Palacký University Olomouc

Faculty of Science

Department of Cell Biology and Genetics

Moses Nyine

Genomic Selection to Accelerate Banana Breeding: Genotyping by Sequencing of Banana Hybrids.

PhD Thesis

2018

Title: Genomic Selection to Accelerate Banana Breeding: Genotyping by Sequencing of Banana Hybrids.

Keywords: allele dosage SNP, banana, genotyping by sequencing (GBS), genomic estimated breeding value (GEBV), genomic prediction, genomic selection**,** polyploid, predictive ability, single nucleotide polymorphism (SNP)

PhD Thesis: Palacký University Olomouc, Czech Republic, January 2018

By: Mr. Moses Nyine
Supervisor: Prof. Ing. Jaroslav Doležel, DrSc.
Co-supervisor: Prof. Rony Swennen, PhD
Co-supervisor: Dr. Brigitte Uwimana, PhD
Co-supervisor: Dr. Allan Brown, PhD

Acknowledgements

I would like to extend my sincere gratitude to all my supervisors, Prof. Ing. Jaroslav Doležel, DrSc., Prof. Rony Swennen, Dr. Brigitte Uwimana and Dr. Allan Brown for their insights and guidance during the four years of my PhD study. I thank Dr. Jim Lorenzen for introducing this PhD concept to me and for finding wonderful mentors listed above. His advices and comments from Prof. Rodomiro Ortiz (SLU, Sweden) were very useful during the course of this study. Dr. Manpreet S. Katari (New York University, USA), and Dr. Steve D. Rounsley (Genus plc, USA) were very helpful in bioinformatics. I am grateful to Dr. Ismail Y. Rabbi (IITA, Nigeria) for his generosity in coordinating the genotyping work with the Institute of Genomic Diversity, Cornell, USA. Some of the materials used in the study were provided by the National Agricultural Research Organization (NARO), Uganda. It was a great opportunity to be affiliated to both the International Institute of Tropical Agriculture (IITA), Uganda and the Institute of Experimental Botany (IEB), Olomouc, Czech Republic. I thank Dr. Eva Hřibová, Dr. Pavla Christelová, Dr. Nicolas Blavet, MSc. Štěpán Stočes, Dr. Jana Čížková and Mrs. Marie Seifertová from IEB for their technical assistance in laboratory procedures and data analysis. In addition, I thank all colleagues at IEB for providing a friendly environment that enabled me to enjoy my stay in Czech Republic. Similarly, the contribution made by Mr. Michael Batte, Ms. Violet Akech, Ms. Catherine Nasuuna, Ms. Harriet Nabatanzi, Ms. Catherine Nakayenga, Ms. Justine Tumushabe, Ms. Helena Vanrespaille, Mrs. Christine Kivumbi and all the field assistants of IITA during phenotyping is highly appreciated. Finally, special thanks to my family and friends for their unconditional love and moral support.

Declaration

I hereby declare that the work presented in this PhD Thesis is original and has not been presented by any student in any academic institution for an academic award except for the literature cited. The Thesis has been written under the supervision of Prof. Ing. Jaroslav Doležel, DrSc., and co-supervision of Prof. Rony Swennen, Dr. Brigitte Uwimana and Dr. Allan Brown.

…………………………………...

Funding Acknowledgement

This work was supported by donors through their contributions to the CGIAR Fund (http://www.cgiar.org/who-we-are/cgiar-fund/fund-donors-2/), and in particular to the CGIAR Research Program, Roots, Tubers and Bananas (RTB-CRP). Additional, but crucial funding was obtained from the Bill and Melinda Gates foundation (BMGF), investment ID: OPP1093845, under the project: Improvement of banana for smallholder farmers in the great lakes region of Africa. Part of this work was supported by the Czech Ministry of Education Youth and Sports (award LO1204 from the National Program of Sustainability I and INGO II award LG15017) while the computing was supported by the Czech National Grid Infrastructure MetaCentrum (grant No. LM2010005 under the program, Projects of Large Infrastructure for Research, Development, and Innovations).

Abstract

Banana (*Musa* spp.) is an important crop in the African Great Lakes region in terms of income and food security, with the highest per capita consumption worldwide. Pests, pathogens and environmental stress hamper sustainable production of bananas. Effort is being made to improve the East African highland bananas (EAHB) through conventional crossbreeding, but the selection cycle is too long. Improving the efficiency of selection in conventional crossbreeding is a major priority in banana breeding. Marker assisted selection (MAS) has the potential to reduce the selection cycle and increase genetic gain. However, the application of molecular tools has been hampered by the limitations inherent with the classical MAS tools and nature of traits in banana. While genomic selection can address some of the limitations of classical MAS, no report about its utility in banana is available to date. This Thesis provides the first empirical evidence on the performance of six genomic prediction models for 15 traits in a banana genomic selection training population based on genotyping by sequencing (GBS) data. The prediction models tested were Bayesian ridge regression (BRR), Bayesian LASSO (BL), BayesA, BayesB, BayesC and reproducing kernel Hilbert space (RKHS). The aim was to investigate the potential of genomic selection (GS) as a method of selection that could benefit breeding through increased genetic gain per unit time and cost. Trait variation, the correlation between traits and genetic diversity in the training population were analyzed as an essential first step in the development and selection of suitable genomic prediction models for banana traits. A training population of 307 genotypes consisting of EAHB breeding material and its progeny was phenotyped for more than 15 traits in two contrasting conditions for two crop cycles. The population was also genotyped by simple sequence repeats (SSR) and single nucleotide polymorphism (SNP) markers. Clustering based on SSR markers revealed that the training population was genetically diverse, reflecting a complex pedigree background, which was mostly influenced by the male parents. A high level of correlation among vegetative and fruit bunch related traits was observed. Genotype response to crop cycle and field management practices varied greatly with respect to traits. Fruit bunch related traits accounted for 31–35% of principal component variation under low and high input field management conditions. The first two principal components accounted for 50% of phenotypic variation that was observed in the training population. Resistance to black leaf streak (Black Sigatoka) was stable across crop cycles, but varied under different field management depending on the genotype. The best cross combination was 1201K-1 × SH3217 based on selection response (R) of hybrids. The predictive ability of genomic prediction models was evaluated for traits phenotyped over two crop cycles and under different cross validation strategies. The 15 traits were grouped into five categories that included plant stature, suckering behaviour, black leaf streak resistance, fruit bunch and fruit filling. Models that account for additive genetic effects provided better predictions with 12 out of 15 traits. The performance of BayesB model was superior to other models particularly on fruit filling and fruit bunch traits. Reproducing kernel Hilbert space model fitted with pedigree and marker data (RKHS_PM) produced mixed results with the majority of traits showing a decrease in prediction accuracy. Although RKHS models account for dominance and epistasis, heterosis is another non-additive genetic factor that affects prediction accuracy in bananas. Models that included averaged environment data for crop cycle one and two were more robust in trait prediction even with reduced numbers of markers. Accounting for allele dosage in SNP markers (AD-SNP) reduced predictive ability relative to traditional bi-allelic SNP (BA-SNP), but the prediction trend remained the same across traits. Since high correlation in prediction was observed within trait categories, only traits easy to phenotype should be considered for genomic predictions during the breeding phase. Although validation and more optimization of model parameters are still required, the high predictive values observed in this study confirmed the potential of genomic prediction in selection of best parents for further crossing and in the negative selection of triploid hybrids with inferior fruits to reduce the number of progenies to be evaluated in the field.

Abstrakt

Banánovník je důležitou plodinou v oblasti Velkých jezer ve východní Africe, která se vyznačuje nejvyšší spotřebou banánů na hlavu na světě. Banánovník má v této oblasti zásadní význam při zajišťování dostatku potravin a představuje významnou část příjmů místních obyvatel. Produkci banánů však snižují choroby a škůdci, a také abiotické stresy. Klasické šlechtění s cílem získat odrůdy s lepšími vlastnostmi je u této plodiny časově i technicky náročné. Výběr pomoci molekulárních markerů má potenciál šlechtění urychlit a usnadnit, bohužel avšak využití molekulárních metod naráží u banánovníku na řadu překážek. Některé z nich by mohla překonat genomická selekce, ale její využití u této plodiny dosud nebylo popsáno. Tato práce přináší první poznatky o úspěšnosti šesti genomických predikčních modelů pro patnáct vybraných znaků u testovací populace banánovníku. V práci byly testovány Bayesian Bridge Regression (BRR), Bayesian LASSO (BL), BayesA, BayesB, BayesC a Reproducing kernel Hilbert space (RKHS). Hlavním cílem bylo ověřit potenciál genomické selekce jako selekční metody, která by mohla významně urychlit a zlevnit šlechtitelské programy banánovníku. V práci byla také zkoumána variabilita jednotlivých hodnocených znaků a jejich korelace s genetickou diverzitou testovací populace, což byl nezbytný krok před vlastním výběrem vhodného predikčního modelu pro fenotypové znaky banánovníku. Testovací populace čítající 307 jedinců a zahrnující šlechtitelský materiál včetně potomstev byla fenotypována pro 15 znaků při pěstování za dvou kontrastních podmínek a po dvě kultivační období. Tato populace byla také genotypována pomocí SSR a SNP markerů. Analýza pomocí SSR markerů odhalila, že testovací populace je geneticky variabilní, což odráží její komplexní rodokmen, který je do značné míry ovlivněný samčím rodičem. Vysoká míra korelace byla pozorována u vegetativních znaků a vlastnosti trsu plodů. Chování jednotlivých genotypů bylo variabilní v průběhu dvou kultivačních obdobích a při odlišných podmínkách kultivace. Znaky související s vlastnostmi trsu představovaly 31 - 35% variability hlavního komponentu v kontrastních polních podmínkách. První dva hlavní komponenty byly odpovědné za 50% fenotypové variability pozorované v testovací populaci. Rezistence vůči chorobě “Black Sigatoka” se v průběhu kultivačních období neměnila, ale lišila se v různých polních podmínkách. Na základě hodnocení vlastností hybridů bylo nejlepší kombinací křížení 1201K-1 × SH3217. Předpovídací schopnost prediktivních genomických modelů byla stanovena pomocí znaků hodnocených po dvě kultivační období a pomocí různých validačních strategií. Patnáct fenotypových znaků bylo sdruženo do pěti kategorií, které zahrnovaly vzrůst rostliny, odnožování, rezistenci k chorobě Black Sigatoka, vlastnosti trsu a vlastnosti plodu. Modely zohledňující aditivní genetické efekty dávaly lepší předpovědi pro 12 z 15 znaků. Model BayesB dopadl nejlépe, zejména pro znaky ovlivňující trs a plod. Model Reproducing kernel Hilbert space, který zohledňoval rodokmen a data získaná analýzou markerů (RKHS_PM) měl sníženou prediktivní hodnotu. Ačkoli RHKS model zohledňoval dominanci a epistazi, heteroze je dalším neaditivním genetickým faktorem, který ovlivňuje přesnost predikce modelů. Modely, které zahrnovaly zprůměrovaná environmentální data za obě kultivační období byly ve svých předpovědích přesnější a to přesto, že se opíraly o méně markerů. Přihlédnutí k dózi alel u SNP markerů (AD-SNP) snižovalo prediktivní hodnotu oproti klasické bi-alelické metodě (BA-SNP), ale trendy jednotlivých predikcí zůstaly stejné pro všechny znaky. S ohledem na vysokou korelaci predikcí u kategorií jednotlivých znaků by během šlechtění měly být do genomických predikcí zahrnuty pouze takové znaky, které jsou jednoduše fenotypovatelné. Ačkoli je nutná další validace a optimalizace parametrů modelu, vysoké prediktivní hodnoty pozorované v této práci potvrdily potenciál genomické selekce při výběru nejvhodnějších rodičů pro křížení. Zároveň umožňují negativní selekci triploidních hybridů s podřadnými vlastnostmi plodů a umožní tak snížení rozsahu potomstva, které musí být hodnoceno v polních podmínkách.

Table of contents

1 General introduction

1.1 Origin of banana

Bananas and plantains are large perennial herbaceous monocotyledonous plants collectively known as bananas. They belong to the order *Zingiberales*, family *Musaceae* and genus *Musa.* The genus *Musa* has about 70 confirmed species, which include edible, ornamental types and their wild relatives. It was previously divided into five sections: *Australimusa* (2n = 2x = 20), *Callimusa* (2n = 2x = 20), *Eumusa* (2n = 2x = 22), *Rhodochlamys* (2n = 2x = 22) and *Ingentimusa* (2n = 2x = 14) (Swennen and Vuylsteke 2001; Daniells *et al.* 2001; Wong et al. 2002). However, the recent revision by Häkkinen (2013) recognizes only section *Callimusa,* which combines *Australimusa* and *Callimusa,* and section *Musa*, which combines *Eumusa* and *Rhodochlamys*. Section *Ingentimusa* was considered as part of section *Callimusa*. This revision is supported by evidence from molecular studies (Hřibová et al. 2011).

Cultivated bananas are believed to have arisen by intra- and inter-specific hybridization between *Musa acuminata* (AA genome) and *Musa balbisiana* (BB genome) species at the area of origin (INIBAP, 1995). The two species belong to section *Musa* (formerly *Eumusa*). They are wild diploid bananas endemic in the Asia and Pacific regions, which includes: India, Southeast Asia, Malaysia, Indonesia, Philippines and Papua New Guinea (Sharrock et al. 2001). Most diversity is found in *M. acuminata*, which has several subspecies including, for example, *M. a.* ssp. *burmannica, M. a.* ssp. *siamea, M. a*. ssp. *malaccensis, M. a. ssp. truncata, M. a.* ssp. *microcarpa, M. a*. ssp. *zebrina, M. a.* ssp. *errans* and *M. a.* ssp. *banksii* (Fig 1). Bats (*Glossophaga soricina*) are one of the natural pollinators that could have facilitated the hybridization and seed dispersal process in the wild (Buddenhagen 2008). Later, female sterility developed such that even pollinated flowers produced seedless fruits (Simmonds, 1962). It is also likely that erratic meiosis within improved diploids followed by backcrossing gave rise to parthenocarpic triploids (De Langhe et al. 2010; Perrier et al. 2011). Human intervention accelerated the process of banana evolution and domestication. Hybrids that were seedless (parthenocarpic), palatable and had good agronomic traits were selected and grown near human settlements.

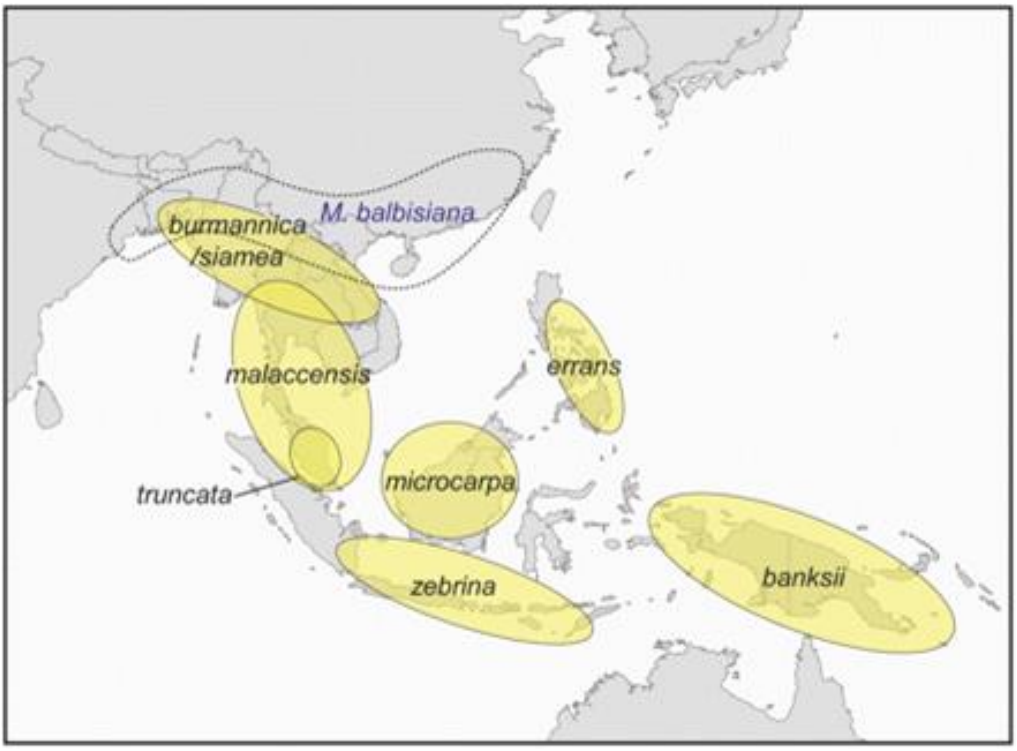


Fig 1. Geographical distribution of banana domestication areas is Southeast Asia (Perrier et al. 2011)

The wide spread of many popular cultivated seedless bananas could have occurred by traders from Arabia, Persia, India and Indonesia who navigated the Indian Ocean from Southeast Asia (INIBAP 1995) (Fig 2). As they moved, they carried along with them suckers of different cultivars with a broad mixture of genomic combinations between *M. acuminata* (AA) and *M. balbisiana* (BB), and ploidy levels. These included diploid (AA, AB), triploid (AAA, AAB, ABB) and tetraploid (AAAB, AABB, ABBB) that were delivered to the coastal areas. Within these genomic combinations, we have East African highland cooking (matooke) and beer bananas (both AAA), dessert bananas (AAA and AAB), plantains (AAB), cooking bananas (ABB) and Mshare, or Mchare bananas (AA). Likewise, the Portuguese and Spaniards between 16th and 19th century, carried bananas to all over tropical America (INIBAP 1995). However, several domestication pathways have been proposed (Perrier et al. 2011).

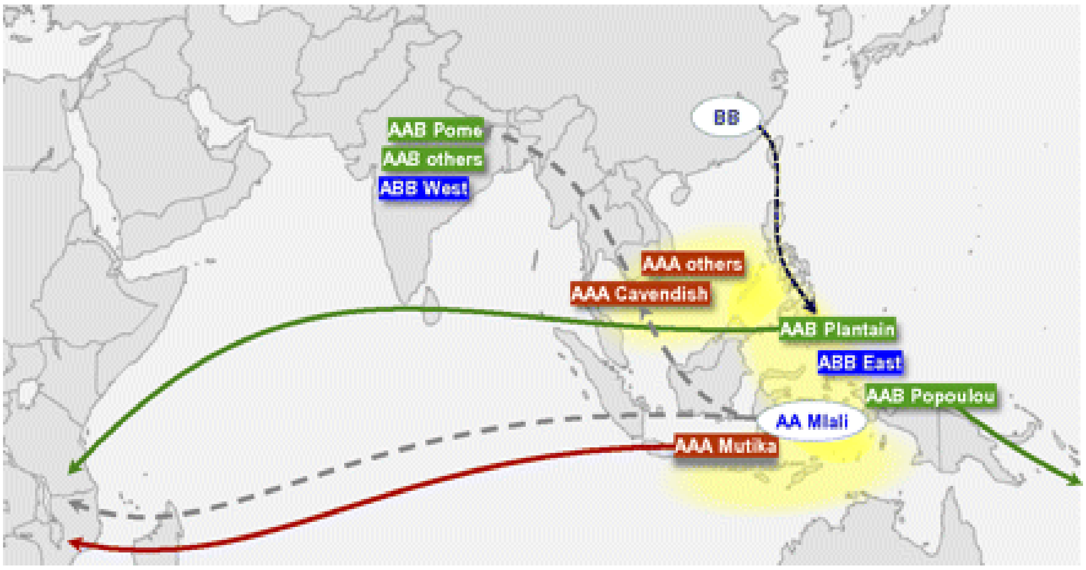


Fig 2. Distribution pathways of domesticated bananas from Asia Pacific to Africa and other tropical areas (Perrier et al. 2011)

1.2 Importance of banana

For several centuries, bananas have been an integral part of the farming systems especially in the tropics and sub-tropics. The crop is grown in 130 countries worldwide (Workman 2006; Evans and Ballen 2012). Bananas contribute tremendously to the livelihood of resource-poor populations especially in the sub-Saharan Africa by providing food security and income (FAO, 2010). Sub-Saharan Africa produces nearly a third of global banana production. The utility of banana depends on the genotypes and area. In the temperate countries, the most commonly consumed bananas are the dessert type (Cavendish, AAA). Cavendish banana is grown for export and it is a cash crop, thus a source of income for the exporting countries. The fruit are eaten when ripe yellow. However, in other countries, Pome, Silk, Mysore and Sukali Ndizi (AAB bananas) are also consumed as dessert bananas. Plantains are AAB bananas with high starch content and the fruit remains very firm even after ripening. They are mostly eaten after roasting and they make good chips as well.

In East Africa, there are two main groups of bananas that are endemic in the region. The EAHB (AAA) and the Mchare (AA). They are grown in areas around Lake Victoria, the highlands and part of the rift valley where severe drought periods are not experienced during the year (Karamura 1998). In Uganda, Rwanda and Burundi, the per capita consumption of banana is estimated at 400-600 kg per year, the highest in the world, indicating that the crop is a major staple in the region (Karamura et al. 1998). EAHB are divided into cooking (matooke) and beer bananas. The term matooke is synonymous to food in Uganda and these bananas are cooked when fresh green in different forms. However, during peak harvesting seasons, the surplus matooke is used for wine production in Western Uganda. The beer bananas are very astringent due to high tannin content (http://www.promusa.org/Uganda). They are allowed to ripen, juice is squeezed out of the pulp and fermented to make beer, hence the name beer banana, also known as Mbidde. The Mchare bananas have high starch content with firm pulp texture and are mostly roasted before eating.

India is the highest producer of ABB cooking bananas. These bananas have starchy fruits and sometimes are cooked when ripe for example, Saba and Bluggoe. In East Africa, about 85% of produced bananas are consumed locally due to high demand and only a small percentage is exported (Ortiz and Swennen 2014). Bananas provide about 25% of food energy requirements for around 90 million people in East, West and Central Africa (Sharrock et al. 2001).

1.3 Main banana production areas

The highest production of bananas occurs in India followed by China and East Africa. Uganda in particular produces about 10 million metric tons per year (De Buck and Swennen 2016). East Africa is considered a secondary centre of banana genetic diversity harboring over 84 cultivars that are not found elsewhere in the world. It is believed that EAHB are a product of single hybridization event that were introduced by Arab traders at the East African coast way back in 600 A.D (Karamura 1998) and over the time, several somatic mutations and selection pressure led to the origin of many distinct cultivars grown in the region (Kitavi et al. 2016). The EAHB subgroup (AAA) was named Lujugira-Mutika (Shepherd 1957). The accessions in Uganda have been grouped into five clone sets (Nfuuka, Nakitembe, Nakabululu, Musakala and Mbidde) based on end-use and morphological distinctiveness (Karamura 1998). The Mbidde clone set is used for beer production due to the astringency of fruit when fresh green while the rest of the clone sets are used as matooke.

Banana plants grow with varying degrees of success in diverse climatic conditions, but commercial banana plantations are primarily found in equatorial regions comprising of the humid tropics and subtropics. In the primary centre of genetic diversity (Asia and Pacific), several hundreds of different banana cultivars are grown alongside other wild uncultivated genotypes. In West Africa, especially in Nigeria and Cameroon, large fields of plantain cultivars are maintained (Ortiz and Vuylsteke 1994) as well as in Latin America. The Caribbean countries mostly grow the Cavendish bananas, which are mainly exported to Europe and United States, accounting for 13% of export banana (FAO 2014).

1.4 Production challenges

Reductions in productivity of landrace banana fields in various countries have been reported (Macharia et al. 2010). The causes are pests, pathogens and environmental stress (Jones, 2000; Biruma et al. 2007; Kumar et al. 2011, van Asten et al. 2011, Swennen et al. 2013). The major pests include banana weevils (*Cosmopolites sordidus*, Gold et al. 2004; Sadik et al. 2010) and the parasitic nematodes (Fig 3). Many nematode species have been associated with banana yield decline and amongst them are *Radopholus similis*, *Helicotylenchus multicinctus* and *Pratylenchus goodeyi* (Dochez 2004). These infect and damage banana roots that leads to toppling of plants due to poor anchorage.

Bacterial, fungal and viral diseases affect bananas, causing varying degrees of yield loss (Jones, 2000). For instance, banana bacterial wilt caused by *Xanthomonas campestris* pv. *musacearum* reduces crop yield by up to 100% (Biruma et al. 2007). Black leaf streak also known as Black Sigatoka is a disease caused by a fungus *Pseudocercospora fijiensis* previously known as *Mycosphearella fijiensis*, that affects banana leaves (Pro*Musa* 2002) reducing yield by 30-50% (Rowe and Rosales, 1996). Fusarium wilt also known as Panama disease is a soil borne disease caused by a fungus *Fusarium oxysporum* f. sp. *cubense*. It caused significant losses in the banana export industry when large plantations of cv. 'Gros Michel' were wiped out in the 1940-1960s (Stover 1962; Ploetz 2000). The export industry was revived when a banana cultivar called Cavendish was discovered to grow in areas where cv. 'Gros Michel' had been wiped out (Simmonds 1954). It was tested to be resistant to *F. oxysporum* f. sp. *cubense* (*Foc*) race 1 and race 2 and it replaced the cv. 'Gros Michel' as a commercial cultivar for global export markets.

Foc is divided into four races that include race 1, race 2, race 3 and race 4. However, *Foc* race 3 does not affect banana, but *Heliconia* species, which belongs to the same order as bananas thus, only three races are important to banana (Czislowski et al. 2017). The order of races reflects the increasing pathogenicity of *Foc,* hence all cultivars that are susceptible to race 1 and 2 are susceptible to race 4. Race 4 is further subdivided into the tropical race 4 (TR4) and sub-tropical race 4 (STR4). In East Africa, *Foc* race 1 affects ABB (Pisang Awak) and AAB (Sukali Ndizi) banana varieties but not the AAA (EAHB). The tropical race 4 (TR4) affects the commercial banana (cv. Cavendish), which replaced cv. 'Gros Michel' despite its resistance to other *Foc* races. Incidences of TR4 have been reported in Indonesia and Mozambique (Ploetz 2015), but it is not yet known if the EAHB and other cultivars will resist, or succumb to TR4. Of late, banana bunchy top virus (BBTV) transmitted by *Pentalonia nigronervosa* (banana aphid), though first reported in 1889 in many Asian banana growing countries, is reported to affect areas of Rwanda, Burundi and parts of Democratic Republic of Congo including many other banana-growing areas. It is said to be more significant on plantains than EAHB (Kumar et al. 2011), causing significant yield decline in those areas.

Among the abiotic constraints, limited rainfall (drought stress) reduces banana production especially in rain-fed agricultural systems. Since bananas are mostly grown in tropics and sub-tropics, taking a global and long-term view, the availability of water is thought to be the most critical limiting factor for photosynthesis on dry land, and hence for agricultural production (van Asten et al. 2011). Bananas require more than 1500 mm/year of rainfall for optimal growth and yield, but in many areas the average annual rainfall is ≤ 1200 mm/year (Taulya 2015). Drought stress causes stomatal closure and has deleterious effects on numerous physiological processes. It reduces photosynthesis and damages the photosynthetic machinery of chloroplasts through a process known as photo-oxidation (Audran et al. 1998). Hence, the most productive plant communities are the ones best supplied with water (Öpik et al. 2005). Under situations of mild drought stress, production has been shown to increase if potassium supply is sufficient (Taulya 2015), but not many farmers in developing countries use fertilizers, or apply sufficient mulch in the banana fields.

Cultivated bananas are vegetatively propagated, which limits gene flow and recombination, and hampers their potential to evolve and adapt to the changing environmental (biotic and abiotic) pressures (Myles 2013). Although the improvement of agronomic practices can lead to higher yield (Ndabamenye et al. 2012), sustainability is limited. Breeding for resistant cultivars is the only sustainable solution to banana production constraints (Simmonds 1986; Rowe 1990).

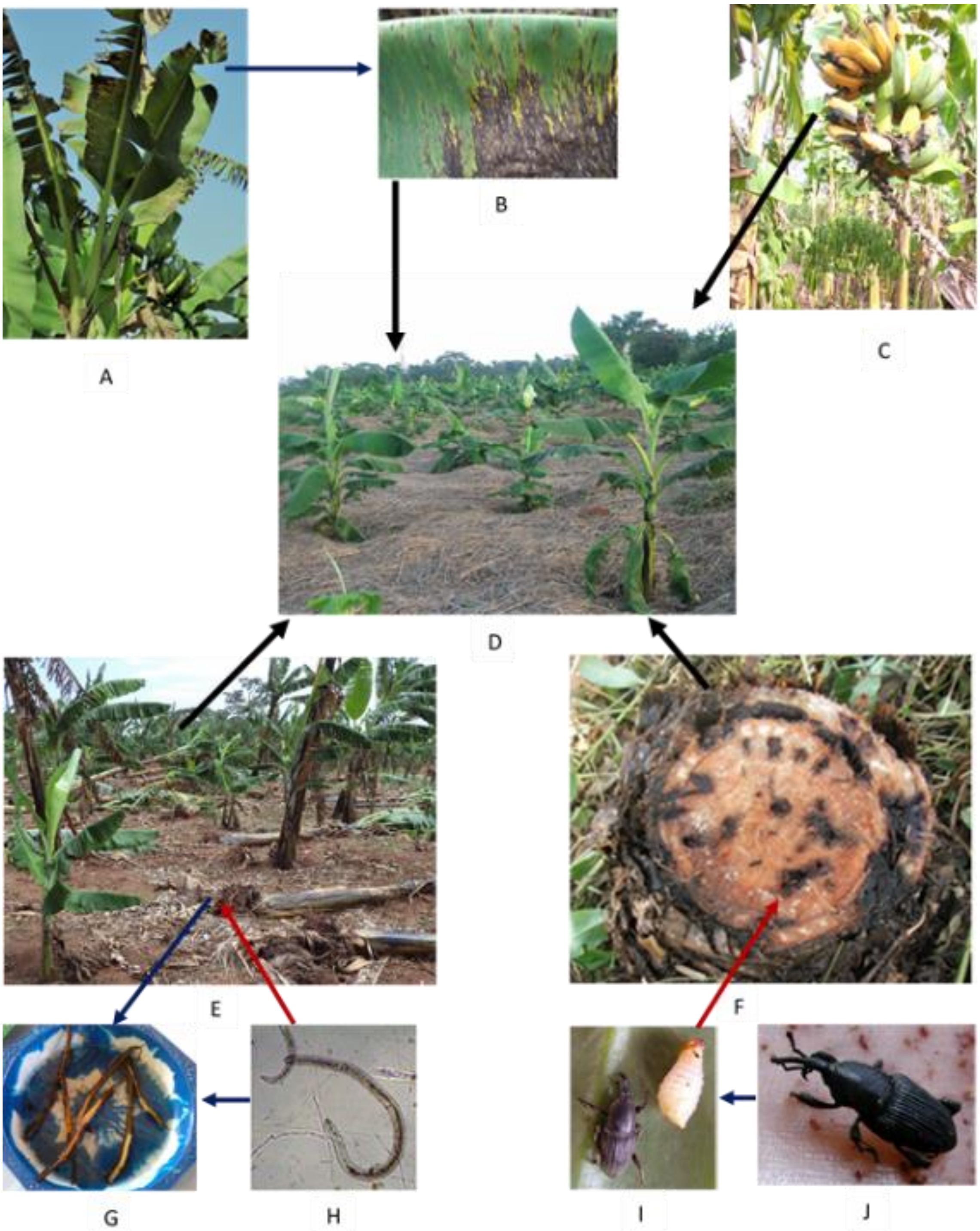


Fig 3: Main production constraints affecting East African highland banana. Banana field (A) infected with black leaf streak disease spreads spores from infected leaves (B) to a healthy plantation (D). Photosynthetic area is reduced by increasing leaf senescence, which affects yield. Banana fruit from plants infected by bacterial wilt (C) are rotten and not edible. The inoculum from infected plants is transmitted to the young health plants through farm tools and insects. *R. similis* (H) burrows into the banana roots causing necrosis (G). Plant anchorage into the soil and nutrient uptake are reduced, which lead to toppling (E). The adult banana weevil (J) lays eggs into the banana pseudostem, which hatch into larvae (I). The larvae make tunnels into the corm (F) that impede nutrient movement and weaken the attachment of pseudostem to the corm, resulting in plant snapping.

1.5 History of banana breeding programs

Several inter- and intra-specific hybridization events that took place in the wild were facilitated by natural pollinators. They gave rise to hybrids that had lost many of the wild characteristics and had attributes attractive to humans such as high yield, plant vigour, seedlessness and palatability of fruits (Simmonds 1962). The ability of man to select and domesticate the best hybrids was the most primitive and by far the most successful method of banana breeding. The selected cultivars were clonally propagated and spread over a wide area across the world (Perrier et al. 2011). Rapid evolution for better adaptation of the selected cultivars has been limited under nature's dynamic forces because of three main reasons: (i) most of the selected hybrids are sterile/partially sterile (Heslop-Harrison and Schwarzacher 2007), (ii) banana propagation and distribution is by vegetative means (Zohary 2004), and (iii) male fertile diploids are not grown in farmers' fields. Changes in the environment have increased pests and pathogens pressure making most cultivars susceptible.

The first breeding program was initiated in 1922 in Trinidad and later in 1924 in Jamaica. However, the first successful breeding program to release improved, farmer-acceptable hybrids was in Honduras, founded in 1984 called Fundación Hondureña de Investigación Agrícola (FHIA). In addition to FHIA, relatively few crossbreeding programs have been established in the world that are active and these include the International Institute of Tropical Agriculture (IITA) in Nigeria where research on banana/plantains started in1976, but the actual breeding started in 1987 (Ray 2002). In Uganda, IITA breeding work to improve the EAHB was initiated in 1994 by the late Dirk Vuylsteke (Vuylsteke, 2001). This is done in collaboration with the National Agricultural Research Organization (NARO). Since 2011, IITA extended its breeding activities to Arusha in Tanzania where breeding of Mchare bananas is ongoing. In Brazil, the Empresa Brasilliera de Pesquisa Agropecuaria (EMBRAPA) was established in 1982 with the main focus on improving the Pome and Silk, 'AAB' bananas. In France, the Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD) began in 1983. It has stations in the Caribbean (Guadeloupe and Martinique) and Cameroon with their main offices in Montpellier. They focus on plantains and other banana types with the exception of EAHB and Mchare. Another active breeding program that was initiated as part of the agreement between the Ministers of Research and Development for West and Central African countries in 2001 is the Centre Africain de Recherches sur Bananiers et Plantains (CARBAP) in Cameroon, which focuses more on plantain improvement. Other institutions such as Bioversity International hosted by Katholieke Universiteit Leuven, Belgium, support the activities of these major breeding programs by maintaining the world’s banana germplasm collection, called the International Musa Germplasm Transit Centre - ITC (Ray 2002; Dochez 2004; Lorenzen et al. 2010).

1.6 Breeding strategies

Three main strategies are used in banana improvement. When natural sources of resistance are available within the germplasm pool, conventional crossbreeding is used (Persley and George 1996, Ortiz and Swennen 2014). This strategy is simple and requires skills in phenotypic variation, taxonomy and genetics, but it is costly, labour intensive and time consuming due to the long selection cycle. Use of doubled haploids (Umber et al. 2016) and autotetraploids from chromosome doubled diploids (do Amaral et al. 2015) to fix important traits and reduce the selection cycle are practiced in conventional crossbreeding, but on a small scale.

The second strategy is marker assisted selection (MAS). In this approach, mapping populations from two parents with contrasting phenotypes are developed. The population is genotyped and phenotyped to identify DNA loci and markers that co-segregate in the presence, or absence of the trait. If the markers and loci controlling the trait are in linkage disequilibrium, then the breeder can use these markers to track the trait of interest in breeding populations (Collard et al. 2005). This approach is sometimes limited by the cost of marker development, high cost of assays for large populations, lack of good mapping populations for agronomically and economically important traits, and the need for technical capacity and modern infrastructure.

When natural sources of resistance are not available or have not been identified yet in the species germplasm, then genetic transformation remains the only strategy of choice (Tripathi et al. 2010). This involves the introduction of foreign genes into the target organism. However, this technology is limited to traits that are controlled by a single gene, or few genes with major genetic effects. Use of genetic engineering approaches to quantitative traits has not been done in banana.

1.6.1 Constraints in conventional crossbreeding of bananas

Conventional banana crossbreeding starts with the identification of the right parents to cross. At flowering, hand pollination is done. Pollen from a fertile diploid is rubbed onto stigma of newly opened female flowers every morning. It is hard to predict the outcome of crosses because of limited knowledge about the genetics of parental clones and how traits are inherited. The success of conventional crossbreeding relies on large numbers of hybrids from which selection is made (Ortiz and Swennen 2014). The hybrids generated come from several cross combinations of parental clones that differ in ploidy (Fig 4). The erratic meiosis of polyploids causes production of gametes with unpredictable chromosome constitution. While some gametes are haploid, others carry unreduced number of chromosomes, or additional chromosomes leading to variable ploidy levels and aneuploids in hybrids. Flow cytometry has been used to ascertain the ploidy level in bananas (Doležel 1997). A flow cytometer is used to measure the fluorescence intensity of cell nuclei stained by a DNA fluorochrome such as propidium iodide (PI), or 4′,6-diamidino-2-phenylindole (DAPI). As the fluorescence is proportional to DNA amount, the assay is suitable for ploidy estimation.

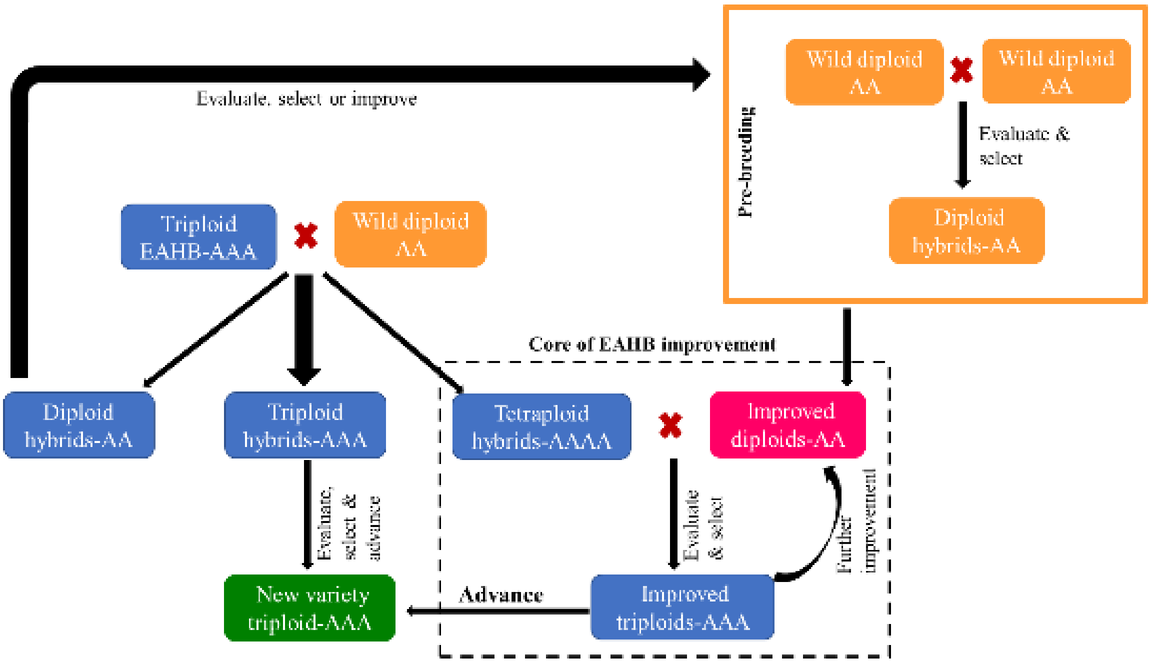


Fig 4. Crossbreeding scheme for improvement of East African Highland bananas showing crosses involving parents of different ploidy levels

Getting many hybrids in banana breeding is a challenge due to partial, or complete sterility of cultivars that have to be improved (Ssebuliba et al. 2006). This is further complicated by low embryo germination (Ssebuliba et al. 2005). Banana seeds do not readily germinate when planted directly in the soil, except for the wild species. Thus, banana breeding programs use *in vitro* embryo rescue techniques to increase the germination rate to about 30-40 % of seed embryos using artificial medium (Vuylsteke and Swennen 1992) (Fig 5). Despite the difficulty of getting seeds from banana crosses and having embryos germinate *in vitro*, about 90 % of hybrids are never selected and advanced from early evaluation trial (EET) to a preliminary yield trial (PYT) because a majority do not bear edible fruit, or show other shortcomings. This problem is not unique to banana only, but has been encountered in other crops, for example, 99.99 % of the 52,000 apple seedlings were discarded after 26 years of evaluation by Dresden-Pillnitz, a Germany apple-breeding program (Ignatov and Bodishevskaya 2011).

