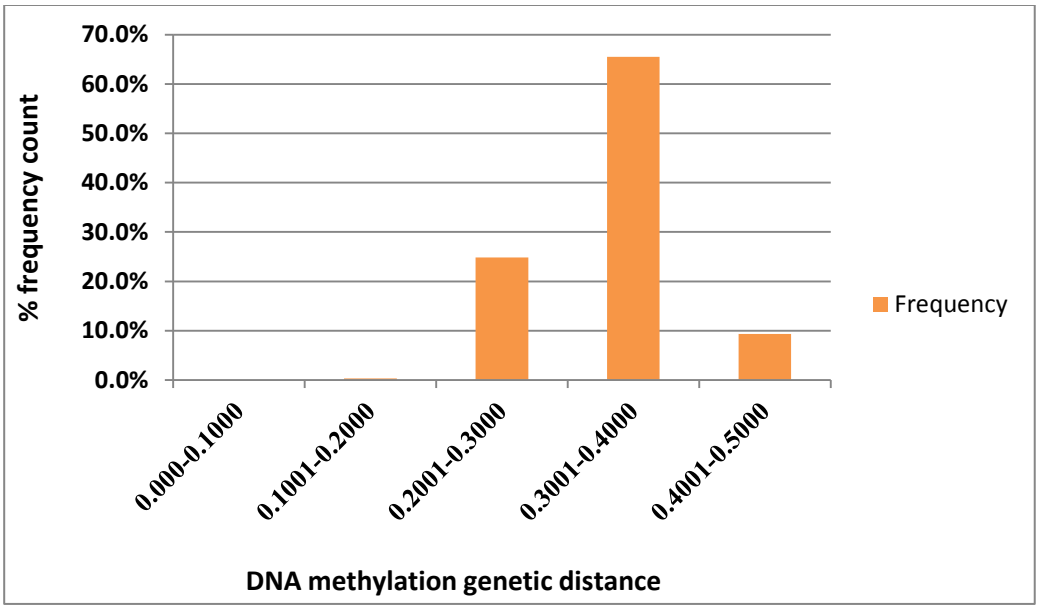


Figure 34: Frequency counts of pairwise (epi)-genetic distance observed in 90 EAHB cultivars from two regions. EAHB cultivars have a higher epigenetic distance showing they are epigenetically different.

5.3.3 Variation in methylation diversity within and between groups

The numbers of various fragments attributed to non-methylated (+/+), CHG methylated (+/-), CH methylated (-/+), and uninformative (-/-) respectively were calculated for each cloneset based on MSAP profiles. The total 5′-CCGG-methylation level among the groups ranged from 40.36 % in Musakala to 43.75 % in Mbidde, with a mean of 42.21 % (Figure 35). Difference in the genome wide methylation level (CHG and CG methylation) was significant (Kruskal-Wallis x^2= 1022.07, *df* = 4, P=0.001); however, a non-significant among group difference was found in the level of CG (Kruskal-wallisx^2 = 3751.75, *df* = 4, P=>0.99) and CHG (Kruskal-wallisx^2 =267.36, *df* =4, p=>0.99) methylation patterns. CG methylation was higher than CHG methylation patterns.

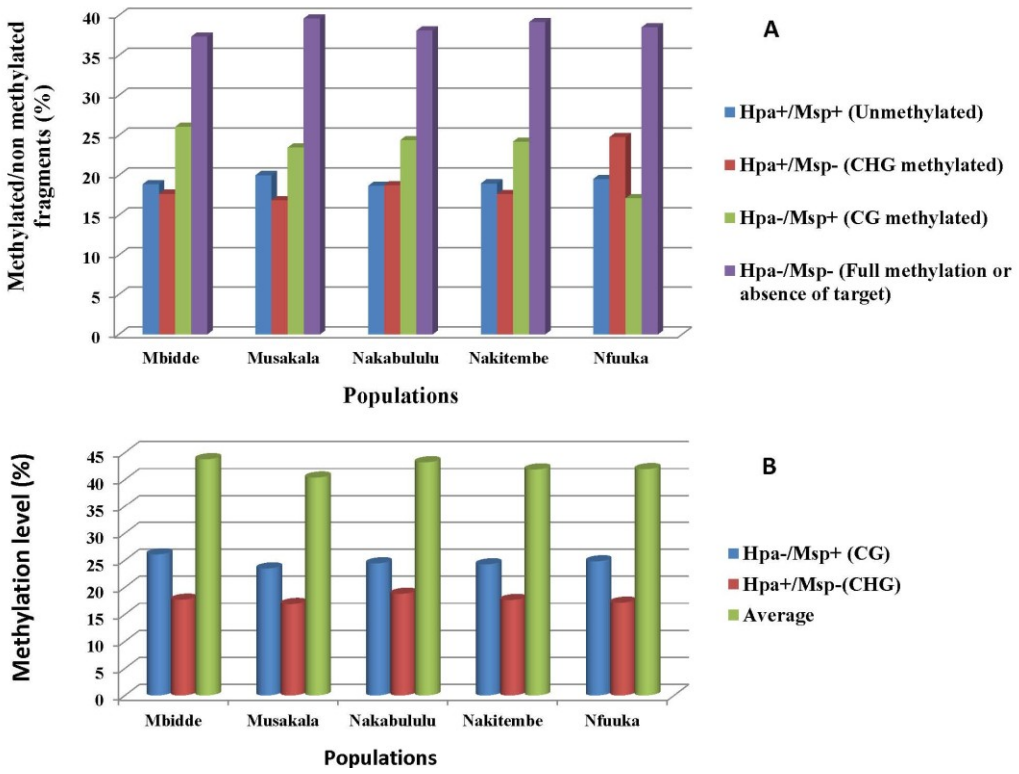


Figure 35: Relative methylation/non-methylation levels in five EAHB morphological groups (A) CHG, CG, and full methylation and non-methylation and (B) methylation level of each group. Significant differences between relative CHG and CG methylation levels within each population was examined using a Wilcoxon rank sum test with P-values of P = 0.001 (Mbidde), P = 0.018 (Musakala), P = 0.012 (Nakabululu), P = 0.001 (Nfuuka), P = 0.007. Also, significant differences between relative total methylation and non-methylation levels within each population was also examined using the Wilcoxon rank sum test, P values in all the populations were significant, P<0001.

The within-population epigenetic Shannon indices calculated for the five populations were high between 0.039 and (Mbidde) and 0.528 (Nfuuka) (Table 20) and were significantly different (Kruskal-Wallis x^2 = 70.982, df = 4, P<0.001) indicating high epigenetic variation. However, no significant differences (students t-test, P>0.008) were found in other calculated diversity indices.

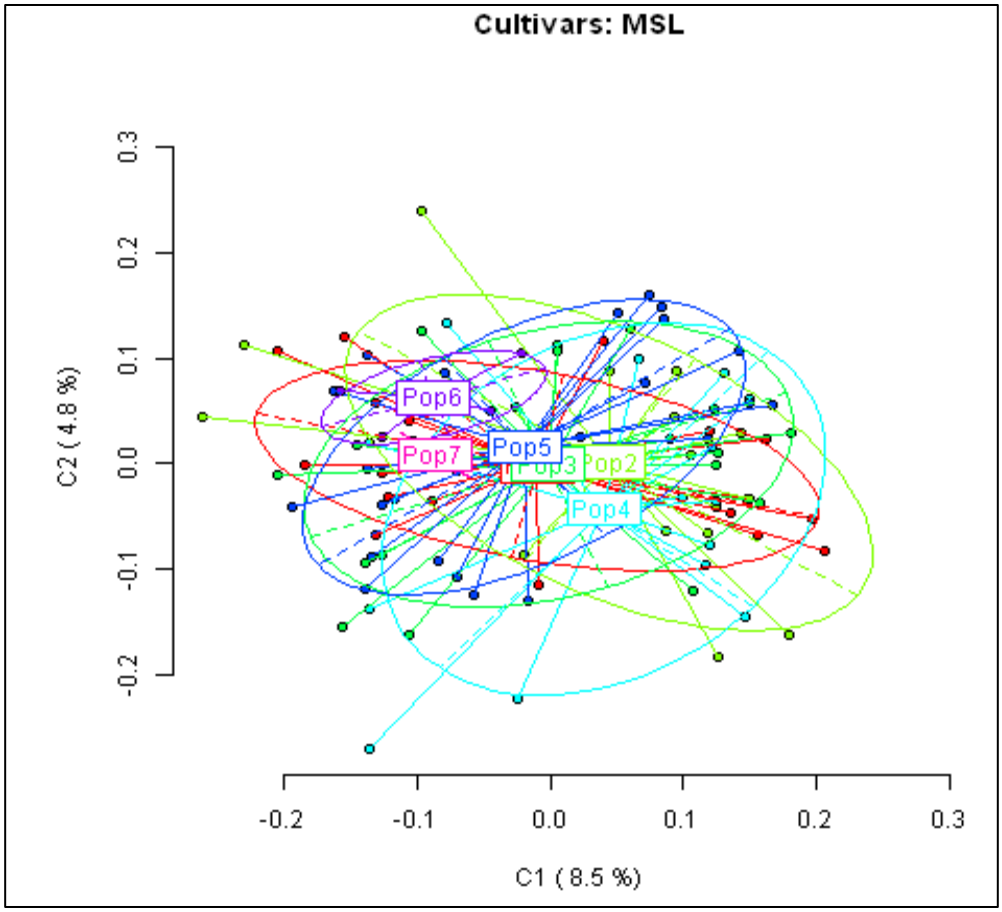
Table 20: Indices ± standard error calculated to estimate epigenetic diversity within the EAHB groups

Cloneset	Mbidde	Musakala	Nakabululu	Nakitembe	Nfuuka
Na	1.667±0.04	2.000±0.04	2.000±0.04	2.333±0.04	2.667±0.04
Ne	1.520±0.02	1.456±0.02	1.769±0.02	1.771±0.02	1.768±0.02
I	0.339±0.01	0.359±0.01	0.425±0.01	0.499±0.01	0.528±0.01
Private alleles	0.000±0.01	0.000±0.01	0.000±0.01	0.333±0.02	0.667±0.01
h	0.203±0.01	0.193±0.01	0.233±0.01	0.272±0.01	0.280±0.01
uh	0.214±0.01	0.206±0.01	0.246±0.01	0.293±0.01	0.292±0.01

The epigenetic diversity was assessed by Shannon's diversity index: I=– g Pi log2 (Pi). The difference of the index among morphological populations was tested using a Kruskal–Wallis H test with the chi-square value = 1039.017 (P, 0.001) that was adjusted by the sequential Bonferroni correction.

The among-population epigenetic Shannon's indices calculated using MSL for the five populations was high (I=0.49, SD 0.16) compared to NML Shannon's index, (I=0.18, SD 0.07), implying higher variation at the MSL, these values were highly significant (Wilcoxon rank sum test with continuity correction, W = 29895.5, P < 0.0001). The overall Shannon's Index (using combined MSL and NML) observed among the populations was 0.33±0.01.

Between-population analyses (BPCA) plot based on covariance matrix of the methylation profile showed no obvious separation of the five EAHB morphological groups and the out-groups but the groups were intertwined (Figure 36; cultivar MSL) and the first two axes summarized 13.3% of the total inertia. However in the NML-PCA, the out-group populations were distinctly clustered (first two Eigen summarized 17.6% of the total inertia) (Figure 36 B). The morphological groups showed very low epigenetic and genetic differentiation, although, epigenetic differentiation was higher compared to latter (Table 21).


Cultivars: MSL
C2 (4.8 %)
C1 (8.5 %)
-0.2
-0.1
0.0
0.1
0.2
0.3
Pop6
Pop7
Pop5
Pop2
Pop4

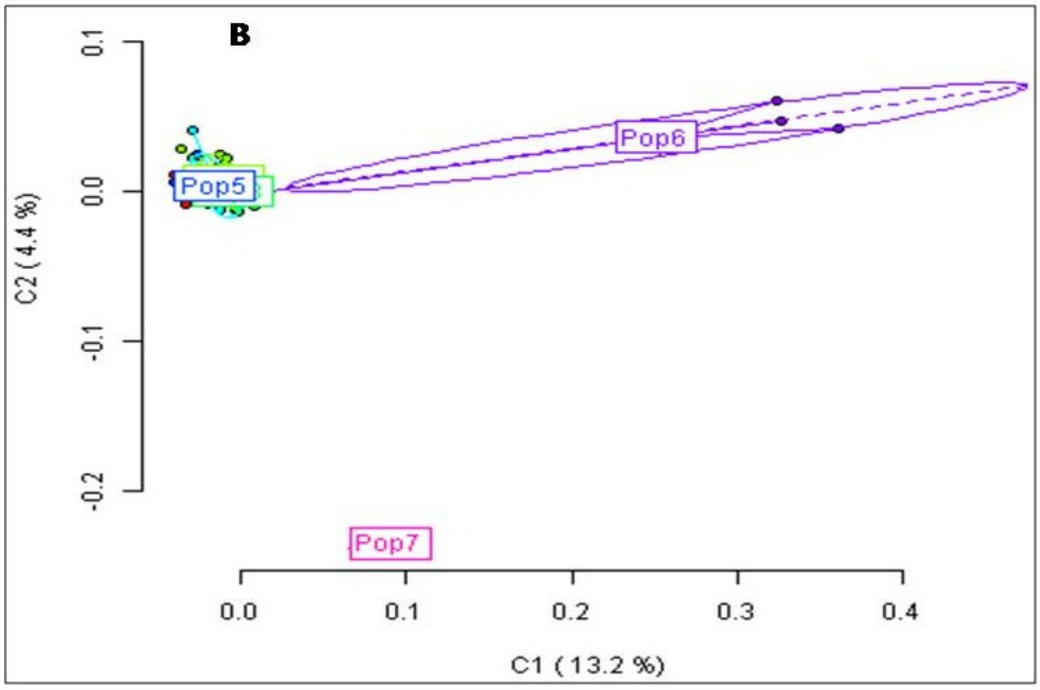


Figure 36: Between-group Eigen analysis (BPCA). Cultivars: MSL represent PCA analysis of the five EAHB morphological groups (Pop 1-5) and out-groups (pop 6 and 7) based on the epigenetic covariance matrix (MSP). (B) PCA of the EAHB and out-groups based on the genetic covariance matrix (NML). C1 and C2 values show the contribution of the two principal components summarizing the total variance of each data set. The labels pop from 1 to 5 represent the five populations: Mbidde, Musakala, Nakabululu, Nakitembe and Nfuuka and pop 6 and 7 are AAA-dessert bananas and AAB-plantains respectively.

Table 21: Within epigenetic and genetic differentiation of the EAHB morphological groups. Pairwise Phi-ST values of MSL and NML between the EAHB morphological groups

Cloneset A	Cloneset B	MSL	NML
Nakabululu	Nfuuka	0.004	-0.001
Musakala	Nakitembe	0.006	0.001
Nakitembe	Nfuuka	0.007	0.006
Mbidde	Nfuuka	0.007	0.004
Musakala	Nfuuka	0.010	0.004
Mbidde	Nakabululu	0.010	0.003
Nakabululu	Nakitembe	0.014	0.006
Musakala	Nakabululu	0.018	0.000
Mbidde	Musakala	0.019	0.003
Mbidde	Nakitembe	0.019	0.000

5.3.4 Correlation between epigenetic and genetic profiles

Co-inertia analysis was used to evaluate the contribution of both genetic and epigenetic profiles to the EAHB population structure. The first two axes explained 72.6 % of the total co-variation between the epigenetic and genetic profiles, and this association was significantly different from the value expected for random association ($P<0.001$). Epigenetic and genetic profiles displayed similar distributions in the co-inertia subspace (Figure 37) and were significantly correlated based on Dice coefficients, r = 0.89, P=0.0001 (permutations 10^4); ßST coefficients, r = 0.90 P=0.0001; RV coefficient = 0.89, P=0.001 (Monte Carlo test). These results suggested a significant correlation between epigenetic variance and nucleotide sequence variation. We performed a partial Mantel`s test to remove the effects of environmental variables, performed partial Mantel's tests while controlling geographic distance and climate variables, and the results still showed significant correlation (Dice coefficients: r =0.901, P = 0.002; ßST: r = 0.943, P = 0.027).

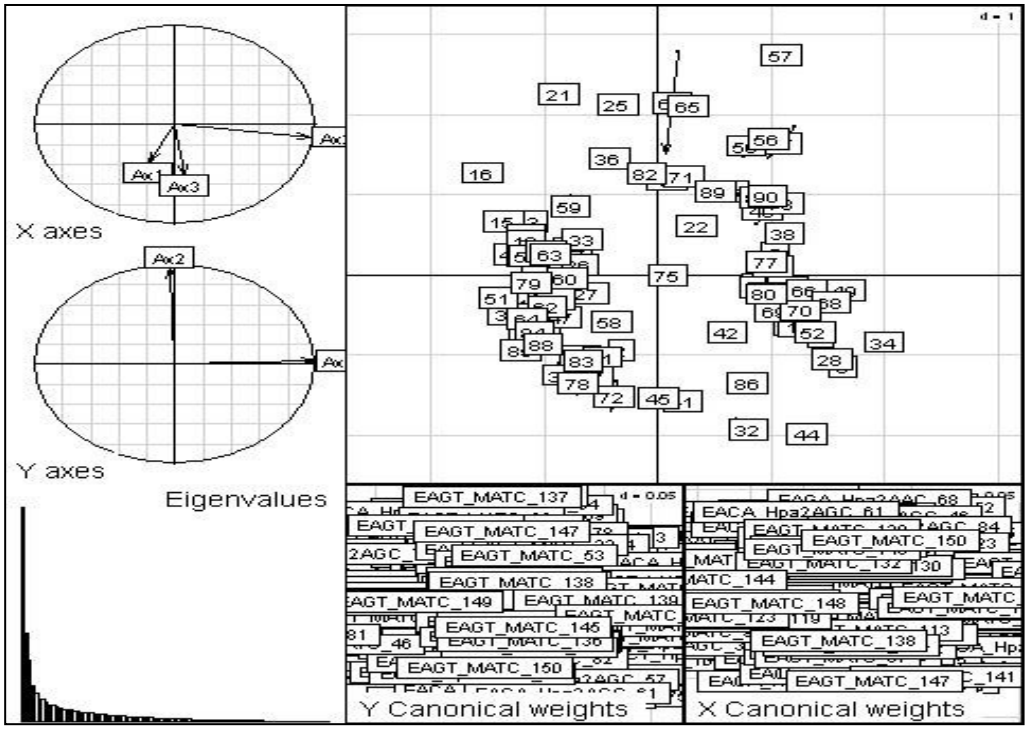


Figure 37: Co-inertia analysis (COIA) of the EAHB population based on genetic (MIP) and epigenetic (MSP) covariance matrices show equal contributions of the epigenetic and genetic matrix to the co-inertia space. X and Y are the two continuous variables measured on the same individuals. X and Y axes are correlation circles showing projections of the PCA axes (from the MSP and MIP data respectively) and both represent a view of the rotation needed to associate the two datasets. Eigenvalues gives the eigen values of the co-inertia analysis. Canonical weights scatter plots represent the coefficients of combinations of varaibles for each table to define the coinertia axes. Scatter plot with arrow is specific to coinertia analysis and represents the cultivars.

5.3.5 Epigenetic EAHB population structure and relationships

Clustering of the EAHB cultivars on the PCA is based on the geographical origin (A and B for Kenya and Uganda respectively) of the cultivars but not morphological groups (Figure 38). Structure analysis classified the EAHB in two clusters and a third cluster being the out-groups; K=3 (S Table 9), but had nothing to do with the morphological grouping (S Figure 9). Interestingly, if probability of ancestry P<0.5 is considered the Kenya and Uganda clusters are distinct (Figure 38, S Figure 8). However, if probability of ancestry P>0.8 is

considered each of the two EAHB regional based clusters split into two revealing two no admix (clusters, a, and b) and two admix clusters (c and d) with a higher probability of ancestry from their region of collection (Figure 38). Epigenetic relationships pictured using the Neighbour- joining (UPGMA) tree supports PCA and structure results, a clusters d and e are admixed cultivars from the two regions (Figure 39).

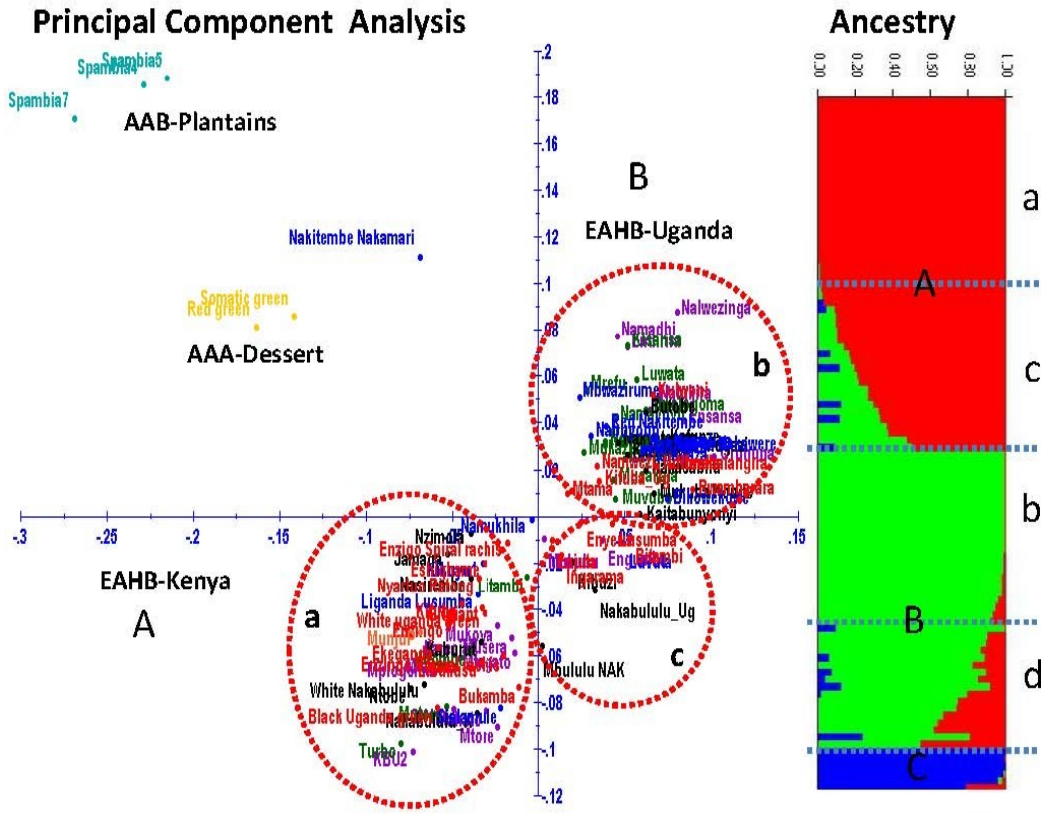


Figure 38: PCA and Structure analysis depicting the epigenetic structure and ancestry of the EAHB. PCA was generated using Modalities dissimilarity coeffient (Sokal & Mitchener, 1958) and ancestry of the EAHB evaluated using Structure, the number of clusters identified were K=3 using combined methylation susceptible loci (MSL) and no methylation loci (NML) datasets.

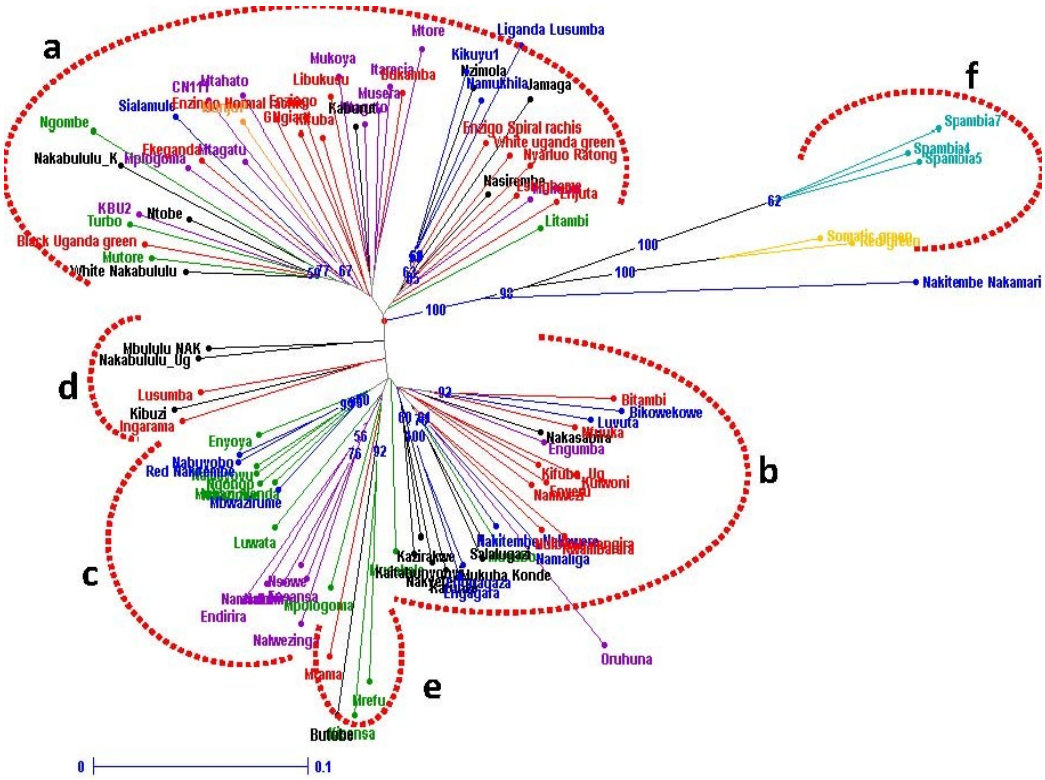


Figure 39: Neighbour-Joining tree and PCA showing epigenetic relationships of the EAHB. Generated using Modalities dissimilarity coefficient (Sokal & Mitchener, 1958). The different colour codes represent morphological groups. Cluster (a) is predominantly composed of cultivars from Kenya, (b) and (c) has cultivars from Uganda, (f) cluster represents the out-groups while (d) and (e) are admixed clusters of Kenya and Uganda cultivars. Only bootstraps values >50 are shown.

Analysis of the genetic population structure calculated using Analysis of Molecular variance (AMOVA) revealed low, but significant ($P< 0.002$) overall epigenetic population differentiation, Phi_{PT} 0.019. Epigenetic variance was divided in distinct within- (98%) and among- (2%) population components (Table 22).

Table 22: Analysis of Molecular variance (AMOVA). Partitioning of epigenetic variation within and between the populations of EAHB (PhiPT = AP / (WP + AP) = AP / TOT (AP = Est. Var. Among Pops, WP = Est. Var. Within Pops).

Source	df	SS	MS	Est. Var.	%
Among groups	4	393.024	98.256	1.397	2%
Within groups	85	6234.265	73.344	73.344	98%
Total	89	6627.289	171.600	74.742	100%

Groups refer to the morphological clonesets; PhiPT = AP / (WP + AP) = AP / TOT: AP = Est. Var. Among Pops, WP = Est. Var. Within Pops

5.4 DISCUSSION

Previous chapters in this PhD study have demonstrated that the EAHB cultivars contain relatively low levels of genetic variation. Despite this, high level of methylation polymorphisms were observed. This underscores the potential significance of methylation polymorphisms within and among EAHB populations. This together with stability of methylation patterns within a given EAHB cultivar, makes it possible that heritable methylation polymorphisms may serve as useful epigenetic markers for a certain populations or cultivars.

The MSAP technique detects methylation only when the one of the methylation sensitive enzymes (*Hpa*11/*Msp*1) has cut and cannot discriminate between methylation and fragment absence when both cytosines are hypermethylated; this may cause underestimation of the level of genomic DNA methylation. Bearing in mind this intrinsic limitation of the technique, our study investigated the level and pattern of genome wide 5′-CCGG-methylation in 90 cultivars of five triploid EAHB populations. Our results show moderate methylation levels and high levels of methylation polymorphism (MP). Considering only the CG and CHG context, the

populations show high levels of genomic methylation at 5´-CCGG-3´ sites, on average CG methylation was higher (24.59%) compared to CHG (17.62%) methylation and a total of 42.21% methylation. Total DNA methylation level is lower than those found in triploid loquat, watermelons and pear but higher than those of poplar and salvia (Ai *et al.*, 2011). However, CG methylation in our study was comparable to that observed in Arabidopsis, while CHG methylation is higher than observed in Arabidopsis (Law & Jacobsen, 2010). Widespread methylation polymorphisms have also been observed in *Gossypium hirsutum* L. (Keyte *et al.*, 2006) and maize (Candaele *et al.*, 2014). It can be concluded that in EAHB cultivars, full methylation of the internal cytosine (CG) occurs more frequently than hemi-methylation of the external cytosine (CHG) of the 5'-CCGG-3'sequence, also is observed in other plant genomes (Lister *et al.*, 2008; Candaele *et al.*, 2014; Osabe *et al.*, 2014).

In our study, epigenetic diversity (Shannons index MSL = 0.49) was higher than genetic diversity (Shannons index NML=0.18). Higher DNA methylation diversity than the genetic diversity has also been demonstrated in cotton (Osabe *et al.*, 2014). Substantial epigenetic diversity was observed in natural populations whose experimental genotypes were propagated from roots and planted in the same conditions (Ma *et al.*, 2013). Our study demonstrates that the DNA methylation diversity remains high even in EAHB genotypes that were grown in the same environment over many generations. Similar results have been reported in other cultivated plants that suggest the possible involvement of epigenetic variation compensating for the lack of genetic variation (Osabe *et al.*, 2014).

In the PCA for MSL the accessions from EAHB, plantains and AAA-desert failed to cluster and were interspersed with one another. This shows that methylation polymorphisms is not genotype related (i.e same genotypes may have different methylation profiles and vice versa) among the cultivars studied, as was also found in Arabidopsis (Cervera *et al.*, 2002) and rice (Ashikawa, 2001). The clear and significant correlation between epigenetic and genetic

variations in the EAHB population suggests that methylation based epigenetic variance might be associated with control of genetic instability as suggested by Liu *et al.* (2012). In addition, Messeguer *et al.* (1991) proposed that methylcytosine could be inherited through meiosis in a Mendelian fashion, suggesting that epigenetic variation is under genetic control and/or their correlation was caused by neutral drift (Liu *et al.*, 2012; Ma *et al.*, 2013). Furthermore, genetic variation in the form of a transposon could directly affect the epigenetic state of the retro element, which could induce varied phenotype in genetically identical offspring (Liu *et al.*, 2012). This shows that methylation polymorphisms is not related to genetic relatedness among the cultivars studied, as was also found in Arabidopsis (Cervera *et al.*, 2002) and rice (Ashikawa, 2001).

The cultivars in the same morphological groups do not seem to be related epigenetically, besides clustering was majorly based on geographical region. This could be because DNA methylation epialleles can produce continuous variation in phenotypes rather than producing discrete phenotypic classes. Continuous variation exists in the degree of radial symmetry of L. vulgaris flowers, where the degree of radial symmetry increases with increasing methylation density of CYCLOIDEA gene (Cubas *et al.*, 1999). No relationship was established between the variant phenotype (27 variants studied) and a particular MSAP pattern in *Coffea Arabica* plants (Landey *et al.*, 2013).

5.5 CONCLUSION

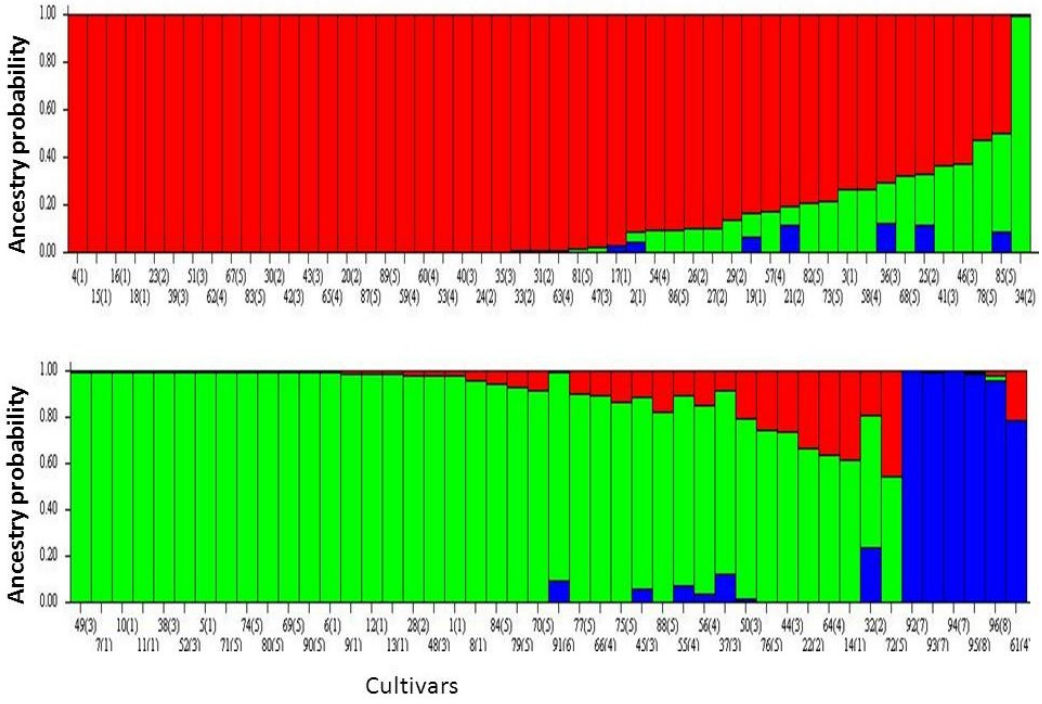
The DNA methylation patterns (of leaf samples of EAHBs) did not show any correlation to morphological groups. It could be that the number of MSAP markers used in this study was too small to identify an association between DNA methylation and discrete morphology of cultivars. Or it can also be the case that there is no relationship between morphology based classification and DNA methylation based classification of the EAHBs. It is important to note that DNA methylation is only one level of a multi-layered epigenetic regulation (including also histone modifications and many different types of epi-marks). Hence, the DNA methylation diversity measured in this study may be underestimating the epigenetic diversity and/or analyzing only one form of epigenetic diversity that has no association with the morphological classification. Further work investigating DNA methylation and other epigenetic profiles in EAHB cultivars will provide a better understanding of the epigenetic association with phenotypic variation in EAHBs. More importantly, in the near-term the high DNA methylation polymorphism in EAHBs may provide sufficient diversity for epi-genome based breeding, even in crops such as EAHB with limited genetic diversity, with further potential to develop epi-markers that are linked to traits of interest.

5.6 SUPPLEMENTARY MATERIAL

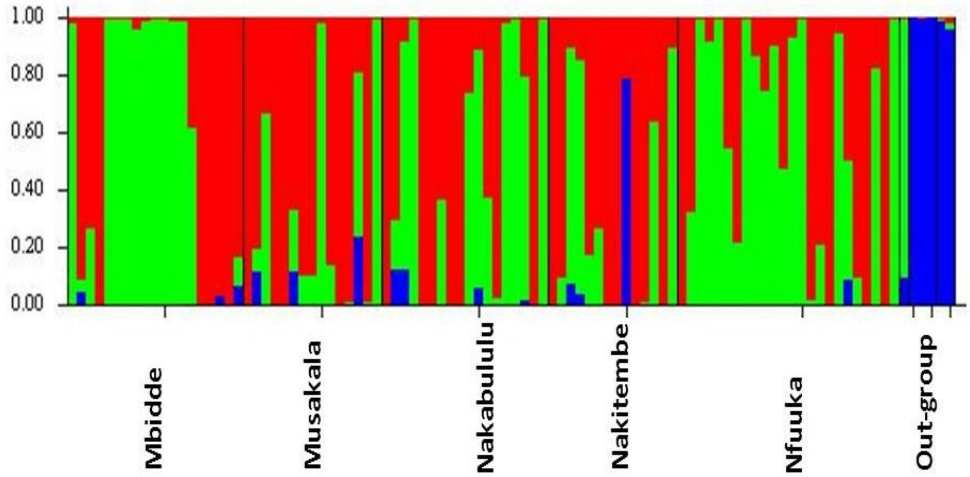
S Table 9: Optimal value of K, the highest Delta K value was K=3 (in bold) obtained from Msap data of the EAHB cultivars using admixture model without priori population assignment.

K	Reps	Mean LnP(K)	StdevLnP (K)	Ln'(K)	Ln''(K)	Delta K
2	3	-53272.53	34.40	—	—	—
3	**3**	**-49253.70**	**15.70**	**4018.83**	**2486.10**	**158.39**
4	3	-47720.97	13.05	1532.73	1097.90	84.13
5	3	-47286.13	276.99	434.83	1218.93	4.40
6	3	-45632.37	494.77	1653.77	961.73	1.94
7	3	-44940.33	565.35	692.03	190.15	0.34
8	3	-44058.15	217.82	882.18	—	—

S Figure 8: STRUCTURE bar plots of genetic membership proportions (K=3) of cultivars. Each cultivar is represented by a vertical line divided into K colors. Cultivar numbers correspond to names in Appendix Table 1 while the number in parenthesis represents the morphological group in which the cultivar belongs to; 1=Mbidde, 2= Musakala, 3=Nakabululu, 4=Nakitembe, 5=Nfuuka, 6=munju P(unknown group), 7=AAB-plantains and 8=AAA_Dessert.



S Figure 9: STRUCTURE bar plots of genetic membership proportions (K=3) of cultivars shown in morphological groups. Each Morphological group is represented by a vertical line divided into K colours



CHAPTER 6

TRANS-GENERATIONAL INHERITANCE OF DNA METHYLATION PATTERNS IN SEXUAL GENERATED HYBRIDS AND VEGETATIVE CLONES OF THE EAST AFRICAN HIGHLAND BANANAS

Abstract

Background

Banana breeding has traditionally focused on exploiting DNA sequence and chromosomal diversity. However, epigenetic variation such as DNA methylation can also potentially contribute to phenotypic variation and has potential to be considered more widely in banana breeding. The heritability of DNA methylation polymorphisms can differ according the transmission efficacy of the methylation polymorphism via meiosis (e.g. sexual reproduction) or mitosis (vegetative propagation).

Materials and Results

In this study, the genome-wide distribution of DNA methylation was assessed in 13 parental lines used in breeding of the EAHB, F_1s and F_2s generated by crosses (sexual) and vegetative (asexual) propagation to identify heritable differentially methylated regions that may contribute to stable epigenetic variation between generations. The results revealed a wide diversity in DNA methylation among siblings and vegetative clones. Using four sets of hybrids and their inbred parents, we investigated the level and pattern of cytosine methylation in each of the hybrids and their corresponding parental inbred lines using the methylation-sensitive amplified polymorphism (MSAP) method. We reveal that differences in DNA methylation found within a family can also be observed in unrelated individuals. Whereas a great majority of cytosine methylation sites displayed high-fidelity epigenetic inheritance, 26.07% and 34.08% of the sites showed altered parental patterns in vegetative clones and hybrids respectively. The methylated susceptible loci (MSL) was highly polymorphic (1468/1638 polymorphic loci) but less differentiated (0.04675, $P<0.0001$) in sexual families compared to the vegetative families (1262/1638 polymorphic loci; 0.0958 P=1e-04). However a high correlation of the methylated susceptible loci and No methylated susceptible loci (NML) was observed in vegetative clones ($r =0.484$; P=9.9e-04) vs sexual families ($r =0.0219$;P=0.3619).

Conclusion

Some methylated DNA loci were highly stable and inherited in the subsequent generation, regardless of whether they were asexually produced by vegetative propagation or sexually produced via hybrid-crossing. By investigating the inheritance of the differential methylation in near-isogenic progeny of the EAHB cultivars, it is possible to demonstrate stable inheritance of DNA methylation variation, even in the absence of genetic differences. This study provides insights into the DNA methylation patterns in EAHB, and suggests the potential for harnessing epigenetic diversity (epimarkers) in banana breeding. This study provides a strong proof of principle for the integration of epigenetic research approaches in EAHB breeding programs.

Key words: Sexual reproduction, vegetative propagation, epigenetic variation, inheritance, mitosis, meiosis

6.1 INTRODUCTION

Recent research indicates that heritable variation in agriculturally important traits may not only be due to genetic variation but can also be caused by underlying epigenetic variation (Richards, 2011). Therefore, studies on transmission genetics of cytosine methylation in plants are important for elucidating the biological roles of this epigenetic modification (Zhao *et al.*, 2007). Phenotypic diversity exhibited within species of many organisms is often attributed to genetic variation (Li *et al.*, 2013), but there is a large part of this diversity that cannot be explained by genetic polymorphisms alone (Manolio *et al.*, 2009). It has been proposed that epigenetic variation could be one component of this missing heritability (Petronis, 2010).

Epigenetic states in plants, once established, can be inherited through the transmission of epigenetic alleles (epialleles) over many generations (Kakutani, 2002; Hofmann, 2012). Such heritable epigenetic alleles can be considered as a new source of polymorphism and may produce novel phenotypes. This could have significant implications in plant breeding. DNA methylation variation can affect plant phenotypes and impact some important agricultural traits, such as plant height and yield (Becker *et al.*, 2011). The genetic causes of phenotypic variation are attributable to mutations that create allelic variation and recombination that alters the genetic structure in which alleles are expressed, offering new backgrounds for epistatic interactions (Tsaftaris *et al.*, 2005). In addition to mutations that create genetic variation underlying phenotypic traits, epialleles have been found to produce a new source of variation for selection.

Heritable phenotypic variation within populations is the basis for selection and breeding. Although DNA methylation is reset in mammals, its resetting during meiosis in plants remains controversial, and several lines of evidence support trans-generational inheritance of DNA methylation (Jullien & Berger, 2010).

Moreover, the cooperative association between genome DNA methylation and genome transcription can be transferred effectively to offspring by sexual or asexual approach and exhibit different phenotypic traits in offspring (Zhang *et al.*, 2008). Stable differences in DNA methylation levels between two genotypes can be the result of differences in epigenetic state that are faithfully propagated to offspring either sexually (via meiosis) or asexually (e.g. vegetative propagation or apomeiosis) (Eichten *et al.*, 2011). Although methylation may occur at various sequence motifs including CG, CHG and non-symmetrical sequence contexts in plants, only the cytosines in CG and CHG contexts allow transmission of the methylation patterns based on the parental strand information.

Recent data shows that epigenetically diverse populations of *Arabidopsis thaliana* produce up to 40% more biomass than epigenetically uniform populations (Latzel *et al.*, 2013). Furthermore, the accelerated evolution and enhanced fitness of allopolyploids may have an epigenetic influence, majorly manifested by genomic cytosine methylation changes (Xiao *et al.*, 2013). Analysis of different allotetraploid sibling orchid taxa demonstrated that ecological divergence and adaptation were largely due to epigenetic effects, which modulate gene expression under an environmental influence (Paun *et al.*, 2010). Hence, we need to incorporate epigenetics into basic breeding research, by quantifying natural epigenetic diversity and investigating its consequences across many different genotypes in the breeding schemes.

6.1.2 Implications of DNA methylation in plant breeding

Creation of favorable variation that will enable the selection of superior genotypes is at the core of a successful breeding program. In the past, rare or induced chance mutations and their shuffling through meiosis and recombination were considered as the major source of variation and formed the basis for selection. Currently, epigenetic information systems (e.g. DNA

methylation) could constitute epigenetic variation that had never been considered in plant breeding as a source of phenotypic variation. Epigenetic changes of genes could generate different epialleles. A clearer view of the role and significance of DNA methylation as a new source of variants in plant breeding can be obtained by a systematic study of DNA methylation, taking into consideration the evolution of plants, the mode of their reproduction, their genotype (inbred line, hybrid, clone, their ploidy level), the degree of isolation during domestication, as well as the time and intensity of breeding effort. Assessing the importance of methylated epialleles in plant breeding requires the determination of: (i) the extent of variation in methylation patterns among individuals within the selection population; (ii) the degree to which methylation patterns affect phenotypes; and (iii) the extent to which methylation variants (i.e. epi-markers) potentially linked to superior phenotypes are stably inherited.

Here, we investigated the genome-wide cytosine methylation pattern in F_1 and F_2 individuals of the sexual crosses, and another parallel set of vegetative 1st cycle offspring, Both sets (i.e. the sex vs asex sets) were compared with the original parents of both, using MSAP analysis. We determined that in EAHBs, the DNA methylation level in sexually generated hybrids was decreased relative to that in the parents, but DNA methylation was increased in the 1st cycle vegetative offspring generated asexually. We also report that significant levels of DNA methylation patterns are faithfully transmitted from the parents to the offspring using both propagation methods. Our study also demonstrates that extensive variation in DNA methylation identified among nearly genetically identical lines could in principle generate functional diversity similar that observed among different genotypes and species. The results of this study are important for demonstrating the heritability of DNA methylation polymorphisms between EAHB generations whether generated sexually (via meiosis) or asexually (via mitosis). Heritable DNA methylation epi-markers

have potential for inclusion in EAHB plant breeding programs and for analysis of performance in farmers fields.

6.2 MATERIALS AND METHODS

6.2.1 Plant materials

Meiotic crosses were made between the three EAHB cultivars (triploids) with a wild (diploid) *Musa acuminata burmaniciodes* (Calcutta 4) to obtain F_1 hybrids (tetraploid). The F_1s were then used as female parents and crossed with wild or improved diploids (Appendix 3) with multiple resistances to production constraints as male parents to generate the secondary F_1s (triploids). Fifty two, 2º F_1s, which also were secondary triploids, are hybrids that do not readily produce seeds and show variation in resistance to various production constraints were selected from the pool for this study. Since propagation of EAHB practiced by farmers is majorly practice vegetative (asexual via mitosis, for comparison of inheritance of DNA methylation patterns between sexual hybrids and vegetative offspring, nine cultivars with unique phenotypes and their vegetative 1st cycle offspring were used in the study.

6.2.2 MSAP technique and allele calling

The method was adapted from Reyna-Lopez *et al.* (1997), who modified the protocol for AFLP (Vos, P *et al.*, 1995) to incorporate the use of methylation-sensitive restriction enzymes. The modifed protocol involved the use of the isoschizomers *Hpa*II and *Msp*I in place of *Mse*I as the frequent cutter, while the rare cutter *EcoR*I was unchanged. The adapter and the basic primer sequences for the *EcoR*I and *Hpa*II-*Msp*I adapters are in S Table 10. MSAP reactions were performed for each genotype, with a modified Vos *et al.* (1995) protocol. Modification made were: 500 ng of template DNA, double-digest

using combinations of *EcoR*I and *Hpa*II (New England Biolabs, Arundel, Australia) or *Msp*I (New England Biolabs, Gold Coast, Australia), PCR using Taq polymerase (New England Biolabs), fluorescently labelled reactions (FAM, NED) were mixed, and peaks were separated on an ABI 3730XL DNA analyzer (Applied Biosystems) with GeneScan 500 LIZ size standard (Applied Biosystems). Adaptors and six oligonucleotides (primer pairs) used in this study are listed in S Table 1. The pre-amplification and selective amplification cycling conditions were performed as manufacturer's instruction with 40 cycles for the selective amplification. To ensure consistency among electrophoretic runs, a control comprising the same sample amplified with the same primer pair was included in every run, and the fragment profile from this control was compared across runs by eye. Fragments were scored manually using GeneMapper 4.0 (Applied Biosystems). The advanced peak detection algorithm was used, with light smoothing turned on and all other settings left at defaults. As suggested by Holland *et al.* (2008) we explored scoring with various bin widths, namely 0.5, 0.7, and 0.9. All other scoring parameters were left at default settings. In preliminary analyses, bin widths of 0.5 produced topologies with the best resolution, so we used this bin width for our final analyses.

In molecular markers such AFLPs in which scoring is binary, allele scoring within the in group (EAHB) can be affected by similarly sized but homoplasics peaks in the out-groups. Because of the tendency for homoplasic peaks to be scored as homologues error is expected to increase with narrower bin widths, and the degree of homoplasy is expected to increase with increasing distance to the out-groups, we scored the in group both with and without out-groups. To reduce the potential impact of size homoplasy, only unambiguous and intense bands, ranging in size from 150 to 500 bp, were scored (Caballero & Quesada, 2010; Liu *et al.*, 2012). A panel with peaks that were strong and consistent was constructed for each oligonucleotide pair based on all methylation sensitive *EcoR*I/*Hpa*II and *EcoR*I/*Msp*I and was applied to all genotype

samples to produce binary data for genetic analysis. We manually checked the quality of each msAFLP fingerprint and bin using the method described by Whitlock *et al.* (2008) and Markert *et al.* (2010) with slight modifications and restricted our analyses to fragments with relative florescence units greater than 100 and larger to reduce background noise. We scored AFLP fragments manually for their presence (denoted as 1) and absence (denoted as 0).

6.2.3 MSAP data analysis

Six primer pairs were used to generate 1868 bands that could be scored reliably across the samples. The percentage polymorphism in each primer combination was calculated by total number of DNA methylation polymorphic sites identified, divided by the total number of sites analyzed. DNA methylation level was quantified for each parent and subsequent generation using the MSAP binary data. Presence of peaks in the *EcoR*I/*Msp*I and absence in *EcoR*I/*Hpa*II was considered CG methylated site, presence of peaks in *EcoR*I/*Hpa*II and absence in *EcoR*I/*Msp*I was considered CHG methylation site. However, when CHG is methylated on both strands *Hpa*II and *Msp*I cannot cleave the site, and is represented by presence or absence of both fragments (Hpa^{+}/Msp^{+} or Hpa^{-}/Msp^{-}). The CG and CHG methylation level was assessed using the number of present/absent peaks in the simultaneous digest (*EcoR*I/*Hpa*II and *EcoRI*/*Msp*I). Visual inspection was done on the number of CHG (only *Hpa*II fragments present) and CG (only *Msp*1 fragments present) methylated fragments in the parents and following generations' off-spring and percentages of methylated fragments were calculated. Significant differences between parent(s) and subsequent generations' methylation level determined by MSAP were analyzed by One-way ANOVA and Tukey's test.

Statistical analysis for MSAP data was done using *Msap* (version 1.1.8) package in R (Pérez-Figueroa, 2013). Individual fragments (loci) were classified as 'methylation-susceptible loci' (MSL) or 'non-methylated loci'

(NML), depending on whether the observed proportion of discordant HPA/MSP scores suggestive of methylation (i.e. number of individuals with contrasting info HPA/MSP scores for the fragment divided by the total number of individuals assayed) exceeded a user-defined threshold (0.05 by default).

To evaluate the diversity level of MSL and NML, the Shannon index of phenotypic diversity, S, derived from the Shannon-Weaver index (Shannon, 1948) was calculated as $S = \sum_{i-1}^{n} pi\ log2\ pi$ where pi is the frequency of the band presence at the ith marker within the population. This index gives more weight to the presence than to the absence of bands. This has no real biological support, although it might account for the occurrence of homoplasic absences of bands (Bonin et al., 2007). The PCoAs` and Neighbour-joining tree for the MSL and NML were constructed using the msap v2.0 package in R. The dendrograms generated from MSAP were supported by Mantel's test with 1000 permutations.

6.3 RESULTS

6.3.1 Methylation level among sexual crosses and vegetative 1st cycle offspring

In this study, three EAHB cultivars; Nakawere, Entukura and Enzirabrahima were crossed with one diploid male (Calcutta4) to generate four F_1 (tetraploid) hybrids (1201K-1, 1438K-1, 660K-1 and 917K-1) which were subsequently crossed with different improved males (C.V rose, SH-3217, Kokopo, Long Tavoy, Malaccensis, SH-3362, SH-3142, 5610s-1 and 9128-3, see Appendix 3) to generate F_2 which also were secondary triploids, are hybrids that do not readily produce seeds and show variation in resistance to various production constraints were selected from the pool for this study (triploid). These were termed as sexual crosses/families arising via meiosis. An additional nine families (Appendix 3) of vegetative cycle (mother plant and 1st cycle offspring)

plants were used to represent the vegetative families generated asexually via vegetative propagation (i.e. mitosis). A total of 1805 loci were analyzed in 6 primer combinations (error rates per primer combination, 0.05). A high number of loci were methylated, the number of Methylation-Susceptible Loci (MSL) and No Methylated Loci (NML) was 1663 and 142. The number of polymorphic MSL was 1562 (94 % of total MSL) and number of polymorphic NML was 124 (87 % of total NML). Diversity of MSL was significantly (Wilcoxon rank sum test with continuity correction; W = 175680.5 ($P < 0.0001$) higher (Shannon's Diversity Index (I) = 0.5066008; SD 0.1496436) compared to the NML (Shannon's Diversity Index I = 0.2393447; SD 0.1129952).

A high level of polymorphism of MSL and NML was observed in both in sexual (90% and 82% MSL and NML respectively) and vegetative families (MSL and NML; 78% and 81% in respective). The sexual MSL was more polymorphic compared to the vegetative MSL. A significant difference of Shannon's Diversity Index (I) was observed in MSL and NML of both sexual (Wilcoxon rank sum test with continuity correction, W = 186745.5; $P < 0.0001$) and vegetative families (Wilcoxon rank sum test with continuity correction: W = 171844 ($P < 0.0001$). However, Shannon's Diversity Index (I) in MSL and NLM was not different when the two propagated families were compared (Table 23). Interestingly, a highly significant differentiation (Phi_ST) level of the MSL and NML was seen in vegetative versus the sexual group (Table 23). Furthermore, a moderate but highly significant correlation ($r = 0.484$, $P = 9.9e\text{-}0^4$) of the MSL (epigenetic loci) and NML (genetic loci) was observed in the vegetative group. Low slightly significant correlation ($r = 0.0219$; $P = 0.03619$) of MSL and NML loci was reported in sexual families.