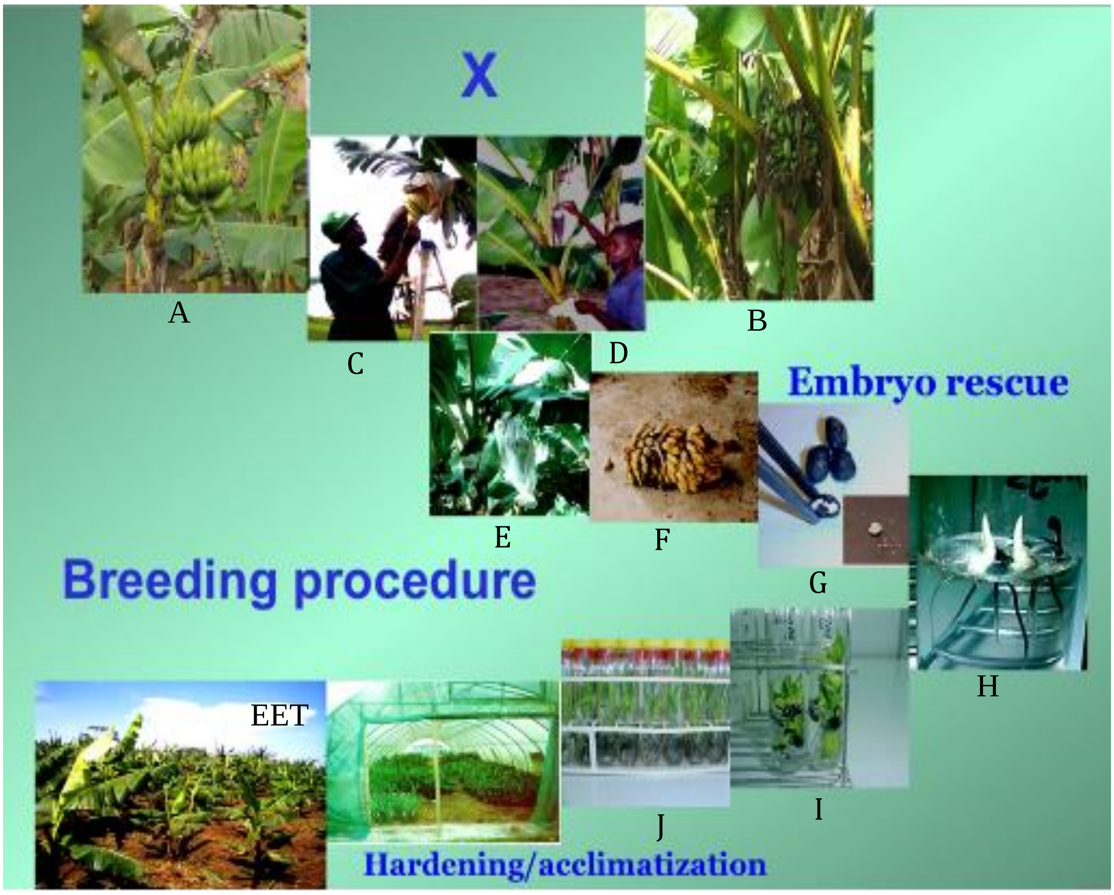


Fig 5. Conventional cross breeding steps of EAHB. The disease susceptible triploid EAHB (A) is crossed with a disease resistant wild diploid (B). At flowering, the female flowers of EAHB are hand pollination (C) by mature pollen from the male flowers (D). The pollinated fruit bunch (E) is covered with a polyethene bag to exclude other sources of pollen. After all the hands are pollinated, the bunch cover is removed and the fruit bunch is allowed to mature, harvested and ripened before seed extraction (F). The seeds are cracked to extract the embryos (G), which are germinated on artificial medium (H). The germinated embryos are transferred onto the proliferation medium

(I) after which they are cloned and transferred onto the rooting medium (J). The resulting plantlets are hardened in a screenhouse after weaning before they are planted in the early evaluation trial (EET).

Banana improvement progress is assessed by phenotypic evaluation of hybrids at various levels (Ortiz and Vuylsteke 1995b; Ortiz 2016) (Fig 6). The evaluation levels include EET, PYT, advanced yield trial (AYT) and multilocational evaluation trial (MET). Newly generated hybrids are first planted in the EET and the selection is based on the ability of a genotype to produce a good fruit bunch and host plant resistance to black leaf streak for at least two crop cycles. The number of replications per genotype in EET range from one to three. Usually, less than 10 % of the genotypes are selected from EET. The selected genotypes are multiplied so that each genotype is planted in two, or three single row plots of five replicates in a PYT (Ortiz and Vuylsteke 1995b). Data on both yield and agronomic traits are collected for at least two crop cycles. The quality of fruits is also used to select genotypes that are advanced to AYT, or MET. Unlike EET and PYT, which are on-station trials, AYT and MET are off-station trials, that involve more replications, blocks and different agroecological zones. The purpose of AYT and MET is to evaluate the stability of genotype performance under different environmental conditions because the genotype by environment (G × E) interaction affects trait expression (Taghouti et al. 2010; Manrique and Hermann, 2000). These trials are done in collaboration with farmers and the selection of best genotypes is more farmer-centred as acceptability of hybrids is the key in the final step of cultivar release. Each banana plant occupies an area of 6 m^2, or 9 m^2, depending on the spacing (Tushemereirwe et al. 2015). Hence, going through all these steps requires a lot of land. The many evaluation steps make the time required for cultivar development to be very long (Tenkouano et al. 1999).

1.6.2 Achievements and improvement strategies

To date, conventional banana crossbreeding has delivered a few improved cultivars to farmers from different breeding programs, but the rate is too low to cope with the demand. For example, the FHIA breeding program released some hybrids that have been widely distributed due to their high yield and resistance to Fusarium wilt, EMBRAPA in Brazil released some Pome and Silk hybrids resistant to Fusarium wilt and some are currently being tested in East Africa. The IITA-NARO breeding program has also released a few cooking banana hybrids and about 26 more hybrids (NARITA) are still under regional evaluation (Tushemereirwe et al. 2015).

Methods that can increase seed set and germination, and speed up the selection process are required to improve the breeding efficiency in banana.

To increase selection speed in conventional breeding, the genetic breeding values of parents should be known so that target crosses are made. A reliable and cost-effective selection system should be used to select the best hybrids with targeted traits prior to field evaluation. Marker assisted selection, MAS (Choudhary et al. 2008) is one way to improve conventional breeding efficiency. Reports on the use of MAS in banana breeding are limited because of two major challenges: (1) many traits, especially those of agronomic and economic importance may be controlled by many quantitative trait loci (QTL), each having a small effect on the phenotype (Asíns 2002; Collard et al. 2005; Choudhary et al. 2008), and (2) the difficulty to identify all markers across the entire genome that are linked to QTL (Guo et al. 2011) due to the cost, labour involved in marker assays and complexity of polyploid genomes. Details on these issues and how markers have been used in banana research are discussed in the next section.

In Uganda, the IITA-NARO collaboration is focused on improving the EAHB that are susceptible to both biotic and abiotic stress (Lorenzen et al. 2010). The choice of breeding parents currently used was based mostly on field and screen-house phenotypic characterization of available germplasm to identify sources of host plant resistance in diploids and female fertility within the different clone sets of the EAHB (Ssebuliba et al. 2005, 2006; Karamura 1998). Since then, several hybrids have been generated from crosses involving Calcutta 4 (wild diploid), improved parthenocarpic diploids, EAHB and tetraploids with EAHB background. Due to partial sterility, polyploidy and the low percentage of germinating embryos in tissue culture, few segregating populations have been generated from a single set of parents to allow molecular characterization and mapping of all important traits (Mbanjo et al. 2012a; Pillay et al. 2012; Xu 2010), but efforts are being made to generate more mapping populations. However, many hybrids with related background are generated that can constitute a training population for genomic predictions.

Application of molecular markers to assess breeding progress is still limited in the program although simple sequence repeat markers are used in genotyping. The new developments in genotyping such as genotyping by sequencing (Elshire et al. 2011) and MAS such as genomic selection (GS) (Meuwissen et al. 2001), should be explored to reduce selection cycle and increase product output in a cost-effective way. This Thesis therefore focuses on the development and evaluation of genomic prediction models based on SNP markers derived from genotyping by sequencing approach and phenotypic data from related hybrids of mixed ploidy levels and their parents as a training population. The training population was chosen to mirror the breeder's population so that inferences can easily be made as opposed to the classic bi-parental diploid mapping populations commonly used in QTL analysis (Heffner 2009). The population consisting of 307 genotypes was phenotyped under low input and high input field management conditions for two crop cycles. Results of experiments are summarized in publications under section six. It is expected that the information provided in this Thesis will be useful in improving the efficiency of banana breeding.

2 Role of molecular markers in banana research

2.1 General overview

Cultivated bananas are susceptible to pests, pathogens and environmental stresses, causing yield reduction that leads to food insecurity (Stover 1962; Ploetz 2000; Gold et al. 2004; Biruma et al. 2007; Tenkouano et al. 2012; Tripathi et al. 2015). Whereas chemical intervention is possible to some extent, it is not a sustainable solution, given the risk of environmental pollution and the economic burden on small-scale farmers. Thus, breeding for resistant banana cultivars is the most sustainable solution (Rowe and Rosales 1993).

Molecular markers play a significant role in identification of genomic loci controlling important traits in plant breeding (Brown et al. 2017). Markers that are linked to traits of interest are determined by linkage and association analysis. Estimation of genetic diversity facilitates gene introgression by choosing parents that are likely to give better genetic gain. The introgression process is quickened by marker assisted selection. Markers are also helping in taxonomic validation, cultivar identification, and characterization of evolutionary and speciation events. Molecular markers reduce the selection cycle in conventional cross breeding as compared to the classic phenotypic selection (Fig 6). The use of molecular markers shows promise in improving the efficiency of plant breeding (Ortiz and Swennen 2014), but in banana breeding programs, their utility is currently limited.

The release and improvement of a draft genomic sequence of the double haploid *M. acuminata* cv. Pahang, A genome (D'Hont et al. 2012; Martin et al. 2016) and a draft sequence of *M. balbisiana* cv. 'Pisang Klutuk Wulung', B genome (Davey et al. 2013) made a significant contribution to marker development in banana. Numerous gene transcript data consisting of 46,665 expressed sequence tags (EST) and 35,752 annotated genes associated with *M. acuminata* and *M. balbisiana* are publicly available (Li et al. 2013; Wang et al. 2012a; https://www.ncbi.nlm.nih.gov/gquery/?term=Musa [retrieved on 14 August 2017]). Several papers have reported on the utility of molecular markers in banana research and these are summarized in Table 1.

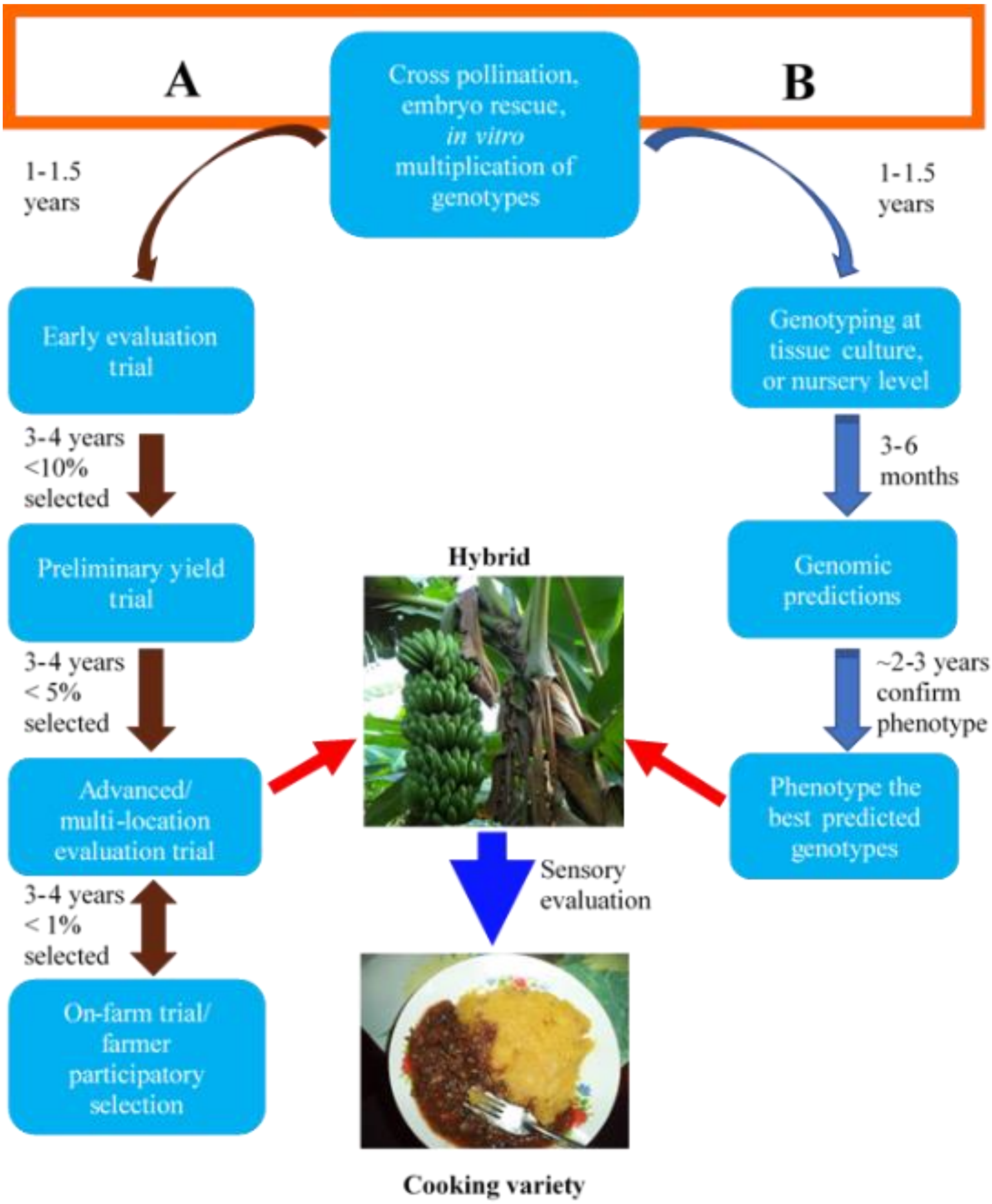

A
B
Cross pollination, embryo rescue, *in vitro* multiplication of genotypes
1-1.5 years
1-1.5 years
Early evaluation trial
Genotyping at tissue culture, or nursery level
3-4 years <10% selected
3-6 months
Preliminary yield trial
Genomic predictions
Hybrid
3-4 years < 5% selected
~2-3 years confirm phenotype
Advanced/ multi-location evaluation trial
Phenotype the best predicted genotypes
Sensory evaluation
3-4 years < 1% selected
On-farm trial/ farmer participatory selection
Cooking variety

Fig 6: Approaches to hybrid selection in banana breeding program. (A) the classical phenotypic selection of banana hybrids and (B) integrated genomic selection and phenotypic selection approach being investigated.

Table 1. Summary of molecular markers that have been used in banana research.

Marker application	Marker type	Reference
Molecular systematics	Isozymes, SSR, DArT, RFLP, ETS and ITS	Simmonds (1966); Bhat et al. (1992); Janssen and Bremer (2004); Kress and Specht (2005, 2006); Boonruangrod et al. (2009); Perrier et al. (2011); Hřibová et al. (2011); Christelová et al. (2011b); Čížková et al. (2015)
Genetic diversity studies	Isozymes, RAPD, SSR, AFLP, RFLP, SRAP, DArT and MSAP	Bhat et al. (1992); Jarret et al. (1993); Bhat et al. (1995); Kaemmer et al. (1997); Tenkuoano et al. (1999); Crouch et al. (1999); Crouch et al. (2000); Pillay et al. (2001); Ude et al. (2002); Ude et al. (2003); Creste et al. (2004); Noyer et al. (2005); Wang et al. (2007); Risterucci et al. (2009); Opara et al. (2010); Onyango et al. (2010); Wei et al. (2011); Nyine and Pillay (2011); Valdez-Ojeda et al. (2014); Kitavi et al. (2016); Karamura et al. (2016); Christelová et al. (2017)
Detection of mutant clones	RAPD	Newbury et al. (2000); Martin et al. (2006)
Genome characterization	RAPD, RFLP, ITS, dCAPS, IRAP and SCAR	Pillay et al. (2000); Nwakanma et al. (2003); Nair et al. (2005); de Jesus et al. (2013); Noumbissié et al. (2016); Mabonga and Pillay (2017)
Cultivar identification and pedigree tracking	Isozymes, RFLP, SSR, RAPD, EST-SSR and ISSR	Horry (1988); Howell et al. (2004); Raboin et al. (2005); Venkatachalam et al. (2008); Horry (2011); Hippolyte et al. (2012); Mbanjo et al. (2012a)

Linkage analysis	Isozyme, RAPD, RFLP, AFLP, SSR AS-PCR and DArT	Fauré et al. (1993); Hippolyte et al. (2010); Mbanjo et al. (2012b)
Genome-wide association studies and marker-assisted selection	Isozymes, dCAPS and SNP	Umber et al. (2016); Noumbissié et al. (2016) Sardos et al. (2016);

AFLP – amplified fragment length polymorphism, AS-PCR – allele specific-polymerase chain reaction, DArT – diversity array technology, dCAPS – derived cleaved amplified polymorphic sequences, EST – expressed sequence tags, ETS – external transcribed spacer, MSAP – methylation-sensitive amplified polymorphism, IRAP – inter retrotransposon amplified polymorphism, , ISSR – inter simple sequence repeats, ITS – internal transcribed spacer, RAPD – randomly amplified polymorphic DNA, RFLP – restriction fragment length polymorphism, SCAR – Sequence characterized amplified region, SNP – single nucleotide polymorphism, SRAP – sequence-related amplified polymorphism, SSR – simple sequence repeats

2.2 Gene markers

Useful markers for molecular breeding are those that are tagged to genes having significant contribution to traits of interest (Collard et al. 2008). When a gene and a marker are in linkage disequilibrium, it allows for the screening of plant germplasm, or hybrid lines at the earliest stages of plant improvement. The association of these markers with important traits can be identified through classical linkage analysis, genome-wide association studies, or candidate gene approaches. For example, Miller et al. (2008) identified 50 distinct nucleotide binding site leucine rich repeats (NBS-LRR) linked to resistance gene analogs in cv. ‘Calcutta 4’. Based on these findings, Emediato et al. (2009) were able to design degenerate primers that could amplify sequence analogs for resistance genes to black leaf streak disease in *M. acuminata* cv. ‘Calcutta 4’ (resistant) and *M. acuminata* cv. ‘Pisang Berlin’ (susceptible).

Similarly, Wang et al. (2012b) identified randomly amplified polymorphic DNA (RAPD) markers that could distinguish between cultivars resistant and susceptible to *Foc* TR4 using pooled DNA from resistant and susceptible cultivars. Two RAPD markers were converted to sequence characterized amplified regions (SCAR) markers, which could be amplified in *Foc* TR4-resistant banana genotypes, but not in the susceptible genotypes. This work continues at the National banana program in Brazil (EMBRAPA) and shows a great promise in providing an early screen for resistance to *Foc* TR4 (Silva et al. 2016).

M. balbisiana (B genome) is a good source of resistance, or tolerance to biotic and abiotic stresses (Vanhove et al. 2012; Ravi et al. 2013). However, it harbors endogenous banana streak virus (eBSV), which is activated when plants are stressed, or upon hybridization (Harper et al. 1999; Lheureux et al. 2003). This causes the limited use of any B genome containing accession in banana breeding. Lheureux et al. (2003) mapped the eBSV-expressed locus on a linkage group using amplified fragment length polymorphism (AFLP) markers. In a different study, Noumbissié et al. (2016) used simple sequence repeat (SSR) markers and eBSV-specific PCR markers to identify hybrids containing the B genome that were free of eBSV. These hybrids resulted from crossing a tetraploid accession (AABB) with a diploid accession (AA). They found that chromosome translocation and recombination had produced 24 offspring (13% of the population) that did not contain eBSV. Using derived cleaved amplified polymorphic sequences (dCAPS), Umber et al. (2016) identified the existence of infectious and non-infectious BSV alleles. By chromosome doubling a haploid plant with B genome (homozygosity checked using SSR markers), they produced lines, which were free of the infectious BSV alleles. The two studies give a hope for the possibility of using diagnostic markers and producing eBSV-free B genome hybrids that could be useful in banana breeding.

2.3 Linkage and association mapping

Linkage and association mapping are the basis of MAS in plant breeding, but have not gained significant practical application in banana breeding. This could be attributed in part to limitations inherent with the marker technologies themselves (Foolad 2007; Pillay et al. 2012), polyploid nature of banana, and the difficulty in developing and maintaining banana genetic mapping populations. Earlier attempts in linkage and association mapping used F_1 and F_2 diploid populations, which limited the resolution and accuracy of mapping quantitative trait loci (QTL) affecting important traits (Asíns 2002). Efforts should be made to develop double haploid populations, or recombinant inbred lines to facilitate QTL mapping in banana (Pollard 2012).

Genetic linkage maps are useful in gene identification and understanding the inheritance pattern of traits (Korte and Farlow 2013). Linkage maps are derived from genotyping bi-parental segregating populations. An important prerequisite is that the two parents from which the segregating population is derived are significantly different in the trait of interest. Moreover, the markers used to genotype the population should show the segregation and population structure, and should be distributed on all chromosomes. Proper and accurate collection of phenotype data is critical if linkage maps are to be of any value. To avoid bias in phenotyping, data from multiple years and locations should be collected.

To date, a limited number of *Musa* genetic linkage maps have been reported (Table 2). This is because cultivated bananas are mostly triploid and partially, or completely sterile (Ssebuliba et al. 2006), which makes it difficult to generate adequate study populations. Indeed, often they lack genetic variability for the most important traits, which hinders construction of genetic linkage maps. All genetic linkage maps reported so far are from diploid segregating population.

Table 2: Summary of banana genetic linkage maps currently publicly accessible

Reference	**Popn type**	**Popn size**	**No. of markers**	**Linkage groups**	**Type of markers**	**Segregation distortion (%)**
Fauré et al. (1993)	F_2 (SF265 × *banksii*)	92	77	15	RFLP, isozyme and RAPD	36
Hippolyte et al. (2010)	F_1 (Borneo × P. Lilin)	180	489	11	SSR and DArTs	22
Mbanjo et al. (2012)	F_1 (half-sib, 6142-1 × 8075-7 and 6142-1-S × 8075-7)	139	316	15	SSR, DArTs and AS-PCRs	41

The first genetic mapping population (Fauré et al. 1993) consisted of an F_2 population of 92 individuals derived from selfing an F_1 hybrid (SFB5) that resulted from a cross between SF265 and *M. acuminata* ssp. *banksii*. Seventy-seven loci consisting of RAPDs, Isozymes and RFLPs were placed on 15 linkage groups and covered 606 cM. Segregation distortion was 36% of the mapped loci and was biased towards *M. acuminata* ssp. *banksii*. Hippolyte et al. (2010) published the most saturated map to date using an F_1 diploid (AA) population created from a cross between *M. acuminata* cv. Borneo and *M. acuminata* cv. ‘Pisang Lilin’. The map was constructed using 489 markers (including, SSR and diversity array technology, DArT) distributed across 11 linkage groups and covering 1197 cM. The segregation distortion of alleles was 22%. The most recent genetic linkage map is that of Mbanjo et al. (2012b). They used an F_1 population consisting of two half sibs derived from crosses between *M. acuminata* hybrids and these were 6142-1 × 8075-7 and 6142-1-S × 8075-7. Two maternal (6142-1 and 6142-1-S) and one paternal (8075-7) maps were generated using DArT, SSR and AS-PCR markers. The most inclusive map was the paternal map with 316 markers that were distributed on 15 linkage groups covering 1004 cM. However, 41% of the allele loci showed segregation distortion.

Association mapping (genome-wide association study, GWAS) offers the opportunity to link genetic markers and their location on genetic maps to phenotypic differences (Korte and Farlow 2013). The advantage of GWAS is the non-reliance on bi-parental populations and the ability to capture both recent and historical recombination events (Borevitz and Nordborg 2003; Korte and Farlow 2013). Whereas linkage mapping requires recombinant inbred lines to achieve a good resolution, GWAS utilizes a panel of genotypes from unrelated population, or a population with known genetic substructure to identify associations between molecular markers that are in linkage disequilibrium with genetic loci affecting phenotypes.

Genome-wide molecular markers such as SNP are preferred for GWAS. For example, Sardos et al. (2016) performed GWAS for parthenocarpy in banana. A panel of 104 diploid (AA) accessions was genotyped by sequencing (GBS) and 5,544 SNP markers were derived. The SNP markers were associated with the publicly available phenotypic data on parthenocarpy. Thirteen genomic loci were identified to be associated with parthenocarpy and female sterility. The genes identified in these regions were mostly related to growth regulators such as auxin, gibberellin and abscisic acid, whereas the others were involved in gametophyte development and one histidine kinase implicated in female sterility. Such studies need to be extended to other traits using more objective and empirical phenotypic data.

In GWAS, the effect of each marker on the trait is estimated and markers with the smallest probability values (P-values) are considered to have a strong significant association with the trait (Korte and Farlow 2013). In order to limit the number of false associations between markers and traits, a Bonferroni correction is used. For example, if the confidence level is set at 95%, Bonferroni correction = 0.05 divided by the number of SNP markers analyzed. GWAS results are presented on Manhattan plots generated by qqman-package in R (R core team, 2017), or in trait analysis by association, evolution and linkage (TASSEL) pipeline. The Bonferroni correction line on a Manhattan plot is placed at -$\log_{10}$ × Bonferroni correction value (Fig 7). All markers that are above the Bonferroni correction line are considered to be significantly associated with the trait.

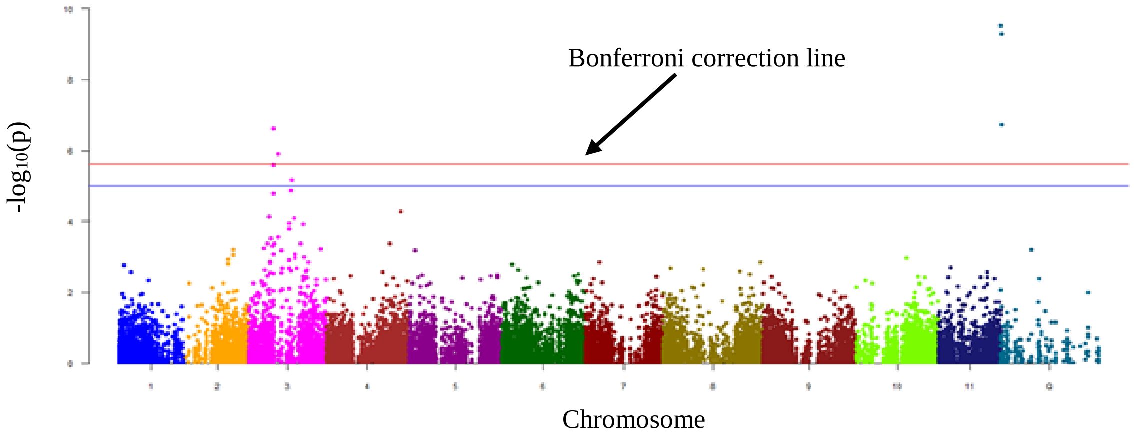


Fig 7. Manhattan plot generated in R using qqman package showing the Bonferroni correction line (red) and the location of markers associated with the trait under study.

2.4 Genetic diversity studies

Genetic diversity is indispensable in breeding and is perhaps the single most limiting factor to plant improvement. It is upon which breeders base their decisions to choose the parents to cross. Conventionally, phenotypic, or morphological characters associated with vegetative and floral structures of banana have long been used to estimate diversity and distinguish among cultivars (Karamura 1998). However, phenotypic characteristics are greatly influenced by genotype, environment and the interaction between genotype and environment (Batte et al. 2017). This limits the genetic gain achieved from crossbreeding when parents are chosen on the basis of morphological characteristics. Molecular markers have been used to supplement this effort and expand germplasm diversity analysis among various collections and representative populations displaying regional variation. A variety of molecular markers has been used in banana genetic diversity studies and they included isozymes, RAPD, AFLP, SRAP, RFLP DArT, MSAP and SSR (Table 1).

Results from different banana genetic diversity studies cannot be compared. Each study is unique in terms of population composition, type and number of markers used. However, the general consensus is that molecular diversity does not correlate well with phenotypic diversity (Crouch et al. 2000; Kitavi et al. 2016). The genetic variation explaining the substantial morphological variation among regional *Musa* landraces is still lacking despite the availability of numerous molecular markers. EAHB have been classified into five clone sets based on phenotypic characteristics (Karamura 1998). This grouping has not been supported by any of the molecular studies (Pillay et al. 2001; Kitavi et al. 2016; Karamura et al. 2016a). Hence, EAHB are considered to be a product of single hybridization event and the morphological differences observed are most probably a result of several somatic mutations, and selection events that led to many distinct cultivars (Kitavi et al. 2016).

Markers can also be used to identify variation from sources where it has not been previously reported. In plantain landraces of West Africa, RAPD, SSR and AFLP markers showed very low polymorphisms (Crouch et al. 2000; Noyer et al. 2005). However, *Hpa*II and *Msp*I, MSAP profiles revealed three clusters that were not correlated with morphological differences in plantains (Noyer et al. 2005) and a subset of plantains from Cameroon was genetically distinct from others (Ude et al. 2003).

Somaclonal mutation resulting from prolonged sub-culturing of plants in tissue culture and chimerism create diversity within cultivars. Molecular markers have been used to detect such variation. For example, Martin et al. (2006) were able to differentiate somaclonal mutant named CUDBT-B1 from the parent clone cv. 'Grand Naine' using RAPD marker S-20 (5'-GGACCCTTAC-3'). The marker produced a unique 1650 bp band only in mutants. In plantains, analysis of 48 clones derived from a single meristem of cv. Agbagba using RAPD markers showed polymorphism within the clones. Field evaluation of these clones correlated well with their genetic clustering leading to a conclusion that cv. Agbagba comprised of periclinal chimera (Newbury et al. 2000).

2.5 Genomic selection in banana

QTL analysis is quite straightforward once one has a well-saturated linkage map and accurate phenotypic data. However, this applies to qualitative traits, or traits governed by few QTL with major genetic effects such as pest and disease resistance (Asíns 2002; Heffner et al. 2009). For highly quantitative traits such as yield, or drought stress, QTL mapping becomes powerless due to the presence of many loci contributing to the trait, each with small-explained variance (Asíns 2002; Collard and Mackill 2008). Even if these QTL could be identified, introgressing and selecting for them during breeding using MAS would be tedious. To overcome the above challenges, genomic selection (GS) that uses predictive models has been proposed with the prospect to reduce the selection cycle and increase genetic gain per unit time.

Genomic selection (GS) is a form of MAS that utilizes high-density molecular markers such as SNP to estimate the genomic breeding value of a genotype using a statistical model (Meuwissen et al. 2001). The approach used to perform genomic selection is called genomic prediction while the unit of selection is called the genomic estimated breeding value (GEBV). In this approach, identification of individual QTL associated with a trait of interest is not necessary because QTL are assumed to be in linkage disequilibrium with at least one, or more SNP (Desta and Ortiz 2014). Since generation of marker data is increasingly becoming cheaper than phenotyping, it is expected that GS will reduce breeding costs, increase selection intensity and accelerate breeding efficiency. It is a well-established technique in animal breeding (Hayes and Goddard 2010) and it is gaining popularity among plant breeders (Crossa et al. 2010; Lorenz et al. 2011; Ceballos et al. 2015; Crossa et al. 2016) with several publications in cereal breeding and fruit trees. GS has not been applied in bananas yet, but it is currently being investigated. More details on genomic selection are given in section three.

2.6 Characterizing evolutionary and speciation events

Identifying and utilizing progenitors of modern banana cultivars in breeding schemes provides potential sources of improved quality traits associated with important commercial cultivars. This provides bridges for gene transfer of traits such as host plant resistance to pathogens and pests as well as drought tolerance from wild relatives. Understanding how these modern cultivars arose may allow us to reconstruct them while also including source of resistance to major abiotic and biotic sources of stress (Perrier et al. 2011). Therefore, proper identification and classification of bananas both at morphological and more importantly at molecular level is very necessary.

Several studies have utilized isozymes, SSR, DArT, chloroplast (and mitochondria) DNA RFLP, 5´ external transcribed spacer rDNA (5´ETS rDNA) sequence information and various cytological techniques to elucidate the domestication pathways of bananas (Boonruangrod et al. 2009; Perrier et al. 2011). For example, through molecular analysis, the EAHB have been shown to be a product of three subspecies of *M. acuminata* (*M. a*. ssp. *banksii, M. a*. ssp. *zebrina* and *M. a.* ssp. *malaccensis*) while *M. balbisiana* and *M. a.* ssp. *banksii* are the founders of plantains (Boonruangrod et al. 2009; Perrier et al. 2011).

The family *Musaceae* consists of domesticated edible and ornamental species, and their wild relatives. The *Musaceae* family consists of three genera including, *Ensete*, *Musa* and *Musella* (Janssens et al. 2016). Different classification systems in banana have been reported including molecular phylogeny. Isozymes such as esterase, acid phosphatase and catalase were used in the earlier classification of bananas (Simmonds 1966; Bhat et al. 1992). Christelová et al. (2011a, 2017) used 19 informative SSR markers to discriminate different levels of classification of *Musa* accession held at the International Musa Germplasm Transit Centre (ITC), Belgium.

Internal transcribed spacers (ITS) of rDNA show genetic variation despite the evolutionary conservation of rRNA genes. This variation was used to assess the structure and genetic diversity of *Musaceae* family. Analysis of ITS1 and ITS2 sequences revealed that section *Callimusa* and *Australimusa* were in the same clade while *Eumusa* and *Rhodochlamys* formed the second clade of genus *Musa* (Hřibová et al. 2011). Results from intronic sequence analysis of single copy genes from *Musa* accessions supported the merger of *Callimusa* with *Australimusa* and *Eumusa* with *Rhodochlamys* however, the old classification is still widely used. Recent findings by Janssens et al. (2016) based on the analysis of four gene markers (*rps*16, *atpB-rbcL*, *trnL-F* and ITS) using Bayesian inference methods, gave further support for the merger of *Callimusa*, *Astralimusa* and *Ingentimusa* into one clade while *Eumusa* and *Rhodochlamys* formed the second clade. In addition, the divergence time of *Musaceae* family and evolution of genus *Musa* were estimated to be 69 Mya and 51 Mya, respectively (Christelová et al. 2011b). These studies were expanded by using cytogenetics, ITS and SSR markers (Čížková et al. 2015). However, discrepancies in estimates of divergence time of *Musaceae* family and speciation of *Musa* are noted in various publication depending on the analysis method used (Janssen and Bremer 2004; Kress & Specht 2005, 2006; Janssens et al. 2016).

2.7 Genome characterization, cultivar identification and pedigree tracking

Four types of genomes are present in banana and these include A, B, S and T representing *M. acuminata*, *M. balbisiana*, *M. schizocarpa* and *M. textilis*, respectively (Swennen and Vuylsteke, 2001). Many cultivated bananas consist of one, or a combination of two genomes. The most common genomes within the edible bananas are the A and B genomes. Markers specific to these genomes allow determination of genomic composition of allopolyploids and track recombination event between genomes in hybrid progeny. For example, three RAPD Operon primers A17, A18 and D10 were used to distinguish between A and B genome composition in 40 banana accessions (Pillay et al. 2000), thus providing a quick means of genome characterization. Nwakanma et al. (2003) used PCR-RFLP on ribosomal DNA internal transcribed spacer (ITS) and identified markers that were specific for A and B genomes in bananas. Restriction digest of ITS-PCR products revealed a 530 bp fragment that was specific to A genome and two fragments of 350 bp and 180 bp that were specific to B genome and their intensity increased with increasing number of copies of B genomes in the accessions.

de Jesus et al. (2013) used a combination of flow cytometry, PCR-RFLP based on ITS amplification products and SSR markers and confirmed the genomic constitution of 94.6% of the total accessions maintained at the EMBRAPA *ex situ* collection. Their results supported the hypothesis of homeologue recombination between A and B genomes. One inter-retrotransposon amplified polymorphism (IRAP) marker designed from a long terminal repeat (LTR) of *Musa* Ty3- gypsy-like retroelement (*M. acuminata* Monkey retrotransposon, AF 143332) was identified to be specific for the B genome in bananas. The marker was used to classify the AAB and ABB cultivars in South India and clarified the genome composition of some cultivars that had been misidentified (Nair et al. 2005). Howell et al. (2004) developed nine RAPD primers that distinguished banana accessions from ITC based on genome composition and ploidy level following cluster analysis and these improved the precision of *Musa* identification and classification. Mabonga and Pillay (2017), reported a SCAR marker developed from a RAPD amplicon that produced 500 bp and 700 bp fragments in A and B genomes, respectively. They concluded that the two genomes may not be fully differentiated as previously reported.

Germplasm collection centres and breeding programs maintain records of accessions and crosses made, but mistakes arise due to human error either during *in vitro* sub-culturing, or field planting. Molecular markers have proven to be useful in cultivar identification and pedigree tracking. For example, cv. Cavendish and cv. ‘Gros Michel’ are popular dessert bananas that arose from 2n restitution and n gamete donors. RFLP markers showed that the 2n donors could have been cvs. Samba, Chicame, or ‘Akondro Mainty’ because they shared almost the full allele profiles (Raboin et al. 2005). Cv. ‘Akondro Mainty’ was highly linked to cv. Cavendish based on isozyme, ribosomal gene spacer patterns and anthocyanin markers (Horry 1988; Horry 2011), whereas cv. Chicame could have contributed the 2n gametes to cv. ‘Gros Michel’. However, it was not possible to identify a single n gamete donor that crossed with 2n gamete donor to produce the triploid cultivars, but putative candidates were cvs. Sa and ‘Khai Nai On’. Similar observation was made when a set of 22 SSR markers was used to analyze 561 *Musa* accessions (Hippolyte et al. 2012). SSR-based platform for clarifying identity and integrity of accessions conserved by the International Musa Germplasm Transit Centre (ITC) was established. Several accessions have been proven to be true to type while others were misidentified based on SSR and cytological results (Christelová et al. 2011a; Christelová et al. 2017).

Expressed sequence tags-SSR (EST-SSR) markers were used to clarify the genotype identity in a diploid segregating population from hybrid 6142-1 and 8075-7. The analysis revealed two half-sib populations instead of a single full-sib population (Mbanjo et al. 2012a). Venkatachalam et al. (2008) used a combination of RAPD and inter simple sequence repeat (ISSR) markers to identify and classify the South Indian cultivars. The authors were able to separate global cultivars such as cvs. Williams and Robusta from those that had limited geographical distribution and purely endemic to South India.