A visual analysis of the DNA methylation patterns of the parents and the subsequent offspring revealed that while the majority of loci exhibit very similar patterns, there are also examples of altered DNA methylation levels

between the generations (Figure 40). There was a significant general decrease in the average of CG and CHG methylation levels and increased in non-methylated fragments from parents (31.07 % CHG and 26.50% CG) to F_1s (15.75% CHG and 14.75% CG methylation) further decreasing in F_2s (14.05% CHG and 14.03% CG methylation) (Figure 40A). However the F_2 hybrids were not identically methylated, and varied in their CG and CHG levels when compared to their maternal parent (F1s) (S Figure 10A). Conversely, the 1st cycle vegetative offspring showed less variation in CG and CHG methylation levels (Figure 40B) but general increase in levels of methylated cytosines in 1st cycle offspring (CHG and CG methylation; 15.6% and 14.3% respectively) compared to that of the mother plant (13.9% and 13.6% CHG and CG methylation) (Figure 40B). The vegetative families showed higher level of CG methylation while the sexual families had higher CHG methylation. There was variation in both methylation pattern and methylation status within the families in the vegetative clones (S Figure 10B).

Table 23: Comparison results of the methylation level and status of the sexual families versus the vegetative clones. Summarizes results on the number and frequency of variant methylation patterns Shannon's Diversity Index (I) and Phi_ST of Methylation susceptible Loci (MSL) and No methylation Susceptible Loci (NML) found in sexual and vegetative propagated EAHB groups for 1805 loci studied.

Index	**Sexual families**	**Vegetative clones**
Number of samples/individuals	59	34
Number of groups/populations	15	9
Number of Methylation-Susceptible Loci (MSL)	1638	1628
Number of No Methylated Loci (NML)	167	177
Number of polymorphic MSL	1468 (90% of 1638)	1262 (78% of 1628)
Number of polymorphic NML	137 (82% of 167)	144 (81 %177)
Shannon's Diversity Index (I) MSL	0.5247 (SD; 0.1395)	0.5286 (SD; 0.1311)
Shannon's Diversity Index(I) NML	0.2467 (SD; 0.1029)	0.2458 (SD; 0.0978)
Phi_ST MSL	0.04675 (P <0.0001)	0.0958 (P= 1e-04)
Phi_ST NML	0.0728 (P=0.0123)	0.1649 (P= 0.036)
Mantel (r)correlation of MSL/NML	0.0219 (P=0.3619)	0.484 (P=9.9e-04)

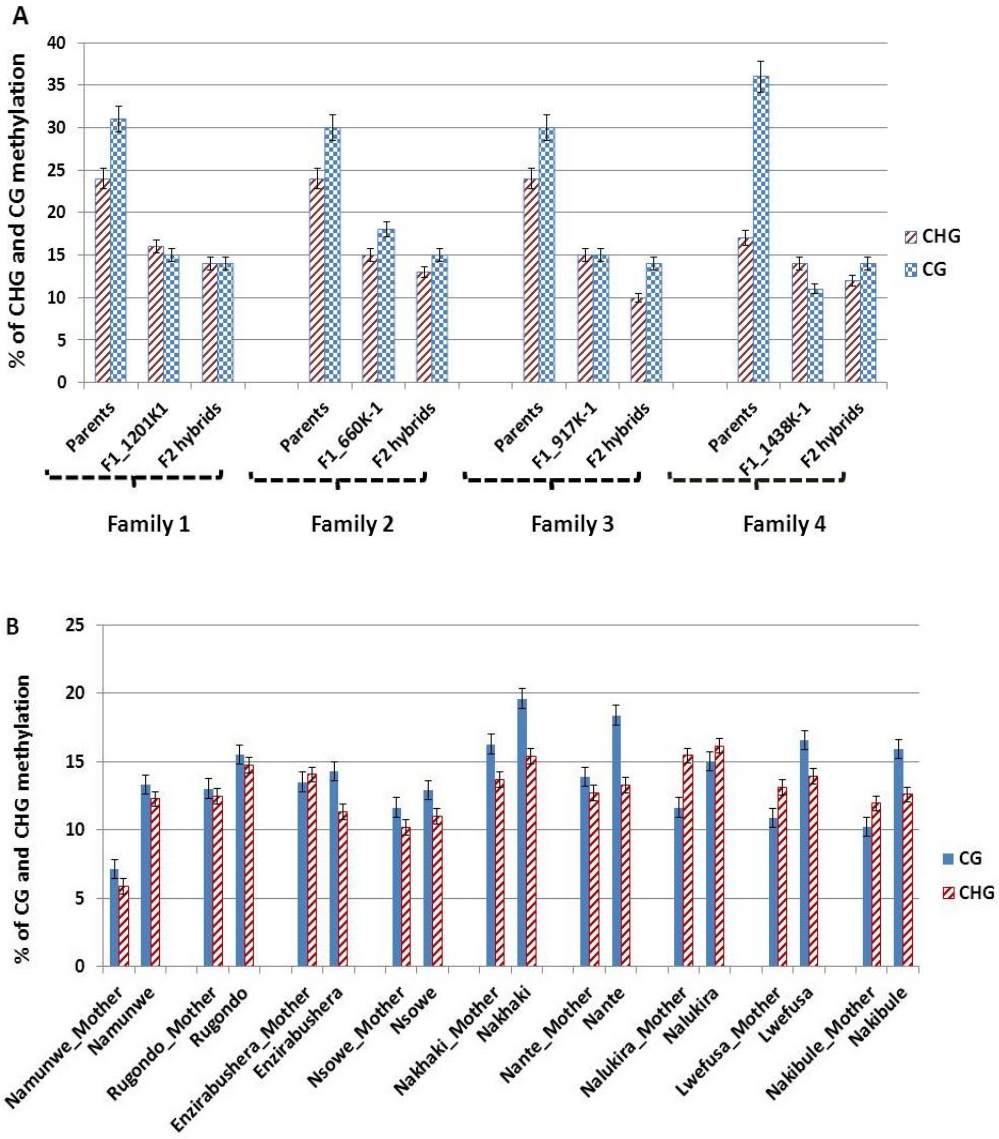


Figure 40: DNA Methylation in Sexual (A) and vegetative (B) parents and their offspring generations. Levels of DNA methylation in CG and CHG context at 1868 methylation-sensitive loci quantified by MSAP. The error bars represent the standard error of mean and the suffix M in B denotes the mother plant. No significant differences were identified between the genotypes in B.

6.3.2 Trans-generational transmission of DNA methylation patterns in EAHB pedigrees

Vegetative propagation of EAHBs is asexual derived from the mother plant. Such vegetative propagation is based on mitosis. While some methylated DNA loci are faithfully transmitted to the next generation via vegetative propagation, DNA methylation at some loci can change from generation to generation generated by vegetative propagation. In the vegetative 1st cycle plants, only 38.66% of 1868 loci exhibited a DNA methylation pattern similar to the mother plant, while 45.34% exhibited a partial gain or loss of DNA methylation and another 26.07% loci had completely different methylation states from that of the mother plant. In comparison, the sexual families generated via meiosis displayed 30.57% complete inheritance from parents, 35.40% similar to either parent and 34.08% displayed a new or acquired methylation state different from both the parents.

On average, the sexual F1s inherited 11.48% (6.02-16.4%) and 16.10% (13.01% -14.4%) of CHG methylation from maternal and paternal parents respectively. In comparison to CHG inheritance, a higher percentage of CG methylation were maternally and paternally inherited, 18.75% (13.4%-20.8%) and 16.5% (11.1%-20.4%). The levels of CHG and CG methylation inherited by the F_2 (2^o F_1s) hybrids were almost similar 15.43% (11.87% -19.06%) and 15.88 % (14.1%-17.4%) (Figure 41A). In the vegetative cycle, 16.1% (6.02%-22.6%) and 17.2% (11.1%-20.8%) of CHG and CG methylation were faithfully transmitted to the 1st cycle generation (Figure 41B).

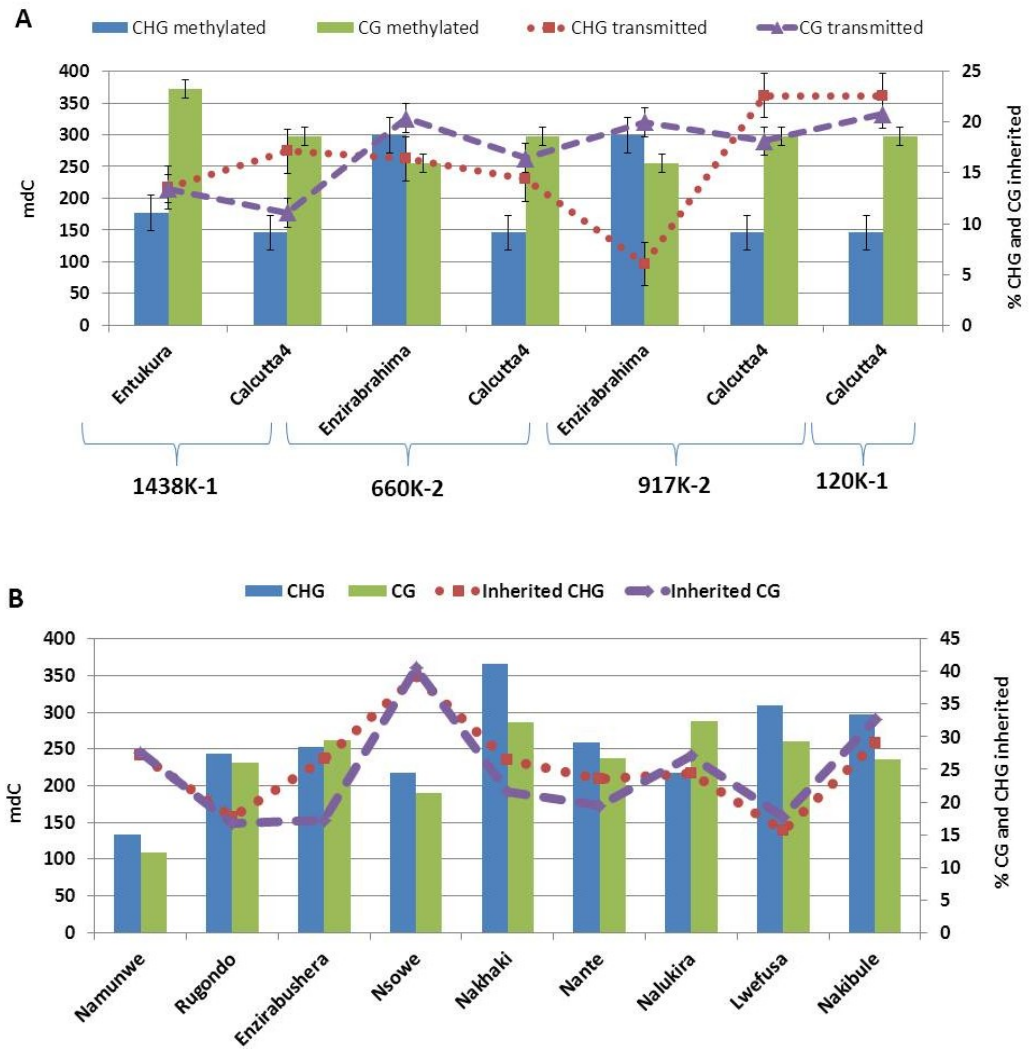


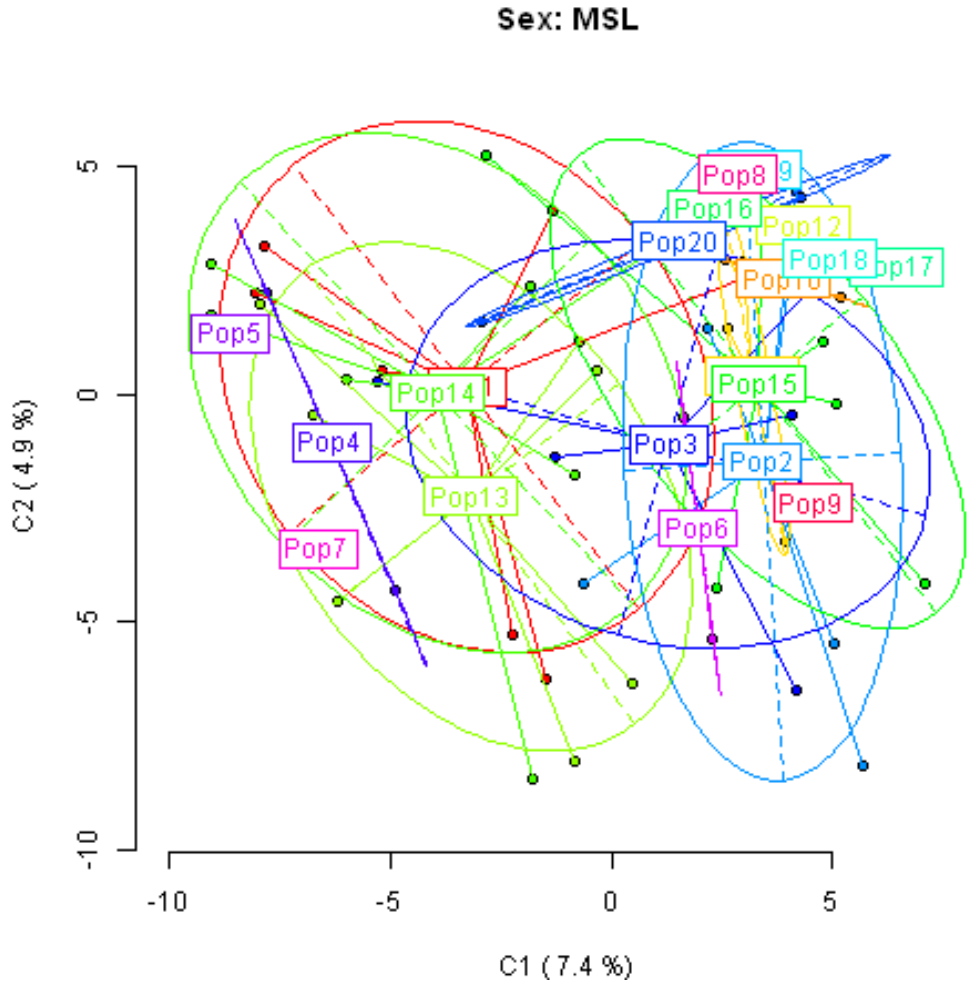
Figure 41: Trans-generational inheritance. (A) transmission of DNA methylation patterns (via meiosis) from parents to F1 generation in sexually generated hybrids. (B) % of CG and CHG DNA methylation patterns passed on from mother plant to 1st cycle plants (via mitosis) in vegetatively propagated families. In both propagation means CG DNA methylation is more inherited compared to CHG methylation.

The percentage of CG and CHG methylation patterns transmitted from sexual F_1s hybrids to F_2 (2º F_1s) hybrids was almost equal 15.55% and 15.86%, respectively. The sexual F_2 were much similar to their maternal than paternal grandparent and inherited 29.66% and 15.25 % CHG methylation respectively.

CG methylation inherited by the F_2s from the grandparents was lower at 17% and 15% maternally and paternallly respectively.

6.3.3 Within - and between- group analyses (BPCA)

The between-population analyses (BPCA) plot based on covariance matrix of the methylation profile for the sexual families showed no obvious epigenetic separation of F_2 genotypes of same maternal parent (e.g. Pop 1 &2, Pop 3 - 9, Pop 10 -13 and Pop 14 & 15) but the families were intertwined (Figure 42). Notably, the F_1s` (Pop 16-Pop 19) were very close to their maternal parents (Pop 20). The first two axes summarized 12.3% of the total inertia. The vegetative MSL-PCA, showed higher differentiation between the families and between the parents and their offspring compared to the sexual group (first two Eigen summarized 14.0% of the total inertia) (Figure 42). However NML loci in both groups was less differentiated and had all families clumped together (S Figure 11). No discrete epigenetic (MSL) or genetic (NML) epi-phylogenetic clusters were observed within sexual families (S Figure 12) and vegetative families S Figure 13).


Sex: MSL
Pop8
Pop16
Pop12
Pop20
Pop18
Pop17
Pop5
Pop14
Pop15
Pop4
Pop3
Pop2
Pop13
Pop9
Pop6
Pop7
C2 (4.9 %)
C1 (7.4 %)
5
0
-5
-10
-10
-5
0
5

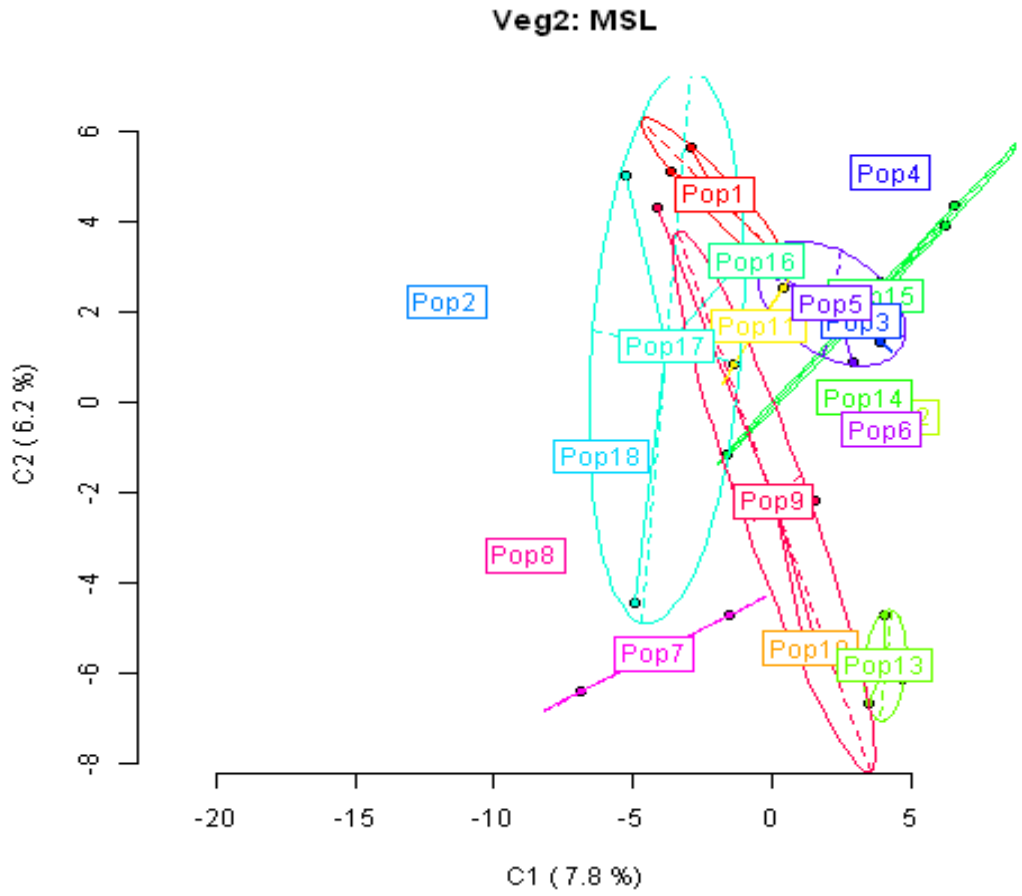


Figure 42: Representation of Principal Coordinate Analysis (PCoA) for epigenetic (MSL) differentiation between the sexual families (Sex: MSL) and vegetative clones (Veg:MSL). The first two coordinates (C1 and C2) are shown with the percentage of variance explained by them. Different point types represent individuals from different groups. Group labels show the centroid for Figure 6.3: Representation of Principal Coordinate Analysis (PCoA) for epigenetic (MSL) differentiation within sexual families (Sex:MSL) groups and vegetative families (Veg: MSL). The points cloud in each group. Ellipses represent the average dispersion of those points around their centre. The long axis of the ellipse shows the direction of maximum dispersion and the short axis, the direction of minimum dispersion.

6.3.4 Epigenetic relationships

Between population PCoA of all genotypes show a separation of the sexual hybrids (F2s`; Pop 1-Pop4, F1`s; Pop 5) from the vegetative clones (Pop 7-Pop 15) however their grandparents (F0; Pop 6) clustered with the vegetative

clones, also triploids, where they initially belonged, suggesting that even though the two groups share some epigenetic patterns, each propagation method evokes specific methylation patterns different from the other (first two Eigen summarized 9.6% of the total inertia, Figure 43). The sexual and vegetative families showed no differentiation of the genetic loci (NML) (S Figure 14).

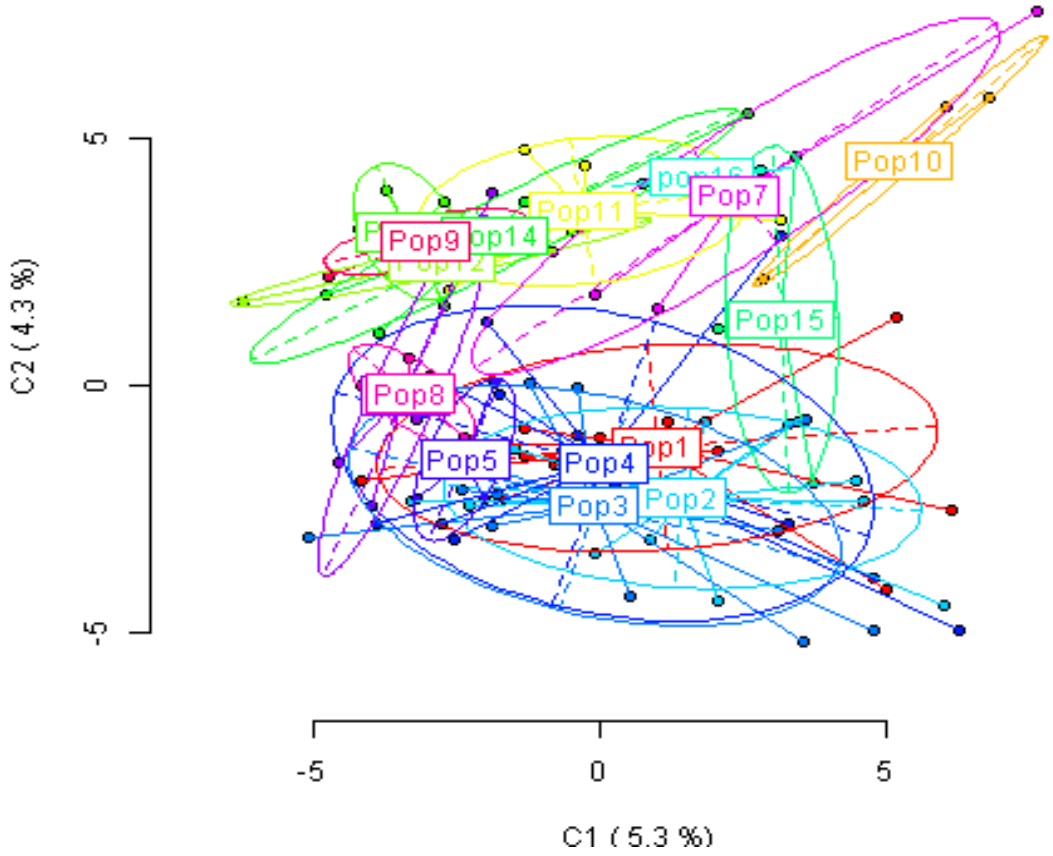


Figure 43: Representation of Principal Coordinate Analysis (PCoA) for epigenetic (MSL) differentiation between the sexual (Pop 1-Pop 5) and vegetative groups (Pop 7-15). The first two coordinates (C1 and C2) are shown with the percentage of variance explained by them. Different point types represent individuals from different groups. Group labels show the centroid for the points cloud in each group. Ellipses represent the average dispersion of those points around their centre. The long axis of the ellipse shows the direction of maximum dispersion and the short axis, the direction of minimum dispersion.

Even though the sexual and vegetative groups of cultivars showed high DNA methylation polymorphism, the similarity coefficient between the cultivars ranged from 0.3640 and 0.7511, suggesting moderate epigenetic diversity. The epi-phylogenetic cluster analysis (MSL) did not show discrete relationships of cultivars based on sexual or vegetative families (Figure 44) nor based on genetic loci (NML) (S Figure 15), however all the vegetative clones formed a single cluster in between the sexual families.

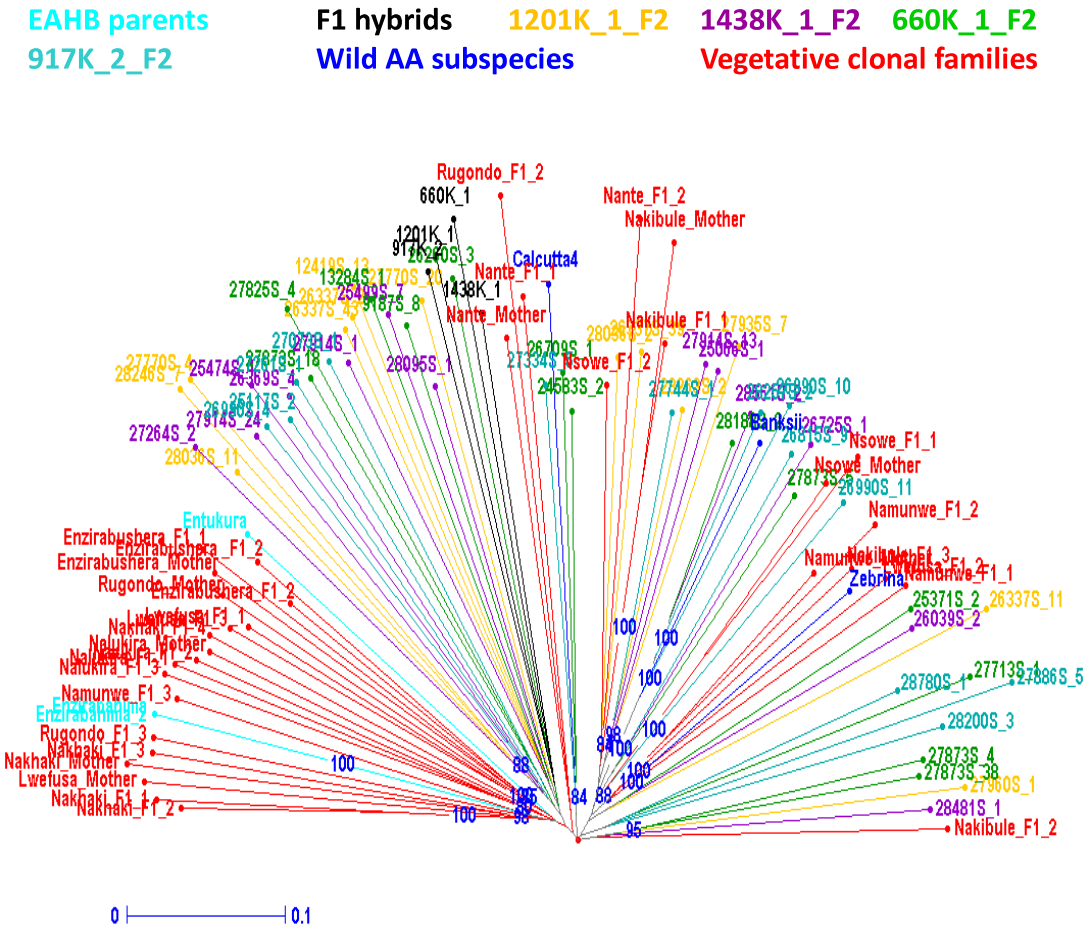


Figure 44: Neighbor-Joining tree generated from epigenetic (MSL) distances of sexual parents, F1 and F2 hybrids samples and vegetative clonal families (mother and 1st cycle plants) for the EAHB. Colors represent different groups/populations as indicated in the Key above. Numbers on branches indicate Bootstrap values; 10,000 bootstraps were done

6.4 DISCUSSION

The integration of DNA methylation diversity and epi-markers into EAHB breeding programs and field studies requires an initial analysis of the heritability of DNA methylation polymorphisms via sexual crosses (meiosis) or vegetative propagation (mitosis). In this study, the extent of methylation at CCGG sites in sexually (crosses) vs asexually (vegetative propagation) derived families were determined. We demonstrate that a significant proportion of methylated DNA CCGG sites remain faithfully methylated in the offspring, whether generate sexually or asexually. The levels of DNA methylation polymorphism observed in this study were high, but not uniquely high. Similar observations has been made in other plants species (Keyte *et al.*, 2006; Salmon *et al.*, 2008; Zhao *et al.*, 2011; Osabe *et al.*, 2014). Extensive polymorphism for DNA methylation has been observed not only among species but among genotypes of a species, individuals belonging to the same genotype, between different organs and tissues of an individual and even among mitotically derived somatic cells of a certain tissue (Tsaftaris & Polidoros, 2000). Silva & Raymond (1988) found loci in homologous chromosomes that were not identically methylated.