3 Genomic prediction

3.1 Overview of genomic selection

Clarification on usage of terms: *Genomic selection is a method of making a decision on which individuals to choose from a population and advance in the breeding process based on the differences in their genomic merit (value). Genomic prediction is a statistical model-based tool that utilizes genomic data to estimate the genomic merit of an individual in a population. Therefore, genomic prediction is a means to genomic selection and the output of genomic prediction that facilitates genomic selection decision is called the genomic estimated breeding value (GEBV).*

Genomic selection (GS) based on genomic prediction models is a form of marker assisted selection (MAS), which allows selection of individuals that have not been phenotyped (Goddard and Hayes 2007; Goddard 2009). It utilizes dense markers that are spread across the genome to predict the genomic breeding value of an individual (Meuwissen et al. 2001; Heffner et al. 2009). As the predictions are based on genomic information, the selection index is called genomic estimated breeding value (GEBV). Genomic selection addresses some limitations of classical MAS and GWAS by simultaneously estimating all marker effects on the trait. Hence, it is suitable for prediction of polygenic traits controlled by many small-effect QTL without a need to identify individual QTL (Heffner et al. 2009) and the associated markers.

Genomic prediction is mostly used for selection of parents for further crossing (Goddard and Hayes 2007). However, Crossa et al. (2014) proposed that genotypic values should also be used to select genotypes with potential for release as new cultivars in maize and wheat breeding. Several modifications to the original genomic selection methodology of Meuwissen et al. (2001) have been proposed and these include: weighted genomic selection, optimal haploid value selection, genotype building selection and optimal population value selection (Goiffon et al. 2017).

Genomic selection has been made possible by high-throughput next generation sequencing technologies that caused a dropdown in genotyping costs and by advances in genotyping methods (Elshire et al. 2011; Deschamps et al. 2012). When dense markers became available through approaches like genotyping by sequencing (Elshire et al. 2011; Poland et al. 2012a), most linear regression models could not handle data where the number of phenotypes, or sample size (n) were less than the number of predictors, or markers (p) (Jannink et al. 2010; de los Campos et al. 2013). To address the issue of small ‘n’ and large ‘p’, Bayesian and kernel methods were developed alongside many other approaches (de los Campos et al. 2009a; Pérez and de los Campos 2014). The Bayesian methods use the Monte Carlo Markov Chain (MCMC) algorithms to sample from a posterior probability distribution (Meuwissen et al. 2001). The posterior distribution of estimates is generated from prior probabilities, which are user defined.

Prior probabilities are very subjective, but can be derived from historical information (Goldstein 2006), like, if one knows the heritability of a trait, or the number of genes controlling the trait. When only prior densities are used, then a non-informative model is generated. The priors are updated when data become available to yield a more realistic posterior probability distribution (Goldstein 2006). Hence, when a lot of data are available, the influence of prior probability on the posterior probability distribution is superseded by the likelihood of the data.

The MCMC algorithms use the Gibbs sampler (Gelfand et al. 1990) and every time a sample is obtained, the model is updated (Meuwissen et al. 2001). The number of iterations that the MCMC must run are pre-set. The user also defines how many iterations should be discarded as burn-in so that the Gibbs sampler does not pick samples from initial values that can bias the mean of estimates. After the burn-in, the interval at which the sampler should collect the samples to update the model is also defined, which is referred to as thin (MacEachern and Berliner 1994). Thinning reduces sample autocorrelation of the Markov chain, which can cause biased Monte Carlo standard errors. It also allows efficient use of computer storage space by reducing the number of posterior samples kept. This means that any number of predictors can be fitted in the model, thus enabling whole-genome regression and prediction (de los Campos et al. 2013). While whole-genome regression is possible, the large amount of data from GBS can still create computational challenges. These have been partly addressed by Bayesian methods that perform variable shrinkage and selection of the linear predictors (de los Campos et al. 2013; Pérez and de Los Campos 2014).

Genomic selection has been successful in dairy cattle for selection of bulls that give female offspring with high milk production (Goddard and Hayes 2007). Traditionally, selection of bulls for milk production depended on the performance of their daughters, which could make the selection cycle very long. In plants, traits such as yield, sensory quality and postharvest qualities can only be determined after harvest, which also increases the selection cycle. The primary advantage of GS is the ability to reduce selection cycle and increase selection intensity that results in faster genetic gain per unit time and cost. Genetic gain (*G*) can be estimated as the product of selection intensity (*i*), prediction accuracy (*r*) and square root of additive genetic variance ($\sqrt{\delta^2_A}$) divided by selection cycle time (*t*). Prediction accuracy is influenced by phenotypic variance (δ_p), which is also influenced by the correlation between the breeder's and farmer's environment while additive genetic variance is influenced by the heritability of the trait. In practice, the breeder can increase genetic gain by increasing the selection intensity (*i*) and by reducing the selection cycle time (*t*) even when the prediction accuracy is low compared to phenotypic selection accuracy (Desta and Ortiz 2014; Bassi et al. 2016).

The predictive abilities of different genomic prediction models have been demonstrated in various crops ranging from cereals to forest trees (Crossa et al. 2010; Heffner et al. 2011; de Oliveira et al. 2012; Kumar et al. 2012; Würschum et al. 2013; Beaulieu et al. 2014; Crossa et al. 2014; Crossa et al. 2016; Onogi et al. 2016; Gezan et al. 2017). However, information concerning use, or performance of genomic prediction models in banana breeding is not available to date. This section of PhD Thesis divulges more of the main developments in the field of genomic predictions to date starting from genotyping by sequencing, then predictive models and computational requirements while putting banana breeding into perspective.

3.2 Genotyping by sequencing: a step towards genomic prediction

Genotyping by sequencing (GBS) is a next generation sequencing-based method that takes advantage of reduced representation libraries to enable high throughput genotyping of large numbers of individuals at a large number of SNP loci (Glaubitz et al. 2014). Advances in sequencing technologies led to reduction in genotyping costs, which caused a rapid growth of sequence databases (Bernardo and Yu 2007). Of all marker types, SNP markers are the most abundant in the genomes of animal and plant species. This makes them the molecular markers of choice for genomic predictions as they satisfy the requirement of dense markers (Bernardo and Yu 2007; Elshire et al. 2011).

To reduce the cost of SNP genotyping without compromising quality, several reduced representation sequencing approaches were developed (Sonah et al. 2013). These include diversity array technology sequencing (DArTseq), restriction site associated DNA (RAD, Baird et al. 2008), genotyping by sequencing (GBS) and reduced representation library (RRL), or complexity reduction of polymorphic sequences (CRoPS) (van Orsouw et al. 2007; Elshire et al. 2011; Beissinger et al. 2013). Of the four, GBS is a low coverage approach, but by far the most advantageous when genotyping large populations. Library construction for GBS is simple and it requires small amounts of starting DNA. The introduction of a barcoding system to samples allows several samples to be multiplexed and sequenced on the same sequencing lane, which reduces the sequencing cost per sample. When a proper choice of restriction enzyme is made, high SNP coverage in gene-rich regions of the genome can be attained in a highly cost-effective manner (Elshire et al. 2011; Sonah et al. 2013). The choice of restriction enzymes for GBS library preparation depends on the number of tags it can generate and the distribution of tags across the genome (Hamblin and Rabbi, 2014). The fewer the tags, the more reads per tag and the better the depth of coverage. However, the tags should be uniformly distributed across the entire genome to get good genomic representation markers. Use of restriction endonuclease *ApeK*I was demonstrated to give good depth of coverage in barley and maize (Elshire et al. 2011).

To improve the robustness of the GBS protocol, Poland et al. (2012a) modified the original GBS protocol by using a two-enzyme approach (*Pst*I/*Msp*I), a rare cutter and a frequent cutter. This approach was used to genotype bi-parental barley and wheat populations and was used to develop a genetically anchored reference map to identify SNP and tags (Poland et al. 2012a). Further studies in wheat were carried out to prove the robustness of GBS in breeding applications (Poland et al. 2012b). Sonah et al. (2013), also improved the standard *ApeK*I protocol by carrying out a final amplification step with selective primers extending across the 3′-*ApeK*I sites by 1 or 2 bases into the insert. With this modification, both the number and depth of coverage of called SNPs were significantly improved. Using the *Pst*I restriction enzyme alone with the standard GBS protocol was also found to give good sequence data. It is a relatively rare cutting enzyme, which generates a moderate number of tags, thus giving more reads with better depth of coverage. The tradeoff is that it gives a lower number of SNP markers (Hamblin And Rabbi, 2014). This is good for genotyping multi-ploidy populations (e.g. banana) that have varying number of alleles at any given locus. In cassava, a combination of *Pst*I and *Taq*I improved the distribution and number of SNP markers (Hamblin and Rabbi, 2014).

Sequence reads from mitochondria DNA (mDNA) and chloroplast DNA (cpDNA) present a problem when mapping reads to a reference genome especially in polyploid plants. For example, in heterozygous autotetraploid potato, cpDNA was shown to represent 60% of total reads (Uitdewilligen et al. 2013). However, in the *M. a.* ssp. *malaccensis* complete chloroplast DNA (cpDNA), only 14 *Pst*I restriction sites were found whereas in the current publicly available banana reference genome (Martin et al. 2016), there are 85714 restriction sites for *Pst*I. This suggests that the number of tags from cpDNA in the sequence library are very few for banana, reducing a possible contamination of nuclear genome sequence reads with organellar DNA sequences even when CTAB DNA extraction protocol is used (Lutz et al. 2011).

Genotyping by sequencing has also some limitations, the main ones being the high level of missing data (Glaubitz et al. 2014), low coverage and non-uniform distribution of sequence reads (Beissinger et al. 2013; Hamblin and Rabbi, 2014). The problem of missing data is usually overcome by imputation methods such as random forest regression, multivariate normal expectation maximum algorithm and impute amongst other methods (Poland et al. 2012b). Proper choice of restriction enzyme during library construction and technical replication during sequencing can also help to improve coverage and reduce missing data.

RAD sequencing in comparison to GBS offers ‘deep-sequencing’ of SNP with a wide range of coverage depending on the requirement of the researcher (Fonseca et al. 2016), while DArT sequencing provides data with a few missing data points both dominant and co-dominant markers (Sansaloni et al. 2011), but the two methods are not yet as cheap as GBS for genotyping large populations.

3.3 Downstream analysis of GBS data

GBS protocol generates millions of short sequences reads, on average 100 bp each using the Illumina sequencing platform. One main requirement for downstream analysis of sequence reads is a reference genome sequence, or DNA contigs from a representative species (Elshire et al. 2011; Perea et al. 2016). Tools such as Burrows-Wheeler alignment (Li and Durbin 2009) and Bowtie 2 (Langmead and Salzberg 3012) work on the principle of Burrows-Wheeler transform (BWT). They were designed to map short reads to the reference sequence in an efficient and accurate manner, but many other read alignment tools exist. Once the reads are aligned to the reference, SNP discovery and genotyping can be done by variant caller tools such as SAMtools, genome analysis toolkit (GATK), or FreeBayes (Clevenger et al. 2015).

The choice of a variant caller depends on the nature of the species under study (Clevenger et al. 2015). Calling SNPs from diploid organisms is straight-forward and also many polyploids with an even ploidy level behave like diploids. However, for autopolyploid species, special considerations must be made (Uitdewilligen et al. 2013). In allopolyploids such as wheat (*T. aestivum*) with three sub-genomes, it is possible to map reads to specific sub-genomes and call SNPs for each genome (Dvorak et al. 2006). Bananas are polyploid and some triploid bananas such as EAHB are composed of three A sub-genomes originating from different subspecies that are not easy to distinguish (Perrier et al. 2011). SNP calling from a banana population comprising individuals of different ploidy levels requires a careful choice of variant caller tools.

Each variant caller has advantages and limitations. For example, SAMtools does not perform well in calling heterozygous SNP, despite being simple to use. In contrast, GATK has many steps and requires special data formats, but it is good for handling species with different ploidy levels and when allele dosage is required. It is also capable of distinguishing true SNP from sequence artifacts. The indel realignment step in GATK improves alignment around indels, which removes frameshifts that usually result in false-positive SNP calls (Polyanovsky et al. 2011; Clevenger et al. 2015).

Bioinformatics workflows and pipelines make SNP calling and genotyping from GBS reads more efficient. Currently, the bioinformatics pipeline that is commonly used is the TASSEL-GBS (Glaubitz et al. 2014). Other bioinformatics pipelines that have been developed include Stacks and next generation sequencing eclipse plugin, NGSEP (Catchen et al. 2011; Perea et al. 2016). They offer flexibility of handling large number of samples with reduced errors. The main characteristics of bioinformatics workflows and pipelines is that they combine the utility of several specific tools and allow the user to specify some parameters although default settings are always provided. Among such tools are the FASTX Toolkit and Picard Tools (http://hannonlab.cshl.edu/fastx_toolkit/; http://broadinstitute.github.io/picard/). Custom requirements are not easy to implement in standard pipelines and this may call for the user to develop a customized workflow to execute specific tasks. The output SNP can be used for GWAS, population structure analysis, genetic diversity studies and genomic predictions (Elshire et al. 2011). Depending on the final use of SNP data, some conversion tools may be required to change the genotype data formats so that the data are compatible with other software. This involves writing Perl scripts, or R functions.

3.4 Genomic prediction models

The basic model commonly used in simple experiments to predict dependent variable given the independent variable data, or vice versa, is the simple linear regression model, or the least squares estimation model given by the formula: $y = \alpha + \beta x + e$, where y is a vector of dependent variables, α is the y intercept, β is the regression coefficient, x is a matrix of independent variable and e is the vector of random residuals. If there are many co-variate factors that influence the outcome variable, then the multiple linear regression model is adopted, which takes the form: $y = \alpha + \beta_1x_1 + \beta_2x_2 + \ldots.\ \beta_nx_n + e$. The utility of these models in genomic selection is limited due to the high number of linear predictors (Crossa et al. 2010).

Animal breeders have for a long time relied on the use of phenotypic data and pedigree information to predict the breeding value of individuals (Goddard and Hayes 2007). Best linear unbiased prediction (BLUP) model has been used to estimate random effects (genetic merit). It is a linear model of the form: $y = X\beta + Zu + e$, where y is a vector of phenotypic observations, β and u are vectors of fixed and random effects, respectively, X and Z are design matrices, e is a vector of random residuals (Robinson, 1991). The advent of next generation sequencing technologies increased access to genotypic data. Integrating these data into prediction models showed an increase in genetic gain per unit time (Meuwissen et al. 2001; Goddard and Hayes 2007; Legarra et al. 2008; Hayes et al. 2009).

Meuwissen et al. (2001) incorporated SNP markers as random variables in BLUP equation in their simulation study. They made an assumption that marker effects were normally distributed and that all loci had equal variance, thus the genetic variance of an individual locus could be obtained by dividing the total genetic variance, V_g by the total number of loci, n (V_g/n). However, in some cases a few loci with major effects, or many loci with varying effects control the trait, and are not uniformly distributed across the genome. This makes the assumption of equal genetic variance unrealistic and leads to model over-parameterization (Resende et al. 2012). Parametric and semiparametric models based on Bayesian principles that perform shrinkage and variable selections were developed as alternatives for use in genomic prediction (de los Campos et al. 2013; Pérez and de Los Campos 2014).

3.4.1 Implementation of genomic prediction

Genomic prediction is implemented in three phases, which include training, validation and breeding (Jannink et al. 2010; Nakaya and Isobe 2012). In the training phase, a panel of genotypes representing the genetic diversity within a breeding program is phenotyped and genotyped. The marker variance and their effect on the trait (regression coefficient) at each locus are estimated and the population's trait mean is obtained from the phenotypic data. This yields a model of the form "predicted phenotype ($\hat{y}$) = general phenotype-mean in the population (intercept, μ) + GEBV ($\sum X\beta$) + residual error (ε)". This can be expressed as $\hat{y} = \mu + \sum X\beta + \varepsilon$, where X is a matrix of independent linear predictors such as SNP markers and β is the regression coefficients of the independent linear predictors. The residual errors could be environmental or spatial errors. When ε is assumed to be random and normally distributed, that is, $\varepsilon = \sim N(0, \delta_{\varepsilon}^{2})$, where δ_{ε}^{2} is the variance of random residuals, then GEBV = $\hat{y} - \mu$ (Pérez and de Los Campos 2014).

The complexity of the above genomic prediction model can be increased by adding a relationship information. This information can be in the form of a genomic relationship matrix (G-matrix), or pedigree matrix (VanRaden 2008). The G-matrix (G) can be calculated from SNP data (X) consisting of score for minor alleles that take the form of 0, 1 and 2 for diploid organisms, where 0 and 2 are homozygous major and minor allele states, respectively, while 1 represents the heterozygous state of a locus. Hence, $G = XX'$, where X' is the transpose of X, which is a data frame of 'n' individuals and 'p' SNP markers.

Pedigree matrix can be calculated when pedigree records are available using the pedigreemm R-package (Vazquez et al. 2010). The choice as to whether a pedigree matrix is added to the model, or not depends on the relationship of individuals in the GS population. When there is a weak relationship, addition of pedigree matrix distorts the relationship based on genomic data causing a reduction in performance of genomic prediction models (Zhong et al. 2009). However, in some cases a combination of pedigree information with marker data was shown to improve the prediction accuracy of genomic prediction models (Crossa et al. 2014).

3.4.2 Estimate of model performance

In genomic selection, predictive ability is a measure of performance of a genomic prediction model and is determined by cross validations. Predictive ability of a model is the correlation between the predicted and observed value of a trait, or the correlation between GEBV and observed phenotype (Crossa et al. 2010). Usually, the correlation between GEBV and predicted phenotype is approximately 1.0. Most studies used five-fold (K=5) and ten-fold (K=10) cross validation (Jannink et al. 2010). However, other strategies are also used. For example, 90 % of the genotypes are used as training set while 10 % as cross validation (testing) set, but there are many other approaches (Crossa et al. 2016). The average correlation of these cross validations is reported as the predictive ability, or prediction accuracy of that model for a trait (Crossa et al. 2014; Crossa et al. 2016). It is important that during cross validation there is no overlap between genotypes in the training set and testing set.

Cross validation is a convenient way of evaluating the accuracy of genomic prediction models. In order to use the genomic prediction model, the accuracy of prediction is first confirmed at the validation phase for breeders to have confidence in the model (Nakaya and Isobe 2012). The validation population should consist of genotypes that are different from those used in the training population. This population is genotyped to allow prediction of the GEBV, then phenotyped preferably in an environment other than that in which the training population was phenotyped (Ly et al. 2013). The correlation between the observed phenotype and GEBV gives the prediction accuracy of the model. To maintain a good performance of the model, the validation and breeding populations must be related to the training population and genomic prediction models have to be updated over time because of linkage disequilibrium decay (Nakaya and Isobe 2012). The data collected from breeding and validation populations can be used to update the genomic prediction model to improve its accuracy (reviewed by Varshney et al. 2013; Ly et al. 2013).

In genomic prediction, the predictive ability value is the proportion of genetic variance explained by marker data. It is often misinterpreted as the proportion of genotypes correctly selected by genomic prediction versus phenotypic selection. As discussed by Bassi et al. (2016), a prediction accuracy of 0.5 does not mean that 50% of the top selected individuals will actually be phenotypically selected. In many cases the percentage of individuals correctly selected based on GEBV has been above the prediction accuracy. For example, Beaulieu et al. (2014) reported that with predictive values between 0.33 and 0.44, they were able to achieve 90 % of traditionally estimated breeding values during validation. Similarly, Heffner et al. (2011) reported a 95 % prediction accuracy of genomic prediction compared to phenotypic selection in a multi-family wheat population even if the predictive values ranged from 0.22 to 0.76. The tradeoff between genomic selection and phenotypic selection is that genomic selection can afford faster genetic gain per unit time, although it is not 100 % accurate as phenotypic selection (Desta and Ortiz, 2014; Bassi et al, 2016).

During the breeding phase, new hybrids from the breeding program are genotyped and the genotype data are fed into a validated genomic prediction model to predict the GEBV. The breeder uses these GEBV to make a decision on which hybrids to select for further crossing, or phenotyping. The model also predicts the likely phenotypic outcome for each hybrid (Pérez and de Los Campos 2014). Selection can be done at the nursery stage so that only hybrids with a good combination of traits are taken to the field for evaluation and the rest are discarded before wasting resources on them. It is important for the breeders to develop the ‘selection index’ of GEBV so that selection is product focused.

‘Selection index’ of GEBV means that among the traits the breeder is predicting, a priority order is set as a way of eliminating hybrids that do not meet product requirements. It is an efficient way of simultaneously selecting for all traits that define a best parent, promising candidate cultivar, or best cross combination (Bassi et al. 2016). If the selection is intended to eliminate hybrids with low genetic value, this can be referred to as negative selection that reduces the phenotyping burden. For example, in banana, most hybrids are triploid and majority show poor fruit filling characteristics. When selecting candidate cultivars, fruit filling trait such as fruit circumference should be given top priority in the ‘selection index’ of GEBV. Once the number of hybrids to phenotype is reduced, more replications can be planted without much strain on financial resources (Heffner et al. 2009), or some evaluation stages such as EET and PYT can be skipped so that hybrids are evaluated faster than usual in multiple locations to reduce the selection cycle. This allow the identification of high performing hybrids with stable traits in a much shorter time.

3.4.3 Types of genomic prediction models

Different studies in both animals and plants have tested the predictive ability, or accuracy of different genomic prediction models (Legarra et al. 2008; Heffner et al. 2011; Kumar et al. 2012; Würschum et al. 2013; Crossa et al.2014; Weng et al. 2016; Momen et al 2017). These models include ridge regression best linear unbiased prediction (rrBLUP), genomic best linear unbiased prediction (GBLUP), best linear unbiased prediction method including a trait-specific relationship matrix (TABLUP), least absolute shrinkage and selection operator (LASSO), Bayesian ridge regression (BRR), Bayesian LASSO (BL), BayesA, BayesB, BayesC, BayesCπ, BayesDπ, elastic net (EN), reproducing kernel Hilbert Space (RKHS), Bayesian neural networks (BNN) and Bayesian regularization for feed-forward neural networks (BRNN) (Robinson 1991; Tibshirani 1996; Meuwissen et al. 2001; Park and Casella 2008; Zhang et al. 2010; Pérez and de Los Campos 2014).

The difference in these models largely lies in how they estimate the marker variance and how they generate the posterior distribution of marker effects (Table 3). They also differ in the assumptions made about traits. Some assume that the traits are controlled by additive genetic effects, while other account for non-additive genetic effects such as dominance and epistasis (e.g. RKHS). The characteristics of these models have been summarized in various publications (Meuwissen et al. 2001; Habier et al. 2011; Pérez and de Los Campos 2014; Desta and Ortiz 2014). In this Thesis, the predictive ability of six models was investigated using different cross validation strategies and these included BRR, BL, BayesA, BayesB, BayesC and RKHS models and a summary of their characteristics is given in Table 3.

The above prediction models were developed and optimized for diploid organisms. However, they have been extended to polyploid organisms (Crossa et al. 2014; Gezan et al. 2017) where a balanced distribution of alleles is assumed to exist as in diploids. Banana is unique in that breeding populations are generated by crossing parents of different ploidy levels, which results in a mixture of diploid, triploid and tetraploid hybrids. The generation of a prediction model with a population consisting of genotypes of different ploidy levels is usually a challenge due to (i) uncertainty of allele frequency in that population and (ii) uncertainty of allele dosage at the loci. Blischak et al. (2015) attempted to address the problem of allele dosage uncertainty in a simulated autopolyploid population. They treated the genotypes as latent variables in a hierarchical Bayesian model and sequence reads as random samples. They concluded that uncertainty of allele dosage in polyploids in addition to number of individuals sampled and sequencing coverage affected the calculation of allele frequencies. Yet, allele frequency is key in population genetics models for understanding allele inheritance patterns.

Table 3: Main characteristics of the six genomic prediction models evaluated in this Thesis

Model characteristics	**BRR**	**BL**	**BayesA**	**BayesB**	**BayesC**	**RKHS**
Parametric	Yes	Yes	Yes	Yes	Yes	
Semiparametric						Yes
Additive genetic effects	Yes	Yes	Yes	Yes	Yes	
Non-additive genetic effect						Yes
Distribution of marker effects	Gaussian	Fixed, Gamma, or Beta	Scaled t	Scaled t	Gaussian	
Distribution of marker variance	X^{-2}	Double exponential	X^{-2}	X^{-2}	X^{-2}	
Uniform shrinkage	Yes					
Nonuniform shrinkage		Yes	Yes	Yes	Yes	
No marker selection	Yes					
Variable marker selection		Yes	Yes	Yes	Yes	
Prior probability of marker effect				Yes	Yes	

Source: Desta and Ortiz (2014) and Pérez and de Los Campos (2014)

In bananas, the expected level of heterozygosity varies with ploidy level. For example, if a bi-allelic SNP, A/G is segregating at locus *i*, then, one, two and three possible heterozygotes are expected in diploids (AG), triploids (AAG and AGG) and tetraploids (AAAG, AAGG and AGGG), respectively. Determining the level of heterozygosity at a locus depends on how well the sequencing reads represent the true genotype and the choice of bioinformatics tools used during SNP calling. Picard tools allow normalization of sequencing reads by marking and removing duplicates so that genomic regions with fewer reads that are uniquely mapped are not excluded during SNP calling. In addition, GATK has an option of setting the ploidy level during SNP calling with UnifiedGenotyper that allows heterozygosity to vary according to ploidy level (Clevenger et al. 2015). This is very useful when dealing with populations of mixed ploidy levels. Genomic prediction models use marker data in a numeric form. In order to maintain allele dosage status of the SNP data, careful choice of tools that convert SNP data to numeric format is important. R-based script named AlleleDosage R function was developed as part of this Thesis to address this issue. The script can be accessed from the link.

http://olomouc.ueb.cas.cz/system/files/users/public/scripts/AlleleDosage_R_function.docx

Other than allele dosage and mixed ploidy population, several factors influence the predictive ability of genomic prediction models. They include size and composition of the training population, the relationship between training and breeding populations, differences in linkage disequilibrium between markers and QTL across training and breeding populations, number of markers used, the interaction between genotype and environment, and heritability of the trait (Crossa et al. 2016; Bassi et al. 2016). In order to reach high predictive ability, the population should be large enough to capture most of the segregating alleles in the breeding gene pool. As noted by Bassi et al. (2016), no ideal population size exists for all species and traits. Hence, attention should be paid to how related the individuals are, the heritability of the trait, whether the population is bi-parental, or a mixture of several families and the cost involved in phenotyping the training population. The breeding population should come from genotypes that were involved in the training phase. The number of markers should be large enough so that at least one, or more markers are in linkage disequilibrium with the QTL controlling the trait (Myles 2013; Desta and Ortiz 2014). GBS gives many SNP markers that improves the prediction accuracy of the genomic prediction models compared to other platforms that give fewer markers with less missing data (Heslot et al. 2013).

Increasing the size of a training population has been shown to increase prediction accuracy and most studies have used training populations ranging from 200 to 10,000 individuals (Lorenz et al 2011). The gain in prediction accuracy due to increase in population size has a threshold beyond which it plateaus, or makes no economic sense. Banana populations are expensive to phenotype as each banana plant occupies 6 m^2 of field space for at least two, or three crop cycles (Tushemereirwe et al. 2015). To obtain representative phenotypic data, each clone has to be replicated within the experimental plot and in several locations. Phenotyping thousands of banana clones requires very large fields and the cost would be exorbitant in comparison to phenotyping the same number of genotypes in cereals like wheat, which require about 0.054 m^2 (18 cm between plants x 30 cm between rows) per plant, i.e. 0.9 % of what is needed for banana.

Therefore, the effective size of a training population required to achieve a high accuracy of the genomic prediction model depends on the population under study (Goddard 2009) and the heritability of trait of interest (Lorenz et al. 2011). Many breeding programs, including animal breeding use a small number of parental lines that constitute the effective breeding population. Animal breeders, however, keep phenotypic and genotypic records from many progenies around the world and these constitute an effective training population, which makes genomic prediction relatively easy to implement at no substantial cost (vanRaden et al. 2009).

In barley, the effective breeding population size is reported to be less than 50 lines (reviewed by Lorenz et al. 2011). Regardless of the number of parental lines used in a breeding program, data from many progenies resulting from crosses between parents is beneficial in genomic prediction. Unlike QTL mapping, the training population for genomic selection is not necessarily derived from bi-parental crosses, but is rather a collection of representative genotypes from a breeding program where genomic prediction is to be applied (Heffner et al. 2009; reviewed by Mammadov et al. 2012). This makes it convenient to investigate the utility of genomic prediction in banana where the effective breeding population is small, and segregating populations for different traits are limited, or completely missing.

During genomic prediction model development, consideration for the interaction of genotype by environment should be made because it leads to differences in phenotypic expressions of some trait (Manrique and Hermann, 2000). Traits that are strongly controlled by the genotype are more stable across different environments as compared to those controlled by environment. The G × E interaction effect analysis is useful in studying trait heritability and stability in breeding materials (Taghouti et al. 2010). Generally, genomic prediction models that use average environment data have been shown to be more robust than those based on a single environment (Burgueño et al. 2012). The challenge in banana is that we do not know what traits are stable across environments due to lack of systematic research.

3.5 Computational and software requirements for genomic prediction

With the fast progress in DNA sequencing technologies, computation challenges arose to cope with the massively generated sequence data (Metzker 2010). The main challenges include efficient storage, retrieval and processing of such huge data with reduced error at reduced cost (Wang et al. 2009). Most breeding programs do not have funds and technical capacity do establish such facilities. However, these services can be outsourced from private service providers. The challenge comes when standard protocols cannot deliver all the breeder needs to answer certain questions. Customizing a protocol for a onetime user, or a few users is very expensive. This means that the breeder should have the capacity to perform these specialized analyses. This is possible if several breeding programs come together and establish a synergy that helps to improve the breeding process even in small, financially less privileged breeding programs (Hickey et al. 2017).

Numerous bioinformatics tools have been developed to perform individual tasks such as alignment of short reads to the reference genome, *de novo* assembly of reads into contigs for organisms without reference genome, SNP calling tools, diversity analysis software and much more. Some of these tools are in the form of bioinformatics kits, or pipelines and freely available to the public, or commercialized. For example, Galaxy tools from galaxyproject.org and the genomic association and prediction integrated tool (GAPIT) from Cornell University (Lipka et al. 2014) are freely available while other are commercialized like for example, CLC genomic workbench and others. Bioinformatics pipelines such as TASSEL-GBS have been developed to help circumvent problems associated with handling GBS data (Glaubitz et al. 2014).

In genomic prediction, statistical modeling is crucial, yet GBS presents a lot of missing data and accurate imputations are still a challenge for polyploid crops. In order for genomic selection to be embraced by breeders, flexible statistical software that allows breeders to analyze massive genomic data in real time and requires less sophisticated computer systems is of importance. The R environment from www.r-project.org provides many packages that facilitate statistical modeling of biological data. Through integrative packages in R, genomic and phenotypic data can be analyzed together to generate genomic prediction models and to test their accuracy. One example is the Bayesian generalized linear regression (BGLR) R package used for genomic predictions (Pérez and de Los Campos 2014).

3.6 Prospects of genomic prediction

Molecular markers have contributed enormously to the understanding of genetic diversity within banana germplasm. They have been used to clarify taxonomic classification, identify cultivars and track pedigrees in breeding populations. However, little progress has been made in using DNA markers for routine breeding and selection of candidate cultivars and breeding parents. With advances in molecular marker technology, it is expected that genomic selection as a form of MAS will play a major role in improving the efficiency of conventional crossbreeding.

Breeding recalcitrant crops and ensuring timely delivery of hybrids to farmers that address issues of food security and income through sustainable production is the dream of every banana breeder. Application of genomic predictions in banana breeding is quite timely as resources are always small to support long-term programs. However, more is yet to be understood about this field of applied biology in crop breeding. In the initial stages, resources need to be directed in developing efficient, accurate and cost-effective phenotyping technologies as well as building necessary capacities in breeding teams to implement genomic prediction.

Banana breeding requires multidimensional and interdisciplinary approaches involving breeders, floral biologists, molecular biologists, geneticists, cytogeneticists, bioinformaticians, biostatisticians, agronomists and farmers/consumers (Hickey et al. 2017). Therefore, there is ultimate need to establish a banana interactive resource database (Musabase) to maintain global *Musa* genotypes and phenotypic information with easy to use bioinformatics pipelines and statistical packages for breeders. Although this may be farfetched, once achieved the benefits could be remarkable. A recent publication by Ruas et al. (2017), which shows effort to link different databases for banana information resources is a good starting point, but more is still required. Trait based models need to be developed and validated for routine use in banana breeding programs to increase genetic gain.

4 Goals of the Thesis

The main goal of this Thesis is to present empirical evidence on the performance of genomic prediction models in banana breeding based on SNP marker data obtained by the genotyping by sequencing approach. The Thesis summarizes the current knowledge about bananas, including production constraints, breeding strategies, use of molecular markers in banana research and the need to accelerate conventional crossbreeding by using genome-wide markers through genomic predictions. Special emphasis was directed towards developing and understanding the predictive ability of six genomic prediction models (BRR, BL, BayesA, BayesB, BayesC and RKHS) and how factors such as field management and crop cycle affect trait variation in genotypes and the predictive ability of the prediction models for a set of 15 traits. The working hypothesis was that field management and crop cycle had no influence on trait expression and predictive ability of genomic prediction models. To achieve the above objective, the following specific objectives were pursued through experimental analysis and the results obtained are summarized in publications:

1. To assess the variation and correlation of traits in the genomic selection training population with respect to crop cycles and field management.
2. To determine the genetic diversity of the genomic selection training population.
3. To compare the predictive ability of a set of six models with marker, pedigree and both pedigree and marker information for fifteen traits scored in the training population and select the best genomic prediction model for each trait, or a group of traits.
4. To determine the predictive ability of models with a training population grown under two different field management practices (Genotype × Environment interaction).
5. To determine the predictive ability of the best model for prediction of traits within and across crop cycle 1 / mother plants and crop cycle 2 / first ratoons/first suckers (Genotype × Cycle interaction)
6. To determine the effect of accounting for allele dosage on the predictive ability of the best genomic prediction model for each trait.
7. To determine the effect of using genomic prediction models fitted with averaged environment data and allele dosage SNP markers in the prediction of genotype performance in particular environments.
8. To determine the accuracy of selection achieved based on GEBV relative to phenotypic data within the training population.

5 General conclusion and recommendations

The aim of this Thesis was to develop and evaluate the predictive ability of genomic prediction models in a banana genomic selection training population. Among all models tested ((BRR, BL, BayesA, BayesB, BayesC and RKHS), BayesB was superior in prediction for most traits, hence, breeders could use it on all traits tested. Fruit filling and fruit bunch traits were predicted quite well in all cross-validation strategies. This implies that negative selection could be applied in breeding program to reduce the burden of phenotyping hybrids with inferior fruits. Although the training population was composed of genotypes of different ploidy levels, accounting for allele dosage in SNP markers (AD-SNP) reduced predictive ability relative to traditional bi-allelic SNP (BA-SNP), but the prediction trend remained the same across traits. However, for some traits, accounting for allele dosage may be necessary. A script to account for allele dosage (AlleleDosage R function) was developed and can be customized depending on the user's requirements and it could be applicable on all polyploid species. The R-script can be accessed from the following link:
http://olomouc.ueb.cas.cz/system/files/users/public/scripts/AlleleDosage_R_function.docx

The high correlation observed between traits within trait categories (plant stature, suckering behaviour, black leaf streak resistance, fruit bunch and fruit filling) during phenotypic analysis was confirmed by the predictive values. Hence, breeders do not need to predict all traits in order to make a decision on which hybrids to select as parents for further crossing, or as promising candidate cultivars. Focus should be on one, or two traits that are easy to phenotype in each trait category. Finally, phenotype data from all field trials should be used to train the prediction model so that the model is robust enough to predict the performance of new hybrids in the phenotyping environment.

The immediate application of the prediction models is to select against triploid hybrids without edible fruits because they constitute the biggest percentage of hybrids in banana breeding and yet, they have no further use in breeding. In the diploids and tetraploids, genomic predictions will help in identifying the best parents for crosses. It is expected that when banana breeding increases the number of hybrids produced, genomic prediction will be a valuable tool during the selection process to improve the genetic gain per unit time and cost.

When implementing genomic selection at the breeding phase, the best parental clones and the best promising candidate, or new cultivar should be the first priority. In order to maximise genetic diversity, two alternatives are proposed. (1) After selecting the top 5 %, the best genotype in each family should be also selected for phenotyping. (2) After selecting the best genotypes, include about 5 % of genotypes with median and worst GEBV for phenotyping as well. Since the genotypic data will be already available, these data sets will be important for updating the models once prediction accuracies decrease due to changes in allele frequencies. Also, it will help in maintaining some rare alleles that could be totally lost if selection focuses on the top best.

If genomic predictions are to be employed in breeding Mchare bananas and Plantain, separate training populations have to be assembled, phenotyped and genotyped because of differences in allele frequencies, trait expression and linkage disequilibrium. Selection of genotypes for the training population should aim at multiple families. Hybrids that show segregation for various traits within each family should be included in order to capture the additive and non-additive genetic effects like heterosis very well. A minimum of 20 genotypes per family is recommended for 15 to 25 families. However, if the cross combinations are many and involve many half-sib families the number may be reduced so that a target training population of 300-500 is achieved.

For EAHB, Mchare and Plantain breeding programs, routine screening of ploidy level using flow cytometry should be done while the plants are still in the nursery. This will help during selection process as the genomic selection criterial for triploids would be slightly different from diploids and tetraploids based on the ‘selection index’ of GEBV.

Given the high prediction of fruit filling, genome-wide association studies should be conducted to identify the loci and SNP markers associated with this trait. This could facilitate development of PCR-based markers alongside genomic prediction for routine diagnosis of the trait by breeding programs.

Sensory and postharvest quality traits should be recorded on the training population so that genomic prediction models are developed for such traits before terminating the trials. Also, the fertility of improved triploids should be tested with other male parents that are not in their pedigree so that progressive breeding is practiced in banana. This could allow the secondary triploids to serve a pathway for gene pyramiding.

6 Publications

6.1 Moses Nyine, Brigitte Uwimana, Nicolas Blavet, Eva Hřibová, Helena Vanrespaille, Michael Batte, Violet Akech, Allan Brown, Jim Lorenzen, Rony Swennen, Jaroslav Doležel (2017) Genomic Prediction in a Multiploid Crop: Genotype by Environment Interaction and Allele Dosage Effects on Predictive Ability in Banana. The Plant Genome [accepted on 19 December 2017]

Abstract

Improving the efficiency of selection in conventional crossbreeding is a major priority in banana (Musa spp.) breeding. Routine application of classical marker assisted selection (MAS) is lagging in banana due to limitations in MAS tools. Genomic selection (GS) based on genomic prediction models can address some limitations of classical MAS, but the use of GS in banana has not been reported to date. The aim of this study was to evaluate the predictive ability of six genomic prediction models for 15 traits in a multi-ploidy training population. The population consisted of 307 banana genotypes phenotyped under low and high input field management conditions for two crop cycles. The single nucleotide polymorphism (SNP) markers used to fit the models were obtained from genotyping by sequencing (GBS) data. Models that account for additive genetic effects provided better predictions with 12 out of 15 traits. The performance of BayesB model was superior to other models particularly on fruit filling and fruit bunch traits. Models that included averaged environment data were more robust in trait prediction even with a reduced number of markers. Accounting for allele dosage in SNP markers (AD-SNP) reduced predictive ability relative to traditional bi-allelic SNP (BA-SNP), but the prediction trend remained the same across traits. The high predictive values (0.47 – 0.75) of fruit filling and fruit bunch traits show the potential of genomic prediction to increase selection efficiency in banana breeding.