Recent evidence suggests that, even in the absence of DNA sequence variation within-species variation in functional traits can be created by epigenetic variation (Latzel *et al.*, 2013). In this study, Methylation-Susceptible Loci diversity (MSL) was significantly higher than diversity of the No Methylated Loci (NML), in both the sexual and vegetative families.

We observed a subsequent decrease in level of methylated cytosines in the F_1s (tetraploids) and F_2 (triploids) in the sexual families compared with their parents, and an increase in methylation level in 1st cycle offsprings compared to the parents. Comparison of the level of methylation in triploid and tetraploid watermelon found a higher methylation levels in tetraploids versus triploids, though no correlation was found between ploidy level and methylation level

(Wang *et al.*, 2009). Similarly, alterations in cytosine methylation have been observed either in F_1 hybrids or in allopolyploids in other species (Aversano *et al.*, 2012; Xiao *et al.*, 2013). The differences in DNA methylation levels between sexually versus asexually and their respective parents, could reflect a differential fidelity of DNA methylation via meiosis versus mitosis.

Data have been presented by others showing that F1 hybrids are in general less methylated than their parental inbreds (Tsaftaris *et al.*, 2005; Zhao, J *et al.*, 2011; Xiao *et al.*, 2013). Contradicting reports have identified mechanisms responsible for genome-wide DNA demethylation in female gametes (Hsieh *et al.*, 2009; Jullien & Berger, 2009) suggesting that DNA methylation patterns might undergo some degree of reprogramming in plants as well. However, there is evidence for such a mechanism and it is proposed that during sexual reproduction a genome-wide decrease of DNA methylation takes place and is compensated by de novo methylation. Such a mechanism could be compatible with the trans-generational inheritance of DNA methylation marks (Jullien & Berger, 2010).

Three classes of patterns of cytosine methylation characterized by differences in degree of methylation between the sexually-generated hybrid and the parental lines were identified in this study: (1) the same level of methylation in the parents and hybrid offspring; (2) an increased level of methylation in the hybrid compared to the parents, and (3) a decreased level of methylation in the hybrid. The first case of banding patterns, appeared to follow simple Mendelian inheritance, while in the latter two cases, the banding patterns were not inherited in a Mendelian fashion. Such increased or decreased methylation in the hybrid compared to the parents could provide an explanation for parent specific and/or hybrid-specific differential gene expression and form a basis for some of the morphological differences observed between parents and offspring.

This indicates that the sexually-generated offspring inherited some of the methylation characteristics from their parent(s) and stably maintained the pattern from one generation to another. Variable epigenetic differences with relatively stable trans-generational inheritance has been observed in other species of diploids (Xiong *et al.*, 1999; Eichten *et al.*, 2011) and polyploids (Xiao *et al.*, 2013). The level of transmission, losses and gains of methylation patterns from parents to the F_1s (crosses) and 1st cycle offspring (vegetative) observed in this study may reflect instability of DNA methylation patterns at some loci, but generally reflect relatively stable inheritance with examples of both gains and losses of DNA methylation. In addition, transmission of methylation patterns in the sexual and vegetative propagation reported in this study confirms that epigenetic variation is not only mitotically stable but also could persist through meiosis in the next generation, in concert with genetic variation (Jablonka & Lamb, 1998; Tsaftaris & Polidoros, 2000).

6.5 CONCLUSION

Until recently, within-species diversity effects have been attributed to underlying variation in DNA sequence. However, some within-species phenotypic differences, and thus potentially functional diversity, could also be due to epigenetic variation. This study identifies DNA methylation epi-markers in EAHBs via msAFLP and tests for transmission of these epimarkers to offspring via sexual (meiotic) crosses versus vegetative propagation (mitosis). The study reveals that a significant proportion of DNA methylation epi-markers can be stably transmitted to offspring vis sexual crosses or asexual vegetative propagation. In addition, other methylated or unmethylated loci behave as metastable epialleles which can display polymorphism between generations and also across families. Such metastable epialleles could provide a basis for functional effects to allow for diversification of phenotypes between genetically similar or identical EAHBs. This finding heralds promise for the

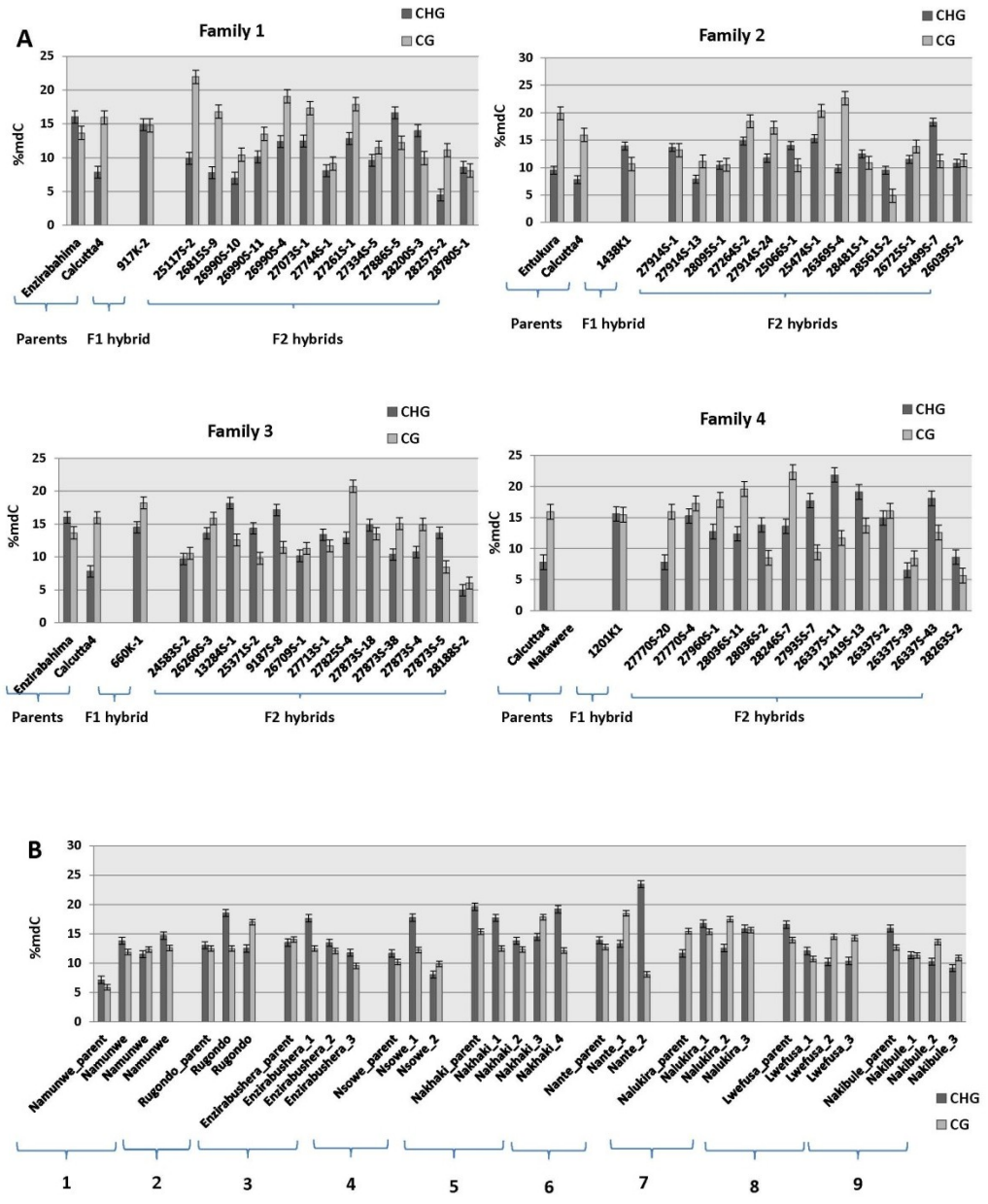
integration of epigenetic markers into EAHB breeding programs and field studies.

6.6 SUPPLEMENTARY MATERIAL

S Table 10: Primers and adaptors used in this chapter

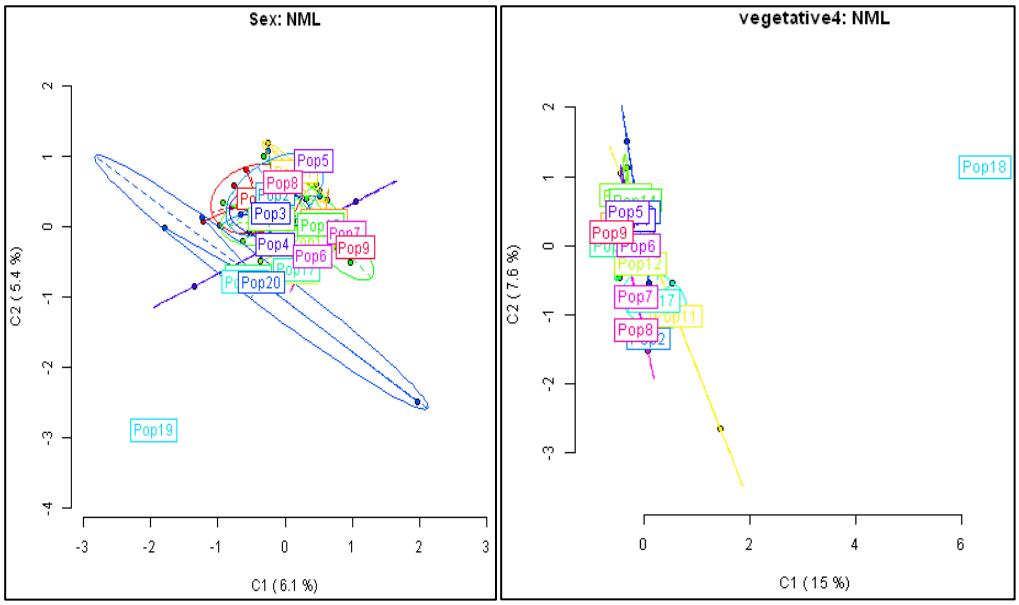
	Sequence
Adaptor	F-GACGATGAGTCTAGAA
*Hpa*II/*Msp*I-adaptor	R- CTACTCAGATCTTGC
	F- CTC GTA GAC TGC GTA C
*Eco*RI adaptor	R -AAT TGG TAC GCA GTC TAC
Preselective primer	
*Eco*RI	GTA GAC TGC GTA CCA ATT CA
*Hpa*II/*Msp*I	ATC ATG AGT CCT GCT CGG T
Selective primer	
E_ACA-NED	Ned-GACTGCGTACCAATTCACA
HPA2AGC	ATCATGAGTCCTGCTCGGAGC
E_AGT-FAM	Fam-GACTGCGTACCAATTCAGT
HPA2ATC	ATCATGAGTCCTGCTCGGATC
E_AGG-FAM	6FAM-GACTGCGTACCAATTCAGG
HPA2AGT	ATCATGAGTCCTGCTCGGAGT
E_AGC-NED	Ned-GACTGCGTACCAATTCAGC
HPA2ATT	ATCATGAGTCCTGCTCGGATT
E_ACG-FAM	6FAM-GACTGCGTACCAATTCACG
HPA2ACA	ATCATGAGTCCTGCTCGGACA
E_ACC-NED	Ned-GACTGCGTACCAATTCACC
HPA2ACT	ATCATGAGTCCTGCTCGGACT

S Figure 10: Methylation levels of sexual (A) and vegetative (B) families quantified by MSAP. The kindred's and vegetative clones showed differential methylation patterns even within families. The error bars represent the standard error of mean. No significant differences were identified between the genotypes

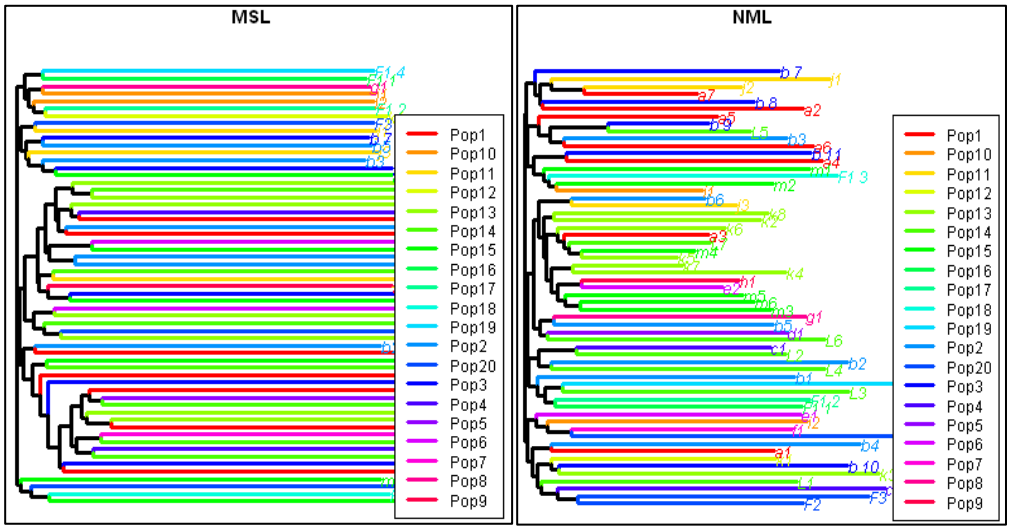


S Figure 11: Representation of Principal Coordinate Analysis (PCoA) for within genetic (NML) differentiation in sexual families (Sex:NML) versus vegetative clones (Vegetative 4: NML). The first two coordinates (C1 and C2)

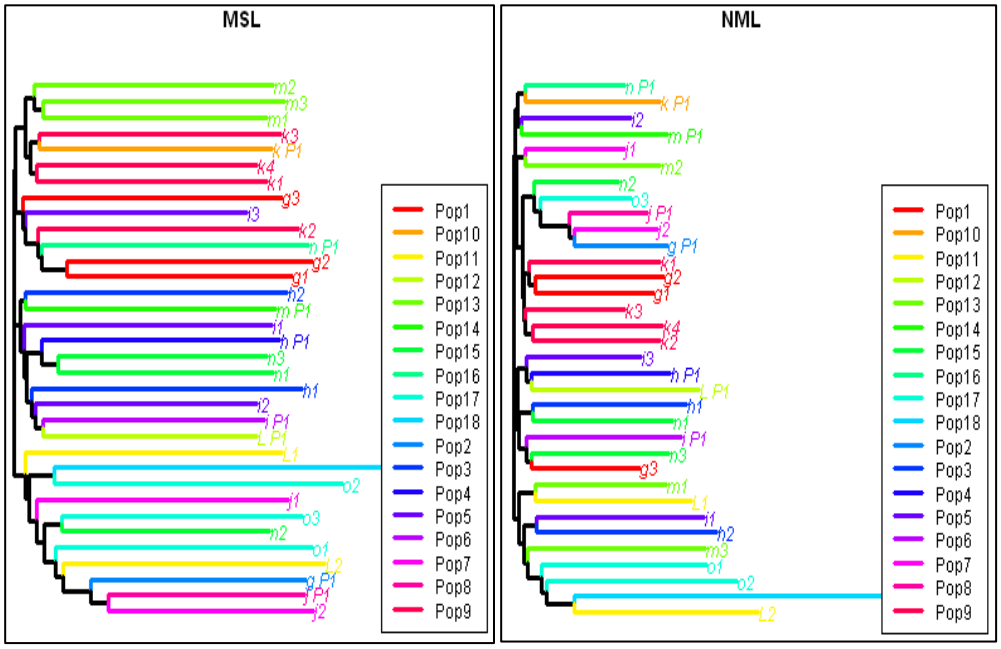
are shown with the percentage of variance explained by them. Different point types represent individuals from different groups. Group labels show the centroid for the points cloud in each group. Ellipses represent the average dispersion of those points around their centre. The long axis of the ellipse shows the direction of maximum dispersion and the short axis, the direction of minimum dispersion.



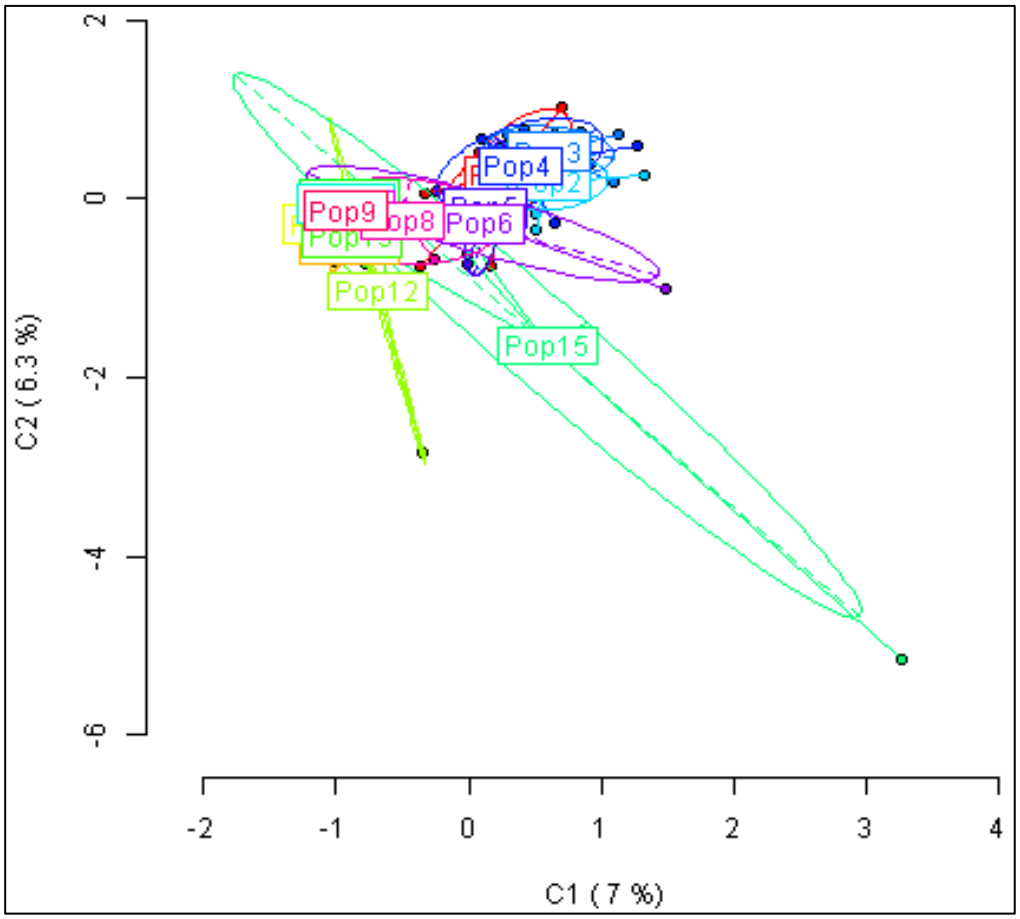
S Figure 12: Within epigenetic and genetic relationships of the sexual families based on MSAP data. Neighbor-Joining tree of all samples (numbered labels at the tips) for epigenetic (MSL) and genetic (NML) distances. Colors represent different families (populations).



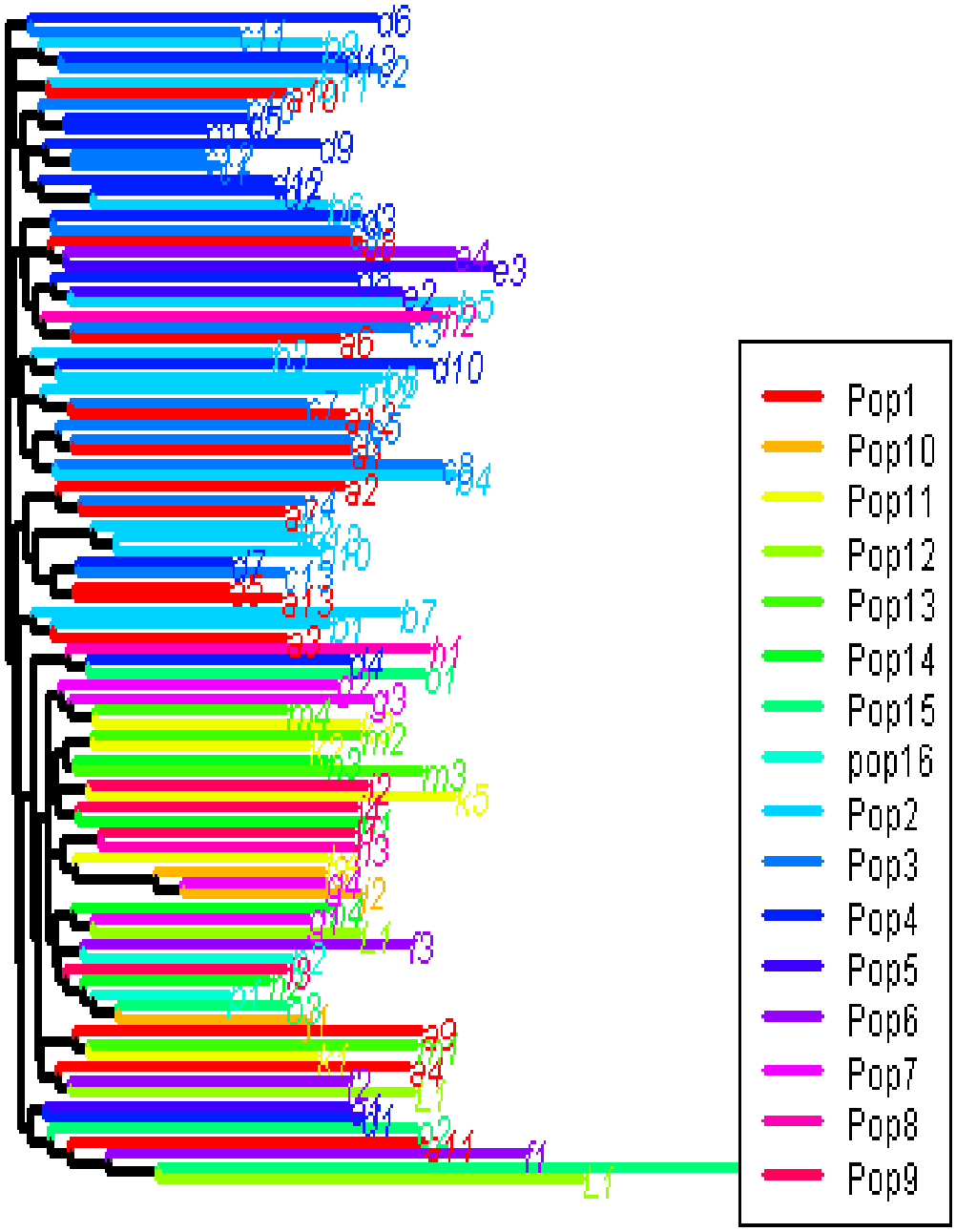
S Figure 13: Within relationships of the Vegetative families based on MSAP data. Neighbor-Joining tree of all samples (numbered labels at the tips) for epigenetic (MSL) and genetic (NML) distances. Colors represent different families (populations).



S Figure 14: Representation of Principal Coordinate Analysis (PCoA) for genetic (NML) differentiation between the sexual and vegetative groups. The first two coordinates (C1 and C2) are shown with the percentage of variance explained by them. Different point types represent individuals from different groups. Group labels show the centroid for the points cloud in each group. Ellipses represent the average dispersion of those points around their centre. The long axis of the ellipse shows the direction of maximum dispersion and the short axis, the direction of minimum dispersion.



S Figure 15: Relationships between the sexual families and vegetative clones based on the No methylated Loci (NML). Neighbour joining tree of the EAHB parental lines (pop 6) the F1 EAHB hybrids (Pop 1-Pop4) from EAHB x Calcutta cross and the F2 hybrids (where the F1 cross was used as maternal parent). The vegetative parents and their 1st cycle generation are represented by Pop 7 to Pop 15) and Pop 16 are wild cultivars Zebrina and Banksii. Colors represent different families (populations), lower clade has most vegetative families while the upper clade contains sexual families.



GENERAL DISCUSSIONS, CONCLUSIONS AND FUTURE DIRECTIONS

7.1 Discussions

7.1.1 Genetic diversity and evolutionary analysis of the East African Highland bananas (Chapters 2, 3 & 4)

A species genetic variation is a product of its long-term evolution and represents its evolutionary potential for survival and development (Feng *et al.*, 2014). Less genetically diverse populations have a reduced ability to buffer the effects of poor environmental conditions or competition (Pluess & Cklin, 2004). While genetic diversity studies have included cultivars from this subgroup and postulated a single seed origin with human aided selection and propagation (Tugume *et al.*, 2002; Ude *et al.*, 2003; Noyer *et al.*, 2005), intra-population structure and phylogenetic relationships *per se* are still unknown. This study was undertaken to provide better understanding of the EAHB cultivars' genetic diversity and genetic structure and the reasons for the tremendous phenotypic diversity. We provide a comprehensive report of the current status of genetic diversity, population structure and evolutionary history/forces that may be influencing genetic diversity of this subgroup.

We found that EAHB subgroup has low genetic diversity (Table 3,8, 14) compared with other *Musa acuminata* subgroups by using similar markers e.g., an average value of PIC in 16 SSR loci was 0.20, He =0.624 (de Jesus *et al.*, 2013), Nei`s gene diversity=0.12 (Changadeya *et al.*, 2012), but their samples included a number of distantly related wild *Musa* species or mixed genomes. Banana subgroups are typical of genotypes that share similar agronomic and fruit quality traits (Creste *et al.*, 2003; de Jesus *et al.*, 2013) and are believed to originate from a common ancestor, meaning, one single meiotic event and the total lack of a sexual stage in the evolution of these subgroups (Noyer *et al.*,

2005), which justifies the small genetic differences. High levels of genetic variation would be expected to have accumulated during a long evolutionary history (Feng *et al.*, 2014). The observed low diversity in this subgroup concurs with short evolutionary history also supported by the low average genetic distance between cultivars and little population differentiation (Phi_{PT} (ϕ_{ST}=0.049 (SNPs), F_{ST} = 0.0271 (AFLP) and Phi_{PT}= 0.036 (SSR) values).

A clear genetic structure according to the geographic regions of the studied population was observed. Even within the same clonesets, cultivars in the two regions were significantly differentiated, but to a small degree. Molecular diversity within EAHB population was irrespective of cloneset origin. Contrary to our expectation, clonesets were only minimally differentiated and diversity was found within- as opposed to between clonesets. Geographic structuring of the EAHB population was supported by the resulting patterns from PC/PCO analysis (Figure 9, 14 and 22), UPGMA clustering dendrogram, (Figure 10, 16 and 23) and STRUCTURE (K = 2) (S Table 1; Figure 8, 16 and 23) clearly reveal the difference in structure between the two geographic regions. Cultivars sampled from the same region are invariably more closely related to each other than to cultivars from other region.

The knowledge of how selection shapes molecular diversity which in turn facilitates the development of phenotypes in heterogeneous environments has become a key endeavour of modern evolutionary biology (Rhode *et al.*, 2013). We consider how these new genomic analyses provide insights into the evolution and have implications on genetic diversity of the EAHB. Using polymorphism-based tests we compared the frequency of alleles with their expectations under the neutral model. Tests for departures from neutrality (Tajima's D, Fu and Li's D*, Ramos-Onsins and Rozas (R^2) statistics and F*, Fu's W and FS and McDonald–Kreitman and the Hudson, Kreitman and Aguade (or HKA) were significant for detection of on-going or recent selection

(Walsh, 2007; Stinchcombe & Hoekstra, 2008). A typical departure showing an excess of common alleles and a deficiency of rare alleles is expected under directional selection, when the coalescent times have been shrunk by a selective sweep. This pattern is also generated by a population bottleneck and/or recent expansion (Walsh, 2007).

During domestication many cultivated species have gone through a bottleneck resulting from founder effects and continued human selection (Cruse-Sanders *et al.*, 2013). Bottleneck analysis of SSRs showed heterozygosity deficit relative to expectations under mutation drift equilibrium, a phenomenon observed in populations facing a rapid expansion after a bottleneck event (Cornuet & Luikartt, 1996). AFLP analysis however, showed heterozygosity excess that is indicative of past bottlenecks (Cornuet & Luikartt, 1996; Cruse-Sanders *et al.*, 2013).

High-throughput biotechnology facilitates next-generation sequencing and high-throughput experimental and bioengineering approaches, large-scale surveys of genome diversity and genome-wide association studies (GWAS) (Karlsson *et al.*, 2014). Genome-wide association study (GWAS) has become an obvious general approach for studying the genetics of natural variation and traits of agricultural importance in higher plants, especially crops (Wang *et al.*, 2012). In this study significant associations were detected for 612 out of 14121 SNPs (Figure 32) in all 13 traits (S Table 7). Although further genetic study is required to confirm the discovery, the present finding highlights the feasibility of high resolution mapping with GWAS in *Musa* species. Significant associations have been observed in other GWAS studies in model and non-model crops (Atwell *et al.*, 2010; Wang *et al.*, 2012). MLM method is the most promising for analyzing traits for GWAS of plant populations (Wang *et al.*, 2012).

7.1.2 Epigenetic analysis in EAHB and inheritance of DNA methylation patterns in sexual generated hybrids and vegetative clones (Chapters 5 & 6)

Even though genetic variation is a major force driving phenotypic variation, there exists great excitement about the potential contribution of epigenetic variation (Richards, 2008). As expected, CG methylation was higher prevalent compared to the CHG methylation in this study. In plants DNA methylation occurs in a CG, CHG and CHH context, with CG methylation occurring in the gene body and transposable elements wherease the CHG methylation occurs only in the gene body (Schmitz *et al.*, 2013).

Methylation patterns did not conform to morphological classification (Figure 36A, 38 and 39) of the studied cultivars. As exemplified by the peloric and colorless non-ripening variants from *Linaria vulgaris* and *Solanum lycopersicum*, respectively epiallele formation in the absence of genetic variation can result in phenotypic variation, which is most evident in the plant kingdom (Schmitz *et al.*, 2013). In this study, the genetic variation seems to be in control of the epigenetic profile, this is supported by the observed high correlation (r =0.90) of the genetic and epigenetic profiles and their equal contribution into the cointeria space (Figure 37). Examples of pure epialleles (methylation variants that form independent of genetic variation) are limited and there are few known examples of DNA methylation variants linked to genetic variants (Lister *et al.*, 2008; Schmitz *et al.*, 2013). It would be therefore necessary to understand how much of the DNA methylation variants are dependent on the underlying genome sequence before we can fully appreciate the extent to which natural epigenetic variation contributes to phenotypic variation. This can be determined by MethylC-sequencing of epigenomes for genotypically distinct EAHB cultivars.

This study had demonstrated high inheritance of DNA methylation patterns in sexual hybrids and vegetative clones. Calarco *et al.* (2012) found out that

symmetric CG and CHG methylation are largely retained in the germline and may account for the prevalence of epigenetic inheritance in plants, compared with mammals. Epigenetic trans-generational inheritance of disease and phenotypic variation has also been demonstrated (Skinner & Guerrero-Bosagna, 2014). While reprogramming of DNA methylation patterns in vegetative offspring in unknown, in sexually reproducing plants, DNA methylation reprogramming in the plant embryo creates a cycle of fluctuation of DNA methylation levels between somatic cells and gametes. This involves loss of CG methylation and gain of CHH methylation through *de novo* DNA methylation and the alternation is predicted to cause slight changes in DNA methylation pattern, as *de novo* methylation has the potential to create new sites of methylation that did not exist in the parents (Kawashima & Berger, 2014). This may explain the lower number of methylated fragments (30.57%) in F_1s similar to the parents versus the high newly or acquired DNA methylation (34.08%; patterns not in parents) observed in sexual hybrids in this study. Through a genomic survey of DNA methylation profiles across individual plants representing a lineage of up to 30 generations, such fluctuations of genome-wide DNA methylation patterns between generations were observed (Becker *et al.*, 2011; Kawashima & Berger, 2014).

7.2 CONCLUSIONS

7.2.1 Genetic diversity and evolutionary analysis of the East African Highland bananas (Chapters 2, 3 and 4)

- Single seed origin, no evidence for multiple origins/ admixture
- Genetic variation did not reflect the large variation in morphological variation suggesting that these particular morphological traits in EAHB may not be genetically controlled
- Past occurrence of a genetic bottleneck and low historical effective population
- Recent population expansion
- Signatures of balancing and negative selection
- Significant GWAS show prospects of genomic selection (GS) in EAHB breeding

7.2.2 Epigenetic analysis in EAHB and inheritance of DNA methylation patterns in sexual generated hybrids and vegetative clones (Chapters 5 and 6)

- Moderate methylation levels but extensive polymorphism
- CG occur more frequently than CHG methylation
- Involvement of epigenetic variation compensating for the lack of genetic variation
- DNA Methylation polymorphisms is not genotype related but occurs randomly
- Methylation based epigenetic variance might be associated with control of genetic instability
- Methylation alleles produce continuous variation in phenotype rather than discrete phenotypic groups

- Significant proportion of methylated DNA CCGG sites remain faithfully methylated in off-spring

- CG methylation is more passed on both meiotically and mitotically compared to CHG methylation

- Decrease in methylation level in meiotic off -springs and increase in methylation level in mitotic off-springs

- Evidence of genome wide decrease in DNA methylation and compensation by de novo methylation in EAHB sexual families

- Differential fidelity in DNA methylation via meiotic vs mitotic

- Mendelian and non-Mendelian inheritance of DNA methylation patterns

7.3 Recommendations and future work

- Reclassification of the EAHB based on genetic relationships

- Banana breeders to use landraces from elsewhere in breeding programme to broaden the genetic diversity of their working collection and revive the lost diversity

- Follow up of the GWAS analysis and genomic selection on economically important traits

- Study of other epigenetic marks that may be causing the phenotypic variation in EAHB

- Further experiments to determine the impact of epigenetics on morphology, e.g. re-setting methylation status and determining impact on few representatives of divergent clone sets

- Genome-wide methylome sequencing of the EAHB and association studies of epiSNPs with important agronomic traits and phenotypes

- Integration of epigenetic markers into EAHB breeding programs as a source of variation

- Study of differential expression of certain structural genes may help uncover the source of phenotypic variations.

REFERENCES

Abdel-Mawgood L 2012. DNA Based Techniques for Studying Genetic Diversity Ahmed In Faculty of Agriculture El-Minia University E.

Abdullah N, Saleh GB, Putra ETS, Wahab ZB. 2012. Genetic relationship among Musa genotypes revealed by microsatellite markers. *African Journal of Biotechnology* **11**(26): 6769-6776.

Ahmad F, Megia R, Poerba YS. 2014. Genetic Diversity of Musa balbisiana Colla in Indonesia Based on AFLP Marker. *HAYATI Journal of Biosciences* **21**(1): 39-47.

Ai L, Bao-Quan H, Zhen-Yi X, Li C, Wei-Xing W, Wen-Qin S, Cheng-Bin C, Chun-Guo W. 2011. DNA Methylation in Genomes of Several Annual Herbaceous and Woody Perennial Plants of Varying Ploidy as Detected by MSAP. *Plant Molecular Biology Reporter* **29**(4): 784-793.

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**(6803): 513-516.

Amorim E, Reis RVd, Santos-Serejo JAd, Amorim VBdO, Silva eSdOe. 2008. Variabilidade genética estimada entre diplóides de banana por meio de marcadores microssatélites. *Pesq. agropec. bras., Brasília* **43(8)**: 1045-1052.

Amos W, Hoffman JIW. 2010. Evidence that two main bottleneck events shaped modern human genetic diversity. *Proc Biol Sci* **277**(1678): 131-137.

Anderson M, Ter-Braak CJF. 2003. Permutation Tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation* **73(2**: 85-113.

Appleby N, Edwards D, Batley J. 2009. New technologies for ultra-high throughput genotyping in plants. *Methods Mol Biol* **513**: 19-39.

Arif IA, Bakir MA, Khan HA, Al Farhan AH, Al Homaidan AA, Bahkali AH, Al Sadoon M, Shobrak M. 2010. A Brief Review of Molecular Techniques to Assess Plant Diversity. *Int J Mol Sci* **11**(5): 2079-2096.

Ashikawa I. 2001. Surveying CpG methylation at 5'-CCGG in the genomes of rice cultivars. *Plant Mol Biol* **45**(1): 31-39.

Asíns M, Monforte AJ, Mestre PF, E.A.Carbonell. 1999. Citrus and Prunus copia-like retrotransposons. *Theor Appl Genet* **99**: 503-510.

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JD, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**(7298): 627-631.

Aversano R, Ercolano MR, Caruso I, Fasano C, Rosellini D, Carputo D. 2012. Molecular tools for exploring polyploid genomes in plants. *Int J Mol Sci* **13**(8): 10316-10335.

Bagamba F, Burger K, Tushemereirwe WK. 2010. Banana (Musa spp.) Production Characteristics and Performance in UgandaEds:Proc. IC on Banana & Plantain in Africa: T. Dubois et al. *Acta Hort. 879, ISHS 2010*.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS One* **3**(10): e3376.

Baneh H, Mohammadi SA, Mahmoudzadeh H, Mattia Fd, Labra M. 2009. Analysis of SSR and AFLP Markers to Detect Genetic Diversity Among Selected Clones of Grapevine (Vitis vinifera L.) cv. Keshmeshi. *S. Afr. J. Enol. Vitic.* **30(1)**.

Barigozzi C 1986. The Origin and Domestication of Cultivated Plants Symposium Proceedings. Oxford: Elsevier Science.

Barkley NA, Krueger RR, Federici CT, Roose ML. 2009. What phylogeny and gene genealogy analyses reveal about homoplasy in citrus microsatellite alleles. *Plant Systematics and Evolution* **282**(1-2): 71-86.

Baurrens F, Nicolleau J, Legarve T, Verdeili J-L, monteuuis O. 2004. Genomic DNA methylation of juvenile and mature Acacia mangium micropropagated in vitro with reference to leaf morphology as a phase change marker. *Tree Physiology* **24**: 401-407.

Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K, Weigel D. 2011. Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. *Nature* **480**(7376): 245-249.

Bhat KV, Jarret RL, Liu ZW. 1994. RFLP characterization of Indian Musa germplasm for clonal identification and classification. *Euphytica* **80**(1-2): 95-103.

Bird A. 2007. Perceptions of epigenetics. *Nature* **447**(7143): 396-398.

Blignaut M, Ellis AG, Roux JJL. 2013. Towards a Transferable and Cost-Effective Plant AFLP Protocol. *PLoS One* **8**(4).

Blouin M, Thuillier V, Cooper B, Amarasinghe V, Cluzel L, Araki H, Grunau C, Taylor E. 2010. No evidence for large differences in genomic methylation between wild and hatchery steelhead (Oncorhynchus mykiss). *Canadian Journal of Fisheries and Aquatic Sciences* **67**(2): 217-224.

Bonin A, Ehrich D, Manel S. 2007. Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. *Mol Ecol* **16**(18): 3737-3758.

Bossdorf O, Arcuri D, Richards CL, Pigliucci M. 2010. Experimental alteration of DNA methylation affects the phenotypic plasticity of ecologically relevant traits in Arabidopsis thaliana. *Evolutionary Ecology* **24**(3): 541-553.

Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. *Am JHum Genet* **32**: 314-331.

Bourguiba H, Audergon J-M, Krichen L, Trifi-Farah N, Mamouni A, Trabelsi S, D'Onofrio C, Asma BM, Santoni S, Khadari B. 2012. Loss of genetic diversity as a signature of apricot domestication and diffusion into the Mediterranean Basin. *BMC Plant Biology* **12**(49).

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**(6470): 455-457.

Brachi B, Morris GP, Borevitz JO. 2011. Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol* **12(10):232**.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**(19): 2633-2635.

Bryant D, Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* **21**(2): 255-265.

Butler D. 2013. Fungus threatens top banana:Fears rise for Latin American industry as devastating disease hits leading variety in Africa and Middle East. *Nature* **504**: 195-196.

Buwa R. 2009. *Using SSR markers to fi ngerprint the East African Highland banana cultivars.* Makerere University.

Caballero A, Quesada H. 2010. Homoplasy and distribution of AFLP fragments: an analysis in silico of the genome of different species. *Mol Biol Evol* **27**(5): 1139-1151.

Calarco JP, Borges F, Donoghue MT, Van Ex F, Jullien PE, Lopes T, Gardner R, Berger F, Feijo JA, Becker JD, Martienssen RA. 2012. Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. *Cell* **151**(1): 194-205.

Caldwell KS, Russell J, Langridge P, Powell W. 2006. Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, Hordeum vulgare. *Genetics* **172**(1): 557-567.

Candaele J, Demuynck K, Mosoti D, Beemster GT, Inze D, Nelissen H. 2014. Differential methylation during maize leaf growth targets developmentally regulated genes. *Plant Physiol* **164**(3): 1350-1364.

Cedar H, Bergman Y. 2012. Programming of DNA methylation patterns. *Annu Rev Biochem* **81**: 97-117.

Cervera MT, Ruiz-Garcia L, Martinez-Zapater JM. 2002. Analysis of DNA methylation in Arabidopsis thaliana based on methylation-sensitive AFLP markers. *Mol Genet Genomics* **268**(4): 543-552.

Changadeya W, Kaunda E, Ambali AJD. 2012. Molecular characterization of Musa cultivars in malawi using SSR. *African Journal of Biotechnology* **Vol. 11**(18): 4140-4157.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**(3): 195-205.

Charlesworth D, Wright SI. 2001. Breeding systems and genome evolution. *Current Opinion in Genetics & Development* **11:**: 685-690.

Cheptou PO, Donohue K. 2013. Epigenetics as a new avenue for the role of inbreeding depression in evolutionary ecology. *Heredity (Edinb)* **110**(3): 205-206.

Christelova P, Valarik M, Hribova E, Van den Houwe I, Channeliere S, Roux N, Dolezel J. 2011. A platform for efficient genotyping in Musa using microsatellite markers. *AoB Plants* **2011**: plr024.

Christelova´ P, rik MV, Hrˇibova´ E, houwe IVd, Channelie`re Sp, Roux N, Dolezˇel J. 2011. A platform for efficient genotyping in Musa using microsatellites. *AoB Plants* **2011**(plr024).

Coart E, Van Glabeke S, Petit RJ, Van Bockstaele E, Roldan-Ruiz I. 2005. Range wide versus local patterns of genetic diversity in hornbeam (Carpinus betulus L.). . *Conserv. Genet.* **6**: 259-273.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**(7184): 215-219.

Cooper H, Hodgkin T, Spillane C. 2000. Broadening the Genetic Base of Crop Production. *International Plant Genetic Resources Institute (IPGRI)*.

Cornuet J, Luikart G. 1996. Description and power analysis of two tests for detecting population bottlenecks from allele frequency data. *Genetics* **144**: 2001-2014.

Creste S, Neto AT, Silva SdO, Figueira A. 2003. Genetic characterization of banana cultivars (Musa spp.) from Brazil using microsatellite markers. *Euphytica* **132**: 259-268.

Crouch J, Vuylsteke D, Ortiz R. 1998. Perspectives on the application of biotechnology to assist the genetic enhancement of plantain and banana (Musa spp.). *EJB Electronic Journal of Biotechnology* **1(1)**

Cruse-Sanders JM, Parker KC, Friar EA, Huang DI, Mashayekhi S, Prince LM, Otero-Arnaiz A, Casas A. 2013. Managing diversity: Domestication and gene flow in Stenocereus stellatus Riccob. (Cactaceae) in Mexico. *Ecol Evol* **3**(5): 1340-1355.

Cubas P, Vincent C, Coen E. 1999. An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* **401**: 157-161.

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, Da Silva C, Jabbari K, Cardi C, Poulain J, Souquet M, Labadie K, Jourda C, Lengelle J, Rodier-Goud M, Alberti A, Bernard M, Correa M, Ayyampalayam S, McKain MR, Leebens-Mack J, Burgess D, Freeling M, Mbeguie AMD, Chabannes M, Wicker T, Panaud O, Barbosa J, Hribova E, Heslop-Harrison P, Habas R, Rivallan R, Francois P, Poiron C, Kilian A, Burthia D, Jenny C, Bakry F, Brown S, Guignon V, Kema G, Dita M, Waalwijk C, Joseph S, Dievart A, Jaillon O, Leclercq J, Argout X, Lyons E, Almeida A, Jeridi M, Dolezel J, Roux N, Risterucci AM, Weissenbach J, Ruiz M, Glaszmann JC, Quetier F, Yahiaoui N, Wincker P. 2012. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488**(7410): 213-217.

D'Hont A, Denoeud F, Jean-MarcAury, Baurens F-C, Carreel F, OlivierGarsmeur, Benjamin Noel2, Bocs Sp, Droc at, Rouard6 M. 2012. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488**(213).

Daniells J, Jenny C, Karamura D, Tomekpe K. 2001. Musalogue: a catalogue of Musa germplasm. Diversity in the genus Musa (E. Arnaud and S. Sharrock, compil.). International Network for the Improvement of Banana and Plantain,Montpellier, France.

Davey J, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next generation sequencing. *Nature Reviews; Genetics* **12**: 499-510.

de Faria-Tavares JS, Martin PG, Mangolin CA, de Oliveira-Collet SA, Machado MdFPS. 2013. Genetic relationships among accessions of mandacaru (Cereus spp.: Cactaceae) using amplified fragment length polymorphisms (AFLP). *Biochemical Systematics and Ecology* **48**: 12-19.

de Jesus, Silva SdOe, Amorim EP, Ferreira CF, Campos JMSd, Silva GdG, Figueira A. 2013. Genetic diversity and population structure of Musa accessions in ex situ conservation. *BMC Plant Biol* **13:41**.

DeGiorgio M, Degnan JH, Rosenberg NA. 2011. Coalescence-time distributions in a serial founder model of human evolutionary history. *Genetics* **189**(2): 579-593.

Dellaporta S, Wood J, Hicks J. 1983. A plant DNA minipreparation: Version II. *Plant Molecular Biology Reporter* **1**(4): 19-21.

Deschamps S, Llaca V, May GD. 2012. Genotyping-by-Sequencing in Plants. *Biology (Basel)* **1**(3): 460-483.

Dhanapal AP, Crisosto CH. 2013. Association genetics of chilling injury susceptibility in peach (Prunus persica (L.) Batsch) across multiple years. *3 Biotech* **3**(6): 481-490.

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB. 1994. Mutational precesses of simple-sequence repeat loci in human populations. *Proc. Nati. Acad. Sci. USA* **91**: 3166-3170.

Dice LR. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology* **26**(3): 297-302.

Drummond AJ. 2005. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Mol Biol Evol* **22**(5): 1185-1192.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**: 214.

Du J, Zhong X, Bernatavichute YV, Stroud H, Feng S, Caro E, Vashisht AA, Terragni J, Chin HG, Tu A, Hetzel J, Wohlschlegel JA, Pradhan S, Patel DJ, Jacobsen SE. 2012. Dual binding of

chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. *Cell* **151**(1): 167-180.

Duputie A, David P, Debain C, McKey D. 2007. Natural hybridization between a clonally propagated crop, cassava (Manihot esculenta Crantz) and a wild relative in French Guiana. *Mol Ecol* **16**(14): 3025-3038.

Earl DA, vonHoldt BM. 2011. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**(2): 359-361.

Edmeades S, Smale M, Karamura D 2005. "Demand for Cultivar Attributes and the Biodiversity of Bananas on Farms in Uganda", Chapter 7, in: Smale, M. and L. Lipper (Eds.) Valuing Crop Biodiversity: On-Farm Genetic Resources and Economics Change. CAB International.

Edwards D, Batley J. 2010. Plant genome sequencing: applications for crop improvement. *Plant Biotechnology Journal* **8**(1): 2-9.

Ehrlich M, Gama-Sosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA, Gehrke C. 1982. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res* **10**(8): 2709-2721.

Eichten SR, Briskine R, Song J, Li Q, Swanson-Wagner R, Hermanson PJ, Waters AJ, Starr E, West PT, Tiffin P, Myers CL, Vaughn MW, Springer NM. 2013. Epigenetic and genetic influences on DNA methylation variation in maize populations. *Plant Cell* **25**(8): 2783-2797.

Eichten SR, Swanson-Wagner RA, Schnable JC, Waters AJ, Hermanson PJ, Liu S, Yeh CT, Jia Y, Gendler K, Freeling M, Schnable PS, Vaughn MW, Springer NM. 2011. Heritable epigenetic variation among maize inbreds. *PLoS Genet* **7**(11): e1002372.

El-Khishin D, Belatus EL, El-Hamid AA, Radwan KH. 2009. Molecular Characterization of Banana Cultivars (Musa Spp.) From Egypt Using AFLP. *Research Journal of Agriculture and Biological Sciences* **5(3**: 272-279.

Elshire R, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**(5): e19379.

Ercisli S, Kafkas E, Orhan E, Kafkas S, Dogan Y, Esitken A. 2011. Genetic characterization of pomegranate (Punica granatum L.) genotypes by AFLP markers. *Biol Res* **44**: 345-350.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* **14**(8): 2611-2620.

Excoffier L, Hofer T, Foll M. 2009. Detecting loci under selection in a hierarchically structured population. *Heredity (Edinb)* **103**(4): 285-298.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. *Genetics* **131**: 479-491.

Falush D, Stephens M, Pritchard JK. 2003. Inference of Population Structure Using Multilocus Genotype Data:Linked Loci and Correlated Allele Frequencies. *Genetics* **164**: 1567-1587.

Falush D, Stephens M, Pritchard JK. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes* **7**: 574-578.

FAO 2002. United Nations. World Summit on Sustainable Development. August 29, 2002.

FAO. 2010. The Second Report on the state of the world's plant genetic resources for Food and Agriculture.

Feng X, Wang Y, Gong X. 2014. Genetic diversity, genetic structure and demographic history of Cycas simplicipinna (Cycadaceae) assessed by DNA sequences and SSR markers. *BMC Plant Biol* **14**: 187.

Flint-Garcia SA, Thornsberry JM, Buckler ESt. 2003. Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol* **54**: 357-374.

Foll, Gaggiotti. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**(2): 977-993.

Foll M. 2012. BayeScan v2.1 User Manual.

Foxe JP, Slotte T, Stahl EA, Neuffer B, Hurka H, Wright SI. 2009. Recent speciation associated with the evolution of selfing in Capsella. *Proc Natl Acad Sci U S A* **106**(13): 5241-5245.

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences* **89**(5): 1827-1831.

Fu N, Wang PY, Liu XD, Shen HL. 2014. Use of EST-SSR markers for evaluating genetic diversity and fingerprinting celery (Apium graveolens L.) cultivars. *Molecules* **19**(2): 1939-1955.

Fu Y-B. 2014. Genetic Diversity Analysis of Highly Incomplete SNP Genotype Data with Imputations: An Empirical Assessment. *G3 :Genes|Genomes|genetics* **4** 891-900.

Fu Y-B, Cheng B, Peterson GW. 2013. Genetic diversity analysis of yellow mustard (Sinapis alba L.) germplasm based on genotyping by sequencing. *Genetic Resources and Crop Evolution* **61**(3): 579-594.

Fu Y-B, Peterson GW. 2011. Genetic Diversity Analysis with 454 Pyrosequencing and Genomic Reduction Confirmed the Eastern and Western Division in the Cultivated Barley Gene Pool. *The Plant Genome Journal* **4**(3): 226.

Fu Y-X, Chakraborty R. 1998. Simultaneous Estimation of All the Parameters of a Stepwise Mutation Model. *Genetics* **150**: 487-497.

Fu YX. 1997. Statistical Tests of Neutrality of Mutations against Population Growth, Hitchhiking and Background Selection. *Genetics* **147(2)**: 915-925.

Fu' Y-X, Li W-H. 1993. Statistical Tests of Neutrality of Mutations. *Genetics* **133**: 693-709.