6.2 Moses Nyine, Brigitte Uwimana, Rony Swennen, Michael Batte, Allan Brown, Pavla Christelová, Eva Hřibová, Jim Lorenzen, Jaroslav Doležel (2017) Trait variation and genetic diversity in a banana genomic selection training population. PLoS ONE 12(6): e0178734. https://doi.org/10.1371/journal.pone.0178734.

Abstract

Banana (*Musa* spp.) is an important crop in the African Great Lakes region in terms of income and food security, with the highest per capita consumption worldwide. Pests, diseases and climate change hamper sustainable production of bananas. New breeding tools with increased crossbreeding efficiency are being investigated to breed for resistant, high yielding hybrids of East African Highland banana (EAHB). These include genomic selection (GS), which will benefit breeding through increased genetic gain per unit time. Understanding trait variation and the correlation among economically important traits is an essential first step in the development and selection of suitable genomic prediction models for banana. In this study, we tested the hypothesis that trait variations in bananas are not affected by cross combination, cycle, field management and their interaction with genotype. A training population created using EAHB breeding material and its progeny was phenotyped in two contrasting conditions. A high level of correlation among vegetative and yield related traits was observed. Therefore, genomic prediction models could be developed for traits that are easily measured. It is likely that the predictive ability of traits that are difficult to phenotype will be similar to less difficult traits they are highly correlated with. Genotype response to cycle and field management practices varied greatly with respect to traits. Yield related traits accounted for 31–35% of principal component variation under low and high input field management conditions. Resistance to Black Sigatoka was stable across cycles but varied under different field management depending on the genotype. The best cross combination was 1201K-1xSH3217 based on selection response (R) of hybrids. Genotyping using simple sequence repeat (SSR) markers revealed that the training population was genetically diverse, reflecting a complex pedigree background, which was mostly influenced by the male parents.

7 References

Asíns, M.J. (2002) Review: Present and future of quantitative trait locus analysis in plant breeding. Plant Breed 121: 281-291.

Audran, C., C. Borel, A. Frey, B. Sotta, C. Meyer, T. Simonneau and A. Marion-Poll (1998) Expression studies of the zeaxanthin epoxidase gene in *Nicotiana plumbaginifolia*. Plant Physiol 118: 1021-1028.

Baird, N.A., P.D. Etter, T.S. Atwood, M.C. Currey, A.L. Shiver, Z.A. Lewis, E.U. Selker, W.A. Cresko and E.A. Johnson (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. Plos One 3(10): e3376. doi: 10.1371/journal.pone.0003376

Bassi, F.M., A.R. Bentley, G. Charmet, R. Ortiz and J. Crossa (2016) Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). Plant Sci 242: 23-36. doi: 10.1016/j.plantsci.2015.08.021

Batte, M., A. Mukiibi, R. Swennen, B. Uwimana, L. Pocasangre, H.P. Hovmalm, M. Geleta and R. Ortiz (2017) Suitability of existing *Musa* morphological descriptors to characterize East African highland 'matooke' banana. Genet Resour Crop Evol [Acceptd].

Beaulieu, J., T. Doerksen, S. Clément, J. MacKay and J. Bousquet (2014) Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. Heredity 113: 343-352.

Beissinger, T.M., C.N. Hirsch, R.S. Sekhon, J.M. Foerster, J.M. Johnson, G. Muttoni, B. Vaillancourt, C.R. Buell, S.M. Kaeppler and N. de Leon (2013) Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics 193: 1073-1081.

Bernardo, R., and J. Yu (2007) Prospects for genome-wide selection for quantitative traits in maize. Crop Sci 47: 1082-1090.

Bhat K.V., R.L. Jarret and R.S. Rana (1995) DNA profiling of banana and plantain cultivars using random amplified polymorphic DNA (RAPD) and restriction fragment length polymorphism (RFLP) markers. Electrophor 16(1): 1736-1745. doi: 10.1002/elps.11501601287

Bhat, K.V., S.R. Bhat and K.P.S. Chandel (1992b). Survey of isozyme polymorphism for clonal identification in *Mus*a. I. esterase, acid phosphatase and catalase. J Hortic Sci Biotechnol 61(4): 501-508.

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

Blischak, P.D., L.S. Kubatko and A.D. Wolfe (2016) Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. Mol Ecol Resour 16(3): 742-754. doi: 10.1111/1755-0998.12493

Boonruangrod, R., S. Fluch and K. Burg (2009) Elucidation of origin of the present day hybrid banana cultivars using the 5´ETS rDNA sequence information. Mol Breed 24:77-91. doi: 10.1007/s11032-009-9273-z

Borevitz, J.O. and M. Nordbarg (2003) The impact of genomics on the study of natural variations is *Arabidopsis*. Plant Physiol 132(2): 718-725.

Brown, A., R. Tumuhimbise, D. Amah, B. Uwimana, M. Nyine, H. Mduma, D. Talengera, D. Karamura, J. Kuriba, and R. Swennen (2017) Bananas and plantains (Musa spp.). In: Campos, H. and P.D.S. Caligari, (edts) Genetic improvement of tropical crops. 219-240. doi : 10.1007/978-3-319-59819-2

Buddenhagen, I.W. (2008) Bats and disappearing wild bananas: Can bats keep commercial bananas on supermarket shelves? Bats Magazine 26(4): 1-17.

Burgueño, J., G. de los Campos, K. Weigel and J. Crossa (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci 52: 707-719. doi: 10.2135/cropsci2011.06.0299

Catchen, J.M., A. Amores, P. Hohenlohe, W. Cresko and J.H. Postlethwait (2011) Stacks: building and genotyping loci *de novo* from short-read sequences. G3 1(3): 171-182. doi: 10.1534/g3.111.000240

Ceballos, H., R.S. Kawuki, V.E Gracen, G.C. Yencho, and C.H. Hershey (2015) Review: Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. Theor Appl Genet 128(9): 1647-1667.

Choudhary, K., O.P. Choudhary and N.S. Shekhawat (2008) Marker Assisted Selection: A Novel Approach for Crop Improvement. Am-Euras J Agron 1(2): 26-30.

Christelová, P., E. De Langhe, E. Hřibová, J. Čížková, J. Sardos, M. Hušáková, I. Van den houwe, A. Sutanto, A.K. Kepler, R. Swennen, N. Roux and J. Doležel (2017) Molecular and cytological characterization of the global Musa germplasm collection provides insights into the treasure of banana diversity. Biodivers Conserv 26(4): 801-824.

Christelová, P., M. Valárik, E. Hřibová, E. De Langhe, and J. Doležel (2011b) A multi gene sequence-based phylogeny of the *Musaceae* (banana) family. BMC Evol Biol 11: 103. doi: 10.1186/1471-2148-11-103

Christelová, P., M. Valárik, E. Hřibová, I. van den Houwe, S. Channeliére, N. Roux and J. Doležel (2011a) A platform for efficient genotyping in Musa using microsatellite markers. AoB Plants 2011: plr024. doi:10.1093/aobpla/plr024

Čížková, J., E. Hřibová, P. Christeloá, I. van den Houwe, M. Häkkinen, N. Roux, R. Swennen and J. Doležel (2015) Molecular and Cytogenetic Characterization of Wild *Musa* Species. PLoS One 10(8): e0134096. doi: 10.1371/journal.pone.0134096

Clevenger, J., C. Chavarro, S.A. Pearl, P. Ozias-Akins, and S.A. Jackson (2015) Single nucleotide polymorphism identification in polyploids: A review, example and recommendations. Mol Plant 8: 831-846.

Collard, B.C.Y., and B.J. Mackill (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos Trans R Soc B Biol Sci 363: 557-572.

Collard, B.C.Y., M.Z.Z. Jahufer, J.B. Brouwer and E.C.K. Pang (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142:169-196.

Creste, S., A.T. Neto, R. Vencovsky, S-O. Silva and A. Figueira (2004) Genetic diversity of *Musa* diploid and triploid accessions from the Brazilian banana breeding program estimated by microsatellite markers. Genet Resour Crop Evol 51: 723-733.

Crossa J, D. Jarquín, J. Franco, P. Pérez-Rodríguez, J. Burgueño, C. Saint-Pierre, P. Vikram, C. Sansaloni, C. Petroli, D. Akdemir, C. Sneller, M. Reynolds, M. Tattaris, T. Payne, C. Guzman, R.J. Peña, P. Wenzl and S. Singh (2016) Genomic prediction of gene bank wheat landraces. G3 6: 1819-1834. doi: 10.1534/g3.116.029637

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgũen, J. L. Araus, D. Makumbi, R. P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger and H-J. Braun (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713-724.

Crossa, J. P. Pérez, J. Hickey, J. Burgueño, L. Ornella, J. Cerón-Rojas, X. Zhang, S. Dreisigacker, R. Babu, Y. Li, D. Bonnett and K. Mathews (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112: 48-60. doi: 10.1038/hdy.2013.16

Crouch, H.K., J.H. Crouch, S. Madsen, D.R. Vuylsteke and R. Ortiz (2000) Comparative analysis of phenotypic and genotypic diversity among plantain landraces (*Musa* spp., AAB group). Theor Appl Genet 101: 1056-1065.

Crouch, J.H., H.K. Crouch, A. Tenkouano and R. Ortiz (1999) VNTR-based diversity analysis of 2x and 4x full-sib *Musa* hybrids. Electron J Biotechnol 2: 130-139.

Crouch, J.H., R.Ortiz and H.K. Crouch (2000) Utilization of molecular genetic techniques in support of plantain and banana improvement. Acta Hortic 540:185–191.

Czislowski, E., S. Fraser-Smith, M. Zander, W.T. O'Neill, R.A. Meldrum, L.T.T. Tran-Nguyen, J. Batley and E.A.B. Aitken (2017) Investigating the diversity of effector genes in the banana pathogen, *Fusarium oxysporum* f.sp. *cubense*, reveals evidence of horizontal gene transfer. Plant Mol Pathol [Accepted] doi: 10.1111/mpp.12594

da Fonseca, R.R., A. Albrechtsen, G.E. Themudo, J. Ramos-Madrigal, J.A. Sibbesen, L. Maretty, M.L. Zepeda-Mendoza, P.F. Campos, R. Heller and Ricardo J. Pereira (2016) Next-generation biology: Sequencing and data analysis approaches for non-model organisms. Marine Genomics 30: 3-13. doi: 10.1016/j.margen.2016.04.012

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

Davey M. W., R. Gudimella, J.A. Harikrishna, L. W. Sin, N. Khalid and J. Keulemans (2013) A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter- and intra-specific *Musa* hybrids. BMC Genomics 14: 683.

De Buck, R.E.S. and R. Swennen (eds), (2016) Fact series, bananas the green gold of the South. VIB pp 1-55.

de Jesus, O.N., S. de Oliveira e Silva, E. P. Amorim, C.F. Ferreira, J.M. Salabert de Campos, G-G. Silva and A. Figueira (2013) Genetic diversity and population structure of *Musa* accessions in ex situ conservation. BMC Plant Biol 13: 41.

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

de los Campos, G., D. Gianola, and G.J.M. Rosa (2009a) Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J Anim Sci 87(6): 1883-1887.

de los Campos, G., J.M. Hickey, R. Pong-Wong, H.D. Daetwyler and M.P.L. Calus (2012) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193: 327-345.

de Oliveira, E.J., M.D.V. de Resende, V.S. Santos, C.F. Ferreira, G.A.F. Oliveira, M.S. da Silva, L.A. de Oliveira and C.I. Aguilar-Vildoso (2012) Genome-wide selection in cassava. Euphytica 187(2): 263-276. doi: 10.1007/s10681-012-0722-0

Deschamps, S., V. Llaca and G.D. May (2012) Genotyping-by-sequencing in plants. Biology 1(3): 460-483. doi: 10.3390/biology1030460

Desta, Z.A., and R. Ortiz (2014) Review: Genomic selection: Genome-wide prediction in plant improvement. Trends Plant Sci 19(9): 592-601. doi: 10.1016/j.tplants.2014.05.006

do Amaral, C.M., J. de Ameida dos Santos-Serejo, S. de Oliveira e Silva, C.A. da Silva Ledo and E. P. Amorim (2015) Agronomic characterization of autotetraploid banana plants derived from 'Pisang Lilin' (AA) obtained through chromosome doubling. Euphytica 202: 435-443.

Dochez, C. (2004) Breeding for resistance to *Radopholus similis* in East African highland bananas (*Musa* spp.). PhD thesis, KU Leuven, 1-222.

Doležel, J. (1997) Application of flow cytometry for the study of plant genomes. J Appl Genet 38: 285-302.

D'Hont A., F. Denoeud, J. M. Aury, F. C. Baurens, F. Carreel, et al. (2012) The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature Letter 488: 213-219.

Dvorak, J., E.D. Akhunov, A.R. Akhunov, K.R. Deal and M-C.Luo (2006) Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from tetraploid wheat to hexaploidy wheat. Mol Biol Evol 23: 1386-1396.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6(5): e19379. doi: 10.1371/journal.pone.0019379

Emediato F.L., F.A.C. Nunes, C.C. Teixeira, M.A.N. Passos, D.J. Bertioli, G.J. Pappas Jr and R.N.G. Miller (2009) Characterization of resistance gene analogs in *Musa acuminata* cultivars contrasting in resistance to biotic stresses**.** In**:** Q.Y. Shu (ed), Induced plant mutations in the genomics era. Food and Agriculture Organization of the United Nations, Rome, 443-445.

Evans, E. and F. Ballen (2012). Banana markets. IFAS Extension, FE901.

Fauré S., J.L. Noyer, J.P. Horry, F. Bakry, C. Lanaud and D. Goazalez de Lean (1993) A molecular marker-based linkage map of diploid bananas (*Musa acuminata).* Theor Appl Genet 87:517-526.

Foolad M.R. (2007) Review article: Genome mapping and molecular breeding of tomato. Int J Plant Genomics 64358. doi:10.1155/2007/64358

FAO (1998). "Banana exports by country". Retrieved from http://www.NationMaster.com/graph/agr_ban_exp-agriculture-banana-exports.

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

FAO (2014) Banana market review and banana statistics 2012-2013.

Gelfand, A.E., S.E. Hills, A. Racine-Poon, and A.F.M Smith (1990) Illustration of Bayesian inference in normal data models using Gibbs sampling. J Am Stat Assoc 85: 972-985.

Gezan, S.A., L.F. Osorio, S. Verma and V.M. Whitaker (2017) An experimental validation of genomic selection in octoploid strawberry. Hortic Res 4. doi: 10.1038/hortres.2016.70

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q.M. Sun and E.S. Buckler (2014) TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS One 9(2): e90346. doi: 10.1371/journal.pone.0090346

Goddard, M. (2009) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136(2): 245-57. doi: 10.1007/s10709-008-9308-0

Goddard, M.E. and B.J. Hayes (2007) Genomic selection. J Anim Breed Genet 124(6): 323-330. doi: 10.1111/j.1439-0388.2007.00702.x

Goiffon, M., A. Kusmec, L. Wang, G. Hu and P.H. Schnable (2017) Improving response in genomic selection with a population-based selection strategy: Optimal population value selection. Genetics 206: 1675-168. doi: 10.1534/genetics.116.197103

Gold, C.S., G.H. Kagezi, G. Night and P.E. Ragama (2004) The effects of banana weevil, *Cosmopolites sordidus*, damage on highland banana growth, yield and stand duration in Uganda. Ann Appl Biol 145: 263-269.

Goldstein, M. (2006) Subjective Bayesian analysis: Principles and practice. In: Bayesian Anal 1(3): 403-420.

Guo, Z., D.M. Tucker, J. Lu, V. Kishore and G. Gay (2011) Evaluation of genome-wide selection efficiency in maize nested association mapping populations. Theor Appl Genet 124(2): 261-75. doi: 10.1007/s00122-011-1702-9

Habier, D., R.L. Fernando, K. Kizilkaya and D.J. Garrick (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186. doi: 10.1186/1471-2105-12-186

Häkkinen, M. (2013) Reappraisal of sectional taxonomy in *Musa* species (*Musaceae*). Taxon 62(4):809-813.

Hamblin, M.T. and I.Y. Rabbi (2014) The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: A study in cassava (*Manihot esculenta*). Crop Sci 54: 2603-2608. doi: 10.2135/cropsci2014.02.0160

Harper G., J.O. Osuji, J. S. Heslop-Harrison and R. Hull (1999) Integration of banana streak badnavirus into the *Musa* genome: molecular and cytogenetic evidence. Virology 255: 207-13.

Hayes, B. and M. Goddard (2010) Genome-wide association and genomic selection in animal breeding. Genome 53(11):876-883. doi: 10.1139/G10-076

Hayes, B., P. Bowman, A. Chamberlain and M. Goddard (2009) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433-443.

Heffner, E.L., J-L. Jannink and M.E. Sorells (2011) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4: 65-75. doi: 10.3835/plantgenome2010.12.0029

Heffner, E.L., M.E. Sorells, and J-L. Jannink (2009) Review & Interpretations: Genomic selection for crop improvement. Crop Sci 49: 1-12.

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

Heslot, N., J. Rutkoski, J. Poland, J-L. Jannink and M.E. Sorrells (2013). Impact of marker ascertainment on genomic selection accuracy and estimates of genetic diversity. PLoS One 8(9): e74612. doi: 10.1371/journal.pone.0074612

Hickey, J.M., T. Chiurugwi, I. Mackay and W. Powell (2017) Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nat Genet 49(9): 1297-1303. doi: 10.1038/ng.3920

Hippolyte, I., F. Bakry, M. Seguin, L. Gardes, R. Rivallan, A.M. Risterucci, C. Jenny, X. Perrier, F. Carreel, X. Argout, P. Piffanelli, A.I. Khan, R.N.G. Miller, G.J. Pappas, D. Mbéguié-A-Mbéguié, T. Matsumoto, V. De Bernardinis, E. Huttner, A. Kilian, F.C. Baurens, A. D'Hont, F. Cote, B. Courtois and J.C. Glaszmann (2010) A saturated SSR/DArT linkage map of *Musa acuminata* addressing genome rearrangements among bananas. BMC Plant Biol 10: 65.

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

Horry, J.P. (1988) Distribution of anthocynins in wild and cultivated banana varieties. Phytochemistry 27: 2667-2672.

Horry, J.P. (2011) The use of molecular markers in the CIRAD banana breeding programme. Acta Hortic 897.

Howell, E.C., H.J. Newbury, R. Swennen, L.A. Withers and B.V. Ford-Lloyd (1994) The use of RAPD for identifying and classifying *Musa* germplasm. Genome 37: 328-332.

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

Ignatov, A. and A. Bodishevskaya (2011). *Malus.* In: Kole C. (ed), wild crop relatives: genomic and breeding resources temperate fruits. Springer 45-64.

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

INIBAP (1999). Networking banana and plantain: INIBAP annual report 1998, 1-64.

Jannink, J-L., A.J. Lorenz and H. Iwata (2010) Genomic selection in plant breeding: From theory to practice. Brief Funct Genomics 9(2): 167-177. doi: 10.1093/bfgp/elq001

Janssen, T. and K. Bremer 2004. The age of major monocot groups inferred from 800+ rbcL sequences. Bot J Linn Soc 146: 385-398.

Janssens, S.B., F. Vandelook, E. De Langhe, B. Verstraete, E. Smets, I. Vandenhouwe and R. Swennen (2016) Evolutionary dynamics and biogeography of *Musaceae* reveal a correlation between the diversification of the banana family and the geological and climatic history of Southeast Asia. New Phytol doi: 10.1111/nph.13856

Jarret, R.L., D.R. Vuylsteke, N.J. Gawel, R.B. Pimentel and L.J. Dunbar (1993) Detecting genetic diversity in diploid bananas using PCR and primers from a highly repetitive DNA sequence. Euphytica 68(1): 69-76. doi: 10.1007/BF00024156

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

Karamura, D.A., (1998). Numerical taxonomic studies in the East African highland banana (*Musa* AAA-East Africa) in Uganda. Ph D thesis, University of Reading, 1-192.

Karamura, E., E. Frison, D.A. Karamura and S. Sharrock (1998) Banana production systems in eastern and southern Africa. In: Bananas and food security. INIBAP 401.

Kaemmer, D., D. Fischer, R. L. Jarret, F.-C. Baurens, A. Grapin, D. Dambier, J.-L. Noyer C. Lanaud, G. Kahl and P. J. L. Lagoda (1997) Molecular breeding in the genus *Musa*: a strong case for STMS marker technology. Euphytica 96(1): 49-63.

Karamura, D., M. Kitavi, M. Nyine, D. Ochola, S. Muhangi, D. Talengera and E. B. Karamura (2016) Genotyping the local banana landrace groups of East Africa. Acta Hortic 1114. doi: 10.17660/ActaHortic.2016.1114.9

Kitavi, M., T. Downing, J. Lorenzen, D. Karamura, M. Onyango, M. Nyine, M. Ferguson and C. Spillane (2016) The triploid East African Highland Banana (EAHB) genepool is genetically uniform arising from a single ancestral clone that underwent population expansion by vegetative propagation. Theor Appl Genet 129(3): 547-561.

Korte, A. and A. Farlow (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9: 29.

Kress, W.J. and C.D. Specht (2005) Between Cancer and Capricorn: phylogeny, evolution and ecology of the primarily tropical *Zingiberales*. Biol Skr 55: 459-478.

Kress, W.J. and C.D. Specht (2006) The evolutionary and biogeographic origin and diversification on the tropical monocot order *Zingiberales.* Aliso 22: 621-632.

Kumar, P.L., R. Hanna O.J. Alabi, M.M. Soko, T.T. Oben, G.H.P. Vangu, R.A. Naidu (2011) Banana bunchy top virus in sub-Saharan Africa: Investigations on virus distribution and diversity. Virus Res 159: 171-182.

Kumar, S., D. Chagné, M.C.A.M. Bink, R.K. Volz, C. Whitworth and C. Carlisle (2012) Genomic selection for fruit quality traits in apple (Malus×domestica Borkh.). PLoS One 7(5): e36674. doi: 10.1371/journal.pone.0036674

Langmead, B. and S.L. Salzberg (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4): 357-359. doi: 10.1038/nmeth.1923

Legarra, A., Robert-Granie C., Manfredi E., and J.M. Elsen (2008) Performance of genomic selection in mice. Genetics 180: 611-618.

Lheureux, F., F. Carreel, C. Jenny, B. Lockhart and M. Iskra-Caruana (2003) Identification of genetic markers linked to banana streak disease expression in inter-specific *Musa* hybrids. Theor Appl Genet 106: 594-598.

Li, H. and R. Durbin (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14): 1754-1760. doi: 10.1093/bioinformatics/btp324

Li, C., J. Shao, Y. Wang, W. Li, D. Guo, B. Yan, Y. Xia, and M. Peng (2013) Analysis of banana transcriptome and global gene expression profiles in banana roots in response to infection by race 1 and tropical race 4 of *Fusarium oxysporum f. sp. cubense*. BMC Genomics 14(1): 851.

Lipka, A.E., Tian F., Wang Q., Peiffer J., Li M., Bradbury P.J., Gore M., Buckler E.S., and Z. Zhang (2014). GAPIT: genome association and prediction integrated tool. Bioinformatics 28(18): 2397-2399. doi:10.1093/bioinformatics/bts444

Lorenz, A.J., S. Chao, G. Franco, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, P. Kevin, K.P. Smith, M.E. Sorrells and J-L. Jannink (2011) Chap. 2: Genomic Selection in Plant Breeding: Knowledge and Prospects. Adv Agronomy 110: 77-123.

Lorenzen, J., A. Tenkouano, R. Bandyopadhyay, B. Vroh, D. Coyne, and L. Tripathi (2010) Overview of banana and plantain (*Musa* spp.) improvement in Africa: past and future. proceedings of international conference on banana and plantain in Africa. Acta Hortic 879.

Lutz, K.A., W. Wang, A. Zdepski and T. Michael (2011) Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing. BMC Biotechnol 11: 54.

Ly, D., M. Hamblin, I. Rabbi, G. Melaku, M. Bakare, H.G. Gauch Jr., R. Okechukwu, A.G.O. Dixon, P. Kulakow and J-L. Jannink (2013) Relatedness and genotype × environment interaction affect prediction accuracies in genomic selection: A study in cassava. Crop Sci 53: 1-14.

Mabonga, L. and M. Pillay (2017) SCAR marker for the A genome of bananas (*Musa* spp. L.) supports lack of differentiation between the A and B genomes. J Agric Sci 9(6): 64-73.

MacEachern, S.N. and L.M. Berliner (1994) Subsampling the Gibbs Sampler. Am Stat 48: 188-190.

Macharia, I., A.M. Kagundu, E.W. Kimani and W. Otieno (2010) Combating phytosanitary constraints to banana (*Musa* spp.) production: the Kenyan example. Proceedings of International Conference on Banana and Plantains in Africa, 561-565.

Mammadov, J., R. Aggarwal, R. Buyyarapu and S. Kumpatla (2012) SNP markers and their impact on plant breeding. Int J Plant Genomics 2012: 728398. doi: 10.1155/2012/728398

Manrique, K., and M. Hermann (2000). Effect of GxE interaction on root yield and beta-carotene content of selected sweet potato (*Ipomoea batatas* (L) Lam.) varieties and breeding clones. CIP Program Report 1999-2000: 281-287.

Martin, G., F.C. Baurens, G. Droc, M. Rouard, A. Cenci, A. Kilian, A. Hastie, J. Dolezel, J. M. Aury, A. Alberti, F. Carreel, and A. D’Hont, (2016) Improvement of the banana “*Musa acuminata*” reference sequence using NGS data and semi-automated bioinformatics methods. BMC Genomics 17(1): 1-12.

Martin, K.P., S.K. Pachathundikandi, C-L. Zhang, A. Slater and J. Madassery (2006) RAPD Analysis of a variant of banana (*Musa* Sp.) Cv. Grande Naine and its propagation via shoot tip culture. In Vitro Cell Dev Biol Plant 42(2): 188-192.

Mbanjo E.G N., F. Tchoumbougnang, A.S. Mouelle, J.E. Oben, M. Nyine, C. Dochez, M.E. Ferguson and J. Lorenzen (2012a). Development of expressed sequence tags-simple sequence repeats (EST-SSRs) for *Musa* and their applicability in authentication of a *Musa* breeding population. Afr J Biotechnol 11(71): 13546-13559.

Mbanjo, E.G.N., F. Tchoumbougnang, A.S. Mouelle, J.E. Oben, M. Nyine, C. Dochez, M.E. Ferguson and J. Lorenzen (2012b). Molecular marker-based genetic linkage map of a diploid banana population (*Musa acuminata* Colla). Euphytica 188(3): 369-386. doi: 10.1007/s10681-012-0693-1

Metzker, M.L (2010). Sequencing technologies – the next generation. Nat Rev Genet 11: 31-46.

Meuwissen, T.H., B.J. Hayes and M.E. Goddard (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Meuwissen, T.H., T.R. Solberg, R. Shepherd and J.A. Woolliams (2009) A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genet Sel Evol 41: 2.

Miller, R.N.G., D.J. Bertioli, F.C. Baurens, C.M.R. Santos, P.C. Alves, N.F. Martins, R.C. Togawa, M.T. Souza Jr. and G.J. Pappas Jr. (2008) Analysis of non-TIR NBS-LRR resistance gene analogs in *Musa acuminata* Colla: Isolation, RFLP marker development, and physical mapping. BMC Plant Biol 8(1): 15.

Momen, M., A.A. Mehrgardi, A. Sheikhy, A. Esmailizaheh, M.A. Fozi, A. Kranis, B.D. Valente, G.J.M. Rosa and D. Gianola (2017) A predictive assessment of genetic correlations between traits in chickens using markers. Genet Sel Evol 49: 16. doi: 10.1186/s12711-017-0290-9

Myles, S. (2013). Improving fruit and wine: what does genomics have to offer. Trends Genet 29(4): 190-196.

Nair, A.S., C.H. Teo, T. Schwarzacher and P.H Harrison (2005) Genome classification of banana cultivars from South India using IRAP markers. Euphytica 144: 285-290.

Nakaya, A. and S.N. Isobe (2012). Will genomic selection be a practical method for plant breeding? Review: Part of a highlight on breeding strategies for forage and grass improvement. Ann Bot 110: 1303-1316.

Ndabamenye, T., P.J.A. van Asten, N. Vanhoudt, G. Blomme, R. Swennen, J.G. Annandale and R.O. Barnard (2012) Ecological characteristics influence farmer selection of on-farm plant density and bunch mass of low input East African highland banana (*Musa* spp.) cropping systems. Field Crops Res 135: 126-136. doi: 10.1016/j.fcr.2012.06.018

Newbury, H.J., E.C. Howell, J.H. Crouch and B.V. Ford-Lloyd (2000) Natural and culture-induced genetic variation in plantains (*Musa* spp. AAB group). Aust J Bot 48(4): 493-500.

Noumbissié, G.B., M. Chabannes, F. Bakry, S. Ricci, C. Cardi, J-C. Njembele, D. Yohoume, K. Tomekpe, M-L. Iskra-Caruana, A. D'Hont and F-C. Baurens (2016) Chromosome segregation in an allotetraploid banana hybrid (AAAB) suggests a translocation between the A and B genomes and results in eBSV-free offsprings. Mol Breed 36(4): 38.

Noyer J.L., S. Causse, K. Tomekpe, A. Bouet, and F.C. Baurens (2005) A new image of plantain diversity assessed by SSR, AFLP and MSAP markers. Genetica 124: 61-69.

Nwakanma D.C., M. Pillay, B.E. Okoli and A. Tenkouano (2003). PCR-RFLP of the ribosomal DNA internal transcribed spacers (ITS) provides markers for the A and B genomes in *Musa* L. Theor Appl Genet 108: 154-159.

Nyine, M., and M. Pillay (2011) The effect of banana breeding on the diversity of East African highland banana (*Musa*, AAA). Acta Hortic 897: 225-229.

Onyango, M., D. Haymer, S. Keeley and R. Manshardt (2010) Analysis of Genetic Diversity and Relationships in East African 'Apple Banana' (AAB genome) and 'Muraru' (AA genome) Dessert Bananas Using Microsatellite Markers. In: Dubois, T. et al. (eds), Proc. IC on Banana & Plantain in Africa, Acta Hortic 879: 623-636.

Opara, U.L., D. Jacobson and N.A. Al-Saady (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. J Zhejiang Univ Sci B 11(5): 332-341.

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

Ortiz, R (2016) *Musa* interspecific hybridization and polyploidy for breeding banana and plantain. In: Mason, A.S. (ed), polyploidy and hybridization for crop improvement. CRC PRESS, US. pp 490.

Ortiz. R. and D. Vuylsteke (1995b) Recommended experimental designs for selection of plantain hybrids. InfoMusa 4(1): 11-12.

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

Ortiz, R. and R. Swennen (2014) Research review: From crossbreeding to biotechnology-facilitated improvement of banana and plantain. Biotechnol Adv 32: 158-169. doi: 10.1016/j.biotechadv.2013.09.010

Park, T. and G. Casella (2008) The Bayesian LASSO. J Am Stat Assoc 103: 482. doi: 10.1198/016214508000000337

Perea, C., J.F. De La Hoz, D.F. Cruz, J. D. Lobaton, P. Izquierdo, J. C. Quintero, B. Raatz and J. Duitama (2016) Bioinformatic analysis of genotype by sequencing (GBS) data with NGSEP. BMC Genomics 17(5): 498. doi: 10.1186/s12864-016-2827-7

Pérez, P. and G. de los Campos (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198: 483-495. doi: 10.1534/genetics.114.164442

Perrier, X., E. De Langhe, M. Donohuec, C. Lentfer, L. Vrydaghs, F. Bakry, F. Carreel, I. Hippolyte, J-P. Horry, C. Jenny, V. Lebot, A-M. Risterucci, K. Tomekpe, H. Doutrelepont, T. Ball, J. Manwaring, de P. Maret, and T. Denham (2011) Multidisciplinary perspectives on banana (*Musa* spp.) domestication. Proc Natl Acad Sci 108(28): 11311-11318.

Persley, G.J., and P. George (1996) Banana improvement: Research challenges and opportunities. Environmentally sustainable development agricultural research and extension group series. Banana improvement project report; No.1.

Pillay, M., D.C. Nwakanma and A. Tenkouano (2000) Identification of RAPD markers linked to A and B genome sequences in *Musa* L. Genome 43: 763-767.

Pillay, M., E. Ogundiwin, D.C. Nwakanma, G. Ude and A. Tenkouano (2001) Analysis of genetic diversity and relationships in East African banana germplasm. Theor Appl Genet 102: 965-970.

Pillay, M. Ude G. and C. Kole (eds), (2012), Genetics, genomics and breeding of bananas. Chap. 4: Molecular Marker Techniques in *Musa* Genomic Research. Boca Raton, Florida: CRC Press.

Pillay, M., and A. Tenkouano (2010). Mapping and Tagging of Simply Inherited Traits in *Musa (Rodomiro Ortiz). Chapt.6:*109-115.

Pillay, M., A. Ashokkumar, A. James, S. Jabekumar, P. Kirubakaran, R. Miller, R. Ortiz and E. Sivalingam (2012). Genetics, Genomics and Breeding of Bananas. Molecular Marker Techniques in *Musa* Genomic Research. Science Publishers, Chapt.4: 70-90.

Pillay, M., E. Ogundiwin, D.C. Nwakanma, G. Ude, and A. Tenkouano (2001) Analysis of genetic diversity and relationships in East African banana germplasm. Theor Appl Genet 102(5): 965-970.

Pillay, M., D.C. Nwakanma, and A. Tenkouano (2000). Identification of RAPD markers linked to A and B genome sequences in *Musa* L. Genome 43(5): 763-767.

Pillay, M., G. Ude and C. Kole (2012) Genetics, genomics, and breeding of bananas. Science Publishers, Chapt.7: 116-123.

Ploetz, R. C. 2000. Panama disease: A classic and destructive disease of banana. Online. Plant Health Progress doi:10.1094/PHP-2000-1204-01-HM

Ploetz, R., S. Freeman, J. Konkol, A. Al-Abed, Z. Naser, K. Shalan, R. Barakat and Y. Israeli (2015) Tropical race 4 of Panama disease in the Middle East. Phytoparastica 43(3): 283-293.

Poland, J.A., P.J. Brown, M.E. Sorrells and J-L. Jannink (2012a) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7(2): e32253. doi: 10.1371/journal.pone.0032253

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sanchez-Villeda, M. Sorrells and J-L. Jannink (2012b) Genomic selection in Wheat Breeding using Genotypying-by Sequencing. Plant Genome 5: 103-113.

Pollard, D.A. (2012) Design and construction of recombinant inbred lines. Methods Mol Biol 871: 31-39. doi: 10.1007/978-1-61779-785-9_3

Polyanovsky, V.O., M.A. Roytberg and V.G. Tumanyan (2011) Comparative analysis of quality of a global algorithm and local algorithm for alignment of two sequences. Algorithms Mol Biol 6: 25. doi: 10.1186/1748-7188-6-25

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

Raboin, L-M., F. Carreel, J-L. Noyer, F-C. Baurens, J. P. Horry, F. Bakry, H.T. Du Montcel, J. Ganry, C. Lanaud and P.J.L. Lagoda (2005) Diploid ancestors of triploid export banana cultivars: molecular identification of 2n restitution gamete donors and n gamete donors. Mol Breed 16: 333-341.

Ravi, I., S. Uma, M.M. Vaganan and M.M. Mustaffa (2013) Phenotyping bananas for drought resistance. Front Physiol 4: 9. doi: 10.3389/fphys.2013.00009

Ray, P.K. (2002). Breeding tropical and subtropical fruits, banana and plantain, pp 44-83. In: Springer Science & Business Media, Technology & Engineering.

Resende Jr., M.F.R., P. Muñoz, M.D.V. Resende, D.J. Garrick, R.L. Fernando, J.M. Davis, E.J. Jokela, T.A. Martin, G.F. Peter and M. Kirst (2012) Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics 190: 1503-1510.

Risterucci, A-M., I. Hippolyte, X. Perrier, L. Xia, V. Caig, M. Evers, E. Huttner, A. Kilian and J-C. Glaszmann (2009) Development and assessment of Diversity Arrays Technology for high-throughput DNA analyses in Musa. Theor Appl Genet 119: 1093-1103.

Robinson, G.K. (1991) That BLUP is a good thing: The estimation of random effects. Stat Sci 6(1): 15-51.

Rowe, P.R. (1990) Breeding bananas and plantains for resistance to fusarial wilt: the track record. In: Ploetz RC (ed), *Fusarium* wilt of bananas. APS, St. Paul, MN. pp 115-119.

Rowe, P. and F.E. Rosales (1993). Diploid breeding at FHIA and the development of Goldfinger. InfoMusa 2: 9-11.

Rowe, P. and F.E. Rosales (1996). Bananas and Plantains. In: Janick J. and J.N. Moore (eds), Fruit breeding vol.1: Tree and tropical fruit. ISBN 0-471-32014-X.

Ruas, M., V. Guignon, G. Sempere, J. Sardos, Y. Hueber, H. Duvergey, A. Andrieu, R. Chase, et al. (2017) MGIS: managing banana (Musa spp.) genetic resources information and high-throughput genotyping data. Database 1-12. doi: 10.1093/database/bax046

Sadik, K., M. Nyine, and M. Pillay (2010). A screening method for banana weevil (*Cosmopolites sordidus* Germar) resistance using reference genotypes. Afr J Biotechnol 9(30): 4725-4730.

Sansaloni, C., C. Petroli, D. Jaccoud, J. Carling, F. Detering, D. Grattapaglia and A. Kilian (2011) Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. BMC Proceedings 5(7): 54.

Sardos, J., M. Rouard, Y. Hueber, A. Cenci, K.E. Hyma, I. van de Houwe, E. Hřibová, B. Courtois and N. Roux (2016) A genome-wide association study on the seedless phenotype in banana (*Musa* spp.) reveals the potential of a selected panel to detect candidate genes in a vegetatively propagated crop. PLoS One 11(5). doi: 10.1371/journal.pone.0154448

Sharrock, S.L., J-P. Horry and E.A.Frison (2001) The state of use of *Musa* diversity. In: Cooper H.D., C. Spillane and T. Hodgkin (eds), Broadening the genetic base of crop production. Technology & Engineering pp. 223-244.

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

Silva, P.R.O., O.N. de Jesus, C.A.D. Bragança, F. Haddad, E.P. Amorim and C.F. Ferreira (2016) Development of a thematic collection of *Musa* spp accessions using SCAR markers for preventive breeding against *Fusarium oxysporum* f. sp. *cubense* tropical race 4. Genet Mol Res 15 (1): gmr.15017765. doi: 10.4238/gmr.15017765

Simmonds, N.W. (1954) Varietal identification in the Cavendish group of bananas. J Hortic Sci 29(2): 81-88. doi: 10.1080/00221589.1954.11513800

Simmonds, N.W. (1962) Evolution of the bananas. London: Longmans, Green & Co.

Simmonds, NW. (1966) Bananas. 2nd edn. Longmans, London, UK.

Simmonds, NW. (1986) Bananas, *Musa* cvs. In: Simmonds NW (ed), Breeding for durable resistance in perennial crops. FAO Technical Papers 70. Food and Agriculture Organization, Rome. pp. 17-24.

Ssebuliba, R.N., P. Rubaihayo, A. Tenkouano, D. Makumbi, D. Talengera and M. Magambo (2005) Genetic diversity among East African highland bananas for female fertility. Afr Crop Sci J 13(1): 13-26.

Ssebuliba, R., D. Talengera, D. Makumbi, P. Namanya, A. Tenkouano, W. Tushemereirwe and M. Pillay (2006) Reproductive efficiency and breeding potential of East African highland (*Musa* AAA-EA) bananas. Field Crops Res 95: 250-255.

Stover, R.H. (1962) Studies on Fusarium wilt of bananas: IX. Competitive saprophytic ability of *F. oxysporum f. sp. cubense.* Can J Bot 40(11): 1473-1481.