Fujimoto R, Sasaki T, Ishikawa R, Osabe K, Kawanabe T, Dennis ES. 2012. Molecular mechanisms of epigenetic variation in plants. *Int J Mol Sci* **13**(8): 9900-9922.

Fumagalli M, Vieira FG, Korneliussen§ TS, Linderoth T, Huerta-S´anchez E, Albrechtsen A, Nielsen R. 2013. Quantifying population genetic differentiation from Next-Generation Sequencing data. *Genetics* **113.154740**.

Galov A, K B, T G, M D, Arbanasić H SM, D M, A K, SM F. 2013. Genetic structure and admixture between the Posavina and croatian coldblood in contrast to Lippizan horse from Croatia. *Czech J. Anim. Sci.* **58**(2): 71-78.

Ganapathy K, Gomashe S, Rakshit S, Prabhakar B, Ambekar S, Ghorade R, Biradar B, Saxena U, Patil J. 2012. Genetic diversity revealed utility of SSR markers in classifying parental lines and elite genotypes of sorghum (Sorghum bicolor L. Moench.). *AJCS* **6**(11): 1486-1493.

Ganapathy KN, Gnanesh BN, Byre Gowda M, Venkatesha SC, Gomashe S, Channamallikarjuna V. 2011. AFLP analysis in pigeonpea (Cajanus cajan (L.) Millsp.) revealed close relationship of cultivated

genotypes with some of its wild relatives. *Genetic Resources and Crop Evolution* **58**(6): 837-847.

Gao L, Geng Y, Li B, Chen J, Yang J. 2010. Genome-wide DNA methylation alterations of Alternanthera philoxeroides in natural and manipulated habitats: implications for epigenetic regulation of rapid responses to environmental fluctuation and phenotypic variation. *Plant Cell Environ* **33**(11): 1820-1827.

Garcı´a-Pereira M, Caballero A, Quesada H. 2010. Evaluating the Relationship between Evolutionary Divergence and Phylogenetic Accuracy in AFLP Data Sets. *Mol Biol Evol* **27**(5): 988-1000.

Garcia A, Benchimol LL, Barbosa AMM, Geraldi IO, Jr. CLS, Souza APd. 2004. Comparison of RAPD, RFLP, AFLP and SSR markers for diversity studies in tropical maize inbred lines. *Genetics and Molecular Biology* **27(4)**: 579-588

Garcia F, Ordonez N, Konkol J, AlQasem M, Naser Z, Abdelwali M, Salem NM, Waalwijk C, Ploetz RC, Kema G. 2013. First Report of Fusarium oxysporum f. sp. cubense Tropical Race 4 associated with Panama Disease of banana outside Southeast Asia. *Plant Disease*.

Garnier-Géré P, Chikhi L 2001. Population Subdivision, Hardy–Weinberg Equilibrium and the Wahlund Effect. *eLS*: John Wiley & Sons, Ltd.

Garris A, McCouch SR, Kresovich S. 2003. Population Structure and Its Effect on Haplotype Diversity and Linkage Disequilibrium Surrounding the xa5 Locus of Rice (Oryza sativa L.). *Genetics* **165:** : 759-769.

Gaudeul M, Till-Bottraud I, Barjon F, Manel S. 2004. Genetic diversity and differentiation in Eryngium alpinum L. (Apiaceae): comparison of AFLP and microsatellite markers. *Heredity (Edinb)* **92**(6): 508-518.

Gebremedhin B, Ficetola GF, Naderi S, Rezaei HR, Maudet C, Rioux D, Luikart G, Flagstad Ø, Thuiller W, Taberlet P. 2009. Combining genetic and ecological data to assess the conservation status of the endangered Ethiopian walia ibex. *Animal Conservation* **12**(2): 89-100.

Gerber S, Mariette S, Streiff R, Bodenes C, Kremer A. 2000. Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis *Mol Ecol* **9**: 1037-1048.

Gibert O, Giraldo A, Uclés-Santos J-R, Sánchez T, Fernández A, Bohuon P, Reynes M, González A, Pain J-P, Dufour D. 2010. A kinetic approach to textural changes of different banana genotypes (Musa sp.) cooked in boiling water in relation to starch gelatinization. *Journal of Food Engineering* **98**(4): 471-479.

Goldstein D, Andres Ruiz bares, Cavalli-Sforzaf LL, Feldman MW. 1995. An Evaluation of Genetic Distances for Use With Microsatellite Loci. *Genetics 1* **139**: 463-471.

Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES. 2009. A first-generation haplotype map of maize. *Science* **326**(5956): 1115-1117.

Guerrero-Bosagna CM, Skinner MK. 2009. Epigenetic transgenerational effects of endocrine disruptors on male reproduction. *Semin Reprod Med* **27**(5): 403-408.

Hajjar R, Jarvis DI, Gemmill-Herren B. 2008. The utility of crop genetic diversity in maintaining ecosystem services. *Agriculture, Ecosystems & Environment* **123**(4): 261-270.

Häkkinen M, Teo CH, Othman YR. 2007. Genome constitution for Musa beccarii (Musaceae) varieties. *Acta Phytotaxonomica Sinica* **45 (1)**: 69-74

Hamblin MT, Salas Fernandez MG, Casa AM, Mitchell SE, Paterson AH, Kresovich S. 2005. Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass Sorghum bicolor. *Genetics* **171**(3): 1247-1256.

Harpending H. 1994. Signature of ancient population growth in a low resolution mitochondrial DNA mismatch distribution. *Human Biology* **66(4)**: 591-600.

He F, Wu D-D, Kong Q-P, Zhang Y-P. 2008. Intriguing Balancing Selection on the Intron 5 Region of *LMBR1* in Human Population. *PLoS One* **3**(8): e2948.

He XJ, Chen T, Zhu JK. 2011. Regulation and function of DNA methylation in plants and animals. *Cell Res* **21**(3): 442-465.

Heled J, Drummond AJ. 2008. Bayesian inference of population size history from multiple loci. *BMC Evol Biol* **8**: 289.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol Biol Evol* **27**(3): 570-580.

Heled J, Drummond AJ. 2012. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst Biol* **61**(1): 138-149.

Hendry AP, Bolnick DI, Berner D, Peichel CL. 2009. Along the speciation continuum in sticklebacks. *J Fish Biol* **75**(8): 2000-2036.

Herrera C, Bazaga P. 2013. Epigenetic correlates of plant phenotypic plasticity:DNA methylation differs between prickly and nonprickly leaves in heterophyllous Ilex aquifolium (Aquifoliaceae) trees. *Botanical Journal of the Linnean Society* **171**: 441-452.

Herrera CM, Bazaga P. 2010. Epigenetic differentiation and relationship to adaptive genetic divergence in discrete populations of the violet Viola cazorlensis. *New Phytol* **187**(3): 867-876.

Heslop-Harrison JS, Schwarzacher T. 2007. Domestication, genomics and the future for banana. *Ann Bot* **100**(5): 1073-1084.

Hippolyte I, Bakry F, Seguin M, Gardes L, Rivallan R, Risterucci AM, Jenny C, Perrier X, Carreel F, Argout X, Piffanelli P, Khan IA, Miller RN, Pappas GJ, Mbeguie AMD, Matsumoto T, De Bernardinis V, Huttner E, Kilian A, Baurens FC, D'Hont A, Cote F, Courtois B, Glaszmann JC. 2010. A saturated SSR/DArT linkage map of Musa acuminata addressing genome rearrangements among bananas. *BMC Plant Biol* **10**: 65.

Hippolyte I, Jenny C, Gardes L, Bakry F, Rivallan R, Pomies V, Cubry P, Tomekpe K, Risterucci AM, Roux N, Rouard M, Arnaud E, Kolesnikova-Allen M, Perrier X. 2012. Foundation characteristics of edible Musa triploids revealed from allelic distribution of SSR markers. *Ann Bot* **109**(5): 937-951.

Hoban S, Gaggiotti OE, Bertorelle G. 2013. The number of markers and samples needed for getecting bottlenecks under realistic scenarios, with and without recovery: A simulation -based study. *Mol Ecol* **22**: 3444-3450.

Hofmann NR. 2012. A global view of hybrid vigor: DNA methylation, small RNAs, and gene expression. *Plant Cell* **24**(3): 841.

Holland BR, Clarke AC, Meudt HM. 2008. Optimizing automated AFLP scoring parameters to improve phylogenetic resolution. *Syst Biol* **57**(3): 347-366.

Hopkins, Khalid AM, Chang P-C, Vanderhoek KC, Lai D, Doerr MD, SJ L. 2013. De novo genetic variation revealed in somatic sectors of single Arabidopsis plants. *F1000Research* **2:5**.

Hornemanna G, Weissb G, Durkaa W. 2012. Reproductive fitness, population size and genetic variation in Muscari tenuiflorum (Hyacinthaceae): The role of temporal variation. *Flora* **207**: 736- 743.

Hribova E, Cizkova J, Christelova P, Taudien S, de Langhe E, Dolezel J. 2011. The ITS1-5.8S-ITS2 sequence region in the Musaceae: structure, diversity and use in molecular phylogeny. *PLoS One* **6**(3): e17863.

Hsieh TF, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, Fischer RL, Zilberman D. 2009. Genome-wide demethylation of Arabidopsis endosperm. *Science* **324**(5933): 1451-1454.

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang QF, Li J, Han B. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* **42**(11): 961-967.

Hudson R, Kreitman' M, Aguade` M. 1997. A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics* **116** (153-159).

Hudson R, Slatkin M, Maddison WP. 1992. Estimation of Levels of Gene Flow From DNA Sequence Data. *Genetics* **132**: 583-589

Hudson RR. 1987. Estimating the recombination parameter of a finite population model without selection. *Genetics Research* **50**(03): 245-250.

Hudson RR, Kaplan NL. 1988. The coalescent process in models with selection and recombination. *Genetics* **120**(3): 831-840.

Huff DR, Peakall R, Smouse PE. 1993. RAPD variation within and among natural populations of outcrossing buffalograss [Buchloë dactyloides (Nutt.) Engelm.]. *Theoretical and Applied Genetics* **86**(8): 927-934.

Hufford K, Mazer SJ, Hodges SA. 2013. Genetic variation among mainland and island populations of a native perennial grass used in restoration. *AoB Plants*.

Hughes A. 1999. Adaptive Evolution of Genes and Genomes. *New York: Oxford University Press; 1999.* **225**.

Hurtado M, Vilanova S, Plazas M, Gramazio P, Fonseka HH, Fonseka R, Prohens J. 2012. Diversity and relationships of eggplants from three geographically distant secondary centers of diversity. *PLoS One* **7**(7): e41748.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**(2): 254-267.

Husson F, Josse J, Le S, Mazet J. 2014. FactoMiner: Multivariate Exploratory Data Analysis and Data Mining with R.

Hyten DL, Choi IY, Song Q, Shoemaker RC, Nelson RL, Costa JM, Specht JE, Cregan PB. 2007. Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* **175**(4): 1937-1944.

Hyten DL, Song Q, Zhu Y, Choi IY, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB. 2006. Impacts of genetic bottlenecks on soybean genome diversity. *Proc Natl Acad Sci U S A* **103**(45): 16666-16671.

Innan H. 2006. Modified Hudson-Kreitman-Aguade test and two-dimensional evaluation of neutrality tests. *Genetics* **173**(3): 1725-1733.

Innan H, Stephan W. 2000. The Coalescent in an Exponentially Growing Metapopulation and Its Application to Arabidopsis thaliana. *Genetics* **155**: 2015-2019

IPGRI-INIBAP-Bioversity. 2003. IPGRI-INIBAP Bioveristy Report.

Jablonka E, Lamb MJ. 1998. Epigenetic inheritance in evolution. *J. evol. biol.* **11**: 159-183.

Jablonka E, Raz G. 2009. Transgenerational epigenetic inheritance: Prevalence, mechanisms, and implications for the study of heredity and evolution. *Q. Rev. Biol. 84:* **84 (2)**: 131-176.

Jaenisch R, Bird A. 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **33 Suppl**: 245-254.

Josefina Ines M-B, José Roberto K-C, Rosa Maria E-G. 2013. Annotation of Differentially Expressed Genes in the Somatic Embryogenesis of Musa and Their Location in the Banana Genome. *The Scientific World Journal* **2013**.

Jullien PE, Berger F. 2009. Gamete-specific epigenetic mechanisms shape genomic imprinting. *Curr Opin Plant Biol* **12**(5): 637-642.

Jullien PE, Berger F. 2010. DNA methylation reprogramming during plant sexual reproduction? *Trends Genet* **26**(9): 394-399.

Kaemmer D, Fischer D, Jarret RL, Baurens F-C, Grapin A, Dambier D, Noyer J-L, Lanaud C, G.Kahl, Lagoda PJL. 1997. Molecular breeding in the genus Musa A strong case for STMS marker. *Euphytica* **96**: 49-63.

Kakutani T. 2002. Epi-alleles in plants: inheritance of epigenetic information over generations. *Plant Cell Physiol* **43**: 1106-1111.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**(4): 348-354.

Kannan B, Senapathy S, Bhasker Raj AG, Chandra S, Muthiah A, Dhanapal AP, Hash CT. 2014. Association Analysis of SSR Markers with Phenology, Grain, and Stover-Yield Related Traits in Pearl Millet (Pennisetum glaucum (L.) R. Br.). *The Scientific World Journal* **2014**: 1-14.

Kaplan N, Hudson RR, Langley CH. 1989. The "Hitchhiking Effect" Revisited. *Genetics* **123**: 887-899

Karamura D. 1998. Numerical taxonomic studies of the East African Highland Bananas (Musa AAA-East Africa) in Uganda. *PhD Thesis. The University of Reading*.

Karamura D, Karamura E, Tinzaara W 2012. Banana cultivar names, synonyms and their usage in Eastern Africa.In Karamura D.A. KEBaTWe: Bioversity International, Uganda.

Karamura D, Karamura E, Tushemereirwe W, Rubaihayo PR, Markham R. 2010. Somatic Mutations and Their Implications to the Conservation Strategies of the East African Highland Bananas (Musa spp.) *ACTA HORTICULTURAE* **2(879)**: 615-622

Karlsson EK, Kwiatkowski DP, Sabeti PC. 2014. Natural selection and infectious disease in human populations. *Nat Rev Genet* **15**(6): 379-393.

Kawashima T, Berger F. 2014. Epigenetic reprogramming in plant sexual reproduction. *Nat Rev Genet* **15**(9): 613-624.

Kempinski C, Crowell SV, Smeeth C, Barth C. 2013. The novel Arabidopsis thaliana svt2 suppressor of the ascorbic acid-deficient mutant vtc1-1 exhibits phenotypic and genotypic instability. *F1000Research* **2:6**.

Kennedy K. 2008. Pacific Bananas: Complex Origins, Multiple Dispersals? *ASIAN PERSPECTIVES* **47(1)**.

Keyte AL, Percifield R, Liu B, Wendel JF. 2006. Infraspecific DNA methylation polymorphism in cotton (Gossypium hirsutum L.). *J Hered* **97**(5): 444-450.

Kimura M. 1983. The neutral Theory of molecular evolution and the world view of the neutralists. *National Institute of Genetics, Japan* **411**.

Kimura M, Crow JF. 1964. The number of alleles that can be mainatined in a finite population. *Genetics* **49**: 725-738.

Kolb A, Durka W. 2013. Reduced genetic variation mainly affects early rather than late life-cycle stages. *Biological Conservation* **159**: 367-374.

Koopman WJ. 2005. Phylogenetic signal in AFLP data sets. *Syst Biol* **54**(2): 197-217.

Krauss S. 2000. Accurate gene diversity estimates from amplified fragment length polymorphism (AFLP) markers. *Mol Ecol* **9**: 1241-1245.

Kropff MJ 2001-2005. "Project: Enhanced biodiversity and weed suppression in agro-ecosystems." Crop and Weed Ecology Group (WUR), METIS Wageningen University (2001-2005).

Kwon S, Simko I, Hellier B, Mou B, Hu J. 2013. Genome-wide association of 10 horticultural traits with expressed sequence tag-derived SNP markers in a collection of lettuce lines. *The Crop Journal* **1**(1): 25-33.

Lande. 1996. Statisticsband partitioning of species Diversitry and similarity among multiple communities. *Oikos* **76(1)**: 5-13.

Landey R, Cenci A, Fre´de´ ric Georget, Bertrand1 Bt, Gloria Camayo, Dechamp E, Herrera JC, Santoni S, Lashermes P, Simpson J, Etienne H. 2013. High Genetic and Epigenetic Stability in Coffea arabica Plants Derived from Embryogenic Suspensions and Secondary Embryogenesis as Revealed by AFLP, MSAP and the Phenotypic Variation Rate. *PLoS One* **8(2)**(e56372).

Langhe ED, Pillay M, Tenkouano A, Swennen R. 2005. Integrating morphological and molecular taxonomy in Musa: the African plantains (Musa spp. AAB group). *Plant Systematics and Evolution* **255**(3-4): 225-236.

Langmead B. 2010. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* **11.7**.

Latzel V, Allan E, Bortolini Silveira A, Colot V, Fischer M, Bossdorf O. 2013. Epigenetic diversity increases the productivity and stability of plant populations. *Nat Commun* **4**.

Laurentin H. 2009. Data analysis for molecular characterization of plant genetic resources. *Genetic Resources and Crop Evolution* **56**(2): 277-292.

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11**(3): 204-220.

Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, Andolfatto P, Przeworski M. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol* **10**(9): e1001388.

Lejju B, Robertshaw P, Taylor D. 2006. Africa's earliest bananas? *Journal of Archaeological Science* **33** (102-113).

Lepage T, Bryant D, Philippe H, Lartillot N. 2007. A general comparison of relaxed molecular clock models. *Mol Biol Evol* **24**(12): 2669-2680.

Lewontin R. 1972. The Apportionment of Human Diversity. *Evolutionary Biology* **6**: 381-398.

Lewontin R. 1974. The genetic basis of Evolutionary change *COLUMBIA UNIVERSITY PRESS.*

Ley AC, Hardy OJ. 2013. Improving AFLP analysis of large-scale patterns of genetic variation--a case study with the Central African lianas Haumania spp (Marantaceae) showing interspecific gene flow. *Mol Ecol* **22**(7): 1984-1997.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Li L, Wang HY, Zhang C, Wang XF, Shi FX, Chen WN, Ge XJ. 2013. Origins and domestication of cultivated banana inferred from chloroplast and nuclear genes. *PLoS One* **8**(11): e80502.

Li Q, Eichten SR, Hermanson PJ, Springer NM. 2013. Inheritance patterns and stability of DNA methylation variation in maize near-isogenic lines. *Genetics*: 1-33.

Li Y, Haseneyer G, Schon CC, Ankerst D, Korzun V, Wilde P, Bauer E. 2011. High levels of nucleotide diversity and fast decline of linkage disequilibrium in rye (Secale cereale L.) genes involved in frost response. *BMC Plant Biol* **11**: 6.

Liebl AL, Schrey AW, Richards CL, Martin LB. 2013. Patterns of DNA methylation throughout a range expansion of an introduced songbird. *Integr Comp Biol* **53**(2): 351-358.

Lira-Medeiros C, Parisod C, Fernandes RA, Mata CS, Cardoso MA, mail PCGF. 2010. Epigenetic Variation in Mangrove Plants Occurring in Contrasting Natural Environment *PLoS One* **5(4)**(e10326).

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**(3): 523-536.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**(7271): 315-322.

Liu, Muse. 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* **21**(9): 2128-2129.

Liu K, Muse SV. 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics Application Note* **21** (9): 2128-2129.

Liu S, Feng J. 2012. Natural epigenetic variation and its influencing factors in bat populations.

Liu S, Sun K, Jiang T, Ho JP, Liu B, Feng J. 2012. Natural epigenetic variation in the female great roundleaf bat (Hipposideros armiger) populations. *Mol Genet Genomics* **287**(8): 643-650.

Lott TJ, Frade JP, Lockhart SR. 2010. Multilocus sequence type analysis reveals both clonality and recombination in populations of Candida glabrata bloodstream isolates from U.S. surveillance studies. *Eukaryot Cell* **9**(4): 619-625.

Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, Buckler ES, Costich DE. 2013. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet* **9**(1): e1003215.

Lu Y, Rong T, Cao M. 2008. Analysis of DNA methylation in different maize tissues. *Journal of Genetics and Genomics* **35**(1): 41-48.

Lu Y, Xin Zhang, Jinji Pu, Yanxian Qi, Xie Y. 2011. Molecular assessment of genetic identity and genetic stability in banana cultivars (Musa spp.) from China using ISSR markers. *AJCS* **5(1)**: 25-31.

Lu Y, Xin Zhang, Pu J, Qi Y, Xie Y. 2011. Molecular assessment of genetic identity and genetic stability in banana cultivars (Musa spp.). *AJCS* **5**(1): 25-31.

Luikart G, Allendorf FW, Cornuet J-M, Sherwin WB. 1998. Distortion of Allele Frequency Distributions Provides a Test for Recent PopulationBottlenecks. *Heredity (Edinb)* **89**(3): 238-247.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet* **4**(12): 981-994.

Lynch M, Milligan B. 1994. Analysis of population-genetic structure using RAPD markers. *Molecular Ecology* **3**: 91-99.

Ma K, Song Y, Yang X, Zhang Z, Zhang D. 2013. Variation in genomic methylation in natural populations of chinese white poplar. *PLoS One* **8**(5): e63977.

Maldonado-Borges J, Ku-Cauich JR, Escobedo-GraciaMedrano R. 2013. Annotation of Differentially Expressed Genes in the somatic embyrogenesis of Musa and their location in the banana genome. *The Scientific World Jornal* **535737**: 7.

Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S. 2012. SNP Markers and Their Impact on Plant Breeding. *Int J Plant Genomics* **2012**: 11.

Manning K, Tor M, Poole M, Hong Y, Thompson AJ, King GJ, Giovannoni JJ, Seymour GB. 2006. A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat Genet* **38**(8): 948-952.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM. 2009. Finding the missing heritability of complex diseases. *Nature* **461**(7265): 747-753.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *cancer Res* **27**(2): 209-220.

Mariette S, Corre L, Austerlitz F, Kremer A. 2002. Sampling within the genome for measuring within-population diversity: trade-offs between markers. *Mol Ecol* **11**: 1145-1156.

Mariette S, Tavaud M, Arunyawat U, Capdeville G, Millan M, Salin F. 2010. Population structure and genetic bottleneck in sweet cherry estimated with SSRs and the gametophytic self-incompatibility locus. *BMC Genet* **11**: 77.

Markert JA, Champlin DM, Gutjahr-Gobell R, Grear JS, Kuhn A, McGreevy TJ, Jr., Roth A, Bagley MJ, Nacci DE. 2010. Population genetic diversity and fitness in multiple environments. *BMC Evol Biol* **10**: 205.

Martin G, Baurens FC, Cardi C, Aury JM, D'Hont A. 2013. The complete chloroplast genome of banana (Musa acuminata, Zingiberales): insight into plastid monocotyledon evolution. *PLoS One* **8**(6): e67350.

Martinez-Castillo J. 2008. "Genetic erosion and in situ conservation of Lima bean (Phaseolus lunatus L.) landraces in its Mesoamerican diversity center." *Genetic Resources and Crop evolution.* **55**(7): 1065-1077.

Maruyama T, Fuerst PA. 1985. Population Bottlenecks and nonequilibrium Models in Population Genetics. 11. Number of Alleles in a Small

Population that was formed by a Recent Bottleneck. *Genetics* **111**: 675-689.

Mather KA, Caicedo AL, Polato NR, Olsen KM, McCouch S, Purugganan MD. 2007. The extent of linkage disequilibrium in rice (Oryza sativa L.). *Genetics* **177**(4): 2223-2232.

Mbanjo E, Tchoumbougnang F, Mouelle AS, Oben JE, Nyine M, Dochez C, Ferguson ME, Lorenzen J. 2012. Development of expressed sequence tags-simple sequence repeats (EST-SSRs) for Musa and their applicability in authentication of a Musa breeding population. *Biotechnology* **11**(71): 13546-13559.

McClelland M, Nelson M, Raschke E. 1994. Effect of site-specific modification on restriction endonucleases and DNA modification methyltransferases. *Nucleic Acids Res* **Vol. 22, No. 17**: 3640-3659.

McEachern MB, Vuren DHV, Floyd CH, May B, Eadie JM. 2011. Bottleneck and rescue effects in fluctuating population of golden mantle ground squirrels. *Conserv Genet* **12**: 285–296.

McGrath M 2012. Bananas could replace potatoes in warming world. *News Science & Environment* Science reporter, BBC World Service: BBC.

McKinnon G, Vaillancourt RéE, Steane DA, Potts BM. 2008. An AFLP marker approach to lower levels systematics in Eucalyptus (Myrtaceae). *American Journal of Botany* **95(3)**: 368-380.

Messeguer R, Ganal MW, Steffens JC, Tanksley SD. 1991. Characterization of the level, target sites and inheritance of cytosine methylation in tomato nuclear DNA. *Plant Mol Biol* **16**(5): 753-770.

Metzker ML. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* **11**(1): 31-46.

Meudt HM, Clarke AC. 2007. Almost forgotten or latest practice? AFLP applications, analyses and advances. *Trends Plant Sci* **12**(3): 106-117.

Michalakis Y, Excoffier L. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics* **142**(3): 1061-1064.

Miller M, Haig SM, Thomas D. Mullins, Poppe KJ, Green M 2012. Evidence for Population Bottlenecks and Subtle Genetic Structure in the Yellow Rail. *The Condor*. 100-112.

Miller MR, Atwood TS, Eames BF, Eberhart JK, Yan YL, Postlethwait JH, Johnson EA. 2007. RAD marker microarrays enable rapid mapping of zebrafish mutations. *Genome Biol* **8**(6): R105.

Miura K, Agetsuma M, Kitano H, Yoshimura A, Matsuoka M, Jacobsen SE, Ashikari M. 2009. A metastable DWARF1 epigenetic mutant affecting plant stature in rice. *Proc Natl Acad Sci U S A* **106**(27): 11218-11223.

Miura K, Ikeda M, Matsubara A, Song X-J, Ito M, Asano K, Matsuoka M, Kitano H, Ashikari M. 2010. OsSPL14 promotes panicle branching and higher grain productivity in rice. *Nat Genet* **42**(6): 545-549.

Mohammadi SA, Prasanna BM. 2003. Analysis of Genetic Diversity in Crop Plants—Salient Statistical Tools and Considerations. *Crop Sci.* **43**(4): 1235-1248.

Moran P, Perez-Figueroa A. 2011. Methylation changes associated with early maturation stages in the Atlantic salmon. *BMC Genet* **12**: 86.

Morcroft G 2013. Climate Change Watch: The Banana's Days May Be Numbered As Warming Prompts Insect Infestations *International Bussiness Times*.

Morrell PL, Toleno DM, Lundy KE, Clegg MT. 2005. Low levels of linkage disequilibrium in wild barley (Hordeum vulgare ssp. spontaneum) despite high rates of self-fertilization. *Proc Natl Acad Sci U S A* **102**(7): 2442-2447.

Mueller U, Wolfenbarger LL. 1999. AFLP genotyping and fingerprinting. *Tree Genetics & Genomes* **14**.

Nath V, Senthil M, Hegde VM, Jeeva ML, Misra RS, Veena SS, Raj M. 2013. Genetic diversity of Phytophthora colocasiae isolates in India based on AFLP analysis. *Biotech* **3**: 297-305.

Nautiyal S, Kaechele H. 2007. Conservation of crop diversity for sustainable landscape development in the mountains of the Indian Himalayan region. *Management of Environmental Quality* **18 (5)**: 514-530.

Nei M. 1973. Analysis of Gene Diversity in Subdivided Populations ((population structure/ genetic variability/heterozygosity/gene differentiation). *Proc. Nat. Acad. Sci. USA* **70(12)**(1): 3321-3323.

Nei M. 1983. Genetic distance and Molecular Phylogeny. 196-223.

Nei M. 1987. *Molecular evolutionary genetics*.

Nei M, Chesser RK. 1983. Estimation of fixation indices and gene diversities. *Annals of Human Genetics* **47**(3): 253-259.

Nei M, Gojobori T. 1986. Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions. *Mol. Biol. Evol.* **3(5)**: 418-426.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A* **76(10)**: 5269-5273.

Ness RW, Wright SI, Barrett SC. 2010. Mating-system variation, demographic history and patterns of nucleotide diversity in the Tristylous plant Eichhornia paniculata. *Genetics* **184**(2): 381-392.

Ng K, Stephen B. 2013. Deserts of Cyanobacteria from Extreme Cold Circadian Gene in Closely Related Species. *Appl. Environ. Microbiol* **79(5)**: 1516.

Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, Indap A, Bustamante CD, Clark AG. 2009. Darwinian and demographic forces affecting human protein coding genes. *Genome Res* **19**(5): 838-849.

Ning S-P, Xu L-B, Lu Y, Huang B-Z, Ge X-J. 2007. Genome composition and genetic diversity of Musa germplasm from China revealed by PCR-RFLP and SSR markers. *Scientia Horticulturae* **114**(4): 281-288.

Nordborg M. 2000. Linkage Disequilibrium, Gene Trees and Selfing: An Ancestral Recombination Graph With Partial Self-Fertilization. *Genetics Society of America*.

Nordborg M, Tina T Hu, Yoko Ishino, Jhaveri J, al CTe. 2005. The Pattern of Polymorphism in Arabidopsis thaliana. *PLoS Biol* **3**((7)e196).

Noyer J, Causse S, Tomekpe K, Bouet A, Baurens FC. 2005. A new image of plantain diversity assessed by SSR, AFLP and MSAP markers. *Genetica* **124**(1): 61-69.

Nybom H. 2004. Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Mol Ecol* **13**(5): 1143-1155.

Nyombi K. 2013. Towards sustainable highland banana production in Uganda:Opportunities and Challenges. *African Journal of Food, Agriculture, Nutrition and Development* **13(2)**

Oleksyk TK, Smith MW, O'Brien SJ. 2010. Genome-wide scans for footprints of natural selection. *Philos Trans R Soc Lond B Biol Sci* **365**(1537): 185-205.

Opara U, Jacobson D, Al-Saady NA. 2010. Analysis of genetic diversity in banana cultivars (Musa cvs.) from the South of Oman using AFLP markers and classification by phylogenetic, hierarchical clustering and principal component analyses. *Zhejiang Univ Sci B. 2010 41* **11(5)**: 332-341.

Ortiz R, Vuylsteke D. 1996. Recent advances in Musa genetics, breeding and Biotechnology. *Plant breeding abstracts* **66**: 1355-1363.

Osabe K, Clement JD, Bedon F, Pettolino FA, Ziolkowski L, Llewellyn DJ, Finnegan EJ, Wilson IW. 2014. Genetic and DNA methylation changes in cotton (Gossypium) genotypes and tissues. *PLoS One* **9**(1): e86049.

Pachuau L, Atom AD, Thangjam R. 2014. Genome classification of Musa cultivars from northeast India as revealed by ITS and IRAP markers. *Appl Biochem Biotechnol* **172**(8): 3939-3948.

Parisod C, Bonvin G. 2008. Fine-scale genetic structure and marginal processes in an expanding population of Biscutella laevigata L. (Brassicaceae). *Heredity (Edinb)* **101**(6): 536-542.

Parisod C, Christin P-A. 2008. Genome-wide association to fine-scale ecological heterogeneity within a continuous population of Biscutella laevigata (Brassicaceae). *New Phytologist* **178**(2): 436-447.

Paszkowski J, Grossniklaus U. 2011. Selected aspects of transgenerational epigenetic inheritance and resetting in plants. *Curr Opin Plant Biol* **14**(2): 195-203.

Paun O, Bateman RM, Fay MF, Hedren M, Civeyrel L, Chase MW. 2010. Stable epigenetic effects impact adaptation in allopolyploid orchids (Dactylorhiza: Orchidaceae). *Mol Biol Evol* **27**(11): 2465-2473.

Peakall R, Smouse P. 2006, 2012. GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research – an update. *Bioinformatics*.

Peakall R, Smouse P. 2012. GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research – an update. *Bioinformatics*.

Pellmyr O, Segraves KA, Althoff DM, Balcazar-Lara M, Leebens-Mack J. 2007. The phylogeny of yuccas. *Mol Phylogenet Evol* **43**(2): 493-501.

Pelsy F. 2010. Molecular and cellular mechanisms of diversity within grapevine varieties. *Heredity (Edinb)* **104**: 331-340.

Pérez-Figueroa A. 2013. msap: a tool for the statistical analysis of methylation-sensitive amplified polymorphism data. *Mol Ecol Resour* **13**(3): 522-527.

Perez-Figueroa A, Garcia-Pereira MJ, Saura M, Rolan-Alvarez E, Caballero A. 2010. Comparing three different methods to detect selective loci using dominant markers. *J Evol Biol* **23**(10): 2267-2276.

Perrier X, Bakry F, Carreel F, Jenny C, Horry J-P, Lebot V, Hippolyte I. 2009. Combining Biological approaches to shed light on evolution of edible bananas. *Ethnobotany Research & Applications* **7**: 199-216

Perrier X, De Langhe E, Donohue M, Lentfer C, Vrydaghs L, Bakry F, Carreel F, Hippolyte I, Horry J-P, Jenny C. 2011. Multidisciplinary perspectives on banana (Musa spp.) domestication. *Proceedings of the National Academy of Sciences* **108**(28): 11311-11318.

Petronis A. 2010. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* **465**(7299): 721-727.

Pfender WF, Saha MC, Johnson EA, Slabaugh MB. 2011. Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in Lolium perenne. *Theor Appl Genet* **122**(8): 1467-1480.

Pillay M, Ogundiwin E, Nwakanma DC, Ude G, Tenkouano A. 2001. Analysis of genetic diversity and relationships in East African banana germplasm. *Theor Appl Genet* **102**: 965-970.

Pillay M, Tenkouano A, Hartman J. 2002. Bananas and Plantains: Future Challenges in Musa Breeding. Chapter 8. In: Crop Improvement, Challenges in the Twenty-First Century. . *Food Products Press, New York*: 223-252.

Piry S, Luikart G, Cornuet J-M. 1999. Bottleneck: A Computer program for Dtecting Recent Reductions in the Effective Population size using Allele Frequency data. *Journal of Heredity* **90(4)**.

Pluess A, Cklin JRS. 2004. Population genetic diversity of the clonal plant geum reptans (Rosaceae) in the swiss alps1. *American Journal of Botany* **91(12)**: 2013-2021.

Poland J, Brown PJ, Sorrells ME, Jannink J-L. 2012. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLoS One* **7(2):**(e32253).

Poland J, Rife T. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. *Plant Gen.* **5**(3): 92-102.

Pongratz N, Sharbel TF, W.Beukeboom L, Michiels NK. 1988. Allozyme variability in sexual and parthenocarpic freshwater planarians:evidence for polyphyletic origin of parthenocarpic lineages through hybridization with coexisting sexual. *Heredity (Edinb)* **81**: 38-47.

Porth I, El-Kassaby Y. 2014. Assessment of the Genetic Diversity in Forest Tree Populations Using Molecular Markers. *Diversity* **6**(2): 283-295.

Pritchard J, Stephens M, Donnelly P. 2000. Inference of Population Structure Using Multilocus Genotype Data *Genetics* **155**: 945-959

Pritchard J, Wen X, Falush D. 2010. Documentation for structure software: Version 2.3. *Department of Statistics, University of Oxford*.

Puşcaş M, Taberlet P, Choler P. 2008. No positive correlation between species and genetic diversity in European alpine grasslands dominated byCarex curvula. *Diversity and Distributions* **14**(5): 852-861.

Rabinowicz PD, Citek R, Budiman MA, Nunberg A, Bedell JA, Lakey N, O'Shaughnessy AL, Nascimento LU, McCombie WR, Martienssen RA. 2005. Differential methylation of genes and repeats in land plants. *Genome Res* **15**(10): 1431-1440.

Rakshit S, Rakshit A, Matsumura H, Takahashi Y, Hasegawa Y, Ito A, Ishii T, Miyashita NT, Terauchi R. 2007. Large-scale DNA polymorphism study of Oryza sativa and O. rufipogon reveals the origin and divergence of Asian rice. *Theor Appl Genet* **114**(4): 731-743.

Ramos-Onsins S, Rozas J. 2002. Statistical Properties of New Neutrality Tests Against Population Growth. *Mol. Biol. Evol.* **19(12)**: 2092-2100.

Rao R, Hodgkin T. 2002. Genetic diversity and conservation and utilization of plant genetic resources. *Plant Cell, Tissue and Organ Culture* **68**: 1-19.

Rauf S, Jaime A, Tda TS, Khan A, Naveed A. 2010. Consequense of plant Breeding on Genetic Diversity. *International Journal of Plant breeding* **4**(1): 1-21.

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ESt. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci U S A* **98**(20): 11479-11484.

Resmi L, Kumari R, Bhat KV, AS N. 2011. Molecular Characterization of genetic diversity and structure in South ndian Musa cultivars. *International Journal of Botany)*: **7**(4): 274-282.

Reyna-Lopez GE, Simpson J, Ruiz-Herrera J. 1997. Differences in DNA methylation patterns are detectable during the dimorphic transition of fungi by amplification of restriction polymorphisms. *Mol Gen Genet* **253**(6): 703-710.

Reynolds J, Weir BS, cockerham CC. 1983. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics* **105**: 767-779.

Rhode C, Vervalle J, Bester-van der Merwe AE, Roodt-Wilding R. 2013. Detection of molecular signatures of selection at microsatellite loci in the South African abalone (Haliotis midae) using a population genomic approach. *Mar Genomics* **10**: 27-36.

Rice W. 1989. Analyzing Tables of Statistical test. *Evolution* **43**: 223-2235.

Richards C, Verhoeven KJF, Bossdorf O. 2012. Evolutionary Significance of Epigenetic Variation. *Plant Genome diversity* **1**: 16-20.

Richards E. 2011. Natural epigenetic variation in plant species: a view from the field. *Curr Opin Plant Biol* **14:**: 204-209.

Richards EJ. 2008. Population epigenetics. *Curr Opin Genet Dev* **18**(2): 221-226.

Riddle N, Richards EJ. 2002. The Control of Natural Variation in Cytosine Methylation in Arabidopsis. *Genetics* **162**: 355-363

Roa A, Chavarriaga-Aguirre P, Duque MC, Maya Ma, W.Bonierbale M, Iglesias C, Tohme J. 2000. CROSS-SPECIES AMPLIFICATION OF cassava (*Manihot esculenta*) euphorbiaceae)microsatellites: allelic polymorphism anddegree of relationship. *American Journal of Botany* **87(11)**: 1647-1655.

Robert CP, Chopin N, Rousseau J. 2009. Harold Jeffreys's Theory of Probability Revisited. *Statistical Science* **24**(2): 141-172.

Rohlf F. 2001. NTSYS-pc:Numeral Taxonomy and Multivariate Analysis system, version 2.10j. . *Exeter Publications, New York*.

Roldán-Ruiz I, van Euwijk FA, Gilliland TJ, Dubreuil P, Dillmann C, Lallemand J, De Loose M, Baril CP. 2001. A comparative study of molecular and morphological methods of describing relationships between perennial ryegrass (Lolium perenne L.) varieties. *Theoretical and Applied Genetics* **103**(8): 1138-1150.

Roth C, Liberles DA. 2006. A systematic search for positive selection in higher plants (Embryophytes). *BMC Plant Biol* **6**: 12.

Rout G, Senapati SK, Aparajita S, Palai SK. 2009. Studies on genetic identification and genetic fidelity of cultivated banana using ISSR markers. *Plant Omics Journal Southern Cross Journals* **2(6)**: 250-258

Rozas J, Rozas R. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**(2): 174-175.

Ruiz-Garcia L, Cervera MT, Martinez-Zapater JM. 2005. DNA methylation increases throughout Arabidopsis development. *Planta* **222**: 301-306.

Sadri R, Hornsby PJ. 1996. Rapid analysis of DNA methylation using new restriction enzyme sites created by bisulfite modification. *Nucleic Acids Res* **Vol. 24, No. 24**: 5058-5059.

Sainudiin R, Durrett RT, Aquadro CF, Nielsen R. 2004. Microsatellite mutation models: insights from a comparison of humans and chimpanzees. *Genetics* **168**(1): 383-395.

Salmon A, Clotault J, Jenczewski E, Chable V, Manzanares-Dauleux MJ. 2008. Brassica oleracea displays a high level of DNA methylation polymorphism. *Plant Science* **174**(1): 61-70.

Samarasinghe W, A.L.T. Perera, I.P. Wickramasinghe, S. Rajapakse. 2010. Morphological and Molecular Characterization of Musa Germplasm in Sri Lanka and Selection of Superior Genotypes: Proc. IC on Banana & Plantain in Africa Eds.: T. Dubois et al. . *Acta Hort* **879 ISHS**.

Saze H, Tsugane K, Kanno T, Nishimura T. 2012. DNA methylation in plants: relationship to small RNAs and histone modifications, and functions in transposon inactivation. *Plant Cell Physiol* **53**(5): 766-784.

Scarcelli N, Couderc M, Baco MN, Egah J, Vigouroux Y. 2013. Clonal diversity and estimation of relative clone age: application to agrobiodiversity of yam (Dioscorea rotundata). *Plant Biology* **13**(178): 1-10.

Schellenbaum P, Mohler V, Wenzel G, Walter B. 2008. Variation in DNA methylation patterns of grapevine somaclones (Vitis vinifera L.). *BMC Plant Biol* **8**: 78.

Schilling MP, Wolf PG, Duffy AM, Rai HS, Rowe CA, Richardson BA, Mock KE. 2014. Genotyping-by-sequencing for Populus population genomics: an assessment of genome sampling patterns and filtering approaches. *PLoS One* **9**(4): e95292.

Schmidt D, Pool J. 2002. The effect of population history on the distribution ofthe Tajima's D statistic

Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, Ecker JR. 2013. Patterns of population epigenomic diversity. *Nature* **495**(7440): 193-198.

Schob H, Grossniklaus U. 2006. The first high-resolution DNA "methylome". *Cell* **126**(6): 1025-1028.

Schoebel C, Stewart J, Gruenwald NJ, Rigling D, Prospero S. 2014. Population History and Pathways of Spread of the Plant Pathogen Phytophthora plurivora. *PLoS One* **9**(1): e85368.

Schrey AW, Coon CA, Grispo MT, Awad M, Imboma T, McCoy ED, Mushinsky HR, Richards CL, Martin LB. 2012. Epigenetic Variation May Compensate for Decreased Genetic Variation with Introductions: A Case Study Using House Sparrows (Passer domesticus) on Two Continents. *Genet Res Int* **2012**: 979751.

Scoville AG, Barnett LL, Bodbyl-Roels S, Kelly JK, Hileman LC. 2011. Differential regulation of a MYB transcription factor is correlated with transgenerational epigenetic inheritance of trichome density in Mimulus guttatus. *New Phytol* **191**(1): 251-263.

Sedlackova T, Repiska G, Celec P, Szemes T, Minarik G. 2013. Fragmentation of DNA affects the accuracy of the DNA quantitation by the commonly used methods. *Biological Procedures Online* **15**: 5-5.

Semagn K, Magorokosho C, Vivek BS, Makumbi D, Beyene Y, Mugo S, Prasanna BM, Warburton ML. 2012. Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers. *BMC Genomics* **13**: 113.

Senan S, Kizhakayil D, Sasikumar B, Sheeja TE. 2014. Methods for Development of Microsatellite Markers: An Overview *Not Sci Biol* **6(1)**: 1-13.

Shaibu A. 2012. Genetic diversity analysis of Musa species using amplified fragment length polymorphism and multivariate statistical technique. *International Journal of Biochemistry and Biotechnology* **1(6)**: 175-178.

Sharbel T, B H, T M-O. 2000. Genetic isolation by distance in Arabidopsis thaliana: biogeography and postglacial colonization of Europe. *Mol Ecol* **9**: 2109-2118

Sheffield VC, Stone EM, Carmi R. 1998. Use of isolated inbred human populations for identification of disease genes. *Trends Genet* **14**(10): 391-396.

Shepherd. 1999. *Cytogenetics of the genus Musa; International Network for the Improvement of Banana and Plantain, Montpellier, France.*

Shepherd K. 1957. A survey of major banana cultivars. *Florida State Horticultural Society*: 341-345.

Shifman S. 2003. Linkage disequilibrium patterns of the human genome across populations. *Human Molecular Genetics* **12**(7): 771-776.

Shifman S, Darvasi A. 2001. The value of isolated populations. *Nat Genet* **28**(4): 309-310.

Shreve SM, Mockford EL, Johnson KP. 2011. Elevated genetic diversity of mitochondrial genes in asexual populations of Bark Lice ('Psocoptera': Echmepteryx hageni). *Mol Ecol* **20**(21): 4433-4451.

Simko I, Haynes KG, Jones RW. 2006. Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. *Genetics* **173**(4): 2237-2245.

Simmonds NW. 1966. *Bananas*: Longmans.

Simmonds NW, Shepherd K. 1955. The taxonomy and origins of the cultivated bananas. *Journal of the Linnean Society of London, Botany* **55**(359): 302-312.

Skinner MK, Guerrero-Bosagna C. 2014. Role of CpG deserts in the epigenetic transgenerational inheritance of differential DNA methylation regions. *BMC Genomics* **15**: 692.

Slatkin M. 1985. Gene flow in natural populations. *Ann. Rev. Ecol. Syst.* **16**: 393-430.

Slatkin M, Hudson RR. 1991. Pairwise Comparisons of Mitochondrial DNA Sequences in Stable and Exponentially Growing Populations. *Genetics* **129(2)**: 555-562.

Smale M. 2006. Assessing the impact of crop genetic improvement in sub-Saharan Africa: Research context and highlights. In: Melinda's., Edmeades, S., and De Groote (Eds.). Promising Crop technologies for smallholder farmers in East Africa: Bananas and Maize. *Genetic Resources Policies Briefs 19-2006*.

Smith L 2008. "GMOs – A Crop Technology Whose Time Has Come." Fleishman and Hillard. June 11, 2008.

Smith M, Lautenberger JA, Shin HD, Chretien J-P, Shrestha S, Gilbert DA, O'Brien SJ. 2001. Markers for Mapping by Admixture Linkage Disequilibrium in African American and Hispanic Populations. *Am. J. Hum. Genet.* **69**: 1080-1094.

Sneath P, RR S. 1973. Numerical taxonomy: the principles and practice of numerical classification. *Medical Research Council Microbial Systematics* **1973**: 573

Sokal RR, Michener CD. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* **28**: 1409-1438.

Sonah H, Bastien M, Iquira E, Tardivel A, Legare G, Boyle B, Normandeau E, Laroche J, Larose S, Jean M, Belzile F. 2013. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One* **8**(1): e54603.

Soppe W, Jacobsen SE, Alonso-Blanco C, Jackson kJP, Kakutani T, Koornneef M, Peeters AJM. 2000. The Late Flowering Phenotype of fwa Mutants Is Caused by Gain-of-Function Epigenetic Alleles of a Homeodomain Gene. *Molecular Cell* **6**: 791-802.

Soto-Cerda BJ, Cloutier S. 2013. Outlier Loci and Selection Signatures of Simple Sequence Repeats (SSRs) in Flax (L.). *Plant Mol Biol Report* **31**: 978-990.

Spencer CC, J EN, Leberg PL. 2000. Bottleneck -Spencer et al 2000-Microsatellites. *molecular ecologyl* **9**: 1517-1528.

Ssebuliba R, Rubaihayo R, Tenkouano A, Makumbi D, D T, Magambo M. 2005. Genetic diversity among East African Highland bananas for female fertility. *African Crop Science Journal* **13(1)**: 13-26.

Ssebuliba R, Talengera D, Makumbi D, Namanya P, Tenkouano A, Tushemereirwe W, Pillay M. 2006. Reproductive efficiency and breeding potential of East African highland (Musa AAA-EA) bananas. *Field Crops Research* **95**(2-3): 250-255.

SSI-Review. 2014. Jason Potts, IISD Geneva, January 31, 2014:The State of Sustainability Initiatives Review 2014:Standards and the Green Economy.

Stapley J, Reger J, Feulner PG, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J. 2010. Adaptation genomics: the next generation. *Trends Ecol Evol* **25**(12): 705-712.

Stewart JE, Thomas KA, Lawrence CB, Dang H, Pryor BM, Timmer LM, Peever TL. 2013. Signatures of recombination in clonal lineages of the

citrus brown spot pathogen, Alternaria alternata sensu lato. *Phytopathology* **103**(7): 741-749.

Stinchcombe JR, Hoekstra HE. 2008. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity (Edinb)* **100**(2): 158-170.

Suttada R, Klaus E, Max-Bernhard S, Benchamas S, Jessada D, Kamnoon K. 2007. Molecular phylogeny of banana cultivars from Thailand based on HAT-RAPD markers. *Genetic Resources and Crop Evolution* **54**(7): 1565-1572.

Syvänen A. 2001. Accessing genetic variation:Genotyping Single Nucleotide Polymorphisms. *Nat. Rev. Genet* **2** 930 - 942.

Tajima F. 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123**: 585-595.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**(10): 2731-2739.

Teixeira H, Rodriguez-Echeverria S, Nabais C. 2014. Genetic diversity and differentiation of Juniperus thurifera in Spain and Morocco as determined by SSR. *PLoS One* **9**(2): e88996.

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). *Proceedings of the National Academy of Sciences* **98**(16): 9161-9166.

Tenkouano A, Crouch JH, Crouch HK, Ortiz R. 1998, 1999. Genetic diversity, hybrid performance, and combining ability for yield in Musa germplasm. . *Euphytica,* **102**: 281-288.

Tenkouano A, Swennen R. 2004. Plantains and banana: progress in breeding and delivering improved plantain and banana to African farmers. . *Chronica Horticulturae* **44(1)**: 9-15.

Terwilliger J, Zöllnerc S, Laanc M, Pääboc S. 1998. Mapping Genes through the Use of Linkage Disequilibrium Generated by Genetic Drift: 'Drift Mapping' in Small Populations with No Demographic Expansion. *Hum Hered* **48**: 138-154.

Thioulouse J, D C, S D, J.M O. 1997. ADE-4: a multivariate analysis and graphical display software *Statistics and Computing,* **7(1)**: 75-83.

Till BJ, Jankowicz-Cieslak J, Sagi L, Huynh OA, Utsushi H, Swennen R, Terauchi R, Mba C. 2010. Discovery of nucleotide polymorphisms in

the Musa gene pool by Ecotilling. *Theor Appl Genet* **121**(7): 1381-1389.

Trethowan RM, Mujeeb-Kazi A. 2008. Novel Germplasm Resources for Improving Environmental Stress Tolerance of Hexaploid Wheat *Crop Sci.* **48**(4): 1255-1265.

Tsaftaris A, N.Polidoros A, Koumproglou R, Tani E, Kovacevic N, Abatzidou E. 2005. Epigenetic mechanisms in plants and their implications in plant breeding; Tuberosa R., Phillips R.L., Gale M. (eds.), Proceedings of the International Congress "In the Wake of the Double Helix: From the Green Revolution to the Gene Revolution". *Avenue media, Bologna,*: 157-171.

Tsaftaris AS, Polidoros AN. 2000. DNA Methylation and Plant Breeding. *Plant Breeding Reviews* **18**: 87-176.

Tugume, Lubega GW, Rubaihayo PR. 2003. Genetic diversity of East African Highland bananas. *Taxonomy* **Vol 11** (N° 2).

Tugume A, Lubega GW, Rubaihayo PR. 2002. Genetic diversity of East African Highland bananas using AFLP. *The International Magazine on Banana and Plantain* **11(2)**: 28-32.

Ude G, Pillay M, Nwakanma D, Tenkouano A. 2002a. Analysis of genetic diversity and sectional relationships in Musa using AFLPs. *Theor Appl Genet* **104**: 1239-1245.

Ude G, Pillay M, Nwakanma D, Tenkouano A. 2002b. Genetic Diversity in Musa acuminata Colla and Musa balbisiana Colla and some of their natural hybrids using AFLP Markers. *Theor Appl Genet* **104**: 1246-1252.

Ude G, Pillay M, Ogundiwin E, Tenkouano A. 2003. Genetic diversity in an African plantain core collection using AFLP and RAPD markers. *Theor Appl Genet* **107**(2): 248-255.

Vandepitte K, Roldan-Ruiz I, Jacquemyn H, Honnay O. 2010. Extremely low genotypic diversity and sexual reproduction in isolated populations of the self-incompatible lily-of-the-valley (Convallaria majalis) and the role of the local forest environment. *Ann Bot* **105**(5): 769-776.

Vekemans X. 2002. A program for genetic diversity analysis with AFLP (and RAPD) population data.

Vekemans X, Beauwens T, Lemaire M R-R. 2002. Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. *Mol Ecol.* **11(1):**139-151.

Venkatachalam L, Sreedhar RV, Bhagyalakshmi N. 2008. The use of genetic markers for detecting DNA polymorphism, genotype identification and phylogenetic relationships among banana cultivars. *Molecular Phylogenetics and Evolution* **47**(3): 974-985.

Venkatachalam L, Venkataramareddy SR, Neelwarne B. 2007. Molecular analysis of genetic stability in long-term micropropagated shoots of banana using RAPD and ISSR markers. *Electronic Journal of Biotechnology* **10(1)** (January 15, 2007).

Vergeer P, Wagemaker NC, Ouborg NJ. 2012. Evidence for an epigenetic role in inbreeding depression. *Biol Lett* **8**(5): 798-801.

Vongs A, Kakutani T, Martienssen R, Richards E. 1993. Arabidopsis thaliana DNA methylation mutants. *Science* **260**(5116): 1926-1928.