Sonah, H., M. Bastien, E. Iquira, A. Tardivel, G. Légaré, B. Boyle, E. Normandeau, J. Laroche, S. Larose, M. Jean, and F. Belzile (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS ONE 8(1): e54603. doi: 10.1371/journal.pone.0054603

Swennen, R. and D. Vuylsteke (2001) Banana (*Musa* L.). In: Raemaekers R.H. (ed), Crop production in tropical Africa. pp. 530-552.

Swennen, R., G. Blomme, P. Van Asten, P. Lepoint, E. Karamura, E. Njukwe, W. Tinzaara, A. Viljoen, P. Karangwa, D. Coyne and J. Lorenzen (2013). Mitigating the impact of biotic constraints to build resilient banana systems in Central and Eastern Africa. In: Van Lauwe, B., P. van Asten and G. Blomme (eds), Agro-ecological intensification of agricultural systems in the African highlands Earthscan book, pp 85-104.

Taghouti, M., F. Gaboun, N. Nsarellah, R. Rhrib, M. El-Haila, M. Kamar, F. Abbad-Andaloussi and S. M. Udupa (2010) Genotype x environment interaction for quality traits in *durum* wheat cultivars adapted to different environments. Afr J Biotechnol 9(21): 3054-3062.

Taulya, G. (2015) Ky'osimba onanya: Understanding productivity of East African highland banana. PhD thesis, Wageningen University.

Tenkouano A., J.H. Crouch, H.K. Crouch, D. Vuylsteke and R. Ortiz (1999). Comparison of DNA marker and pedigree-based methods of genetic analysis of plantain and banana (*Musa* spp.) clones. I. estimation of genetic relationships. Theor Appl Genet 98: 62-68.

Tenkouano A., R. Ortiz and D. Vuylsteke (2012) Estimating genetic effects in maternal and paternal half-sibs from tetraploid-diploid crosses in *Musa* spp. Euphytica 185: 295-301.

Tibshirani, R. (1996) Regression shrinkage and selection via the LASSO. J R Stat Soc Series B Stat Methodol 58: 267-288.

Tripathi, L., H. Mwaka, J.N. Tripathi and W.K. Tushemereirwe (2010) Expression of sweet pepper Hrap gene in banana enhances resistance to *Xanthomonas campestris pv. musacearum*. Mol Plant Pathol 11(6): 721-731.

Tripathi, L., A. Babirye, H. Roderick, J. N.Tripathi, C. Changa, P. E. Urwin, W. K. Tushemereirwe, D. Coyne and H.J. Atkinson (2015) Field resistance of transgenic plantain to nematodes has potential for future African food security. Sci Rep 5: 8127. doi: 10.1038/srep08127

Tushemereirwe W., M. Batte, M. Nyine, R. Tumuhimbise, A. Barekye, T. Ssali, D. Talengera, J. Kubiriba, J. Lorenzen, R. Swennen and B. Uwimana (2015) Performance of NARITA banana hybrids in the preliminary yield trial for three cycles in Uganda**.** IITA, NARO, Uganda. 35p. http://www.iita.org/c/document_library/get_file?uuid=1500c39e-7c05-4219-b649-85cdb1bd73f6&groupId=25357 (accessed on July 3, 2017).

Ude G., M. Pillay, D. Nwakanma and A. Tenkouano (2002) 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., M. Pillay, E. Ogundiwin and A. Tenkouano (2003) Genetic diversity in an African plantain core collection using AFLP and RAPD markers. Theor Appl Genet 107(2): 248-255.

Uitdewilligen, J.G.A.M.L., A-M A. Wolters, B.B. D'hoop, T.J.A. Borm, R.G.F. Visser and H.J. van Eck (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PloS One 10(10): e0141940. doi: 10.1371/journal.pone.0141940

Umber M., J-P. Pichaut, B. Farinas, N. Laboureau, B. Janzac, K. Plaisir-Pineau, G. Pressat, F-C. Baurens, M. Chabannes, P-O. Duroy, C. Guiougou, J-M. Delos, C. Jenny, M-L.

Iskra-Caruana, F. Salmon and P-Y. Teycheney (2016) Marker-assisted breeding of *Musa balbisiana* genitors devoid of infectious endogenous Banana streak virus sequences. Mol Breed 36(6): 74.

Valdez-Ojeda, R., A. James-Kay, J.R. Ku-Cauich and R.M. Escobedo-GraciaMedrano (2014) Genetic relationships among a collection of *Musa* germplasm by fluorescent-labeled SRAP. Tree Genet Genomes 10: 465-476. doi: 10.1007/s11295-013-0694-9

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

Vanhove, A.C., W. Vermaelen, B. Panis, R. Swennen and S. Carpentier (2012) Screening the banana biodiversity for drought tolerance: can an in vitro growth model and proteomics be used as a tool to discover tolerant varieties and understand homeostasis. Front Plant Sci 3. doi: 10.3389/fpls.2012.00176

van Orsouw, N.J., R.C.J. Hogers, A. Janssen, F. Yalcin, S. Snoeijers, E. Verstege, H. Schneiders, H. van der Poel, J. van Oeveren, H. Verstegen, M.J.T. van Eijk (2007) Complexity reduction of polymorphic sequences (CRoPS): A novel approach for large-scale polymorphism discovery in complex genomes. PLoS One 2(11): e1172. doi: 10.1371/journal.pone.0001172

VanRaden, P.M. (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91: 4414-4423. doi:10.3168/jds.2007-0980.

VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor and F.S. Schenkel (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92(1): 16-24. doi: 10.3168/jds.2008-1514

Varshney, R.K., S.M. Mohan, P.M. Gaur, N.V. Gangarao, M.K. Pandey, A. Bohra, S.L. Sawargaonkar, A. Chitikineni, P.K. Kimurto, P. Janila, K.B. Saxena, A. Fikre, M. Sharma, A. Rathore, A. Pratap, S. Tripathi, S. Datta, S.K. Chaturvedi, N. Mallikarjuna, G. Anuradha, A. Babbar, A.K. Choudhary, M.B. Mhase, C. Bharadwaj, D.M. Mannur, P.N. Harer, B. Guo, X. Liang, N. Nadarajan and C.L. Gowda (2013) Review: Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics. Biotechnol Adv 31:1120-1134.

Vazquez A.I., D.M. Bates, G.J.M. Rosa, D. Gianola and K.A. Weigel (2010) Technical Note: An R package for fitting generalized linear mixed models in animal breeding. J Anim Sci 88: 497-504.

Venkatachalam, L., R.V. Sreedhar and N. Bhagyalakshmi (2008) The use of genetic markers for detecting DNA polymorphism, genotype identification and phylogenetic relationships among banana cultivars. Mol Phylogenet Evol 47: 974-985.

Vuylsteke, D. (2001) Strategies for utilization of genetic variation in plantain improvement, a tribute to Dirk R. Vuylsteke (1958-2000). PhD thesis, K.U. Leuven pp 213.

Vuylsteke, D. and R. Swennen (1992) Biotechnological approaches to plantain and banana improvement at IITA, pp 143-150. In: Thottapilly, G., L. Monti, D.R. Mohan Raj and A.W. Moore (eds), Biotechnology: enhancing research on tropical crops in Africa. Ibadan, Nigeria: CAT/IITA.

Wang W., Y. Hu, D. Sun, C. Staehelin, D. Xin and J. Xie (2012b) Identification and evaluation of two diagnostic markers linked to Fusarium wilt resistance (race 4) in banana (*Musa* spp.). Mol Biol Rep 39: 451-459.

Wang, X-L., T-Y. Chiang, N. Roux, G. Hao, X-J. Ge (2007) Genetic diversity of wild banana (*Musa balbisiana* Colla) in China as revealed by AFLP markers. Genet Resour Crop Evol 54(5): 1125-1132.

Wang, Z., M. Gerstein, and M. Snyder (2009) RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63.

Wang, Z., J. Zhang, C. Jia, J. Liu, Y. Li, X. Yin, B. Xu and Z. Jin (2012a) *De novo* characterization of the banana root transcriptome and analysis of gene expression *under Fusarium oxysporum* f. sp. *cubense* tropical race 4 infection. BMC Genomics 13: 650.

Wei, J.Y., D.B. Liu, S.X. Wei, Z.S. Xie, G.Y. Xu, and Y.Y. Chen (2011) Analysis of genetic diversity in banana cultivars (*Musa* spp.) using sequence-related amplified polymorphism markers. Acta Hortic 897: 263-265.

Wong, C., R. Kiew, G. Argent, O. Set, S.K. Lee and Y.Y. Gan (2002) Assessment of the Validity of the Sections in *Musa* (*Musaceae*) using AFLP. Ann Bot 90(2): 231-238.

Workman, D. (2006) Top ten banana producing countries. International Trade, Suite 101.

Würschum, T., J.C. Reif, T. Kraft, G. Janssen and Y. Zhao (2013) Genomic selection in sugar beet breeding populations. BMC Genet 14: 85. doi: 10.1186/1471-2156-14-85

Xu, S. (2007) An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics 63: 513-521.

Xu, Y. (2010). Molecular plant breeding. CABI, Chapt.1:1-20.

Zhang, Z., J. Liu, X. Ding, P. Bijma, D-J. de Koning and Q. Zhang (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker derived relationship matrix. PLoS One 5(9): e12648. doi: 10.1371/journal.pone.0012648

Zhong, S., J.C.M. Dekkers, R.L. Fernando and J-L. Jannink (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. Genetics 182: 355-364. doi: 10.1534/genetics.108.098277

Zohary D. 2004. Unconscious selection and the evolution of domesticated plants. Econ Bot 58: 5-10.

8 Abbreviation

BL	Bayesian LASSO
BRR	Bayesian ridge regression
cv.	Cultivar
EAHB	East African Highland banana
GBS	Genotyping by sequencing
GEBV	Genomic estimated breeding value
GS	Genomic selection
LASSO	Least absolute shrinkage and selection operator
RKHS_M	Reproducing kernel Helbert space with marker data
RKHS_P	Reproducing kernel Helbert space with pedigree data
RKHS_PM	Reproducing kernel Helbert space with pedigree and marker data
SNP	Single nucleotide polymorphism

9 Presentations

9.1 Conference abstract

Nyine, M. B. Uwimana, N. Blavet, E. Hřibová, H. Vanrespaille, M. Batte, V. Akech, A. Brown, J. Lorenzen, R. Swennen and J. Doležel (2018) The Benefits, Challenges and Prospects of Genomic Prediction in Polyploid Banana. [Abstract] presented at Plant and Animal Genome Conference XXVI. San Diego, CA (USA) 13-17 Jan. 2018.
https://pag.confex.com/pag/xxvi/meetingapp.cgi/Paper/30796

Abstract

The interploidy breeding approaches practiced in banana limit the application of classical marker assisted selection strategies. Yet, there is an ultimate need to improve the efficiency of conventional crossbreeding and reduce the selection cycle to respond more rapidly to abiotic and biotic stresses. The development of sequencing and genotyping technologies such as genotyping by sequencing (GBS) are leveraging the breeders to explore genomic prediction-based approaches. In this work, the performance of six genomic prediction models was evaluated in banana under different cross validation strategies using data from a genomic selection training population comprising 307 genotypes. The population consisting of diploid, triploid and tetraploid genotypes was phenotyped under two different field management conditions and genotyped using GBS. Sequence data were processed through a bioinformatics workflow and single nucleotide polymorphisms (SNPs) were called using the genomic analysis tool kit (GATK). A custom R script was developed to process the SNP data prior to input into the models. The genotypic data were both bi-allelic SNP and allele dosage SNP markers. The total number of SNP markers varied from 5574 to 10807 depending on cross-validation strategy. Phenotypic data collected for four years on 15 traits under plant stature, suckering behavior, black leaf streak resistance, fruit bunch and fruit filling were used in cross validation. We compared the effect of accounting for allele dosage in SNP markers on the predictive ability of genomic prediction models. The results permit the evaluation of benefits, challenges and prospects of applying genomic prediction in banana, an important polyploid clonally propagated crop.

9.2 Conference abstract

Nyine, M. B. Uwimana, R. Swennen, M. Batte, A. Brown, E. Hřibová and J. Doležel (2016) Genomic breeding approaches for East African bananas. [Abstract] presented at Plant and Animal Genome Conference XXIV. San Diego, CA (USA) 9-13 Jan. 2016. http://hdl.handle.net/10568/78754

Abstract

The polyploidy nature of banana is a limiting factor in the implementation of strategies such as marker assisted selection (MAS) or genome wide association mapping (GWAS). The triploid nature of cultivated varieties complicates conventional breeding strategies and improved varieties can take up to 20 years before they can be released to the public, which necessitates the use of efficient molecular tools to more rapidly respond to abiotic and biotic stresses and to address the needs of growers and consumers. In addition, the high cost of phenotyping perennial large-stature plants such as banana, and the rapidly decreasing cost of genotyping, makes the use of genomic prediction models using single nucleotide polymorphism (SNP) markers extremely attractive to banana breeders. A Genomic Selection (GS) training population consisting of 307 banana genotypes was developed for initial analysis with ploidy levels of the plant material ranging from diploids to tetraploids. Plants were genotyped using the genotyping by sequencing (GBS) approach (Elshire et al. 2011) with PstI as the sole restriction enzyme. Sequence data was processed through a bioinformatics workflow and single nucleotide polymorphisms (SNPs) were called using the genomic analysis tool kit (GATK). Data was filtered for quality and for loci with >50% missing data. Phenotypic data for 25 traits are being collected from two locations since 2012. Yield-related traits (fruit pulp diameter, bunch weight, number of suckers, etc.) are collected at flowering and harvest Analysis of GBS data resulted in 11201 SNP loci. The results of multiple prediction models are discussed and compared.

9.3 Poster presentation

Moses Nyine, Brigitte Uwimana, Rony Swennen, Michael Batte, Allan Brown, Pavla Christelová, Eva Hřibová, Jim Lorenzen, Jaroslav Doležel (2017) Trait variation in a banana training population for genomic selection. Annual Banana Project Meeting, April, Kampala, Uganda.

Poster abstract

Conventional crossbreeding is the main approach used in banana improvement. However, the method requires up to two decades of crossing and field evaluation to develop a new hybrid. This is because selection is carried out at different levels. At every level, plants are evaluated after three crop cycles, each taking about a year. Yield traits can only be scored at harvest while organoleptic traits are recorded after harvesting, making the selection process slow, expensive and labour intensive. New breeding tools with increased crossbreeding efficiency are being investigated to breed for resistant, high yielding hybrids of East African Highland banana (EAHB). These include genomic selection (GS), which will benefit breeding through increased genetic gain per unit time. Understanding trait variation and the correlation among economically important traits is an essential first step in the development and selection of suitable genomic prediction models for banana. In this study, we tested the hypothesis that trait variations in bananas are not affected by cross combination, cycle, field management and their interaction with genotype. A training population created using EAHB breeding material and its progeny was phenotyped in two contrasting conditions. A high level of correlation among vegetative and yield related traits was observed. This could mean that the predictive ability of traits that are difficult to phenotype will be similar to less difficult traits they are highly correlated with. Therefore, genomic prediction models could be developed for traits that are easily measured. Black Sigatoka related traits were not affected by crop cycle, meaning that these could be measured in the first cycle only, to reduce on phenotyping burden. Growth traits such as plant height and girth were the least affected by field input management. Conversely, yield-related traits such as bunch weight, number of hands and number of fingers were significantly affected by both crop cycle and field input management.

9.4 Poster presentation

Nyine, M., B. Uwimana, R. Swennen, M. Batte, A. Brown, P. Christelová, E. Hřibová, J. Lorenzen and J. Doležel (2016) Trait Variation in a Banana Training Population for Genomic Selection. P4D and R4D meeting, November at IITA, Ibadan, Nigeria.

9.5 Poster presentation

Nyine, M., B. Uwimana, T.R. Ssali, J. Kubiriba, E. Amorim, Y. Othman, R. Swennen, M. Batte, E. Hřibová and J. Doležel (2015) Towards marker assisted breeding in banana. R4D meeting, November at IITA, Ibadan, Nigeria.

9.6 Poster presentation

Nyine, M., B. Uwimana, R. Swennen, M. Batte, E. Hřibová, J. Lorenzen and J. Doležel (2015) Genomic selection to accelerate banana breeding. Roots, Tubers and Bananas (RTB) project evaluation, February at IITA, Sendusu, Uganda.

10 Supplementary information

Appendix I

Genomic prediction in a multiploid crop: Genotype by environment interaction and allele dosage effects on predictive ability in banana. The Plant Genome [Accepted on 19 December 2017] doi: 10.3835/plantgenome2017.10.0090

Published online March 2, 2018

Genomic Prediction in a Multiploid Crop: Genotype by Environment Interaction and Allele Dosage Effects on Predictive Ability in Banana

Moses Nyine, Brigitte Uwimana, Nicolas Blavet, Eva Hřibová, Helena Vanrespaille, Michael Batte, Violet Akech, Allan Brown, Jim Lorenzen, Rony Swennen, and Jaroslav Doležel*

Abstract

Improving the efficiency of selection in conventional crossbreeding is a major priority in banana (*Musa* spp.) breeding. Routine application of classical marker assisted selection (MAS) is lagging in banana due to limitations in MAS tools. Genomic selection (GS) based on genomic prediction models can address some limitations of classical MAS, but the use of GS in banana has not been reported to date. The aim of this study was to evaluate the predictive ability of six genomic prediction models for 15 traits in a multi-ploidy training population. The population consisted of 307 banana genotypes phenotyped under low and high input field management conditions for two crop cycles. The single nucleotide polymorphism (SNP) markers used to fit the models were obtained from genotyping by sequencing (GBS) data. Models that account for additive genetic effects provided better predictions with 12 out of 15 traits. The performance of BayesB model was superior to other models particularly on fruit filling and fruit bunch traits. Models that included averaged environment data were more robust in trait prediction even with a reduced number of markers. Accounting for allele dosage in SNP markers (AD-SNP) reduced predictive ability relative to traditional bi-allelic SNP (BA-SNP), but the prediction trend remained the same across traits. The high predictive values (0.47– 0.75) of fruit filling and fruit bunch traits show the potential of genomic prediction to increase selection efficiency in banana breeding.

Core Ideas

- First empirical evidence of genomic prediction in a multi-ploidy banana population is presented.
- The effect of allele dosage single nucleotide polymorphism on prediction accuracy depends on the trait.
- Use of averaged environmental data improves prediction accuracy.
- BayesB model can be used across all traits during genomic prediction in banana breeding.
- The high predictive values show the potential of genomic prediction in banana breeding.

M. Nyine, Faculty of Science, Palacký University, CZ-77146, Olomouc, Czech Republic; M. Nyine, B. Uwimana, M. Batte, V. Akech, International Institute of Tropical Agriculture, 7878, Kampala, Uganda; M. Nyine, N. Blavet, E. Hřibová, J. Doležel, Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, CZ-78371, Olomouc, Czech Republic; H. Vanrespaille, R. Swennen, Lab. of Tropical Crop Improvement, Division of Crop Biotechnics, Katholieke Universiteit 2455, 3001 Leuven, Belgium; J. Lorenzen, International Institute of Tropical Agriculture, 7878, Kampala, Uganda (current address, Bill and Melinda Gates Foundation, 23350, Seattle); R. Swennen, A. Brown, International Institute of Tropical Agriculture, 10, Arusha, Tanzania. Received 17 Oct. 2017. Accepted 19 Dec. 2017. *Corresponding author (dolezel@ueb.cas.cz).

Abbreviations: AD-SNP, allele dosage single nucleotide polymorphism; BA-SNP, bi-allelic single nucleotide polymorphism; BL, Bayesian least absolute shrinkage and selection operator; BLUP, best linear unbiased prediction; BRR, Bayesian ridge regression; cv., cultivar; EAHB, East African highland banana; GATK, genome analysis toolkit; GBS, genotyping by sequencing; GEBV, genomic estimated breeding value; GS, genomic selection; GS1, genomic selection trial one; GS2, genomic selection trial two; IITA, International Institute of Tropical Agriculture; LIP, loss in predictive ability; MAS, marker assisted selection; NPK, nitrogen, phosphorus and potassium; PDP, percentage difference in prediction; PHF, plant height at flowering; PLP, percentage loss in prediction; QTL, quantitative trait loci; RKHS_M, reproducing kernel Hilbert space with markers; RKHS_P, reproducing kernel Hilbert space with pedigree; RKHS_PM, reproducing kernel Hilbert space with pedigree and markers; SNP, single nucleotide polymorphism.

Published in Plant Genome 11:170090.
doi: 10.3835/plantgenome2017.10.0090

BANANAS ARE LARGE, perennial, herbaceous monocots with a majority of cultivated types being triploid ($2n = 3x = 33$). They are a staple food to millions of people in many tropical countries and a source of income for many homesteads. Triploid bananas are mostly sterile although some cultivars have residual fertility that leads to limited seed production when hand pollinated (Ssebuliba et al., 2006). They are vegetatively propagated by means of suckers, a method that limits gene flow and recombination. The lack of genetic variability of bananas grown in particular regions renders all cultivars susceptible to pests, pathogens and environmental stress. This causes reduced productivity of bananas that leads to food insecurity and income loss.

Given the importance of banana, improving the resistance of cultivated bananas is the most sustainable solution to declining production (Simmonds, 1986; Rowe, 1990). This can be achieved by crossing with wild or improved diploids that carry host plant resistance genes for pathogens and pests. The triploid nature of cultivated bananas such as the East African highland banana (EAHB), impedes the breeding process due to low fertility or complete sterility of most cultivars. To overcome this problem, breeders have to develop intermediary improved diploids and tetraploids, which serve as parents to generate secondary triploids that are resistant and high yielding. Unlike a majority of crops, banana breeding involves crossing parents of different ploidy levels (Fig. 1). Partial fertility of polyploids relies on irregular meiosis and progenies consist of individuals with different ploidy. Due to linkage drag of undesirable alleles, several evaluations and phenotypic selection at various stages are implemented making banana breeding (depicted in Fig. 2) expensive and slow. Clearly, the integration of molecular tools into conventional breeding programs is required to increase banana breeding efficiency.

Marker assisted selection (MAS) helps in selection of genotypes carrying the trait of interest at an early stage. However, very few reports on the use of MAS in banana improvement are available. For example, markers have been used to screen for *Fusarium* tropical race 4 resistance and identification of banana hybrids that are devoid of infectious endogenous banana streak virus in the B-genome (Wang et al., 2012b; Umber et al., 2016; Noumbissié et al., 2016). Most MAS technologies aim at identifying molecular markers that are linked to traits through quantitative trait loci (QTL) analysis. Once the markers are identified, the breeder can use them to track the inheritance of the traits of interest. Marker assisted selection has been successfully implemented where traits are controlled by a few QTL with major genetic effects (Asíns, 2002; Collard and Mackill, 2008). However, some traits such as yield, drought tolerance, and some others may be controlled by numerous QTL, each explaining a small portion of the genetic variance (Asíns, 2002). Identifying all QTL controlling such traits and the markers that are in linkage disequilibrium with those QTL becomes a challenge. Even if it would be possible to identify small-effect QTL, their introgression into active breeding programs would be extremely challenging.

A relatively new approach of MAS in plant breeding known as genomic selection (GS) that uses genomic prediction models was proposed by Meuwissen et al. (2001). Several variants of the original GS methodology have also been proposed (Goiffon et al., 2017). In GS, high-density markers spread across the entire genome are utilized to estimate the genetic value of a genotype using statistical models. As this estimate is based on genomic data, it is referred to as genomic estimated breeding value (GEBV). The primary advantage of GS over other forms of MAS is that the identification of individual QTL associated with a trait of interest is not necessary because QTL are assumed to be in linkage disequilibrium with at least one or more SNP (Meuwissen et al., 2001; Desta and Ortiz, 2014). The decrease in genotyping costs by next generation sequencing technologies and the emergence of GBS, which allows SNP discovery in large populations, made genomic prediction possible (Elshire et al., 2011). As the generation of marker data becomes increasingly cheaper than phenotyping, it is expected that GS will reduce breeding costs, increase selection intensity and accelerate the breeding efficiency.

Genomic selection is implemented in three phases that include: training, validation, and breeding (Jannink et al., 2010; Nakaya and Isobe, 2012). In the training phase, a model of the form "predicted phenotype = general phenotype mean in the population (intercept) + GEBV + residual error" is generated from both phenotypic and genotypic data. The predictive ability of a genomic prediction model is determined by cross validation as the correlation between the predicted and observed value of a trait or the correlation between GEBV and observed phenotype (Jannink et al., 2010; Crossa et al., 2014; Crossa et al., 2016).

Genomic selection has been successful in animal breeding (Gorddard and Hayes, 2007). It is also expected to increase genetic gain per unit time and cost in plant breeding especially when applied on traits with low heritability for which phenotypic selection is difficult and for crops with long selection cycle such as fruit trees, or banana (Wong and Bernardo, 2008; Crossa et al., 2010; Beaulieu et al., 2014; Crossa et al., 2014). Different studies in plants and animals have tested the predictive ability, or accuracy of different genomic prediction models (Legarra et al., 2008; Heffner et al., 2011; Kumar et al., 2012; Würschum et al., 2013; Crossa et al., 2016; Weng et al., 2016; Momen et al., 2017). These include best linear unbiased prediction (BLUP) and different Bayesian models (Robinson, 1991; Tibshirani, 1996; Meuwissen et al., 2001; Park and Casella, 2008; Zhang et al., 2010; Pérez and de los Campos, 2014). Characteristics of the models are summarized in numerous publications (Meuwissen et al., 2001; Habier et al., 2011; Desta and Ortiz, 2014; Pérez and de los Campos, 2014). Although these models were originally developed and optimized for diploid organisms, they have then been extended to polyploid organisms (Crossa et al., 2014; Gezan et al., 2017). However, all studies used populations with

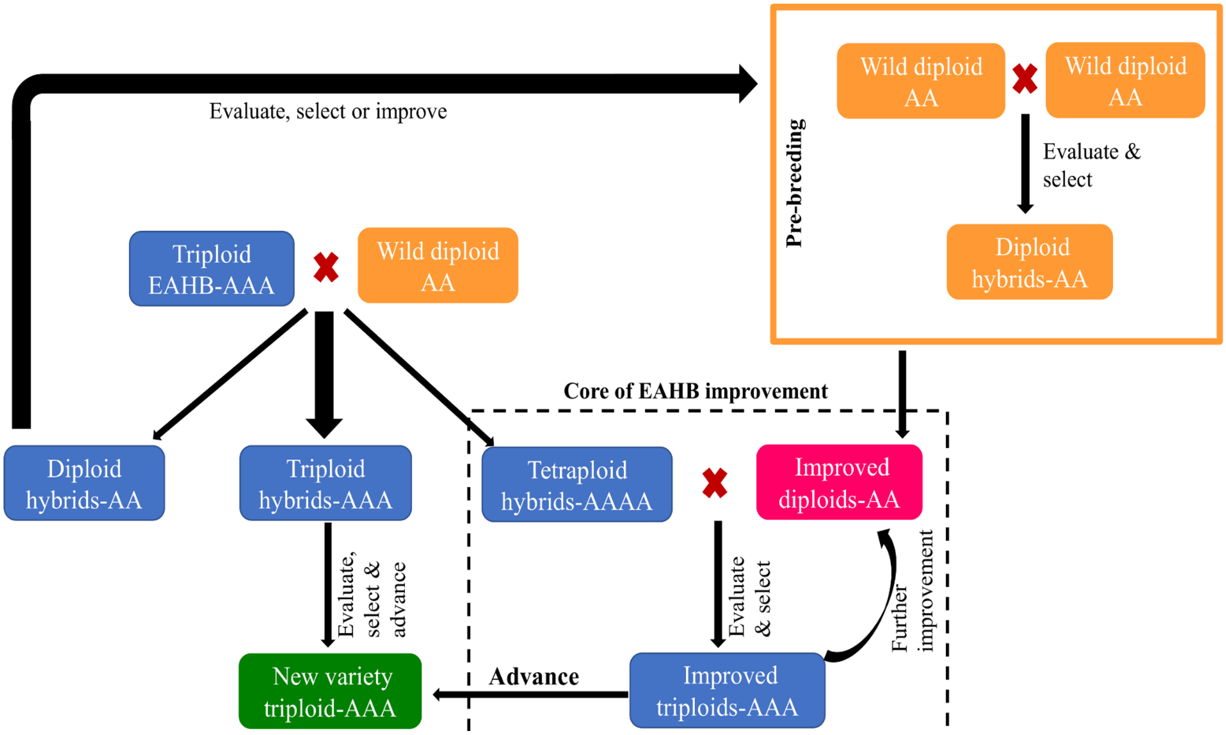


Fig. 1. Conventional crossbreeding of East African Highland bananas (EAHB) starts with crossing a triploid parthenocarpic landrace with a wild, seeded diploid accession or a diploid cultivar showing fruit parthenocarpy. This cross gives diploids, triploids and tetraploid hybrids. Tetraploids are selected and crossed with improved diploid hybrids selected from inter-diploid crosses. The resulting secondary triploids are evaluated, selected and advanced as promising improved genotypes aiming at new cultivars. The diploid and triploid (if fertile) hybrids can be further improved by crossing with other wild or improved diploids.

organisms of the same ploidy level. Polyploid organisms are challenging to model due to (i) uncertainty of allele frequency in the population and (ii) uncertainty of allele dosage at the loci (Blischak et al., 2016).

For bananas, besides the polyploid nature, there is a small effective breeding population. Yet the accuracy of genomic prediction depends on the size of training population. It should be large enough to capture all the segregating alleles in the breeding genetic pool (Crossa et al., 2014; Bassi et al., 2016). However, as noted by Bassi et al. (2016), no ideal population size exists for all species and traits. The parameters that need to be considered include relatedness of the individuals, the heritability of the trait, differences in linkage disequilibrium between markers and QTL across training and breeding populations, whether the population is bi-parental, or a mixture of several families and the cost involved in phenotyping the training population. For example, Beaulieu et al. (2014) used 1694 open pollinated genotypes of white spruce with 6385 SNP markers and obtained different accuracies of prediction depending on the trait and the relationship between the training and validation data sets. The highest predictive ability observed was 0.44 for cell radial diameter. In contrast, Crossa et al. (2010) used a maize population of less than 300 individuals with less than 1200 markers and obtained a predictive ability as high as 0.79 for male flowering under well-watered conditions.

This study explored the potential of genomic prediction in banana, a polyploid crop for which the population was composed of individuals with different ploidy levels, but mostly triploids (~85%) derived from EAHB. The objectives were to (i) compare the predictive ability of a set of six models with marker, pedigree, and both pedigree and marker information for 15 traits scored in the training population, and select the best genomic prediction model for each trait or a group of traits, (ii) determine the predictive ability of models with a training population grown under two different field management practices (i.e., studying genotype × environment interaction), (iii) determine the predictive ability of the best model for prediction of traits within and across crop cycle 1/mother plants and crop cycle 2/first ratoons/first suckers (i.e., genotype × cycle interaction), (iv) determine the effect of accounting for allele dosage on the predictive ability of the best genomic prediction model for each trait, (v) determine the effect of using genomic prediction models fitted with averaged environment phenotype data and allele dosage SNP (AD-SNP) markers on the prediction of genotype performance in particular environments and (vi) determine the accuracy of selection based on GEBV relative to phenotypic data within the training population. To achieve these objectives, a training population of 307 banana genotypes consisting of breeding clones and hybrids was phenotyped and genotyped.

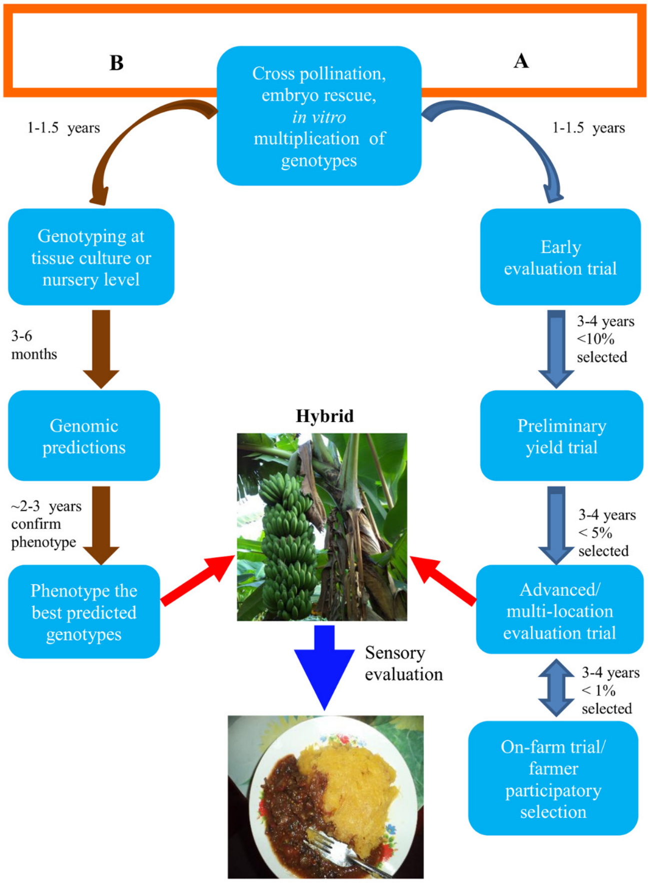

B
A
Cross pollination, embryo rescue, in vitro multiplication of genotypes
1-1.5 years
1-1.5 years
Genotyping at tissue culture or nursery level
Early evaluation trial
3-6 months
3-4 years <10% selected
Genomic predictions
Hybrid
Preliminary yield trial
~2-3 years confirm phenotype
3-4 years < 5% selected
Phenotype the best predicted genotypes
Advanced/ multi-location evaluation trial
Sensory evaluation
3-4 years < 1% selected
On-farm trial/ farmer participatory selection

Fig. 2. Approaches to hybrid selection in banana breeding program. (A) The classical phenotypic selection of banana hybrids and (B) integrated genomic selection and phenotypic selection approach being investigated.

MATERIALS AND METHODS

Phenotyping

The banana genomic selection training population used in this study and the traits measured were described in detail by Nyine et al. (2017). Briefly, the training population consisted of 307 genotypes that included diploid (11%), triploid (85%), and tetraploid (4%) plants. The core breeding clones (parents) accounted for 12% of the population. The triploid parents were EAHB some of which were crossed with cultivar (cv.) Calcutta 4 to generate tetraploid hybrids, which are used as breeding clones (Supplemental Table S1). The diploid parents consisted of both wild and improved parthenocarpic genotypes. The rest were hybrids from early evaluation trials and advanced clones that had been selected over time during the 20 year of banana breeding by the International Institute of Tropical Agriculture (IITA) and the National Agricultural Research Organization of Uganda. In total, 77 families (cross combinations) of variable sizes were represented in this population. Phenotyping was done at IITA research station located at Sendusu in Namulonge, 0.53° N 32.58° E, 1150 m above sea level with rainfall of about 1200 mm/year split into two rainy seasons, March-June and September-December, and an average annual temperature of 22°C.

Two phenotyping fields were established to mimic different agronomic practices that farmers use, thus creating a difference in growth environment. A completely randomized design with three replications per genotype was used to establish the fields. Sword and maiden suckers were used as planting materials with a spacing of 2 × 3 m. In the genomic selection trial one (GS1), 20 kg of manure was applied at planting, but neither mulching, nor nitrogen, phosphorus and potassium (NPK) fertilizer application was done afterward and this was considered a low input field management. The genomic selection trial two (GS2) was planted with 20 kg of manure, then mulched, and NPK fertilizer (25:5:5) was added at a rate of 480 g per plant mat per year, and this was considered a high input field management. In both fields, sucker management was done to maintain a maximum of three plants per mat.

Data were collected on two crop cycles in each field between 2013 and 2016. Fifteen traits were considered for genomic prediction modeling and these were categorized as plant stature, suckering behavior, black leaf streak resistance, fruit bunch, and fruit filling. For plant stature, plant height and girth at 100 cm from soil surface were measured at flowering. The total number of suckers and height of tallest sucker were recorded at flowering of crop cycle 1 and height of tallest sucker at harvest to represent suckering behavior. The number of standing leaves and index of non-spotted leaves were determined at flowering to characterize black leaf streak resistance. The index of non-spotted leaves was calculated according to the formula of Craenen (1998) with some modification as reported by Nyine et al. (2017). The fruit bunch traits scored at harvesting included the days to fruit maturity, bunch mass, number of hands, and number of fruits. For fruit filling, fruit length, fruit circumference, fruit diameter, and pulp diameter were measured at harvest. The data were checked for outliers and entry errors prior to use in model fitting. It should be noted that not all traits had full data sets because some genotypes had not completed the second cycle through harvest by the time of these analyses.

Genotyping

The population was genotyped by sequencing as described by Elshire et al. (2011). The restriction enzyme *Pst*I was used in the genome complexity reduction during sequencing library preparation. Barcodes containing adaptors were ligated to the genomic DNA fragments. Ninety-six samples were multiplexed and sequenced on a single Illumina lane at the Institute of Genomic Diversity, Cornell University. Each set of 96 samples was run twice to increase the number of reads per *Pst*I tag. Single-end reads of 100 bp were generated during sequencing. A workflow for the analysis of sequence reads was developed (Supplemental Fig. S1).

Sequence reads were filtered using fastq_quality_filter provided in the module fastx.0.0.13 (-q 20-p 90). Sequence reads were subjected to quality control analysis using fastqc provided in module FastQC.0.10.1. Reads from each lane were de-multiplexed into individual sample reads using fastx_barcode_splitter.pl provided in fastx.0.0.13. The barcodes were trimmed using fastx_trimmer in the module fastx.0.0.13. Any remaining adaptor sequences were removed using fastx_clipper also provided in module fastx.0.0.13. The *Pst*I tag (5'-TGCAG—–3') was retained on each sequence read to act as a reference point during read alignment to the reference genome. Reads of the same genotype were merged into one file for downstream analysis. Bowtie2 was used to align reads to the latest publicly available reference banana genome (Martin et al., 2016). Read groups were added to aligned sample reads after which the duplicate reads were marked and removed using picard-1.100. Indels were realigned and all realigned reads from all samples were merged into one file before SNP calling.

Genome analysis tool kit (GATK) version 2.7.2, UnifiedGenotyper (https://software.broadinstitute.org/gatk/documentation/) was used as the variant caller. First, all genotypes were considered as diploids and as such bi-allelic SNP (BA-SNP) were called. Second, the population was split and grouped according to ploidy level. The respective ploidy levels were set during SNP calling. Preliminary filtering of SNP was performed prior to output of variant call file (VCF). The filters used were QD < 2.0, FS > 60.0, MQ < 40 and Haplotypescore > 13.0. Further stringent filtering was done in R (R core team, 2016) where SNP loci with quality score less than 98 and more than 50% of the banana genotypes having missing data were excluded. Concordant SNP loci across all ploidy levels were selected to generate a file with SNP where allele dosage had been accounted for. The remaining missing data were imputed with impute function in R and SNP converted into numerical data for input into genomic prediction models using a custom R-script. The description of how the script works can be accessed here: http://olomouc.ueb.cas.cz/system/files/users/public/scripts/AlleleDosage_R_function.docx

Comparison of Genomic Prediction Models and the Effect of Field Management and Crop Cycle on their Performance

Bayesian models accounting for additive genetic effects (Bayesian Ridge Regression [BRR], Bayesian LASSO [BL], BayesA, BayesB and BayesC), and reproducing kernel Hilbert space models with pedigree (P), markers (M), pedigree and markers (PM) accounting for non-additive genetic effects (RKHS_P, RKHS_M and RKHS_PM) were compared. All models were implemented in R-package BGLR (Pérez and de los Campos, 2014) using 10807 BA-SNP markers. Since the training population consisted of many small families and genotypes of different ploidy levels, both phenotype and SNP data were completely randomized in the same order. The aim was to minimize the effect of family structure and ploidy level during cross validation.