Vos P, Bleeker RHM, Reijans M, Lee Tvd, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M. 1995. AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Research* **23 (21)**: 4407-4414.

Vuylsteke D, Ortiz R. 1995. Plantain-derived diploid hybrids (TMP2x) with black sigatoka resistance. *HORTSCIENCE* **30**: 147-149.

Walsh B. 2007. Using molecular markers for detecting domestication, improvement, and adaptation genes. *Euphytica* **161**(1-2): 1-17.

Wang CG, Gu Y, Chen CB, Jiao DL, Xue ZY, Song WQ. 2009. Analysis of the level and pattern of genomic DNA methylation in different ploidy watermelons by MSAP (Citrullus lanatus). *Fen Zi Xi Bao Sheng Wu Xue Bao* **42**(2): 118-126.

Wang M, Jiang N, Jia T, Leach L, Cockram J, Comadran J, Shaw P, Waugh R, Luo Z. 2012. Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. *Theor Appl Genet* **124**(2): 233-246.

Wang N, Zhang D, Wang Z, Xun H, Ma J, Wang H, Huang W, Liu Y, Lin X, Li N, Ou X, Zhang C, Wang M-B, Liu B. 2014. Mutation of the RDR1 gene caused genome-wide changes in gene expression, regional variation in small RNA clusters and localized alteration in DNA methylation in rice *BMC Plant Biol* **14:177**.

Wang T, Chen G, Zan Q, Wang C, Su YJ. 2012. AFLP genome scan to detect genetic structure and candidate loci under selection for local adaptation of the invasive weed Mikania micrantha. *PLoS One* **7**(7): e41310.

Wang X-L, Chiang T-Y, Roux N, Hao G, Ge X-J. 2007. Genetic diversity of wild banana (Musa balbisiana Colla) in China as revealed by AFLP markers. *Genetic Resources and Crop Evolution* **54**(5): 1125-1132.

Watt F, Molloy PL. 1988. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes & Development* **2**(9): 1136-1143.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**(2): 256-276.

Wendel J. 2000. Genome evolution in polyploids. *Plant Mol Biol* **42**: 225-249.

Whitlock R, Hipperson H, Mannarelli M, Butlin RK, Burke T. 2008. An objective, rapid and reproducible method for scoring AFLP peak-height data that minimizes genotyping error. *Mol Ecol Resour* **8**(4): 725-735.

Wong C. 2001. Genetic Diversity of the Wild Banana Musa acuminata Colla in Malaysia as Evidenced by AFLP. *Annals of Botany* **88**(6): 1017-1025.

Wong C, Kiew R, Argent G, Set O, Lee SK, Gan YY. 2002. Assessment of the valididty of the sections in Musa (Musaceae) using AFLP. *Annals of Botany* **90**: 231-238.

Wright SI, Ness RW, Foxe JP, Barrett SCH. 2008. Genomic Consequences of Outcrossing and Selfing in Plants. *International Journal of Plant Sciences* **169**(1): 105-118.

Xiao J, Song C, Liu S, Tao M, Hu J, Wang J, Liu W, Zeng M, Liu Y. 2013. DNA methylation analysis of allotetraploid hybrids of red crucian carp (Carassius auratus red var.) and common carp (Cyprinus carpio L.). *PLoS One* **8**(2): e56409.

Xiong L, Xu CG, Maroof MAS, Zhang Q. 1999. Patterns of cytosine methylation in an elite rice hybrid and its parental lines, detected by a methylation-sensitive ampli®cation polymorphism technique. *Mol Gen Genet* **261**: 439-446.

Xiong Z, Laird PW. 1997. COBRA: a sensitive and quantitative DNA methylation assay. *Nucleic Acids Res* **25**(12): 2532-2534.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *TREE* **15(12)**: 496-502.

Yares K. 2007. "What Country Consumes the Most Bananas?" eHow Articles and Online News. April 2007. Updated January, 2009.

Youssef M, James AC, Rivera-Madrid R, Ortiz R, Escobedo-GraciaMedrano RM. 2011. Musa genetic diversity revealed by SRAP and AFLP. *Mol Biotechnol* **47**(3): 189-199.

Zhang C, Hsieh T-F. 2013. Heritable Epigenetic Variation and its Potential Applications for Crop Improvement. *Plant Breeding and Biotechnology* **1**(4): 307-319.

Zhang H, Hare M. 2012. Identifying and reducing AFLP genotyping error: an example of tradeoffs when comparing population structure in broadcast spawning versus brooding oysters. *Heredity (Edinb)* **108**: 616-625.

Zhang M, Xu C, von Wettstein D, Liu B. 2011. Tissue-specific differences in cytosine methylation and their association with differential gene expression in sorghum. *Plant Physiol* **156**(4): 1955-1966.

Zhang X, Su D, Ma L. 2007. Analysis of Genetic Diversity in Buffalograss Determined by Random Amplified Polymorphic DNA Markers *HORTSCIENCE* **42(3)**: 474-477.

Zhang Y, Liu Z, Liu C, Yang Z, Deng K, Peng J, Zhou J, Li G, Tang Z, Ren Z. 2008. Analysis of DNA methylation variation in wheat genetic background after alien chromatin introduction based on methylation-sensitive amplification polymorphism. *Chinese Science Bulletin* **53**(1): 58-69.

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**(4): 355-360.

Zhao J, Zhang JS, Wang Y, Wang RG, Wu C, Fan LJ, Ren XL. 2011. DNA methylation polymorphism in flue-cured tobacco and candidate markers for tobacco mosaic virus resistance. *J Zhejiang Univ Sci B* **12**(11): 935-942.

Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR. 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nat Commun* **2**: 467.

Zhao X, Chai Y, Liu B. 2007. Epigenetic inheritance and variation of DNA methylation level and pattern in maize intra-specific hybrids. *Plant Science* **172**(5): 930-938.

Zhivotovsky L. 1999. Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology Notes* **8**: 907-913.

APPENDICES

Appendix 1: Names of cultivars, genomic classification and region of collection used in this study for Chapters 2, 3, 4 and 5 (shown by ticking).

S/No	Genotype	Genomic classification	Region of collection	Chapter 2- SSR	Chapter 3- AFLP	Chapter 4 -GBS	Chapter 5 MsAFLP-analysis
1	Muture_1	EAHB-AAA	Kenya	√	√	√	√
2	Musera	EAHB-AAA	Kenya	√	√	√	√
3	Itarecia	EAHB-AAA	Kenya	√	√		√
4	CN111	EAHB-AAA	Kenya	√	√	√	√
5	Mtagatu	EAHB-AAA	Kenya	√	√		√
6	Mplogoma	EAHB-AAA	Kenya	√	√	√	√
7	Mtahato	EAHB-AAA	Kenya	√	√		√
8	Mukoya	EAHB-AAA	Kenya	√	√		√
9	KBU2	EAHB-AAA	Kenya	√	√		√
10	Mukubu	EAHB-AAA	Kenya	√	√		√
11	Mtagato	EAHB-AAA	Kenya	√	√		√
12	Ishighame	EAHB-AAA	Kenya	√	√		√
13	Jamaga	EAHB-AAA	Kenya	√	√		√
14	GNgiant	EAHB-AAA	Kenya	√	√		√
15	Libukusu	EAHB-AAA	Kenya	√	√		√
16	Bukamba	EAHB-AAA	Kenya	√	√		√
17	Black Uganda green	EAHB-AAA	Kenya	√	√		√
18	White Uganda green	EAHB-AAA	Kenya	√	√		√
19	Nasirembe	EAHB-AAA	Kenya	√	√	√	√
20	Kiffuba	EAHB-AAA	Kenya	√	√		√
21	Nzimola	EAHB-AAA	Kenya	√	√	√	√
22	Enzingo_S(with spiral rachis)	EAHB-AAA	Kenya	√	√	√	√
23	Enzingo_N(with normal rachis)	EAHB-AAA	Kenya	√	√	√	√
24	Kaburut	EAHB-AAA	Kenya	√	√	√	√
25	Mbululu Nak	EAHB-AAA	Kenya	√	√	√	√
26	Enjuta	EAHB-AAA	Kenya	√	√		√
27	Nyarluo Ratong	EAHB-AAA	Kenya	√	√		√
28	Ekeganda	EAHB-AAA	Kenya	√	√		√
29	Mtama	EAHB-AAA	Kenya	√	√		√
30	Litambi	EAHB-AAA	Kenya	√	√	√	√
31	Turbo	EAHB-AAA	Kenya	√	√		√
32	Mrefu	EAHB-AAA	Kenya	√	√	√	√
33	Ngombe	EAHB-AAA	Kenya	√	√	√	√
34	Namukhila	EAHB-AAA	Kenya	√	√	√	√
35	Kikuyu_1	EAHB-AAA	Kenya	√	√	√	√
36	Liganda Lusumba	EAHB-AAA	Kenya	√	√	√	√
37	Sialamule	EAHB-AAA	Kenya	√	√	√	√
38	Ntobe	EAHB-AAA	Kenya	√	√	√	√
39	Nakabululu	EAHB-AAA	Kenya	√	√		√
40	White Nakabululu	EAHB-AAA	Kenya	√	√	√	√
41	Enzingo	EAHB-AAA	Kenya	√	√		√
42	Mpologoma	EAHB-AAA	Uganda	√	√	√	√
43	Luwata	EAHB-AAA	Uganda	√	√	√	√
44	Ngongo	EAHB-AAA	Uganda	√	√	√	√
45	Namayovu	EAHB-AAA	Uganda	√	√	√	√
46	Mukazi Alanda	EAHB-AAA	Uganda	√	√	√	√

47	Namunwe	EAHB-AAA	Uganda	√	√	√	√
48	Muvubo	EAHB-AAA	Uganda	√	√	√	√
49	Kisansa	EAHB-AAA	Uganda	√	√	√	√
50	Musakala	EAHB-AAA	Uganda	√	√	√	√
51	Enyoya	EAHB-AAA	Uganda	√	√		√
52	Nabuyobo	EAHB-AAA	Uganda	√	√	√	√
53	Nakitembe Red	EAHB-AAA	Uganda	√	√	√	√
54	Kibagampera	EAHB-AAA	Uganda	√	√	√	√
55	Mbwazirume	EAHB-AAA	Uganda	√	√	√	√
56	Nakitembe Nakawere	EAHB-AAA	Uganda	√	√	√	√
57	Nakitembe Nakamali	EAHB-AAA	Uganda	√	√	√	√
58	Namaliga	EAHB-AAA	Uganda	√	√	√	√
59	Bikowekowe	EAHB-AAA	Uganda	√	√	√	√
60	Luvuta	EAHB-AAA	Uganda	√	√	√	√
61	Entaragaza	EAHB-AAA	Uganda	√	√		√
62	Engagara	EAHB-AAA	Uganda	√	√	√	√
63	Kibuzi	EAHB-AAA	Uganda	√	√	√	√
64	Nakyetengu	EAHB-AAA	Uganda	√	√	√	√
65	Mukuba_Konde	EAHB-AAA	Uganda	√	√	√	√
66	Salalugazi	EAHB-AAA	Uganda	√	√		√
67	Butobe	EAHB-AAA	Uganda	√	√	√	√
68	Kaitabunyonyi	EAHB-AAA	Uganda	√	√	√	√
69	Nakasabira	EAHB-AAA	Uganda	√	√	√	√
70	Nakabululu	EAHB-AAA	Uganda	√	√	√	√
71	Kafunze	EAHB-AAA	Uganda	√	√		√
72	Kazirakwe	EAHB-AAA	Uganda	√	√	√	√
73	Ndibwabalangira	EAHB-AAA	Uganda	√	√	√	√
74	Rwambarara	EAHB-AAA	Uganda	√	√	√	√
75	Bitambi	EAHB-AAA	Uganda	√	√	√	√
76	Nfuuka	EAHB-AAA	Uganda	√	√	√	√
77	Lusumba	EAHB-AAA	Uganda	√	√		√
78	Ingarama	EAHB-AAA	Uganda	√	√	√	√
79	Namwezi	EAHB-AAA	Uganda	√	√	√	√
80	Kiffuba_Ug	EAHB-AAA	Uganda	√	√	√	√
81	Enyeru	EAHB-AAA	Uganda	√	√		√
82	Kulwoni	EAHB-AAA	Uganda	√	√	√	√
83	Endirira	EAHB-AAA	Uganda	√	√	√	√
84	Engumba ye embire	EAHB-AAA	Uganda	√	√	√	√
85	Namadhi	EAHB-AAA	Uganda	√	√	√	√
86	Nsowe	EAHB-AAA	Uganda	√	√	√	√
87	Nalukira	EAHB-AAA	Uganda	√	√	√	√
88	Nalwezinga	EAHB-AAA	Uganda	√	√	√	√
89	Ensansa	EAHB-AAA	Uganda	√	√	√	√
90	Oruhuna	EAHB-AAA	Uganda	√	√	√	√
91	MunjuP	unknown	Kenya	√	√		√
92	Spambia (4)	Plantain (AAB)	Kenya	√	√		√
93	Spambia (6)	Plantain (AAB)	Kenya	√	√		√
94	Spambia (7)	Plantain (AAB)	Kenya	√	√		√
95	Somatic green	AAA_dessert	Kenya	√	√	√	√
96	Red green	AAA_dessert	Kenya	√	√		√
97	Bukomo	EAHB-AAA	Uganda			√	
98	Enyabakazi_Green	EAHB-AAA	Uganda			√	
99	Enyabakazi_Red	EAHB-AAA	Uganda			√	
100	Enyamashari	EAHB-AAA	Uganda			√	
101	Enzirabushera	EAHB-AAA	Uganda			√	
102	Kaitabunyonyi	EAHB-AAA	Uganda			√	
103	Kibidebidde	EAHB-AAA	Uganda			√	

104	Kibungo	EAHB-AAA	Uganda	√
105	Lumenyamagali	EAHB-AAA	Uganda	√
106	Luwuna	EAHB-AAA	Uganda	√
107	Lwefusa	EAHB-AAA	Uganda	√
108	Mbirambire	EAHB-AAA	Uganda	√
109	Mukazi_mugumba	EAHB-AAA	Uganda	√
110	Murure	EAHB-AAA	Uganda	√
111	Muvubo_Variant	EAHB-AAA	Uganda	√
112	Nakabinyi	EAHB-AAA	Uganda	√
113	Nakayonga	EAHB-AAA	Uganda	√
114	Nakhaki	EAHB-AAA	Uganda	√
115	Nakibule	EAHB-AAA	Uganda	√
116	Nakinyika	EAHB-AAA	Uganda	√
117	Nakitembe	EAHB-AAA	Uganda	√
118	Nakitembe_Omunyoro	EAHB-AAA	Uganda	√
119	Nakyetengu_Tall	EAHB-AAA	Uganda	√
120	Namulondo	EAHB-AAA	Uganda	√
121	Nante	EAHB-AAA	Uganda	√
122	Nasala	EAHB-AAA	Uganda	√
123	Nshule	EAHB-AAA	Uganda	√
124	Nusu_Ngombe	EAHB-AAA	Kenya	√
125	Rugondo	EAHB-AAA	Uganda	√
126	Siira	EAHB-AAA	Uganda	√
127	Calcutta	Wild-AA	Uganda	√
128	Zebrina	Wild-AA	Uganda	√
129	Ornata	Wild-AA	Uganda	√

Appendix 2: CTAB protocol for *Musa* DNA extraction using the genogrinder strip tubes. A combined and modified protocol based on the methods of Mace *et al.* (2006) and Dellaporta *et al.* (1989).

1) Dispense 2-steel metal balls into each strip tube before putting the leaf tissue. Put approx. 0.005g of freeze dried leaf tissue or 100-150 mg fresh tissue / frozen young tender cigar leaf in strip tubes and submerge into a bucket with liquid nitrogen (do not let the leaves thaw).
2) Place your plates onto the genogrinder and make sure they are well balanced and grind into fine powder by setting the genogrinder at full (1x) speed of 500 strokes/min for 2 minutes. Remove samples, exchange the position of samples (outer samples towards inner and vice versa), (for fresh samples-dip in liquid nitrogen), and grind for additional 1 minute. Fresh tissue samples may need longer grinding than lyophilized. [Note: Dipping in liquid nitrogen help to grind the samples into fine powders when samples are not freeze dried.] Spin down tubes until the centrifuge reaches about 1500 RPM to bring the ground tissue the bottom of the tube. Longer centrifugation makes dispersion difficult after adding the extraction buffer.

3) Add 600 µl of freshly prepared modified CTAB extraction and grind for about 30 seconds. The time can be extended to 1-2 minutes if the tissues have not properly ground in step 1. Grinding after addition of extraction buffer serves to grind better (if they have not ground properly in step 1) or to disperse/homogenize the powder/tissue with the extraction buffer. **[Note: prolonged grinding causes DNA degradation].**
4) Incubate the samples at 65°C water bath for 30 minutes with continuous gentle rocking (***set the RPM to 20 to 30; high rotation will result degraded DNA***). Invert or gently tap tubes once in every 10 minutes to properly homogenize the tissue with extraction buffer (***be cautious not to splash the buffer while inverting***).
5) Remove tubes from the water bath and allow them to cool for 5-10 min in a fume hood. Gently mix or tap samples and centrifuge at 3500 rpm for 10 min.
6) Transfer the aqueous phase into new tubes and add 400µl chloroform:isoamylalcohol (24:1) in to the side of the tubes. Mix very gently by gently inverting the tubes for 1 -5 minutes (or about 20 times). ***DNA is very fragile so any stronger force can cause degradation***.
7) Centrifuge at 3500 rpm for 10 min
8) Transfer the upper aqueous layer to fresh strip tubes and **repeat the chloroform: isoamylalcohol** *(step 6 and 7)***.**
9) Transfer the upper aqueous layer into fresh strip tubes. ***Do not transfer the layer containing chloroform (any trace transfer of chloroform will affect PCR).*** *If the aqueous phase of the sample look dirty, repeat chloroform: isoamylalcohol wash for the third time.*
10) Add 500 µl 100% cold (stored at -20 ^{o}C) isopropanol (2-propanol) and mix very gently for about 5 min (or gently invert for about 50 times) to precipitate the nucleic acid. **Optional step:** *Keep tubes in the freezer (-20°C) for about 60 minutes, take out tube from the freezer and gently invert tubes for 2-3 minutes, leave the tubes on the bench for about 10 minutes and again gently invert tubes until you see whitish floating stuff.* (For higher yields keep overnight)

11) Centrifuge at 3500 rpm for 25 min to form a pellet at the bottom of the tube. Discard the supernatant. [Centrifugation while the tubes are still very cold will result either to very small pellet or no pellet at all].
12) Add 300 µl of 70% ethanol; flap the tubes gently to let the pellet float for ease in washing (you can also vortex the tubes for 15-20 seconds to let the pellets float for washing). Centrifuge for 10 min at 3500 rpm and discard ethanol by decantation.
13) Wash the pellet with 70% ethanol once again. Centrifuge for 10 min and discard ethanol by decantation again.
14) Allow pellet to air dry in hood (or using the 37 ^{o}C incubator) until ethanol evaporates completely (until smell for ethanol disappears). This takes 30-60 minutes. DON'T OVER DRY PELLET AS IT WILL BE DIFFICULT TO DISSOLVE IT. Any remaining alcohol smell indicates pellet is not completely dry.
15) Add 200µl of Tris-EDTA (T.E) buffer and digest RNA by adding 4 µl of 10µg/ml RNase and incubate at 37^{o}C (or at room temperature for at least 1 hour).
16) Add 20 µl 3M Sodium acetate followed by 400 µl of cold 99% ethanol and incubate at -20^{o}C for 30 mins -1 hr. Centrifuge mixture at 3500 rpm for 10 minutes. A white precipitate is observed at the bottom of the tube.
17) Decant the supernatant (RNA), wash the pellet twice with 200 µl of 70% ethanol and allow to dry briefly at room temperature.
18) Dissolve DNA in 100µl of T.E buffer (or distilled deionized water) to the pellet to dissolve the DNA.
19) Check DNA quality using 0.8% agarose gel and purity using the nanadrop spectrophotometer

DNA is stored in fridge at 4^{o}C for a short time but for long time storage keep it at -20^{o}C and below

Buffers and solutions

Buffers and solutions

1. Liquid nitrogen
2. β-mercaptoethanol

3. CTAB buffer: 2%CTAB (Cetyltrimethylammonium bromide)
   1.4M NaCl
   100mM Tris-HCl, pH 8.0
   20mM EDTA (Sterilize by autoclaving)
4. Chloroform: Iso-Amyl alcohol (24:1)
5. Isopropanol
6. TE buffer 10mM Tris-HCl
   1.0mM EDTA, pH 8.0
7. RNase A 10µg/ml
8. 3M Ammonium acetate/Sodium acetate, pH 6.8
9. 70% Ethanol
10. Absolute Ethanol

Appendix 3: Samples from Sedusu IITA for DNA methylation heritability study used in chapter 6. Sexual families are represented by S/No 1-59 and vegetative clones are represented by S/No 60-68. The number of 1st cycle offspring in the vegetative clone families is represented by a superscript on each cultivar names

S/No	Genotype name	Fparent	Mparent	Ploidy	Description	Bunch character
1	27770S-20	1201K-1	C.V rose	3x	hybrid	inferior bunch
2	27770S-4	1201K-1	C.V rose	3x	hybrid	good bunch size
3	27935S-1	1201K-1	C.V rose	3x	hybrid	good bunch size
4	28036S-11	1201K-1	C.V rose	3x	hybrid	inferior bunch
5	28036S-2	1201K-1	C.V rose	3x	hybrid	good bunch size
6	28246S-7	1201K-1	C.V rose	3x	hybrid	good bunch size
7	27935S-7	1201K-1	C.V rose	3x	hybrid	inferior bunch
8	26337S-11	1201K-1	SH-3217	3x	Hybrid	good bunch size
9	12419S-13	1201K-1	SH-3217	3x	hybrid	good bunch size
10	26337S-2	1201K-1	SH-3217	3x	hybrid	good bunch size
11	26337S-39	1201K-1	SH-3217	3x	hybrid	inferior bunch
12	26337S-43	1201K-1	SH-3217	3x	hybrid	good bunch size
13	28263S-2	1201k-1	SH-3217	3x	hybrid	good bunch size
14	27914S-1	1438K-1	C.V rose	3x	hybrid	good bunch size
15	27914S-13	1438K-1	C.V rose	3x	hybrid	good bunch size
16	28095S-1	1438K-1	C.V rose	3x	hybrid	inferior bunch
17	27264S-2	1438K-1	C.V rose	3x	hybrid	inferior bunch
18	27914S-24	1438K-1	C.V rose	3x	hybrid	good bunch size
19	25066S-1	1438K-1	Kokopo	3x	Hybrid	inferior bunch
20	25474S-1	1438K-1	Kokopo	3x	hybrid	good bunch size
21	26369S-4	1438K-1	Long tavoy	3x	hybrid	inferior bunch
22	28481S-1	1438K-1	Malaccensis	3x	hybrid	inferior bunch
23	28561S-2	1438K-1	Malaccensis	3x	hybrid	inferior bunch
24	26725S-1	1438K-1	SH-3362	3x	hybrid	good bunch size
25	25499S-7	1438K-1	SH-3142	3x	hybrid	good bunch size
26	26039S-2	1438K-1	SH-3217	3x	hybrid	inferior bunch
27	24583S-2	660K-1	5610S-1	3x	hybrid	inferior bunch

28	26260S-3	660K-1	5610S-1	3x	hybrid	good bunch size
29	13284S-1	660K-1	9128-3	3x	hybrid	good bunch size
30	25371S-2	660K-1	9128-3	3x	hybrid	good bunch size
31	9187S-8	660K-1	9128-3	3x	hybrid	good bunch size
32	26709S-1	660K-1	Calcutta 4	3x	hybrid	inferior bunch
33	27713S-1	660K-1	Malaccensis	3x	hybrid	good bunch size
34	27825S-4	660K-1	Malaccensis	3x	hybrid	good bunch size
35	27873S-18	660K-1	Malaccensis	3x	hybrid	inferior bunch
36	27873S-38	660K-1	Malaccensis	3x	hybrid	good bunch size
37	27873S-4	660K-1	Malaccensis	3x	hybrid	inferior bunch
38	27873S-5	660K-1	Malaccensis	3x	hybrid	inferior bunch
39	28188S-2	660K-1	Malaccensis	3x	hybrid	inferior bunch
40	25117S-2	917K-2	5610S-1	3x	hybrid	good bunch size
41	26815S-9	917K-2	5610S-1	3x	hybrid	inferior bunch
42	26990S-10	917K-2	5610S-1	3x	hybrid	inferior bunch
43	26990S-11	917K-2	5610S-1	3x	hybrid	good bunch size
44	26990S-4	917K-2	5610S-1	3x	hybrid	inferior bunch
45	27073S-1	917K-2	5610S-1	3x	hybrid	good bunch size
46	27744S-1	917K-2	5610S-1	3x	hybrid	good bunch size
47	27261S-1	917K-2	Malaccensis	3x	hybrid	inferior bunch
48	27334S-5	917K-2	Malaccensis	3x	hybrid	inferior bunch
49	27886S-5	917K-2	Malaccensis	3x	hybrid	good bunch size
50	28033S-3	917K-2	Malaccensis	3x	hybrid	good bunch size
51	28257S-2	917K-2	Malaccensis	3x	hybrid	good bunch size
52	28780S-1	917K-2	Malaccensis	3x	hybrid	inferior bunch
[a]53	1438K-1	Entukura	Calcutta 4	4x	parent	good bunch size
[b]54	660K-1	Enzirabahima	Calcutta 4	4x	parent	good bunch size
[c]55	917K-2	Enzirabahima	Calcutta 4	4x	parent	good bunch size
[d]56	1201K-1	Nakawere	Calcutta 4	4x	parent	good bunch size
57	Calcutta4			2x	parent	inferior bunch
58	Entukura			3x	parent	good bunch size
59	Enzirabahima			3x	parent	good bunch size
60	Namunwe [3]			3x	Clone-Family1	good bunch size
61	Rugondo [2]			3x	clone-Family2	good bunch size
62	Enzirabushera [3]			3x	clone-Family3	good bunch size
63	Nsowe [2]			3x	clone-Family4	good bunch size
64	Nakhaki[4]			3x	clone-Family5	good bunch size
65	Nante[2]			3x	clone-Family6	good bunch size
66	Nalukira [3]			3x	clone-Family7	good bunch size
67	Lwefusa[3]			3x	clone-Family8	good bunch size
68	Nakibule[3]			3x	clone-Family9	good bunch size
69	Zebrina GF			2x	out-group	inferior bunch
70	Banksii type Madang			2x	out-group	inferior bunch

[a]1438K-1 Entukura (female) x Calcutta 4 (male)
[b]660K-1 Enzirabahima (female) x Calcutta 4 (male)
[c]917K-2 Enzirabahima (female) x Calcutta 4 (male)
[d]1201K-1 Nakawere (female) x Calcutta 4 (male)