The phenotype data used were the average phenotypic observations per genotype per field. These were calculated using the function 'aggregate' provided in R-package plyr. The training population was divided into five groups and each group was used once as the testing (cross validation) set. The predictive ability of the model was determined as the average correlation between the predicted and observed phenotype of the testing sets from five cross validations. Across field management, cross validation was done so that data from one field were used to generate the model using the training set, and the predicted phenotypes of the genotypes in the testing set were correlated to the observed phenotypes in the second field.

For all models, the priors for parameters such as shape, rate, and counts were estimated from the data. However, for BayesB and BayesC models, the prior probability of a marker having a non-null effect on the phenotype (probIn value) was set at 0.05 and the degrees of freedom were set according to the available phenotype and genotype data. The genetic variance in all models was set at 0.5. For every cross validation, 10,000 iterations were run with a burnIn of 5000 and thin 10.

The fifteen traits mentioned above were predicted with all models to determine the best genomic prediction model for each trait or group of traits. The effect of using models generated with data from low input field management to predict performance of genotypes under high input management and vice versa (G × E effect) was also evaluated.

Next, the effect of crop cycle on trait prediction was evaluated using one of the best identified genomic prediction model. Cross validation across and within crop cycles was done using the 10807 BA-SNP markers and the average phenotype per crop cycle 1 and crop cycle 2 of each field. Five cross validations were performed without overlap of genotypes between the training and testing set in each round. Only a few traits representing the trait categories were considered because of high correlation of traits within trait categories (Nyine et al., 2017). They included plant girth at 100 cm from soil surface, index of non-spotted leaves, bunch mass, and fruit circumference. The total number of suckers was not analyzed because this trait was scored only in crop cycle 1.

Effect of Allele Dosage on Model Performance

The performance of BayesB, BRR, BL, and RKHS_M models fitted with BA-SNP and AD-SNP markers was compared for the 15 traits. Predictions based on BA-SNP markers were used as the baseline for comparison. Equal number of SNP from same loci for both BA-SNP and AD-SNP were used. Combined phenotypic data from the two fields for the two crop cycles (environment averaged data) were used to calculate the mean phenotype of each individual genotype. In this cross-validation strategy, first, genotypes were completely randomized. A five-fold cross validation was performed using similar priors to determine the predictive ability of the model for the trait. Second, the performance of parents' model versus progeny's model was compared using BA-SNP and AD-SNP. Here, the training set consisted of either only parents (parents' model), or progeny (progeny's model). Third, the population was divided into three groups consisting of diploids, triploids, and tetraploids. The training set comprised of any two of the ploidy groups while the testing set consisted of genotypes from one ploidy level. Due to differences in population sizes under different ploidy level, we also used only triploids to compare the effect of accounting for allele dosage.

The effect of using averaged environment model was assessed based on AD-SNP to predict plant girth at 100 cm from soil surface, total number of suckers, index of non-spotted leaves, bunch mass, and fruit circumference under low and high input fields. The percentage difference in prediction (PDP) between low and high input fields was calculated in reference to the prediction in the low input field management.

To understand the variation and trend of predictive ability across traits, both broad (H^2) and narrow (h^2) sense heritabilities were estimated following the methods described by Kruijer et al. (2015). The BA-SNP markers (10,807) and phenotypic means from each field were used to estimate h^2 using R-package heritability while the results from analysis of variance were used to estimate H^2. Type B genetic correlation was also performed based on phenotypic means from GS1 and GS2 to determine the effect of G × E interaction on the trend of trait prediction across fields (Burdon, 1977).

The Accuracy of Genomic Prediction within the Training Population

The GEBV obtained from the models fitted with AD-SNP with best and worst predictive abilities for plant girth, total number of suckers, index of non-spotted leaves, bunch mass and fruit circumference were used to rank the genotypes. The top 100 genotypes were compared with the best 100 genotypes ranked on the basis of the environment averaged phenotypic data. The number of genotypes out of 100 captured by both GEBV and phenotypic data was reported as the estimated accuracy of genomic prediction within the training population. For this analysis, the best genomic prediction model identified above was used.

RESULTS

Genotyping

The discovery of SNP markers from GBS reads for the training population was based on the latest publicly available version of the double haploid *Musa acuminata* cv. Pahang reference genome sequence (Martin et al., 2016). To account for allele dosage in genotypes of different ploidy, a workflow was developed for the analysis of sequence data and GATK, UnifiedGenotyper was used as SNP caller (Supplemental Fig. S1). It produced 52076 BA-SNP after pre-filtering. Less than one percent of the loci had multi-allelic SNP. They were eliminated from the data to avoid potential sequencing artifacts. After further stringent filtering in R (R core team, 2016), 10807 BA-SNP markers that were polymorphic with a minimum minor allele frequency of 0.01 were retained. These were distributed on 11 pseudomolecules as well as on unanchored scaffold of the banana reference genome (Fig. 3). The percentage of imputed missing genotypes was 16%. Accounting for allele dosage within the ploidy groups (diploids, triploids, and tetraploids) reduced the number of SNP markers to 5574.

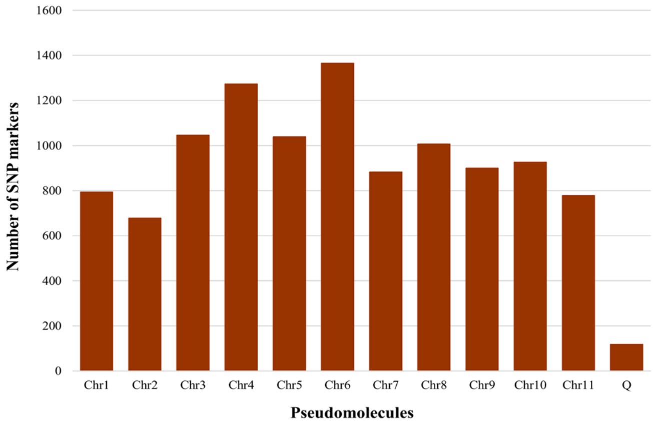


Fig. 3. Distribution of filtered SNP markers on 11 pseudomolecules of the double haploid of *M. acuminata* cv. Pahang (Martin et al., 2016). Q represents the unanchored scaffolds.

Comparison of Genomic Prediction Models and the Effect of Field Management and Crop Cycle on their Performance

The best genomic prediction model for different traits was selected based on congruity of predictive ability results from cross validation between fields using BA-SNP markers. The predictive ability of all models varied across traits (Table 1; Supplemental Table S2). For 12 out of 15 traits, genomic prediction models that account for additive genetic effects gave the highest predictions ranging from 0.2 to 0.72. These were the correlations between the predicted and observed phenotypes for the various traits. Reproducing kernel Hilbert space model combining both pedigree and marker information (RKHS_PM) gave the highest predictions ranging from 0.24 to 0.49 for 3 out of 15 traits and these were the days to fruit maturity, height of tallest sucker at flowering and height of tallest sucker at harvesting. BayesB and BayesC models predicted equally well and better than other models for fruit filling and fruit bunch traits. For example, the predictions of all fruit filling traits by both models ranged from 0.65 to 0.72. For plant stature, suckering behavior and black leaf streak resistance traits, BayesB and BayesC models were not the best, but either had the same predictive ability, or were lower by 5 – 13 % in prediction as compared to other models. The trend of prediction starting from the highest to the lowest trait category was fruit filling, fruit bunch, plant stature, black leaf streak resistance, and suckering behavior. In general, genomic prediction models fitted with phenotypic data from GS1 underpredicted the performance of genotypes in GS2, and vice-versa (Fig. 4), but this did not affect the trend of prediction across

Table 1. Comparison of average correlation (standard errors in parentheses) for five-fold cross validations between the predicted and observed phenotypes across models fitted with data from either low input (GS1) or high input (GS2) fields and 10807 bi-allelic SNP markers.

Trait category	Trait	BRR		BayesB		BayesC		RKHS_M		RKHS_PM	
		GS1	GS2	GS1	GS2	GS1	GS2	GS1	GS2	GS1	GS2
Plant stature	Plant height	0.54 (0.06)	0.46 (0.09)	0.54 (0.06)	0.44 (0.09)	0.54 (0.07)	0.45 (0.09)	0.55 (0.06)	0.44 (0.09)	0.54 (0.05)	0.48 (0.07)
	Plant girth	0.60 (0.06)	0.52 (0.05)	0.60 (0.06)	0.52 (0.06)	0.60 (0.06)	0.51 (0.05)	0.60 (0.06)	0.51 (0.06)	0.55 (0.04)	0.50 (0.05)
Suckering behavior	Total number of suckers	0.16 (0.06)	0.17 (0.06)	0.16 (0.06)	0.1(9 (0.06)	0.15 (0.06)	0.19 (0.07)	0.17 (0.06)	0.18 (0.06)	0.16 (0.04)	0.17 (0.07)
	Height of tallest sucker at flowering	0.28 (0.05)	0.18 (0.09)	0.27 (0.05)	0.20 (0.08)	0.26 (0.05)	0.2 (0.08)	0.28 (0.05)	0.19 (0.09)	0.30 (0.06)*	0.24 (0.09)*
	Height of tallest sucker at harvesting	0.27 (0.05)	0.26 (0.07)	0.28 (0.06)	0.24 (0.06)	0.27 (0.06)	0.25 (0.07)	0.26 (0.05)	0.26 (0.06)	0.29 (0.03)*	0.32 (0.07)*
Black leaf streak	Number of standing leaves at flowering	0.36 (0.08)	0.42 (0.08)	0.43 (0.06)	0.40 (0.08)	0.36 (0.08)	0.41 (0.08)	0.37 (0.08)	0.41 (0.08)	0.29 (0.07)	0.34 (0.04)
	Index of non-spotted leaves	0.35 (0.04)	0.42 (0.06)	0.34 (0.05)	0.43 (0.06)	0.34 (0.05)	0.43 (0.06)	0.35 (0.05)	0.42 (0.06)	0.32 (0.07)	0.36 (0.10)
Fruit bunch	Days to fruit maturity	0.47 (0.07)	0.42 (0.09)	0.47 (0.07)	0.42 (0.09)	0.46 (0.07)	0.42 (0.09)	0.47 (0.07)	0.42 (0.10)	0.49 (0.06)*	0.44 (0.09)*
	Bunch mass	0.63 (0.03)	0.61 (0.03)	0.64 (0.03)*	0.62 (0.03)*	0.64 (0.03)*	0.62 (0.03)*	0.61 (0.03)	0.61 (0.03)	0.52 (0.06)	0.55 (0.04)
	Number of hands	0.60 (0.03)*	0.62 (0.04)*	0.60 (0.02)*	0.62 (0.04)*	0.59 (0.02)	0.62 (0.04)	0.59 (0.03)	0.62 (0.04)	0.48 (0.03)	0.53 (0.02)
	Number of fruits	0.47 (0.03)	0.51 (0.04)	0.47 (0.03)*	0.52 (0.04)*	0.47 (0.02)*	0.52 (0.04)*	0.45 (0.03)	0.52 (0.04)	0.35 (0.04)	0.45 (0.04)
Fruit filling	Fruit length	0.65 (0.04)	0.64 (0.02)	0.67 (0.04)*	0.65 (0.02)*	0.67 (0.03)*	0.65 (0.02)*	0.64 (0.04)	0.64 (0.02)	0.59 (0.07)	0.59 (0.02)
	Fruit circumference	0.67 (0.02)	0.66 (0.01)	0.70 (0.01)*	0.69 (0.01)*	0.70 (0.01)*	0.69 (0.01)*	0.65 (0.02)	0.66 (0.01)	0.57 (0.05)	0.60 (0.02)
	Fruit diameter	0.67 (0.01)	0.63 (0.05)	0.70 (0.01)*	0.71 (0.02)*	0.70 (0.01)*	0.71 (0.02)*	0.65 (0.02)	0.67 (0.03)	0.57 (0.04)	0.59 (0.02)
	Pulp diameter	0.67 (0.02)	0.68 (0.04)	0.70 (0.01)*	0.72 (0.03)*	0.70 (0.01)*	0.72 (0.03)*	0.65 (0.02)	0.67 (0.04)	0.57 (0.04)	0.60 (0.03)

*Highest predictive value observed in both GS1 and GS2 for a trait using same model type. The values under GS1 column are the correlations between predicted and observed phenotype (predictive ability) in GS2 when GS1 data were used to fit the model and vice versa for GS2 column.

traits. Little difference in prediction was observed across all models for traits within the same category.

The performance of RKHS model fitted with marker data (RKHS_M) was comparable to BRR, BL, and BayesA models fitted with marker data. RKHS model fitted with pedigree information alone (RKHS_P) had the least predictive ability that ranged from 0.12 to 0.5 (Supplemental Table S2). There was a 4 to 29% loss in predictive ability (LIP) of most traits when marker and pedigree information were combined in the RKHS_PM model. However, the same model gave a 4 to 21% gain in prediction for plant height, height of tallest sucker at flowering, height of tallest sucker at harvesting and days to fruit maturity.

The effect of crop cycle on trait prediction was tested with BayesB model using BA-SNP markers, because this model either out-performed other models, or performed equally well as noted in Table 1; Supplemental Table S2. The cross-validation strategies used were (a) within crop cycle cross validation for which both the training and testing sets were from the same crop cycle and (b) across crop cycle cross validation where the training and testing sets were selected from different crop cycles within the same field. The predictive ability of BayesB model fitted with crop cycle 1, or crop cycle 2 data in both low input and high input fields yielded mixed results when within and across crop cycle cross validations were performed for different traits (Table 2). Predictive ability of the model for fruit circumference and bunch mass ranged from 0.58 to 0.73, while for plant girth and index of non-spotted leaves ranged from 0.39 to 0.61 and 0.26 to 0.44, respectively, in both fields and crop cycles. Less than 2% variation in prediction across and within crop cycles was observed in both bunch mass and fruit circumference. The highest difference of 20% in prediction across (0.28) and within (0.35) crop cycle was recorded in GS2 for index of non-spotted leaves when crop cycle 2 data were used to fit the model.

Effect of Allele Dosage

The effect of AD-SNP on predictive ability of the best genomic prediction models was evaluated for 15 traits in comparison to predictions based on BA-SNP markers. For both BA-SNP and AD-SNP, 5574 SNP markers from the same loci and combined phenotypic data from the two fields for the two crop cycles (environment averaged data) were used to fit the models. First, genotypes were completely randomized to minimize the effect of family structure and ploidy. Second, the training set consisted of either only parents (parents' model), or progeny (progeny's model). Third, the population was divided into diploids, triploids, and tetraploids. The training set comprised of any two of the ploidy groups while the testing set consisted of genotypes from one ploidy level. Lastly, only triploids were considered during cross validation since 85% of genotypes in the training population were triploids. The aim was to understand what traits and which ploidy level were mostly affected by allele dosage when implementing genomic predictions.

The results of the comparison of the effect of allele dosage on performance of BayesB, BRR, BL, and RKHS_M models are summarized in Table 3. When AD-SNP were used to fit the models, predictive ability of all models was trait dependent, but generally reduced by 15% on average as compared to the traditional BA-SNP markers. When only triploids were considered during the cross validation, predictive ability for fruit circumference fell by 10% from 0.76 to 0.68, while for bunch mass it decreased by 5% from 0.62 to 0.59. The highest loss in prediction (PLP) of 24 to 44% was observed in suckering

Table 2. Average predictive ability (standard errors in parentheses) of BayesB model fitted with either crop cycle 1, or crop cycle 2 phenotype data from low (GS1) and high (GS2) input field management using bi-allelic SNP markers to predict traits across and within crop cycles.

		Low input field management (GS1)				High input field management (GS2)			
		Cycle 1 model		Cycle 2 model		Cycle 1 model		Cycle 2 model	
Trait category	Trait	Across	Within	Across	Within	Across	Within	Across	Within
Plant stature	Plant girth	0.39 (0.04)	0.55 (0.03)	0.51 (0.02)	0.44 (0.05)	0.54 (0.02)	0.59 (0.02)	0.61 (0.02)	0.57 (0.02)
Black leaf streak	Index of non-spotted leaves	0.42 (0.06)	0.44 (0.03)	0.40 (0.04)	0.41 (0.03)	0.30 (0.08)	0.26 (0.04)	0.28 (0.05)	0.35 (0.05)
Fruit bunch	Bunch mass	0.58 (0.03)	0.60 (0.04)	0.60 (0.06)	0.59 (0.03)	0.63 (0.02)	0.65 (0.03)	0.65 (0.02)	0.62 (0.03)
Fruit filling	Fruit circumference	0.72 (0.02)	0.71 (0.03)	0.72 (0.04)	0.72 (0.02)	0.73 (0.02)	0.73 (0.03)	0.71 (0.02)	0.72 (0.02)

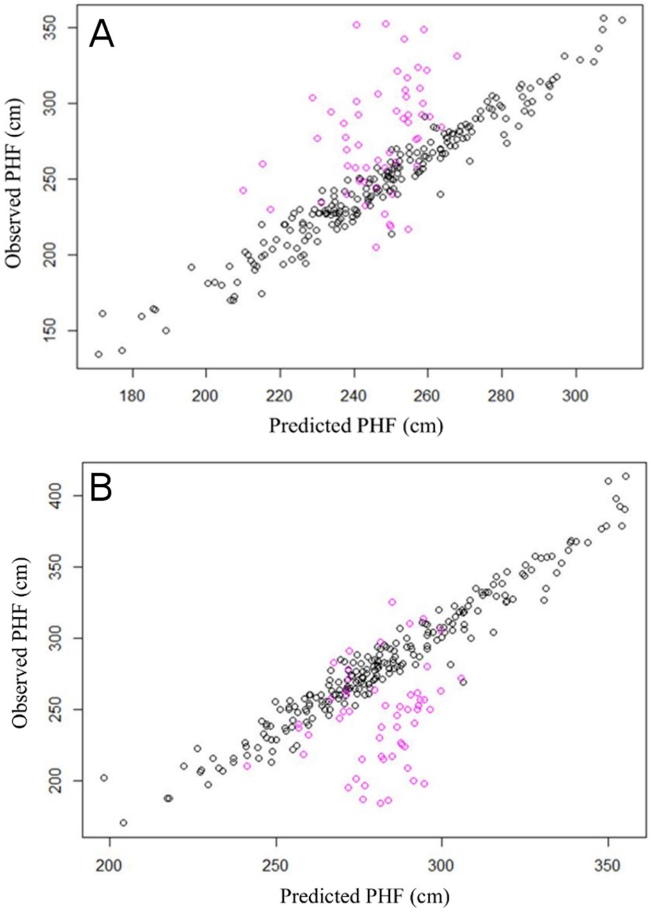


Fig. 4. Prediction of plant height at flowering (PHF) using a Bayesian ridge regression model fitted with phenotype data from low input field (A) and high input field (B). Where A, shows underprediction and B, shows overprediction of PHF. The black and magenta circles represent genotypes in the training and testing sets, respectively.

behavior traits when AD-SNP markers were used to fit model using genotypes from all ploidy levels. However, the trend of prediction within and across trait categories did not change by accounting for allele dosage. Fruit filling traits were the best predicted with the highest predictive ability of 0.68 for pulp diameter. BayesB model maintained its superior prediction accuracy over other models, especially for fruit filling and fruit bunch traits.

Although the number of SNP markers used in this prediction was reduced to 5574 because we wanted to eliminate the bias in predictions due to variable number and location of BA-SNP and AD-SNP, the environment (field management) averaged models with BA-SNP markers gave higher predictions than those obtained with across field cross validation with 10,807 SNP markers for all traits. The highest predictive ability recorded was 0.75 for fruit filling traits with the BayesB model (Table 3).

When only parental data were used to fit BayesB model (parents' model), the predictive ability of traits within the progeny ranged from 0.13 to 0.59 for BA-SNP and from -0.15 to 0.33 for AD-SNP (Supplemental Table S3). The LIP due to accounting for allele dosage was 63% on average (36–179%). Similarly, when progeny data were used to fit BayesB model (progeny's model), the predictive ability of traits within parents ranged from 0.39 to 0.86 with BA-SNP and from -0.03 to 0.77 with AD-SNP markers. The LIP due to accounting for allele dosage was 35% on average (1.5–107%). The highest predictive value obtained with BayesB model fitted with BA-SNP was 0.86 for number of hands. This prediction dropped by nearly 50% (0.48) when AD-SNP markers were used. Prediction accuracy of the same trait in progeny using parents' model was 0.45 with BA-SNP and 0.03 with AD-SNP markers. The prediction of bunch mass in the progeny using a parents' model with AD-SNP was 0.17 while the prediction of the same trait in parents using a progeny's model reduced to 0.08.

Since allele dosage varies with ploidy level, cross validation across ploidy levels was performed. Genotypes from two ploidy levels were used to train the model and only genotypes of same ploidy level were included in the testing set during cross validation. Accounting for allele dosage positively increased the predictive ability of all fruit filling traits in tetraploids with BayesB model, but the results from other trait categories varied greatly (Supplemental Table S4). For example, prediction of pulp diameter increased from -0.39 to 0.60, fruit diameter increased from -0.45 to 0.53 and fruit circumference increased from -0.15 to 0.35. BayesB model fitted with triploid and tetraploid data, and BA-SNP gave the predictions ranging from 0.32 to 0.86 for traits among diploids. Tetraploids and diploids were the least represented in the training population (47 out of 307 genotypes, or 15%) and of which the majority were parents. When their data were used to fit the model to predict traits in triploids the prediction varied from 0.20 to 0.54 and from -0.06 to 0.11 with BA-SNP and AD-SNP, respectively.

When BayesB model was fitted with the environment averaged data (including all ploidy levels) and AD-SNP

Table 3. Effect of accounting for allele dosage on the predictive ability of genomic prediction models using environment averaged phenotype data.

		Bi-allelic SNP				Allele dosage SNP			
Trait category	Trait	BRR	BayesB	BL	RKHS_M	BRR	BayesB	BL	RKHS_M
Plant stature	Plant height	0.54 (0.03)†	0.53 (0.02)	0.52 (0.03)	0.53 (0.03)	0.46 (0.07)	0.45 (0.06)	0.44 (0.07)	0.45 (0.07)
	Plant girth	0.53 (0.04)	0.53 (0.03)	0.52 (0.04)	0.52 (0.04)	0.48 (0.04)	0.47 (0.04)	0.47 (0.04)	0.48 (0.04)
Suckering behavior	Total number of suckers	0.32 (0.06)	0.29 (0.06)	0.33 (0.05)	0.31 (0.06)	0.21 (0.05)	0.16 (0.05)	0.21 (0.05)	0.21 (0.05)
	Height of tallest sucker at flowering	0.37 (0.04)	0.34 (0.04)	0.37 (0.04)	0.38 (0.04)	0.27 (0.06)	0.26 (0.05)	0.27 (0.05)	0.28 (0.05)
	Height of tallest sucker at harvesting	0.35 (0.04)	0.33 (0.03)	0.34 (0.04)	0.35 (0.04)	0.24 (0.03)	0.23 (0.03)	0.23 (0.03)	0.25 (0.03)
Black leaf streak	Number of standing leaves at flowering	0.49 (0.05)	0.48 (0.05)	0.48 (0.05)	0.48 (0.05)	0.48 (0.06)	0.48 (0.06)	0.48 (0.06)	0.49 (0.06)
	Index of non-spotted leaves	0.58 (0.03)	0.59 (0.03)	0.58 (0.03)	0.58 (0.03)	0.53 (0.03)	0.52 (0.03)	0.53 (0.04)	0.53 (0.03)
Fruit bunch	Days to fruit maturity	0.53 (0.05)	0.54 (0.06)	0.53 (0.06)	0.53 (0.06)	0.44 (0.05)	0.43 (0.05)	0.44 (0.05)	0.44 (0.05)
	Bunch mass	0.61 (0.05)	0.62 (0.04)	0.61 (0.05)	0.61 (0.04)	0.54 (0.03)	0.56 (0.03)	0.54 (0.03)	0.54 (0.02)
	Number of hands	0.63 (0.04)	0.62 (0.04)	0.62 (0.04)	0.63 (0.04)	0.56 (0.03)	0.56 (0.03)	0.56 (0.03)	0.56 (0.03)
	Number of fruits	0.49 (0.04)	0.49 (0.04)	0.48 (0.04)	0.50 (0.04)	0.43 (0.03)	0.42 (0.04)	0.42 (0.03)	0.43 (0.04)
Fruit filling	Fruit length	0.69 (0.02)	0.70 (0.02)	0.69 (0.03)	0.69 (0.02)	0.60 (0.03)	0.64 (0.02)	0.60 (0.02)	0.59 (0.03)
	Fruit circumference	0.67 (0.03)	0.75 (0.02)	0.68 (0.03)	0.66 (0.03)	0.59 (0.03)	0.66 (0.03)	0.60 (0.03)	0.59 (0.03)
	Fruit diameter	0.67 (0.03)	0.75 (0.02)	0.68 (0.03)	0.66 (0.03)	0.60 (0.03)	0.67 (0.03)	0.62 (0.02)	0.60 (0.02)
	Pulp diameter	0.68 (0.03)	0.75 (0.03)	0.69 (0.03)	0.67 (0.03)	0.61 (0.03)	0.68 (0.03)	0.63 (0.03)	0.61 (0.02)

†The values in parentheses are the standard errors of predictive ability.

to predict the traits under low and high input fields, there was a 2 to 8% increase in predictive ability under high input field relative to low input field for plant girth, bunch mass and fruit circumference (Table 4). However, for total number of suckers and index of non-spotted leaves, the predictions reduced by 47 and 15%, respectively.

Estimated H^2 and h^2 had positive relationship with predictive ability. However, h^2 varied across fields with some traits having higher h^2 than H^2 (Table 5). A similar trend was observed between Type B genetic correlation and predictive ability. The correlation varied between 0.71 and 0.9 for fruit bunch, fruit filling and plant stature traits. The lowest correlation was recorded on the index of non-spotted leaves (Table 5).

The Accuracy of Genomic Prediction within the Training Population

The first 100 genotypes with the highest GEBV and the first 100 genotypes with the highest environment averaged phenotypic data were compared (Fig. 5). The GEBV used were obtained from BayesB model with best and worst predictive abilities based on AD-SNP markers. The number of genotypes out of 100 captured by both GEBV and phenotypic data was reported as the estimated accuracy of genomic prediction within the training population for the trait. The accuracy of prediction ranged from 76 to 84% for all the traits whereas the prediction values ranged from 0.04 to 0.76. Models that gave high predictive ability values had also the highest prediction accuracy.

DISCUSSION

Genotyping

Genomic selection as a form of marker assisted selection has been investigated in a range of plant species including, for example, maize and wheat (Heffner et al., 2011; Crossa et al., 2014; Crossa et al., 2016; Pérez-Rodríguez et al., 2017), white spruce (Beaulieu et al., 2014), sugar beet (Würschum et al., 2013), apples (Kumar et al., 2012), strawberries (Gezan et al., 2017), and rice (Onogi et al., 2016). In these experiments, genotypes of same ploidy level constituted the training population. The present study on banana is unique in this respect as three ploidy levels were represented in the training population. Within the three ploidy levels, both parents and progeny were represented in varying proportions. The hybrids in the training population arose from 77 cross combinations, mainly involving crosses between tetraploids and diploids (Nyine et al., 2017). Innovative approaches in SNP calling, including custom R-script had to be adopted for such an unconventional population (Supplemental Fig. S1). The script removes loci with monomorphic SNP, eliminates loci with more than two alternative SNP alleles, and converts the SNP file into a numerical format while accounting for allele dosage, and it can be customized to any polyploid plant species. Loci with multi-allelic SNP were eliminated because GBS is a low coverage sequencing technology. This makes it hard to differentiate true rare SNP from sequence artifacts especially when the population is small and the species is clonally propagated due to lower rate of multiple mutations at the same locus. Bowtie2 was used as the sequence alignment tool while GATK, UnifiedGenotyper was the variant caller. However, as indicated by Clevenger et al. (2015), optimal alignment programs and variant callers may vary among species.

GATK (https://software.broadinstitute.org/gatk/documentation/) in particular is useful when handling polyploid species. It allows setting the ploidy level and reduces false positive SNP calls arising from frameshifts by running INDEL realignment step (Clevenger et al., 2015). When Picard tools (https://sourceforge.net/projects/picard/

Table 4. Performance of BayesB model fitted with average phenotype data for all fields (environments) and AD-SNP markers for predictions of five traits representing the trait categories within low and high input fields.

Trait category	Trait	Low input field (GS1)	High input field (GS2)	Percentage loss in prediction (PDP)
Plant stature	Plant girth	0.48 (0.07) †	0.52 (0.08)	8.3
Suckering behavior	Total number of suckers	0.15 (0.05)	0.08 (0.05)	–46.7
Black leaf streak	Index of non-spotted leaves	0.39 (0.06)	0.33 (0.05)	-15.4
Fruit bunch	Bunch mass	0.56 (0.05)	0.57 (0.05)	1.8
Fruit filling	Fruit circumference	0.66 (0.01)	0.69 (0.03)	4.5

†The values in parentheses are the standard errors of predictive ability, PDP is percentage difference in prediction.

Table 5. Estimated broad (H^2), narrow (h^2) sense heritability within low (h^2_GS1) and high (h^2_GS2) input fields and type B genetic correlation (*r*) between GS1 and GS1.

Trait category	Trait	H^2	h^2_GS1	h^2_GS2	r GS1/GS2 (type B)
Plant stature	Plant height	0.89	0.99	0.93	0.79
	Plant girth	0.90	0.93	0.91	0.83
Suckering behavior	Total number of suckers	0.80	0.45	0.36	0.49
	Height of tallest sucker at flowering	0.82	0.70	0.93	0.56
	Height of tallest sucker at harvesting	0.86	0.41	0.84	0.47
Black leaf streak	Number of standing leaves at flowering	0.83	0.63	0.81	0.54
	Index of non-spotted leaves	0.72	0.72	0.63	0.38
Fruit bunch	Days to fruit maturity	0.89	0.65	0.85	0.71
	Bunch mass	0.94	0.96	0.95	0.86
	Number of hands	0.93	0.91	0.91	0.81
	Number of fruits	0.89	0.97	0.94	0.74
Fruit filling	Fruit length	0.96	0.97	0.98	0.84
	Fruit circumference	0.97	0.94	0.96	0.87
	Fruit diameter	0.97	0.93	0.99	0.89
	Pulp diameter	0.97	0.93	0.92	0.90

files/picard-tools/1.100/) are used prior to SNP calling, normalization of sequence reads is possible by marking and removing duplicate reads. This allows regions with low reads coverage, but carrying SNP of interest to be included in the genotype data. Picard tools also allow merging of aligned sample reads by addition of read groups, which help in separating genotypes after SNP calling.

What is the Best Genomic Prediction Model for Each Trait or Group of Traits?

Different genomic prediction models were compared in this work in terms of their predictive ability, or accuracy for different traits as noted in Table 1 and Supplemental Table S2. We compared the performance of models that account for additive genetic effects and those that account for non-additive genetic effects. A good performance of models that account for additive genetic effects suggested that a large proportion of phenotypic variation observed in the training population was due to additive genetic effects. Indeed, traits with high narrow sense heritability (h^2) had higher predictive values. A similar observation was made by Luan et al. (2009). They reported a strong relationship between prediction accuracy and trait heritability in Norwegian red cattle. Differences in h^2 between GS1 and GS2, and H^2 were attributed to bias in residual error variance. Using phenotypic means reduces error variance leading to over estimation of h^2 as compared to replicated phenotypic data used in estimating H^2. Usually, proper estimation of heritability requires balanced phenotypic data (Piepho and Möhring, 2007). However, it is hard to get balanced data for bananas because growth is not synchronized between plants as well as data collection, which causes high variation between genotypes and replicates in the same environment. Generally, H^2 is specific to a given population at a given location and period, but depending on the genetic architecture of the trait correlations might be observed across populations. For example, our H^2 results are comparable to those summarized by James et al. (2012) from various publications on bananas and plantains.

Additive genetic effect models BayesB and BayesC performed better than or equally well as other models. These models perform both shrinkage and variable selection on markers to include in the model (Desta and Ortiz, 2014). The prior probability of a marker having a non-null effect (π) was set at 0.05 in both models because it gave the highest predictive ability values as compared to higher prior settings. It is likely that the same markers were selected and included in both models thus yielding closely related results.

Our results agree with other studies, which indicate that models that perform specific shrinkage and variable selection give better predictive ability values. For example, Crossa et al. (2010) showed that a BL model that shares some characteristics with BayesB outperformed BLUP, which assumes equal variance for each marker. Similarly, Clark et al. (2011) reported the superiority of

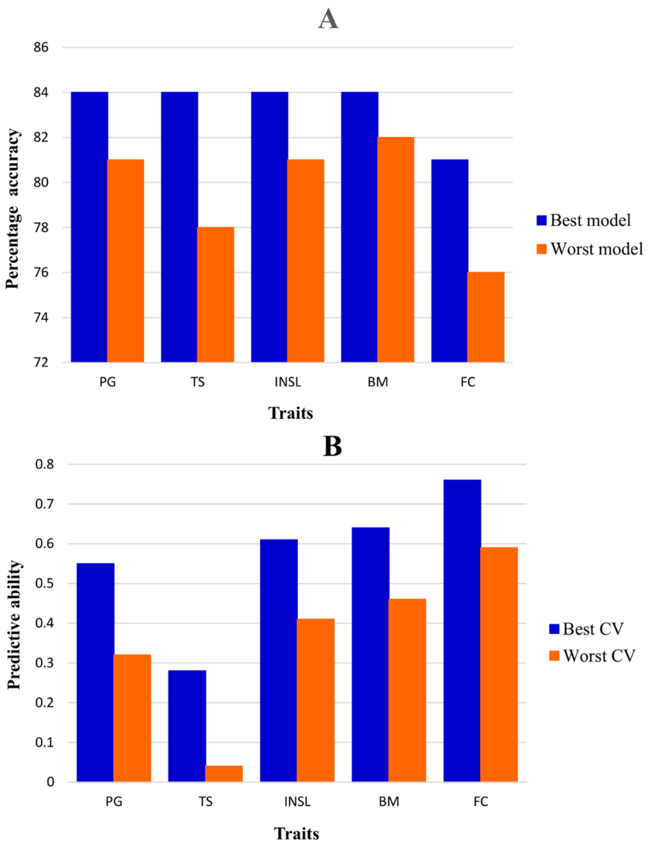


Fig. 5. Accuracy of genomic prediction in the training population. (A) Percentage of genotypes selected by both GEBV and phenotypic data within the first best ranked 100 genotypes. (B) Correlations of the best and worst BayesB models used to generate GEBV. Where, PG is plant girth at 100 cm from soil surface, TS is total number of suckers, INSL is index of non-spotted leaves, BM is bunch mass, FC is fruit circumference and CV is cross validation.

BayesB model over genomic BLUP. They argued that the superiority was highly dependent on the presence of large QTL effects. In relation to this argument, it is likely that even in banana, fruit filling traits could be controlled by large effect QTL that were selected by BayesB model in all cross-validations. However, this remains to be proved by QTL mapping and genome-wide association studies that are out of the scope of this study. Tagging of loci controlling fruit filling with DNA markers and selecting for favorable alleles should also be considered. Fruit filling is a bunch mass component that reflects the sink capacity of a fruit bunch. It was treated separately from other bunch mass components to better describe the proportion of edible part of the fruit. Variation in performance of models that perform shrinkage and variable selection has also been reported. For example, in loblolly pine, BayesCπ (Habier et al., 2011) and BayesA had better prediction of fusiform rust disease-resistance traits than BL (Resende et al., 2012)

The predictive ability of all models varied across traits. Similar predictive values for traits within the same category confirmed the findings of Nyine et al. (2017) who reported a high correlation between these traits and recommended that only traits easier to phenotype should be considered for genomic predictions. The difference in model performance between trait categories suggests that variation in trait architecture, number of QTL controlling the trait and linkage disequilibrium between markers and QTL influence the performance of the models (Clark et al., 2011).

The RKHS_PM model, which accounts for non-additive genetic effects yielded mixed prediction results. While some traits had a slight increase in prediction, a majority showed loss in predictive ability (Table 1; Supplemental Table S2). Previous studies (Crossa et al., 2010) indicated minor improvement in trait prediction in wheat and maize when marker and pedigree information were included in the model. However, Pérez-Rodríguez et al. (2017) reported better prediction with RKHS_P for wheat lines in international environments. The contradictions could be attributed to the training population structure. Our training population consisted of 77 subfamilies (cross combinations) of varying sizes with diverse pedigree background (Nyine et al., 2017). This suggests that when the population consists of many subfamilies, the relationship by pedigree becomes less important. This is reflected by the poor performance of RKHS_P model, which gave the least prediction accuracy for all traits (Supplemental Table S2). A similar trend was observed by Beaulieu et al. (2014). Hence, the estimates of allele distribution within such a population is better performed with marker data, while addition of pedigree information distorts the relationship between the genotypes. Zhong et al. (2009) also highlighted that knowledge of pedigree is less informative in populations where the average genetic relationship is low and homogeneity is high.

What is the Effect of G × E on Model Predictions?

We used a very conservative approach in determining the best genomic prediction model by carrying out across field (environment) cross validations. The purpose was to understand the effect of genotype by field management (G × E) interaction on the model performance. Nyine et al. (2017) performed analysis of variance on the same population and reported a variation in G × E interaction across different traits. However, type B genetic correlations (Table 5) were high for traits related to fruit bunch and fruit filling, which explains why they had high predictive ability values across all cross-validation strategies. When Burdon (1977) proposed the use of type B genetic correlation, he noted that in the analysis of variance, any genetic expression variation between environments can lead to statistical interaction that is not necessarily a true interaction characterized by a change in ranking of genotypes between different environments. The results showed that models fitted with GS1 phenotype data underpredicted the phenotypic expression of genotypes in GS2 while the models fitted with GS2 phenotype data overpredicted genotypes in GS1 (Fig. 4). However, the trend of prediction did not change (Table 1). A similar approach was used by Ly et al. (2013), who observed that across environment cross validations resulted into lower prediction accuracies. However, our prediction values were substantially higher as compared to those reported in other crops.

Trait overprediction in GS1 with models fitted with GS2 data and vice versa indicated a variation in genotype response to environment that influenced the training population trait mean, estimated marker effect and the predictive ability of the genomic prediction models (Crossa et al., 2016). The high correlation between the two fields shows that it is possible to use phenotype data from any of the field management conditions to predict genotypes that have the potential to perform well in other field management conditions. However, the predicted and the actual observed phenotype may differ for a single genotype. For example, plants that had poor fruit filling characteristics under low input field management did not fill under high input field management, as well. However, for genotypes that fill their fruits, there was an increase in fruit size depending on the amount of available nutrients and soil moisture in the field. A similar trend was reported in maize flowering where QTL were consistent across environments and less affected by environment interaction (Buckler et al., 2009). This means that genomic prediction models could be used in 'negative selection' to discriminate the poor fruit filling hybrids from those with potential of fruit filling at an early stage.

In banana breeding, most triploid hybrids are sterile. The application of genomic prediction in its strict sense of selecting best parents for further crossing (Meuwissen et al., 2001; Gorddard and Hayes, 2007) may not be realistic, unless the focus is only on diploid and tetraploid improvement. Since the prediction models give both GEBV and predicted phenotype (Pérez and de los Campos, 2014), these two parameters can be used to eliminate triploid hybrids that are likely to be of no value. Crossa et al. (2014) also proposed that another application of genomic prediction was to predict the genetic values of individuals for potential release as cultivars. Therefore, if the prediction accuracy remains high during the breeding phase, then breeders could save time, space, and money by excluding 90% of hybrids from phenotyping (Fig. 2). To achieve this, breeders have to set priority order of traits, which could serve as the 'selection index' for promising candidate cultivars (i.e., within triploids hybrids) and future parental clones (within diploid and tetraploid hybrids). Also, family based selection should be done to reduce future inbreeding and maximize genetic diversity to ensure increase in genetic gain (Jannink et al., 2010).

Although crop cycle was shown to influence variation in fruit filling, fruit bunch and plant stature, and no effect on black leaf streak resistance traits (Nyine et al., 2017), the predictions within and across crop cycle 1 and crop cycle 2 did not vary much for fruit filling and fruit bunch traits. This is because fruit filling and fruit bunch traits increase in crop cycle 2 relative to crop cycle 1 (Tushemereirwe et al., 2015). However, for black leaf streak resistance, resistant hybrids remain resistant across crop cycles and field management. Variation may be observed among susceptible hybrids depending on the spore density in the field (Tushemereirwe, 1996). Disease expression also depends on vigor of the plant due to available nutrients, seasonal changes and relative humidity in the field (Tushemereirwe, 1996). This probably explains the variation observed in the prediction within and across crop cycle for the index of non-spotted leaves.

In bananas, suckering behavior traits had the lowest prediction accuracy. One possible explanation is the low heritability and poor representation of markers linked to the QTL controlling these traits. Second, scoring total number of suckers at crop cycle 1 from a trial established with suckers, seems to result in biased phenotype data. Two types of suckers are used as planting materials, the sword and maiden suckers. Most maiden suckers are much closer to flowering than sword suckers (Ortiz and Vuylsteke, 1994) and tend to direct most of resources toward the initiation of the inflorescence, and less to the development of lateral buds (future suckers). On the contrary, sword suckers commit most of their resources to lateral bud development. Hence, when a field is established with suckers, the variation in physiological age of suckers likely impacts sucker emergence that causes bias in total number of suckers produced by a genotype at first crop cycle.

When environment averaged models were used to predict the performance of genotypes in a particular environment, the predictions were high (0.75 for fruit filling traits) despite the lower number of SNP markers (Table 3). This indicated that incorporation of data from many environments could make the models more robust (Burgueño et al., 2012). As discussed by Burgueño et al. (2012), breeders either evaluate new breeding lines so that they can select the best to advance, or evaluate the performance stability of new, or old lines in a new environment. In each of these cases, the model should be robust enough to give accurate predictions in the respective environments (Pérez-Rodríguez et al., 2017). Hence, using data from multi-environment trials and crop cycles to fit the model has the advantage of incorporating information due to genetic relationship and the interaction between genotype and environment (Crossa et al., 2014).

Traits that are stable across environments are much easier to predict using data from one environment. However, if there is a proportional change (collinearity) in the trait expression within an environment across genotypes, then selection based on predictions is likely to be efficient (Burgueño et al., 2012). Plant environments vary and may refer to geographical locations with different weather and climatic conditions, difference in seasons within a same location and difference in soil conditions based on the different agronomic practices used. As perennial plants, bananas suffer the consequences of nutrient deficiency and soil moisture variation across seasons and locations depending on field management practices (Ndabamenye et al., 2012; Taulya, 2015). These factors influence phenotypic expression of traits and are likely to affect the predictive ability of prediction models. Although we considered field management and crop cycle as the major environment co-variables, phenotyping of the current training population in a different geographical location is ongoing. Once the data are available, they will be used to update the models to the benefit of the breeding program.

Bi-Allelic SNP vs. Allele Dosage SNP

Whereas many factors have been reported to influence the accuracy of genomic predictions (Crossa et al., 2014), our results showed that allele dosage was another important factor to consider when conducting predictions in multi-ploidy populations (Supplemental Table S4). The loss in predictive ability of the models fitted with AD-SNP relative to those fitted with BA-SNP could be attributed to variation in minor allele frequency across loci, a key factor for determining SNP effects on the traits and the allopolyploid nature of the training population. The negative correlations observed from across ploidy cross validation indicated a weak relationship between the training and testing sets (Crossa et al., 2016). Clearly, not all traits were affected equally by allele dosage (Supplemental Table S4). The effect of allele dosage becomes more important as the ploidy level increases. This suggests that additive genetic effects vary across traits. It is likely that the effect of deleterious recessive alleles is masked by the dominant alleles and the more copies of masking alleles the better the effect (Gu et al., 2003). However, for traits controlled by exclusively recessive alleles, the effect of allele dosage may be different. In cassava, a large proportion of deleterious alleles arising from mutations have not been eliminated by breeding due to limited recombination, but the maintenance of cassava yield through breeding has been attributed to masking of most damaging mutations (Ramu et al., 2017).

Predictions within multi-family population was shown by Heffner et al. (2011) to be accurate and cost effective. It is likely that genomic prediction models trained only on diploid segregating populations would be less efficient in prediction of traits among triploid banana hybrids, yet promising candidate cultivars are selected in this ploidy level. Second, allele dosage could be accounted for in the marker data especially when predicting fruit filling in tetraploids although use of models that assume diploid state of all genotypes still performed better in many cross-validation strategies.

To ensure that good hybrids are not left out, selection based on GEBV should be done with prior knowledge of ploidy level in multi-ploidy populations. Bunch mass and general phenology in bananas tend to increase with increase in ploidy level although in banana hybrids, the trend is not always uniform due to positive and negative heterosis (Tenkouano, 2000). Since banana breeding involves crossing parents of different ploidy levels, prediction of hybrid performance based on parental phenotype data is less accurate due to heterosis. That is why the parents' model prediction accuracies were low. Although we did not measure heterosis in this study, the results of selection differential and response to selection reported by Nyine et al. (2017) show that it exists in this training population.

When the progeny's model was used to predict the parental traits, the predictions were appreciably high (Supplemental Table S3). This indicated that a large size of the training set relative to the testing set improves prediction (Jannink et al., 2010; Clark et al., 2011; Crossa et al., 2014). The lesson learned is that in bananas, when the training population is made up of many diverse hybrids, the segregation of parental alleles is observed. Most of the additive genetic effects, heterosis, dominance, and epistasis that control the phenotype are captured in the model when all these phenotypic variants are available (Lorenz et al., 2011). These results suggest that for plant species with small effective breeding population sizes like banana that show heterosis, increasing the number of progeny from several parental crosses in the training population could improve the predictive ability of the models for future hybrids as compared to using only parental clones.

The Accuracy of Genomic Prediction

The prediction accuracy within the training population based on GEBV was above 75% even with models that had low predictive abilities. The accuracy of genomic prediction model is determined by the correlation between GEBV and the observed phenotype, or the correlation between predicted phenotype and observed phenotype (Jannink et al., 2010; Lorenz et al., 2011). This shows the proportion of genetic variance explained by marker data. It is therefore not surprising that even with low correlations, the accuracy of prediction can be high. Beaulieu et al. (2014) reported that with GEBV accuracies between 0.33 and 0.44, they were able to achieve 90% of traditionally estimated breeding values during validation. Similarly, Heffner et al. (2011) reported a 95% prediction accuracy of genomic prediction compared to phenotypic selection in a multi-family wheat population even when the predictive values ranged from 0.22 to 0.76.

The true accuracy is estimated at the validation stage using the validation population. It depends on the size of the training population, heritability of the trait and the estimated number of effects (Lorenz et al., 2011). Sometimes, it is not possible to explain all the genetic variance due to missing marker data, or failure to capture other QTL affecting the trait. This is further confounded by uncontrolled environmental variable (Buckler et al., 2009; Burgueño et al., 2012). That is why genomic selection is considered less accurate than phenotypic selection but its power lies in increased selection intensity within a much shorter time hence increasing the genetic gain per unit time and cost (Desta and Ortiz, 2014; Lorenz et al., 2011). Our results suggest that even with low predictive values, the accuracy of prediction within the training population was high. It remains to be verified at the validation stage if the accuracy remains high. Given the long selection cycle observed in banana as depicted in Fig. 2, prediction accuracies above 70% could result in accelerated selection efficiency at reduced cost as compared to phenotypic selection.

Conclusion and Practical Implications

Polyploid breeding programs ought to use genomic prediction models that have been fitted with data from genotypes of all ploidy levels otherwise genomic selection will face similar limitations as other MAS techniques, which focus on bi-parental populations for QTL and marker discovery. Fruit filling and fruit bunch traits had the highest predictive ability hence, could be targeted for early selection of hybrids. Accounting for allele dosage in SNP markers (AD-SNP) reduced predictive ability of the models relative to traditional bi-allelic SNP (BA-SNP). Unlike autopolyploid, allele dosage seems to have less influence on genomic prediction in allopolyploid populations. However, if ploidy specific prediction models are required, the R script reported could be used to generate AD-SNP. The heritability of traits estimated in this training population were high and positively correlated with the predictive ability. The results demonstrate that genomic prediction in multi-ploidy population is possible and the prediction accuracy can be improved by using models based on data from many different environments.

To generate prediction models for each ploidy level is expensive in the initial stages of genomic selection, but as the training population keeps growing it becomes possible. To minimize costs, the current models based on multi-ploidy population should be validated and used with the following recommendations: (i) unlike other breeding programs where genomic prediction is used entirely for prediction of best parents for further crossing, in banana, selection among triploids should aim at identifying promising candidate cultivars because a majority of them are sterile and breeding clones should be selected from diploids and tetraploids, (ii) ‘selection index’ is required for efficient selection of new hybrids, i.e., the priority order of traits should be set for promising cultivars and breeding clones, (iii) family-based (cross combination) selection should be considered to avoid reducing genetic diversity, (iv) the lowest GEBV should be targeted for plant height, or else a ratio of plant height to plant girth at 100 cm from soil surface should be used. In the light of genomic selection, a potential area of research would be to investigate the level of fertility in triploid banana hybrids so that they are also selected as parents. This will allow ‘progressive’ breeding to be practiced in banana for faster genetic gain since some traits are already fixed in the triploids.

Supplemental Material

Supplemental Table S1: List of banana genotypes used in genomic predictions.
Supplemental Table S2: Comparison of average correlation for five-fold cross validations between the predicted and observed phenotypes across all models fitted with data from either low input (GS1) or high input (GS2) fields and 10807 bi-allelic SNP markers.
Supplemental Table S3: Comparison of predictive ability of BayesB model fitted with parents’ data and progeny’s data using bi-allelic and allele dosage SNP markers.
Supplemental Table S4: Effect of ploidy level and allele dosage on the predictive ability of BayesB model fitted with environment averaged phenotype data.
Supplemental Fig. S1: Workflow used to analyze the genotyping by sequencing (GBS) reads to generate SNP marker data used in genomic predictions.

Conflict of Interest Disclosure

The authors declare that there is no conflict of interest.

Funding

This work was supported by donors through their contributions to the CGIAR Fund (http://www.cgiar.org/our-funders/), and in particular to the CGIAR Research Program, Roots, Tubers and Bananas. Additional funding was obtained from Bill and Melinda Gates foundation, investment ID: OPP1093845, under the project: Improvement of banana for smallholder farmers in the great lakes region of Africa. Part of this work was supported by the Czech Ministry of Education Youth and Sports (award LO1204 from the National Program of Sustainability I and INGO II award LG15017) while the computing was supported by the Czech National Grid Infrastructure MetaCentrum (grant No. LM2010005 under the program, Projects of Large Infrastructure for Research, Development, and Innovations).

Acknowledgments

We extend our sincere appreciation to the Institute of Genomic Diversity, Cornell for their sequencing services. Special thanks to Ismael Y. Rabbi (IITA, Nigeria) for coordinating the genotyping work, Rodomiro Ortiz (SLU, Sweden) for his useful comments during the experimental process and write-up of this manuscript, Manpreet S. Katari (New York University, USA) for his advices in bioinformatics, Harriet Nabatanzi, Catherine Nasuuna, and Catherine Nakayenga (field staff, IITA, Uganda) for their dedicated effort in phenotypic data collection.

References

Asíns, M.J. 2002. Review: Present and future of quantitative trait locus analysis in plant breeding. Plant Breed. 121:281–291. doi:10.1046/j.1439-0523.2002.730285.x

Bassi, F.M., A.R. Bentley, G. Charmet, R. Ortiz, and J. Crossa. 2016. Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). Plant Sci. 242:23–36. doi:10.1016/j.plantsci.2015.08.021

Beaulieu, J., T. Doerksen, S. Clément, J. MacKay, and J. Bousquet. 2014. Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. Heredity 113:343–352. doi:10.1038/hdy.2014.36

Blischak, P.D., L.S. Kubatko, and A.D. Wolfe. 2016. Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. Mol. Ecol. Resour. 16(3):742–754. doi:10.1111/1755-0998.12493

Buckler, E.S., J.B. Holland, P.J. Bradbury, C.B. Acharya, P.J. Brown, C. Browne, et al. 2009. The genetic architecture of maize flowering time. Science 325:714–718. doi:10.1126/science.1174276

Burdon, R.D. 1977. Genetic correlation as a concept for studying genotype-environmental interaction in forest tree breeding. Silvae Genet. 26:168–175.

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa. 2012. Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci. 52:707–719. doi:10.2135/cropsci2011.06.0299

Clark, S.A., J.M. Hickey, and J.H.J. van der Werf. 2011. Different models of genetic variation and their effect on genomic evaluation. Genet. Sel. Evol. 43:18. doi:10.1186/1297-9686-43-18

Clevenger, J., C. Chavarro, S.A. Pearl, P. Ozias-Akins, and S.A. Jackson. 2015. Single nucleotide polymorphism identification in polyploids: A review, example and recommendations. Mol. Plant 8:831–846. doi:10.1016/j.molp.2015.02.002

Collard, B.C.Y., and D.J. Mackill. 2008. Marker-assisted selection: An approach for precision plant breeding in the twenty-first century. Philos. Trans. R. Soc. B 363:557–572. doi:10.1098/rstb.2007.2170

Craenen, K. 1998. Black Sigatoka disease of banana and plantain: A reference manual, International Institute of Tropical Agriculture, Nigeria, Ibadan. p. 41–45.

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño, L. Araus, et al. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724. doi:10.1534/genetics.110.118521

Crossa, J., D. Jarquín, J. Franco, P. Pérez-Rodríguez, J. Burgueño, C. Saint-Pierre, et al. 2016. Genomic prediction of gene bank wheat landraces. G3 (Bethesda) 6:1819–1834. doi:10.1534/g3.116.029637

Crossa, J., P. Pérez, J. Hickey, J. Burgueño, L. Ornella, J. Cerón-Rojas, et al. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112:48–60. doi:10.1038/hdy.2013.16

Desta, Z.A., and R. Ortiz. 2014. Review: Genomic selection: Genome-wide prediction in plant improvement. Trends Plant Sci. 19(9):592–601. doi:10.1016/j.tplants.2014.05.006

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, et al. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6(5):e19379. doi:10.1371/journal.pone.0019379

Gezan, S.A., L.F. Osorio, S. Verma, and V.M. Whitaker. 2017. An experimental validation of genomic selection in octoploid strawberry. Hortic. Res. 4. doi:10.1038/hortres.2016.70

Goiffon, M., A. Kusmec, L. Wang, G. Hu, and P.S. Schnable. 2017. Improving response in genomic selection with a population-based selection strategy: Optimal population value selection. Genetics 206:1675–1682. doi:10.1534/genetics.116.197103

Gorddard, M.E., and B.J. Hayes. 2007. Review: Genomic selection. J. Anim. Breed. Genet. 124:323–330.

Gu, Z., L.M. Steinmetz, X. Gu, C. Scharfe, R.W. Davis, and W.-H. Li. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature 421:63–66. doi:10.1038/nature01198

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186. doi:10.1186/1471-2105-12-186

Heffner, E.L., J.-L. Jannink, and M.E. Sorells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4:65–75. doi:10.3835/plantgenome2010.12.0029

James, A., R. Ortiz, and R. Miller. 2012. Map-based cloning in *Musa* spp In: Pillay, M., G. Ude, and C. Kole, editors, Genetics, genomics, and breeding of bananas. CRC Press, Boca Raton, FL. p. 124–155.

Jannink, J.-L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: From theory to practice. Brief. Funct. Genomics 9(2):166–177. doi:10.1093/bfgp/elq001

Kruijer, W., M.P. Boer, M. Malosetti, P.J. Flood, B. Engel, R. Kooke, et al. 2015. Marker-based estimation of heritability in immortal populations. Genetics 199(2):379–398. doi:10.1534/genetics.114.167916

Kumar, S., D. Chagné, M.C.A.M. Bink, R.K. Volz, C. Whitworth, and C. Carlisle. 2012. Genomic selection for fruit quality traits in apple (Malus×domestica Borkh.). PLoS One 7(5):e36674. doi:10.1371/journal.pone.0036674

Legarra, A., C. Robert-Granié, E. Manfredi, and J.-M. Elsen. 2008. Performance of genomic selection in mice. Genetics 180:611–618. doi:10.1534/genetics.108.088575

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, et al. 2011. Genomic selection in plant breeding: Knowledge and prospects. In: Advances of Agronomy, 1st ed. 110:77–123. doi:10.1016/B978-0-12-385531-2.00002-5

Luan, T., J.A. Woolliams, S. Lien, M. Kent, M. Svendsen, and T.H. Meuwissen. 2009. The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. Genetics 183(3):1119–1126. doi:10.1534/genetics.109.107391

Ly, D., M. Hamblin, I. Rabbi, G. Melaku, M. Bakare, H.G. Gauch, Jr., et al. 2013. Relatedness and genotype × environment interaction affect prediction accuracies in genomic selection: A study in cassava. Crop Sci. 53:13212–1325. doi:10.2135/cropsci2012.11.0653

Martin, G., F.C. Baurens, G. Droc, M. Rouard, A. Cenci, A. Kilian, et al. 2016. Improvement of the banana "*Musa acuminata*" reference sequence using NGS data and semi-automated bioinformatics methods. BMC Genomics 17(1):243. doi:10.1186/s12864-016-2579-4

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Momen, M., A.A. Mehrgardi, A. Sheikhy, A. Esmailizaheh, M.A. Fozi, A. Kranis, et al. 2017. A predictive assessment of genetic correlations between traits in chickens using markers. Genet. Sel. Evol. 49:16. doi:10.1186/s12711-017-0290-9

Nakaya, A., and S.N. Isobe. 2012. Will genomic selection be a practical method for plant breeding? Ann. Bot. (Lond.) 110:1303–1316. doi:10.1093/aob/mcs109

Ndabamenye, T., P. Van Asten, N. Vanhoudt, G. Blomme, R. Swennen, J.G. Annandale, et al. 2012. Ecological characteristics influence farmer selection of on-farm plant density and bunch mass of low input East African Highland banana (*Musa* spp.) cropping systems. Field Crops Res. 135:126–136. doi:10.1016/j.fcr.2012.06.018

Noumbissié, G.B., M. Chabannes, F. Bakry, S. Ricci, C. Cardi, J.-C. Njembele, et al. 2016. Chromosome segregation in an allotetraploid banana hybrid (AAAB) suggests a translocation between the A and B genomes and results in eBSV-free offsprings. Mol. Breed. 36(4):38. doi:10.1007/s11032-016-0459-x

Nyine, M., B. Uwimana, R. Swennen, M. Batte, A. Brown, P. Christelová, et al. 2017. Trait variation and genetic diversity in a banana genomic selection training population. PLoS One 12(6):e0178734. doi:10.1371/journal.pone.0178734

Onogi, A., M. Watanabe, T. Mochizuki, T. Hayashi, H. Nakagawa, T. Hasegawa, et al. 2016. Toward integration of genomic selection with crop modelling: The development of an integrated approach to predicting rice heading dates. Theor. Appl. Genet. 129:805–817. doi:10.1007/s00122-016-2667-5

Ortiz, R., and D.R. Vuylsteke. 1994. Genetics of apical dominance in plantain (*Musa* spp., AAB group) and improvement of suckering behavior. J. Am. Soc. Hortic. Sci. 119(5):1050–1053.

Park, T., and G. Casella. 2008. The Bayesian LASSO. J. Am. Stat. Assoc. 103:681–686. doi:10.1198/016214508000000337

Pérez, P., and G. de los Campos. 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495. doi:10.1534/genetics.114.164442

Pérez-Rodríguez, P.J., J. Crossa, J. Rutkoski, R. Poland, A. Singh, E. Legarra, et al. 2017. Single-step genomic and pedigree genotype × environment interaction models for predicting wheat lines in international environments. Plant Genome 10(2):1–15. doi:10.3835/plantgenome2016.09.0089

Piepho, H.-P., and J. Möhring. 2007. Computing heritability and selection response from unbalanced plant breeding trials. Genetics 177:1881–1888. doi:10.1534/genetics.107.074229

Ramu, P., W. Esuma, R. Kawuki, I.Y. Rabbi, C. Egesi, J.V. Bredeson, et al. 2017. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. Nat. Genet. 49:959–963. doi:10.1038/ng.3845

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (accessed 6 Feb. 2018).

Resende, M.F.R., Jr., P. Muñoz, M.D.V. Resende, D.J. Garrick, R.L. Fernando, J.M. Davis, et al. 2012. Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics 190:1503–1510. doi:10.1534/genetics.111.137026

Robinson, G.K. 1991. That BLUP is a good thing: The estimation of random effects. Stat. Sci. 6(1):15–32. doi:10.1214/ss/1177011926

Rowe, P.R. 1990. Breeding bananas and plantains for resistance to fusarial wilt: The track record. In: R.C. Ploetz, editor, Fusarium wilt of bananas. APS, St. Paul, MN. p. 115–119.

Simmonds, N.W. 1986. Bananas, *Musa* cvs. In: N.W. Simmonds, editor, Breeding for durable resistance in perennial crops. FAO Technical Papers 70. Food and Agriculture Organization, Rome. p. 17–24.

Ssebuliba, R., D. Talengera, D. Makumbi, P. Namanya, A. Tenkouano, W. Tushemereirwe, et al. 2006. Reproductive efficiency and breeding potential of East African highland (*Musa* AAA-EA) bananas. Field Crops Res. 95:250–255. doi:10.1016/j.fcr.2005.03.004

Taulya, G. 2015. Ky'osimba Onaanya: Understanding productivity of East African highland banana. Ph.D. thesis, Wageningen University.

Tenkouano, A. 2000. Current issues and future directions for *Musa* genetic improvement research at the International Institute of Tropical Agriculture. Advancing banana and plantain R & D in Asia and the Pacific. INIBAP 10:11–23.

Tibshirani, R. 1996. Regression shrinkage and selection via the LASSO. J. R. Stat. Soc. Series B Stat. Methodol. 58:267–288.

Tushemereirwe, W. 1996. Factors influencing the expression of leaf spot diseases of highland bananas in Uganda. Ph. D. thesis. University of Reading, U.K.

Tushemereirwe, W., M. Batte, M. Nyine, R. Tumuhimbise, A. Barekye, T. Ssali, et al. 2015. Performance of NARITA hybrids in the preliminary yield trial for three cycles in Uganda. IITA, NARO, Uganda.

Umber, M., J.P. Pichaut, B. Farinas, N. Laboureau, B. Janzac, K. Plaisir-Pineau, et al. 2016. Marker-assisted breeding of *Musa balbisiana* genitors devoid of infectious endogenous banana streak virus sequences. Mol. Breed. 36(6):74. doi:10.1007/s11032-016-0493-8

Wang, W., Y. Hu, D. Sun, C. Staehelin, D. Xin, and J. Xie. 2012b. Identification and evaluation of two diagnostic markers linked to Fusarium wilt resistance (race 4) in banana (*Musa* spp.). Mol. Biol. Rep. 39:451–459. doi:10.1007/s11033-011-0758-6

Weng, Z., A. Wolc, X. Shen, R.L. Fernando, J.C.M. Dekkers, J. Arango, et al. 2016. Effects of number of training generations on genomic prediction for various traits in a layer chicken population. Genet. Sel. Evol. 48:22. doi:10.1186/s12711-016-0198-9

Wong, C.K., and R. Bernardo. 2008. Genome-wide selection in oil palm: Increasing selection gain per unit time and cost with small populations. Theor. Appl. Genet. 116:815–824. doi:10.1007/s00122-008-0715-5

Würschum, T., J.C. Reif, T. Kraft, G. Janssen, and Y. Zhao. 2013. Genomic selection in sugar beet breeding populations. BMC Genet. 14:85. doi:10.1186/1471-2156-14-85

Zhang, Z., J. Liu, X. Ding, P. Bijma, D.-J. de Koning, and Q. Zhang. 2010. Best linear unbiased prediction of genomic breeding values using a trait-specific marker derived relationship matrix. PLoS One 5(9):e12648. doi:10.1371/journal.pone.0012648

Zhong, S., J.C.M. Dekkers, R.L. Fernando, and J.-L. Jannink. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. Genetics 182:355–364. doi:10.1534/genetics.108.098277

Genomic Prediction in a Multiploid Crop: Genotype by Environment Interaction and Allele Dosage Effects on Predictive Ability in Banana

Moses Nyine, Brigitte Uwimana, Nicolas Blavet, Eva Hřibová, Helena Vanrespaille, Michael Batte, Violet Akech, Allan Brown, Jim Lorenzen,[&] Rony Swennen, Jaroslav Doležel[*]

M. Nyine, Faculty of Science, Palacký University, 77146, Olomouc, Czech Republic; M. Nyine, B. Uwimana, M. Batte, V. Akech, J. Lorenzen, International Institute of Tropical Agriculture, 7878, Kampala, Uganda; M. Nyine, N. Blavet, E. Hřibová, J. Doležel, Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, CZ-78371, Olomouc, Czech Republic; H. Vanrespaille, R. Swennen, Laboratory of Tropical Crop Improvement, Division of Crop Biotechnics, Katholieke Universiteit 2455, 3001 Leuven, Belgium; R. Swennen, A. Brown, International Institute of Tropical Agriculture, 10, Arusha, Tanzania; [&]Current address: Bill & Melinda Gates Foundation, 23350, Seattle, USA. *Corresponding author (dolezel@ueb.cas.cz).

Supplemental Table S1: List of banana genotypes used in genomic predictions (Nyine et al. 2017)

S/No	Genotype name	Female parent	Male parent	Ploidy
1	Enzirabahima			Triploid
2	Kabucuragye			Triploid
3	Tereza			Triploid
4	Enyeru			Triploid
5	Nakayonga			Triploid
6	Namwezi			Triploid
7	Entukura			Triploid
8	Nakasabira			Triploid
9	Nakawere			Triploid
10	Nante			Triploid
11	Kazirakwe			Triploid
12	Nfuuka			Triploid
13	Calcutta 4			Diploid
14	1201K-1	Nakawere	Calcutta 4	Tetraploid
15	917K-2	Enzirabahima	Calcutta 4	Tetraploid
16	660K-1	Enzirabahima	Calcutta 4	Tetraploid
17	1438K-1	Entukura	Calcutta 4	Tetraploid
18	222K-1	Nfuuka	Calcutta 4	Tetraploid
19	376K-7	Nante	Calcutta 4	Tetraploid
20	365K-1	Kabucuragye	Calcutta 4	Tetraploid
21	401K-1	Entukura	Calcutta 4	Tetraploid
22	2180K-6			Diploid
23	8075-7	SH3362	Calcutta 4	Diploid
24	7197-2	SH3362	Long Tavoy	Diploid
25	SH3142	SH1734	Pisang Jari Buaya	Diploid
26	SH3362	SH3217	SH3142	Diploid
27	SH3217	SH2095	SH2766	Diploid
28	5610S-1	Kabucuragye	7197-2	Diploid
29	9128-3	Tjau lagada	Pisang lilin	Diploid
30	1968-2	Who-gu	Calcutta 4	Triploid
31	861S-1	Namwezi	Calcutta 4	Diploid
32	cv. Rose			Diploid

33	Pisang Lilin			Diploid
34	Kokopo			Diploid
35	Long Tavoy			Diploid
36	*M. a. M. a. malaccensis 250*			Diploid
37	28165S-1	1201K-1	1968-2	Triploid
38	25583S-2	1201K-1	5610S-1	Triploid
39	26660S-1	1201K-1	5610S-1	Triploid
40	28434S-9	1201K-1	5610S-1	Triploid
41	17503S-3	1201K-1	7197-2	Triploid
42	16242S-1	1201K-1	8075-7	Triploid
43	12479S-1	1201K-1	9128-3	Triploid
44	12479S-13	1201K-1	9128-3	Triploid
45	26317S-1	1201K-1	9128-3	Triploid
46	27262S-1	1201K-1	9128-3	Triploid
47	27262S-3	1201K-1	9128-3	Triploid
48	27770S-20	1201K-1	cv. Rose	Triploid
49	27770S-4	1201K-1	cv. Rose	Triploid
50	27935S-1	1201K-1	cv. Rose	Triploid
51	27960S-1	1201K-1	cv. Rose	Triploid
52	28036S-11	1201K-1	cv. Rose	Triploid
53	28036S-2	1201K-1	cv. Rose	Triploid
54	28164S-3	1201K-1	cv. Rose	Triploid
55	28246S-4	1201K-1	cv. Rose	Triploid
56	28246S-7	1201K-1	cv. Rose	Triploid
57	27935S-7	1201K-1	cv. Rose	Triploid
58	26363S-1	1201K-1	Kokopo	Triploid
59	26075S-6	1201K-1	Long Tavoy	Triploid
60	26075S-7	1201K-1	Long Tavoy	Triploid
61	26075S-8	1201K-1	Long Tavoy	Triploid
62	27346S-2	1201K-1	*M. a. malaccensis 250*	Triploid
63	27346S-4	1201K-1	*M. a. malaccensis 250*	Triploid
64	27437S-1	1201K-1	*M. a. malaccensis 250*	Triploid
65	27579S-1	1201K-1	*M. a. malaccensis 250*	Triploid
66	27579S-3	1201K-1	*M. a. malaccensis 250*	Triploid

67	28030S-2	1201K-1	*M. a. malaccensis 250*	Triploid
68	28030S-6	1201K-1	*M. a. malaccensis 250*	Triploid
69	28071S-1	1201K-1	*M. a. malaccensis 250*	Triploid
70	28465S-2	1201K-1	*M. a. malaccensis 250*	Triploid
71	28465S-21	1201K-1	*M. a. malaccensis 250*	Triploid
72	28479S-2	1201K-1	*M. a. malaccensis 250*	Triploid
73	26337S-22A	1201K-1	SH3217	Triploid
74	26337S-40	1201K-1	SH3217	Triploid
75	26840S-7	1201K-1	SH3362	Diploid
76	26315S-1	1201K-1	SH3142	Triploid
77	12419S-13	1201K-1	SH3217	Triploid
78	26337S-11A	1201K-1	SH3217	Triploid
79	26337S-2	1201K-1	SH3217	Triploid
80	26337S-34	1201K-1	SH3217	Triploid
81	26337S-37	1201K-1	SH3217	Triploid
82	26337S-39	1201K-1	SH3217	Triploid
83	26337S-43	1201K-1	SH3217	Triploid
84	28263S-2	1201K-1	SH3217	Triploid
85	12618S-1	1201K-1	SH3362	Triploid
86	26316S-7	1201K-1	SH3362	Triploid
87	26840S-10	1201K-1	SH3362	Triploid
88	25328S-3	1438K-1	1537K-1	Triploid
89	24948S-10	1438K-1	5610S-1	Triploid
90	24948S-13	1438K-1	5610S-1	Triploid
91	24948S-24	1438K-1	5610S-1	Triploid
92	24948S-9	1438K-1	5610S-1	Triploid
93	26060S-1	1438K-1	9128-3	Triploid
94	13573S-1	1438K-1	9719-7	Triploid
95	27914S-1	1438K-1	cv. Rose	Triploid
96	27914S-13	1438K-1	cv. Rose	Triploid
97	28095S-1	1438K-1	cv. Rose	Triploid
98	27264S-2	1438K-1	cv. Rose	Diploid
99	27914S-24	1438K-1	cv. Rose	Triploid
100	27914S-26	1438K-1	cv. Rose	Triploid

27914S-3	1438K-1	cv. Rose	Triploid
25474S-1	1438K-1	Kokopo	Triploid
26369S-4	1438K-1	Long Tavoy	Triploid
28481S-1	1438K-1	*M. a. malaccensis 250*	Triploid
28561S-2	1438K-1	*M. a. malaccensis 250*	Triploid
26725S-1	1438K-1	SH3362	Triploid
25499S-7	1438K-1	SH3142	Triploid
26039S-2	1438K-1	SH3217	Triploid
26466S-2	1977K-1	5610S-1	Triploid
26466S-5	1977K-1	5610S-1	Triploid
22598S-2	365K-1	1201K-1	Triploid
14539S-4	365K-1	660K-1	Triploid
9750S-13	401K-1	9128-3	Triploid
25031S-1	5610S-1	2180K-6	Diploid
25031S-15	5610S-1	2180K-6	Diploid
25031S-16	5610S-1	2180K-6	Diploid
25031S-17	5610S-1	2180K-6	Diploid
25031S-19	5610S-1	2180K-6	Diploid
25031S-27	5610S-1	2180K-6	Diploid
25031S-33	5610S-1	2180K-6	Diploid
25031S-34	5610S-1	2180K-6	Diploid
25031S-7	5610S-1	2180K-6	Diploid
24583S-2	660K-1	5610S-1	Triploid
26260S-3	660K-1	5610S-1	Triploid
13284S-1	660K-1	9128-3	Triploid
25371S-2	660K-1	9128-3	Triploid
9187S-8	660K-1	9128-3	Triploid
26709S-1	660K-1	Calcutta 4	Triploid
27713S-1	660K-1	*M. a. malaccensis 250*	Triploid
27825S-4	660K-1	*M. a. malaccensis 250*	Triploid
27873S-18	660K-1	*M. a. malaccensis 250*	Triploid
27873S-38	660K-1	*M. a. malaccensis 250*	Triploid
27873S-4	660K-1	*M. a. malaccensis 250*	Triploid
27873S-5	660K-1	*M. a. malaccensis 250*	Triploid

28188S-2	660K-1	*M. a. malaccensis 250*	Triploid
25623S-11	8817S-1	917K-2	Triploid
28492S-1	917K-2	1968-2	Triploid
26998S-1	917K-2	2180K-6	Triploid
27074S-1	917K-2	2180K-6	Triploid
25117S-1	917K-2	5610S-1	Triploid
25117S-2	917K-2	5610S-1	Triploid
25117S-3	917K-2	5610S-1	Triploid
25508S-1	917K-2	5610S-1	Triploid
25628S-11	917K-2	5610S-1	Triploid
26815S-3	917K-2	5610S-1	Triploid
26815S-8	917K-2	5610S-1	Triploid
26815S-9	917K-2	5610S-1	Triploid
26990S-10	917K-2	5610S-1	Triploid
26990S-11	917K-2	5610S-1	Triploid
26990S-4	917K-2	5610S-1	Triploid
27073S-1	917K-2	5610S-1	Triploid
27744S-1	917K-2	5610S-1	Triploid
12949S-2	917K-2	7197-2	Triploid
25909S-3	917K-2	7197-2	Triploid
25089S-4	917K-2	861S-1	Triploid
19798S-2	917K-2	9128-3	Triploid
24434S-3	917K-2	9128-3	Triploid
25435S-11	917K-2	9128-3	Triploid
25435S-4	917K-2	9128-3	Triploid
25737S-1	917K-2	9128-3	Triploid
26288S-4	917K-2	9128-3	Triploid
26975S-1	917K-2	9128-3	Triploid
26975S-2	917K-2	9128-3	Triploid
7798S-2	917K-2	9128-3	Triploid
27184S-4	917K-2	cv. Rose	Triploid
27885S-9	917K-2	cv. Rose	Triploid
27184S-8	917K-2	cv. Rose	Triploid
27494S-12	917K-2	cv. Rose	Triploid

27494S-4	917K-2	cv. Rose	Triploid
27494S-5	917K-2	cv. Rose	Triploid
28068S-9	917K-2	cv. Rose	Triploid
27184S-6	917K-2	cv. Rose	Triploid
27885S-1	917K-2	cv. Rose	Triploid
24410S-2	917K-2	Kokopo	Triploid
25680S-11	917K-2	Long Tavoy	Triploid
25680S-13	917K-2	Long Tavoy	Triploid
27261S-1	917K-2	*M. a. malaccensis 250*	Triploid
27261S-10	917K-2	*M. a. malaccensis 250*	Triploid
27261S-11	917K-2	*M. a. malaccensis 250*	Triploid
27334S-5	917K-2	*M. a. malaccensis 250*	Triploid
27401S-1	917K-2	*M. a. malaccensis 250*	Triploid
27524S-12A	917K-2	*M. a. malaccensis 250*	Triploid
27524S-12B	917K-2	*M. a. malaccensis 250*	Triploid
27524S-22	917K-2	*M. a. malaccensis 250*	Triploid
27524S-30	917K-2	*M. a. malaccensis 250*	Triploid
27833S-10	917K-2	*M. a. malaccensis 250*	Triploid
27833S-13	917K-2	*M. a. malaccensis 250*	Triploid
27886S-5	917K-2	*M. a. malaccensis 250*	Triploid
28033S-14	917K-2	*M. a. malaccensis 250*	Triploid
28033S-15	917K-2	*M. a. malaccensis 250*	Triploid
28033S-18	917K-2	*M. a. malaccensis 250*	Triploid
28033S-23	917K-2	*M. a. malaccensis 250*	Triploid
28033S-3	917K-2	*M. a. malaccensis 250*	Triploid
28060S-8	917K-2	*M. a. malaccensis 250*	Triploid
28200S-3	917K-2	*M. a. malaccensis 250*	Triploid
28257S-1	917K-2	*M. a. malaccensis 250*	Triploid
28257S-2	917K-2	*M. a. malaccensis 250*	Triploid
28257S-4	917K-2	*M. a. malaccensis 250*	Triploid
28432S-19	917K-2	*M. a. malaccensis 250*	Triploid
28432S-20	917K-2	*M. a. malaccensis 250*	Triploid
28432S-3	917K-2	*M. a. malaccensis 250*	Triploid
28780S-1	917K-2	*M. a. malaccensis 250*	Triploid

26874S-5	917K-2	SH3362	Triploid
12468S-18	917K-2	SH3217	Triploid
12477S-13	917K-2	SH3217	Triploid
8386S-19	917K-2	SH3217	Triploid
13522S-5	917K-2	SH3362	Triploid
25974S-?	917K-2	SH3362	Triploid
25974S-19	917K-2	SH3362	Triploid
25974S-21	917K-2	SH3362	Triploid
25974S-30	917K-2	SH3362	Triploid
25974S-35	917K-2	SH3362	Triploid
26666S-1	917K-2	SH3362	Triploid
28476S-7	917K-2	SH3362	Triploid
9494S-10	917K-2	SH3362	Triploid
16457S-2	Entukura	365K-1	Triploid
26540S-182	Entukura	8075-7	Diploid
28260S-2	Enzirabahima	Calcutta 4	Triploid
21086S-1	Kazirakwe	7197-2	Triploid
28073S-1	Namwezi	7197-2	Triploid
25356S-1	Tereza	7197-2	Triploid
HB	unknown	unknown	Triploid
HJ	unknown	unknown	Triploid
HX	unknown	unknown	Triploid
26337S-11B	1201K-1	SH3217	Triploid
16285S-13	Calcutta 4	660K-1	Diploid
26337S-22B	1201K-1	SH3217	Triploid
16285S-3	Calcutta 4	660K-1	Diploid
26337S-28	1201K-1	SH3217	Triploid
25066S-1	1438K-1	Kokopo	Triploid
16285S-6	Calcutta 4	660K-1	Diploid
25066S-2	1438K-1	Kokopo	Triploid
16285S-8	Calcutta 4	660K-1	Diploid
25974S-11	917K-2	SH3362	Triploid
25974S-15	917K-2	SH3362	Triploid
25457S-1	1438K-1	Kokopo	Triploid

16191S-6	Calcutta 4	917K-2	Diploid
24797S-7	917K-2	Kokopo	Triploid
25102S-1	917K-2	Kokopo	Triploid
28452S-11	Nakasabira	Calcutta 4	Triploid
28033S-9	917K-2	*M. a. malaccensis 250*	Triploid
25974S-13	917K-2	SH3362	Triploid
28256S-1	917K-2	cv. Rose	Triploid
25974S-17	917K-2	SH3362	Tetraploid
12468S-6	917K-2	SH3217	Triploid
27914S-11	1438K-1	cv. Rose	Triploid
27914S-18	1438K-1	cv. Rose	Triploid
27914S-21	1438K-1	cv. Rose	Triploid
27914S-22	1438K-1	cv. Rose	Triploid
27914S-6	1438K-1	cv. Rose	Triploid
27914S-7	1438K-1	cv. Rose	Triploid
27914S-8	1438K-1	cv. Rose	Triploid
27873S-12	660K-1	*M. a. malaccensis 250*	Triploid
27873S-14	660K-1	*M. a. malaccensis 250*	Triploid
27873S-17	660K-1	*M. a. malaccensis 250*	Triploid
27873S-33	660K-1	*M. a. malaccensis 250*	Triploid
27873S-37	660K-1	*M. a. malaccensis 250*	Triploid
27873S-7	660K-1	*M. a. malaccensis 250*	Triploid
26224S-3	1201K-1	SH3362	Triploid
26840S-9	1201K-1	SH3362	Triploid
26316S-14	1201K-1	SH3362	Triploid
26224S-2	1201K-1	SH3362	Triploid
26840S-5	1201K-1	SH3362	Triploid
25653S-3	1201K-1	SH3142	Triploid
26315S-3	1201K-1	SH3142	Triploid
28528S-1	1201K-1	Kokopo	Triploid
26369S-8	1438K-1	Long Tavoy	Triploid
26530S-1	1438K-1	SH3362	Triploid
27528S-1	1438K-1	*M. a. malaccensis 250*	Triploid
27915S-3	1438K-1	*M. a. malaccensis 250*	Triploid

28561S-5	1438K-1	*M. a. malaccensis 250*	Triploid
27915S-2	1438K-1	*M. a. malaccensis 250*	Triploid
28974S-11	1438K-1	*M. a. malaccensis 250*	Triploid
28974S-15	1438K-1	*M. a. malaccensis 250*	Triploid
28974S-22	1438K-1	*M. a. malaccensis 250*	Triploid
28974S-29	1438K-1	*M. a. malaccensis 250*	Triploid
29114S-1	5610S-1	*M. a. malaccensis 250*	Diploid
29114S-14	5610S-1	*M. a. malaccensis 250*	Triploid
29114S-19	5610S-1	*M. a. malaccensis 250*	Triploid
29114S-24	5610S-1	*M. a. malaccensis 250*	Triploid
27873S-26	660K-1	*M. a. malaccensis 250*	Triploid
27873S-31	660K-1	*M. a. malaccensis 250*	Triploid
29165S-5	660K-1	*M. a. malaccensis 250*	Triploid
28506S-1	Entukura	Calcutta 4	Triploid
29364S-2	Namwezi	cv. Rose	Tetraploid
28077S-5	Nfuuka	8075-7	Triploid
28164S-15	1201K-1	cv. Rose	Triploid
29285S-20	1201K-1	cv. Rose	Triploid
26337S-32	1201K-1	SH3217	Triploid
27684S-5	1201K-1	SH3362	Triploid
24948S-12	1438K-1	5610S-1	Triploid
24948S-21	1438K-1	5610S-1	Triploid
24948S-27	1438K-1	5610S-1	Triploid
29586S-4	1438K-1	5610S-1	Triploid
24948S-22	1438K-1	5610S-1	Triploid
24948S-2	1438K-1	5610S-1	Triploid
24948S-29	1438K-1	5610S-1	Triploid
26820S-1	917K-2	1968-2	Triploid
25474S-5	917K-2	861S-1	Triploid
25974S-18	917K-2	SH3362	Triploid
28476S-8	917K-2	SH3362	Triploid
25974S-31	917K-2	SH3362	Triploid
29275S-1	Enzirabahima	*M. a. malaccensis 250*	Tetraploid
29275S-4	Enzirabahima	*M. a. malaccensis 250*	Tetraploid

305	29275S-5	Enzirabahima	*M. a. malaccensis 250*	Tetraploid
306	29636S-1	Tereza	7197-2	Tetraploid
307	28776S-2	Tereza	8075-7	Triploid

Supplemental Table S2: Comparison of average correlation for five-fold cross validations between the predicted and observed phenotypes across all models fitted with data from either low input (GS1) or high input (GS2) fields and 10807 bi-allelic SNP markers

Trait category	Traits	BRR		BL		BayesA		BayesB		BayesC		RKHS_P		RKHS_M		RKHS_PM	
		GS1	GS2	GS1	GS2	GS1	GS2	GS1	GS2	GS1	GS2	GS1	GS2	GS1	GS2	GS1	GS2
Plant stature	Plant height	0.54	0.46	0.55	0.45	0.54	0.45	0.54	0.44	0.54	0.45	0.42	0.40	0.55	0.44	0.54	0.48
	Plant girth	0.60	0.52	0.6	0.51	0.6	0.51	0.60	0.52	0.60	0.51	0.44	0.40	0.60	0.51	0.55	0.50
Suckering behaviour	Total number of suckers	0.16	0.17	0.17	0.20	0.17	0.20	0.16	0.19	0.15	0.19	0.12	0.12	0.17	0.18	0.16	0.17
	Height of tallest sucker at flowering	0.28	0.18	0.30	0.20	0.28	0.18	0.27	0.20	0.26	0.20	0.27	0.24	0.28	0.19	0.30	0.24
	Height of tallest sucker at harvesting	0.27	0.26	0.26	0.28	0.28	0.25	0.28	0.24	0.27	0.25	0.28	0.29	0.26	0.26	0.29	0.32
Black leaf streak	Number of standing leaves at flowering	0.36	0.42	0.37	0.40	0.37	0.42	0.43	0.40	0.36	0.41	0.17	0.19	0.37	0.41	0.29	0.34
	Index of non-spotted leaves	0.35	0.42	0.35	0.42	0.34	0.43	0.34	0.43	0.34	0.43	0.22	0.22	0.35	0.42	0.32	0.36
Fruit bunch	Days to fruit maturity	0.47	0.42	0.47	0.42	0.47	0.42	0.47	0.42	0.46	0.42	0.44	0.41	0.47	0.42	0.49	0.44
	Bunch mass	0.63	0.61	0.62	0.61	0.62	0.62	0.64	0.62	0.64	0.62	0.41	0.43	0.61	0.61	0.52	0.55
	Number of hands	0.60	0.62	0.59	0.62	0.59	0.63	0.60	0.62	0.59	0.62	0.34	0.39	0.59	0.62	0.48	0.53
	Number of fruits	0.47	0.51	0.47	0.53	0.47	0.52	0.47	0.52	0.47	0.52	0.25	0.33	0.45	0.52	0.35	0.45
Fruit filling	Fruit length	0.65	0.64	0.65	0.64	0.65	0.64	0.67	0.65	0.67	0.65	0.50	0.48	0.64	0.64	0.59	0.59
	Fruit circumference	0.67	0.66	0.67	0.67	0.66	0.66	0.70	0.69	0.70	0.69	0.40	0.42	0.65	0.66	0.57	0.60
	Fruit diameter	0.67	0.63	0.67	0.68	0.66	0.67	0.70	0.71	0.70	0.71	0.39	0.40	0.65	0.67	0.57	0.59
	Pulp diameter	0.67	0.68	0.67	0.69	0.66	0.68	0.70	0.72	0.70	0.72	0.39	0.41	0.65	0.67	0.57	0.60

The values under GS1 column are the correlations between predicted and observed phenotype (predictive ability) in GS2 when GS1 data were used to fit the model and vice versa for GS2 column.

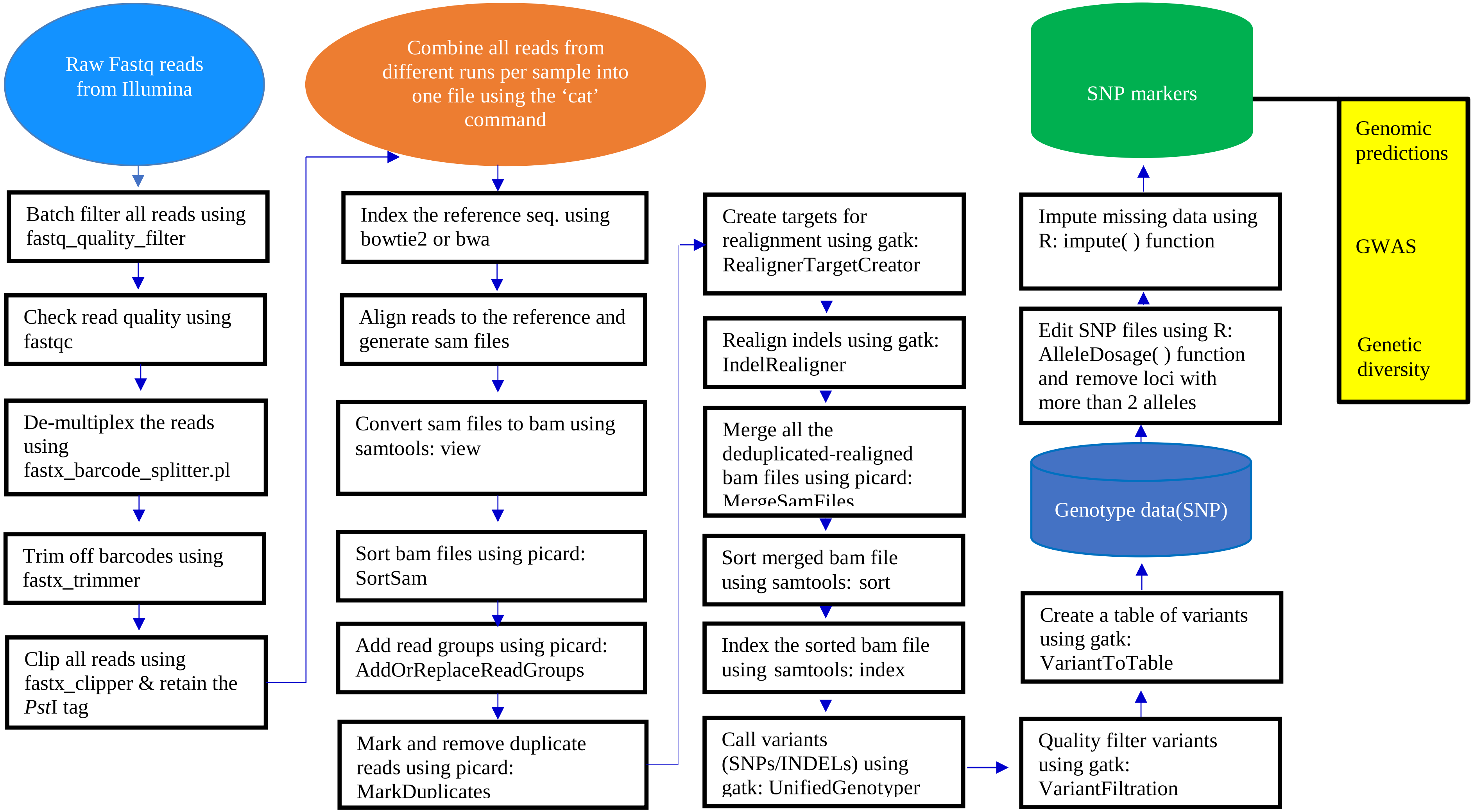
Supplemental Table S3: Comparison of predictive ability of BayesB model fitted with parents' data and progeny's data using bi-allelic and allele dosage SNP markers

		Parents model			Progeny model		
Trait category	Traits	BA-SNP	AD-SNP	LIP	BA-SNP	AD-SNP	LIP
Plant stature	Plant height	0.36	0.18	-50.0	0.77	0.51	-33.8
	Plant girth	0.39	0.05	-87.2	0.80	0.43	-46.3
Suckering behaviour	Total number of suckers	0.13	0.06	-53.8	0.39	0.22	-43.6
	Height of tallest sucker at flowering	0.23	0.12	-47.8	0.50	0.37	-26.0
	Height of tallest sucker at harvesting	0.19	-0.15	-178.9	0.43	-0.03	-107.0
Black leaf streak	Number of standing leaves at flowering	0.31	0.20	-35.5	0.43	0.46	7.0
	Index of non-spotted leaves	0.39	0.33	-15.4	0.85	0.77	-9.4
Fruit bunch	Days to fruit maturity	0.39	0.32	-17.9	0.77	0.66	-14.3
	Bunch mass	0.50	0.17	-66.0	0.66	0.08	-87.9
	Number of hands	0.45	0.03	-93.3	0.86	0.48	-44.2
	Number of fruits	0.31	0.10	-67.7	0.77	0.36	-53.2
Fruit filling	Fruit length	0.59	0.23	-61.0	0.78	0.22	-71.8
	Fruit circumference	0.49	0.17	-65.3	0.65	0.62	-4.6
	Fruit diameter	0.42	0.22	-47.6	0.66	0.65	-1.5
	Pulp diameter	0.49	0.23	-53.1	0.66	0.68	3.0

LIP = 100*((prediction with AD-SNP – prediction with BA-SNP)/prediction with BA-SNP)

Supplemental Table S4: Effect of ploidy level and allele dosage on the predictive ability of BayesB model fitted with environment averaged phenotype data

		Bi-allelic SNP			Allele dosage SNP		
Trait category	Traits	Tetraploid	Triploid	Diploid	Tetraploid	Triploid	Diploid
Plant stature	Plant height	0.04	0.37	0.71	-0.54	0.10	0.09
	Plant girth	0.02	0.38	0.72	-0.02	-0.06	-0.46
Suckering behaviour	Total number of suckers	-0.17	0.07	0.34	-0.48	-0.04	0.01
	Height of tallest sucker at flowering	0.19	0.31	0.32	-0.42	0.01	-0.13
	Height of tallest sucker at harvesting	0.06	0.20	0.57	-0.15	-0.03	-0.17
Black leaf streak	Number of standing leaves at flowering	0.19	0.40	0.38	0.05	0.09	-0.12
	Index of non-spotted leaves	-0.09	0.44	0.70	-0.30	0.12	0.31
Fruit bunch	Days to fruit maturity	0.01	0.46	0.56	0.01	0.06	0.21
	Bunch mass	0.15	0.39	0.73	0.03	0.03	-0.50
	Number of hands	0.33	0.44	0.70	0.48	0.08	0.05
	Number of fruits	0.50	0.37	0.57	-0.21	0.03	0.08
Fruit filling	Fruit length	-0.10	0.54	0.86	0.25	0.06	-0.21
	Fruit circumference	-0.15	0.43	0.79	0.35	0.05	-0.25
	Fruit diameter	-0.45	0.39	0.77	0.53	0.11	-0.15
	Pulp diameter	-0.39	0.41	0.79	0.60	-0.05	-0.23


Raw Fastq reads from Illumina
Batch filter all reads using fastq_quality_filter
Check read quality using fastqc
De-multiplex the reads using fastx_barcode_splitter.pl
Trim off barcodes using fastx_trimmer
Clip all reads using fastx_clipper & retain the *Pst*I tag
Combine all reads from different runs per sample into one file using the ‘cat’ command
Index the reference seq. using bowtie2 or bwa
Align reads to the reference and generate sam files
Convert sam files to bam using samtools: view
Sort bam files using picard: SortSam
Add read groups using picard: AddOrReplaceReadGroups
Mark and remove duplicate reads using picard: MarkDuplicates
Create targets for realignment using gatk: RealignerTargetCreator
Realign indels using gatk: IndelRealigner
Merge all the deduplicated-realigned bam files using picard: MergeSamFiles
Sort merged bam file using samtools: sort
Index the sorted bam file using samtools: index
Call variants (SNPs/INDELs) using gatk: UnifiedGenotyper
Quality filter variants using gatk: VariantFiltration
Create a table of variants using gatk: VariantToTable
Genotype data(SNP)
Edit SNP files using R: AlleleDosage( ) function and remove loci with more than 2 alleles
Impute missing data using R: impute( ) function
SNP markers
Genomic predictions
GWAS
Genetic diversity

Supplemental Fig. S1: Workflow used to analyse the genotyping by sequencing (GBS) reads to generate SNP marker data used in genomic predictions

Appendix II

Trait variation and genetic diversity in a banana genomic selection training population

https://doi.org/10.1371/journal.pone.0178734

PLOS | ONE

RESEARCH ARTICLE

Trait variation and genetic diversity in a banana genomic selection training population

Moses Nyine[1,2,3], Brigitte Uwimana[2]*, Rony Swennen[2,4,5,6], Michael Batte[2], Allan Brown[6], Pavla Christelová[3], Eva Hřibová[3], Jim Lorenzen[2¤], Jaroslav Doležel[3]*

1 Faculty of Science, Palacký University, Olomouc, Czech Republic, **2** International Institute of Tropical Agriculture, Kampala, Uganda, **3** Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Olomouc, Czech Republic, **4** Laboratory of Tropical Crop Improvement, Division of Crop Biotechnics, Katholieke Universiteit Leuven, Leuven, Belgium, **5** Bioversity International, Leuven, Belgium, **6** International Institute of Tropical Agriculture, Arusha, Tanzania

¤ Current address: Bill & Melinda Gates Foundation, Seattle, Washington, United States of America
* B.Uwimana@cgiar.org (BU); dolezel@ueb.cas.cz (JD)

Abstract

Banana (*Musa spp.*) is an important crop in the African Great Lakes region in terms of income and food security, with the highest per capita consumption worldwide. Pests, diseases and climate change hamper sustainable production of bananas. New breeding tools with increased crossbreeding efficiency are being investigated to breed for resistant, high yielding hybrids of East African Highland banana (EAHB). These include genomic selection (GS), which will benefit breeding through increased genetic gain per unit time. Understanding trait variation and the correlation among economically important traits is an essential first step in the development and selection of suitable GS models for banana. In this study, we tested the hypothesis that trait variations in bananas are not affected by cross combination, cycle, field management and their interaction with genotype. A training population created using EAHB breeding material and its progeny was phenotyped in two contrasting conditions. A high level of correlation among vegetative and yield related traits was observed. Therefore, genomic selection models could be developed for traits that are easily measured. It is likely that the predictive ability of traits that are difficult to phenotype will be similar to less difficult traits they are highly correlated with. Genotype response to cycle and field management practices varied greatly with respect to traits. Yield related traits accounted for 31–35% of principal component variation under low and high input field management conditions. Resistance to Black Sigatoka was stable across cycles but varied under different field management depending on the genotype. The best cross combination was 1201K-1xSH3217 based on selection response (R) of hybrids. Genotyping using simple sequence repeat (SSR) markers revealed that the training population was genetically diverse, reflecting a complex pedigree background, which was mostly influenced by the male parents.

Citation: Nyine M, Uwimana B, Swennen R, Batte M, Brown A, Christelová P, et al. (2017) Trait variation and genetic diversity in a banana genomic selection training population. PLoS ONE 12(6): e0178734. https://doi.org/10.1371/journal.pone.0178734

Editor: Marco Buenrostro-Nava, Universidad de Colima Facultad de Ciencias Biologicas y Agropecuarias, MEXICO

Received: October 11, 2016

Accepted: May 18, 2017

Published: June 6, 2017

Data Availability Statement: All relevant data are within the paper and its supporting Information files.

Funding: This work was supported by several donors through their contributions to the CGIAR Fund (http://www.cgiar.org/who-we-are/cgiar-fund/fund-donors-2/), and in particular to the CGIAR Research Program on Roots, Tubers and Bananas. Additional funding was provided by the Bill & Melinda Gates Foundation (BMGF)

(Investment ID: OPP1093845) to the IITA banana breeding program. The genotyping work at the Institute of Experimental Botany was supported by the Ministry of Education, Youth and Sports of the Czech Republic (grant LO1204 from the National Program of Sustainability I). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Introduction

East Africa is considered a secondary center of banana genetic diversity. Uganda in particular is a home to over eighty cultivars of East African Highland banana (EAHB) commonly divided into cooking and beer types [1]. The crop greatly contributes to the income and food security of many smallholder farmers in the region. The significance of the crop in the region is reflected in the per capita consumption that ranges between 250kg and 600kg with an average of 400kg in Uganda [2]. Over 85% of the production is consumed locally due to high demand [3, 4]. Sustainable production of bananas is a challenge because of disease, insect and nematode pressure. This is worsened by abiotic stress arising through factors associated with climate change [5]. Yield reductions in EAHB are caused by pests such as root burrowing nematodes especially *Radopholus similis* and banana weevil (*Cosmopolites sordidus*). Black Leaf Streak (Black Sigatoka), a fungal disease caused by *Mycosphaerella fijiensis* reduces the photosynthetic area of the plant, which decreases yield. Banana bacterial wilt caused by *Xanthomonas campestris* pv. *musacearum* causes 100% yield loss when the banana is attacked [6–8]. Variation in rainfall patterns impacts banana production by causing drought stress because most farmers in the region rely on rain for agricultural production. Although phenotypic variation is observed in EAHB, their genetic variation is low [9, 10] making them all susceptible to biotic and abiotic stress. Adaptation of cultivated banana varieties to changing environment is limited because while some are capable of sexual reproduction, they are all propagated clonally.

In order to meet the food demand for the growing population, breeding for resistance and high yielding varieties is considered to be the most sustainable solution to address banana production constraints [11, 12]. Unlike other crops, banana breeding is complicated by the polyploid nature of the crop characterized by abnormal meiosis in the cultivated triploid varieties that results in reduced fertility or complete sterility [13–15]. Crossing cultivated varieties with resistant wild diploids is possible, but a majority of the generated hybrids are inferior due to linkage drag of unfavorable genes from the wild diploids. However, when tetraploids are obtained, further improvement is possible because they are both male and female fertile (Fig 1). Incorporating resistance while maintaining the unique attributes such as fruit colour, aroma, texture and taste in existing varieties is a big challenge to banana breeders that calls for dedicated effort and careful choice of cross combinations. Crossbreeding is labour-intensive, costly and time consuming. In the last two decades, some success has been registered with new hybrids released to farmers while others are in the advanced stages of evaluation [16]. In order to keep up with the pace at which environmental changes occur and the demand for new varieties that are productive and of good quality, new breeding strategies should be employed to increase breeding efficiency and reduce the lengthy selection period [3].

Marker assisted selection (MAS) has been implemented in many animal and crop breeding programs. The success of MAS greatly depends on the genetic architecture of traits being improved. To date MAS has not been effectively deployed in banana breeding. The possible reasons are polyploidy, important economic and agronomic traits may be controlled by many quantitative trait loci (QTL), each with a small additive effect, and the lack of saturated linkage maps for QTL mapping. It is believed that the application of genomic selection (GS) will improve the efficiency of crossbreeding programs especially for crops with long breeding and selection cycle [17, 18] like banana. GS is a form of MAS where selection is based on the genomic estimated breeding values (GEBV) of superior individuals in the population as determined by a statistical model [19–21]. This technique is well established in animal breeding [22, 23]. In plants, GS has been tested in maize and wheat [24],

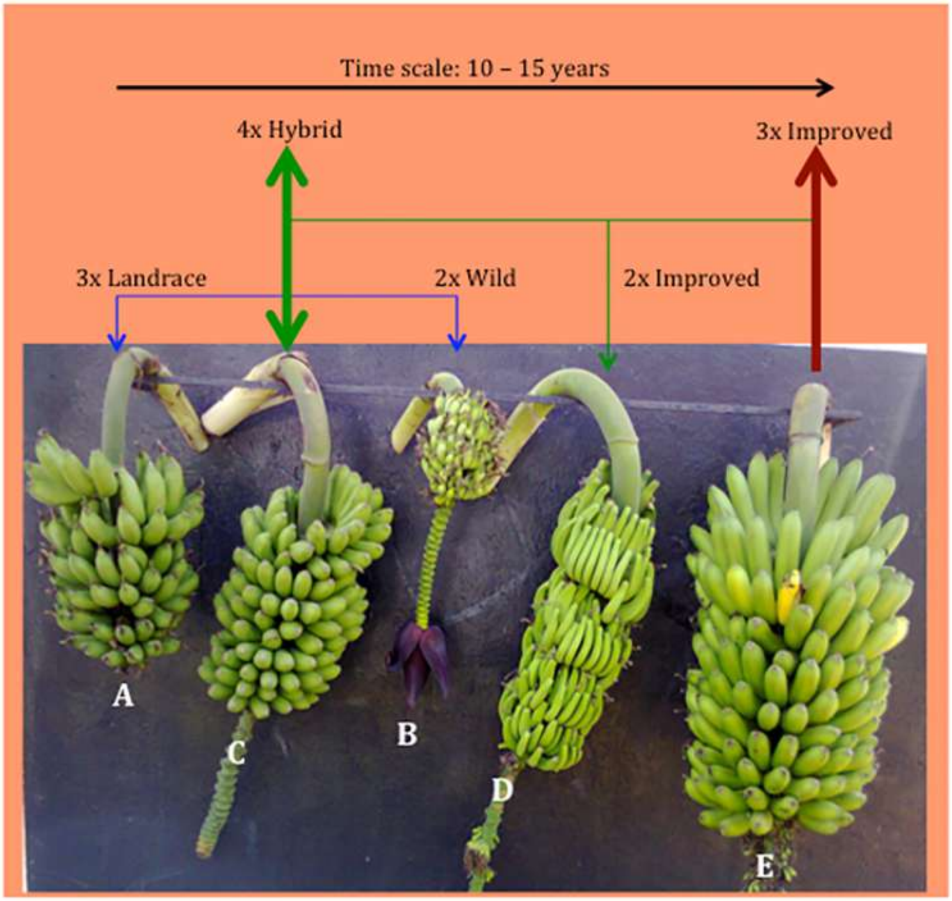


Fig 1. Conventional banana breeding starts with crossing 3x inferior and parthenocarpic landrace varieties (A) with a wild, seeded 2x accession (B). 4x resulting from this cross (C) are selected and crossed with improved 2x hybrids (D). The resulting secondary 3x (E) are selected and evaluated as potential improved varieties. This process takes up to 15 years.

https://doi.org/10.1371/journal.pone.0178734.g001

white spruce [25], rice [26] and cassava [27]. However, in bananas GS is in its infancy. Given that new varieties are selected based on a combination of traits, a selection index of GEBV in bananas is necessary.

GS studies have reported varying accuracies in prediction (predictive ability of GS models) and this has been attributed to differences in trait heritability, number of markers, training population size and genotype x environment interaction [24]. Bananas as perennial plants suffer the consequences of nutrient deficiency and soil moisture variation across seasons and locations depending on field management practices. Breeding generates genotypes from many crosses that are genetically different and respond to growth environment differently and this could affect the accuracy of GS. Therefore, understanding trait variation and the correlation between different traits is essential to guide the development and selection of suitable GS models for banana breeding. In this study we tested the hypothesis that trait variations in bananas are not affected by cross combination, cycle, field management and their interaction with genotype. For this, a training population created using EAHB breeding material and its progeny was phenotyped in two contrasting conditions. Genetic diversity of the training population was assessed using simple sequence repeat (SSR) markers.

Materials and methods

Plant population

Data were collected at the International Institute of Tropical Agriculture, Uganda from a banana genomic selection (GS) training population between 2013 and 2016. The institute is located at Namulonge research station, 0.53˚ N 32.58˚ E, 1150 m above sea level with rainfall of about 1200 mm/y split into two rainy seasons, March-June and September-December and an average annual temperature of 22˚C. The GS population consisted of 307 genotypes that included diploid (11%), triploid (85%) and tetraploid (4%) plants (S1 Table). The ploidy level of the genotypes was determined using flow cytometry [28, 29]. The core breeding lines (parents) accounted for 12% of the entire population. Two fields were established with each genotype replicated three times in a completely randomized design. Suckers were used as planting materials and before planting, 20kg of farmyard manure was applied in each hole. One field (GS1) was managed without mulching, additional manure nor inorganic fertilizer (low input). The second field (GS2) was mulched twice a year. Six months after planting, 480 g of NPK (25:5:5) fertilizer was added and the same amount was added to each mat per year (high input).

Traits

The yield-related traits scored included: days to fruit maturity (DFM) that is, days between flowering and harvesting, bunch weight at full maturity (BWT), number of hands (cluster) (NH) and number of fruit fingers (NF), fruit length (FL), fruit circumference (FC), fruit diameter (FRD), pulp diameter (PLD) and peel thickness (PED), where PED = (FRD—PLD)/2. The vegetative (growth) traits included: number of standing leaves at flowering (NSLF), youngest leaf spotted with Black Sigatoka at flowering (YLSF), index of non-spotted leaves at flowering (INSL), height of tallest sucker at harvesting (HTSH), plant height at flowering (PHF), plant girth at 100 cm from soil surface (PG), height of tallest sucker at flowering (HTSF), total number of suckers at flowering (TS), number of standing leaves at harvesting (NSLH) and youngest leaf spotted with black sigatoka at harvesting (YLSH).

Total number of suckers (TS) was recorded at flowering in cycle 1 only after which each mat was left with a maximum of three plants and these included the flowered plant, follower sucker and the sucker produced by follower sucker if present. A Vernier caliper was used to measure FRD and PLD. Fruit related traits such as FL, FC, FRD and PLD were recorded from the middle finger of the second hand on the bunch. Measurements for FC, FRD and PLD were recorded midway the length of the finger. However, to measure FRD and PLD, a cross-section of the fruit was made to expose the pulp. The INSL was calculated from the formula, INSL = 100*(YLSF-1)/NSLF [30]. This formula should give percentage values ranging from 0–100% to represent completely susceptible (0%) and completely resistant (100%). In order to get 100% INSL for completely resistant genotypes, the YLSF was scored as NSLF +1 thus INSL = 100*((NSLF+1)-1)/NSLF or INSL = 100*NSLF/NSLF

Data analysis

All analyses were performed in R, open source statistical software from www.r-project.org. A combination of Shapiro-Wilk test, boxplots, standard deviations and histograms were used to check for normality and outliers in the data and where necessary the outliers were removed before further analysis. Total number of suckers and bunch weight were transformed by square root. Using the aggregate function from library (plyr), trait means were calculated for every genotype and cross combination (family) in every cycle, and field and these were used in correlation analysis and principal component analysis (PCA).

Correlation analysis and test of significance for the correlations between traits were done using library (Hmisc) and Student's t-test based on cycle 2 data for cross combinations. Coefficient of determination (R^2) was calculated as a square of correlation coefficient between cycle 1 and 2 data. To understand the structure of the population and how different traits influenced that structure, principal component analysis was done using PCA function provided in the library (FactoMineR). Traits (dependent variable), cross combinations and individual genotypes were projected on the first two components (Dim1 and Dim2).

Sources of trait variation were assessed using unbalanced analysis of variance (ANOVA) based on cycle 1 and 2 data. Linear models were constructed for each trait in respect to each cycle, field management practice and their interaction with genotype as model_fit = lm(trait response~clone+cycle+field+clone:field+clone:cycle, data = mydata) where lm = linear model function. Type III SS ANOVA tables were generated using Anova function provided in the library(car) as result = Anova(model_fit, singular.ok = TRUE, type = "III"). In cases where no significant interactions were observed between two independent variables and where one explanatory variable was not significant, then type II or type I SS ANOVA was used for further investigation.

Selection differential (S) and response to selection (R) were used to compare performance of parental cross combinations [31]. S and R were calculated as, S = P—G and R = H—G, where P = average performance of a pair of parents, G is the average performance of all parental lines in the training population and H is the average performance of all hybrid that shared same parental pair. Only cross combinations that had at least five hybrids were compared across all traits using combined data from the two fields.

Genetic diversity

Genetic diversity of the training population was assessed using simple sequence repeat (SSR) markers. Cigar leaf samples were collected from the training population in Uganda and shipped to the Institute of Experimental Botany, Olomouc, Czech Republic under cold chain. Samples were lyophilized prior to DNA extraction. DNA from lyophilized samples was extracted using NucleoSpin Plant II kit, Macherey-Nagel, Germany, following the manufacturer's instructions. The concentration and quality of DNA was assessed by NanoDrop ND-1000 spectrophotometer. Nineteen informative *Musa* SSR primers were used to genotype the GS training population. The list of primers used, polymerase chain reaction (PCR) conditions, and fragment analysis procedure were adopted from Christelová et al. [32].

Two independent rounds of PCR were performed on each sample. The concordance of alleles from each sample were inspected and scored manually in GeneMarker v1.75 (Softgenetics, State College, PA, USA). A third round of PCR was performed only for samples that showed incongruity with the two reactions. Alleles were scored as dominant markers for presence and absence (1/0). Data were imported in R and squared Euclidean distances were generated using the function dist provided in the library(ape). Clustering was done with function hclust based on ward.D method [33, 34]. Polymorphism information content of each marker was estimated by PowerMarker v3.25 software [35].

Results

During data analysis, some genotypes were excluded for some traits due to missing data or extreme outliers. The outliers were mainly recorded on plants that were infected with banana

Xanthomonas wilt before full maturity, plants that snapped due to weevil damage and premature breaking of the peduncle due to windstorm.

Correlation of traits

Significant correlations were observed among and between growth and yield traits (Tables 1 and 2). PHF had significant positive correlation with PG followed by HTSF. PG positively correlated with BWT, NF and HTSF in that respective order. The traits associated with Black Sigatoka resistance (NSLF, YLSF and INSL) also correlated significantly to each other. However, they had significant negative correlations with fruit traits such as FC, FRD and PLD. A positive and significant correlation was observed between BWT and all fruit traits (NH, NF, FL, FC, FRD, PLD), which were similarly significantly and positively correlated to each other. Under conditions of low input field management (GS1), TS, NSLH and NF were not significantly correlated with other traits while under high input field management (GS2), it was INSL, DFM and HTSH that did not have significant correlation with other traits. In both fields, the highest positive correlations were observed among the yield traits. In this population, absolute apical dominance was not observed as all genotypes had at least one sucker at the time of flowering. However, sucker regulation varied among genotypes with a range of 1–25 suckers per plant.

Principal component analysis (PCA)

Principal component analysis showed that in both fields, the yield (fruit) traits contributed to the first component (Dim 1) while the vegetative (growth) traits contributed to the second component (Dim 2) (Fig 2A and 2B). Among the vegetative traits, PHF and PG contributed to Dim 1. Dim 1 accounted for 31.07% of variation in GS1 and 35.86% in GS2. Dim 2 accounted for 21.89% of variation in GS1 and 15.40% in GS2. The traits with the highest negative loading on Dim 1 included FC, FRD and PLD for GS1 while for GS2 it was FC, FRD, PLD and FL. In both GS1 and GS2, the traits with the highest positive loading on Dim 2 were NSLF, YLSF, INSL and NSLH. Both DFM and TS had the least contribution to the two components with completely different orientation in GS1 and GS2. Generally, in both fields the two components accounted for 50% of the variation observed in the genotype cross combinations (Fig 3A and 3B).

For individual genotypes, a similar trend was observed with Dim 1 and Dim 2 accounting for 31.43% and 19.11% of total trait variation, respectively (Fig 4A). Projection of the individual factors (genotypes) on the two components did not reveal any distinct population structure (Fig 4B). The same trend was observed when individual cross combinations were projected on the two components. However, in GS1 cross combinations C35 (917K-2 x Kokopo), C28 (8817S-1 x 917K-2) and C52 (SH2095 x SH2766) and in GS2 cross combinations C35 (917K-2 x Kokopo), C22 (365K-1 x 660K-1) and C29 (8817S-1 x 917k-2) were distinct and clearly separated out from the others (Fig 3a and 3b). When the data were re-examined, genotypes from cross C35 had the least average scores on the yield traits while cross C22, C29 and C52 had the highest average scores on the yield traits. All the four planes of the two components were represented in the population.

Based on Black Sigatoka resistance and fruit filling (indicated by FRD), four main groups were represented in the population: (i) genotypes with high INSL and good fruit filling, (ii) high INSL with poor fruit filling, (iii) low INSL with good fruit filling and (iv) low INSL with poor fruit filling. On average the observed INSL and FRD for the genotypes in the four groups were as follows: (i) 78.1% and 3.0cm, (ii) 80.1% and 1.4cm, (iii) 66.8% and 3.1cm, and (iv) 67.1% and 1.4cm, respectively. Genotypes projected on Dim 2 had high average scores on

Table 1. Pearson's correlation coefficients of traits under low input field management (GS1).

	PHF	PG	NSLF	YLSF	HTSF	TS	INSL	DFM	NSLH	BWT	NH	NF	FL	FC	FRD	PLD	PED
PHF																	
PG	0.807***																
NSLF	0.108	0.254															
YLSF	0.083	0.181	0.926***														
HTSF	0.373**	0.36**	0.319**	0.364**													
TS	−0.229	−0.329**	0.022	0.116	0.3**												
INSL	−0.001	0.021	0.579***	0.834***	0.326**	0.24											
DFM	0.086	0.133	0.282	0.277**	0.109	0.038	0.231										
NSLH	0.042	0.078	0.386	0.352**	0.363**	0.034	0.194	-0.356									
BWT	0.346**	0.554***	−0.083	−0.122	0.213	0.02	−0.133	0.152	−0.22								
NH	0.372**	0.426**	0.19	0.166	0.087	0.041	0.119	0.191	−0.021	0.411**							
NF	0.412**	0.512***	0.226	0.195	0.151	−0.032	0.126	0.221	0.053	0.4**	0.878***						
FL	0.2	0.411**	−0.077	−0.113	0.057	−0.042	−0.123	0.173	−0.266	0.855***	0.168	0.169					
FC	0.191	0.375**	−0.284**	−0.338**	0.025	−0.097	−0.323**	0.005	−0.254	0.807***	0.019	0.008	0.856***				
FRD	0.206	0.359**	−0.357**	−0.415**	0.022	−0.11	−0.395**	−0.055	−0.258	0.782***	0.017	0.02	0.82***	0.987***			
PLD	0.192	0.333**	−0.379**	−0.432**	−0.004	−0.133	−0.41**	−0.099	−0.199	0.717***	0.048	0.057	0.709***	0.9***	0.919***		
PED	−0.26**	0.108	0.143	0.024	0.007	−0.078	−0.114	0.108	−0.013	0.225	−0.182	−0.118	0.359**	0.293**	0.272**	0.179	

https://doi.org/10.1371/journal.pone.0178734.t001

Table 2. Pearson's correlation coefficients of traits under high input field management (GS2).

	PHF	PG	NSLF	YLSF	HTSF	TS	INSL	DFM	NSLH	BWT	NH	NF	FL	FC	FRD	PLD	PED
PHF																	
PG	0.774***																
NSLF	−0.422**	−0.257															
YLSF	−0.197	−0.024	0.75***														
HTSF	0.702***	0.563***	−0.357**	−0.251													
TS	0.358**	0.224	−0.092	0.062	0.45***												
INSL	0.213	0.272	−0.128	0.548***	0.084	0.197											
DFM	−0.007	0.006	−0.026	0.063	−0.218	−0.02	0.152										
NSLH	−0.149	−0.077	0.619***	0.533***	−0.222	−0.156	0.002	−0.194									
BWT	0.37**	0.623***	−0.081	−0.14	0.46***	0.165	−0.132	0.019	−0.173								
NH	0.218	0.424**	0.071	0.119	0.227	0.09	0.095	0.175	−0.068	0.521***							
NF	0.368**	0.582***	0.006	0.11	0.348**	0.169	0.194	0.227	−0.007	0.57***	0.843***						
FL	0.204	0.439**	−0.076	−0.145	0.285**	0.134	−0.151	−0.065	−0.22	0.826***	0.284**	0.27**					
FC	0.327**	0.449**	−0.233	−0.255	0.397**	0.198	−0.146	−0.151	−0.19	0.807***	0.148	0.153	0.85***				
FRD	0.39**	0.478**	−0.254	−0.281**	0.42**	0.28**	−0.156	−0.154	−0.223	0.791***	0.158	0.184	0.803***	0.968***			
PLD	0.389**	0.446**	−0.271	−0.3**	0.398**	0.31**	−0.161	−0.176	−0.22	0.741***	0.114	0.135	0.76***	0.945***	0.991***		
PED	0.005	0.199	0.062	0.022	0.242	−0.171	−0.048	0.022	−0.077	0.513***	0.324**	0.34**	0.464**	0.337**	0.217**	0.1	

*** P-value < 0.001,

** P-value < 0.05 but > 0.001

https://doi.org/10.1371/journal.pone.0178734.t002

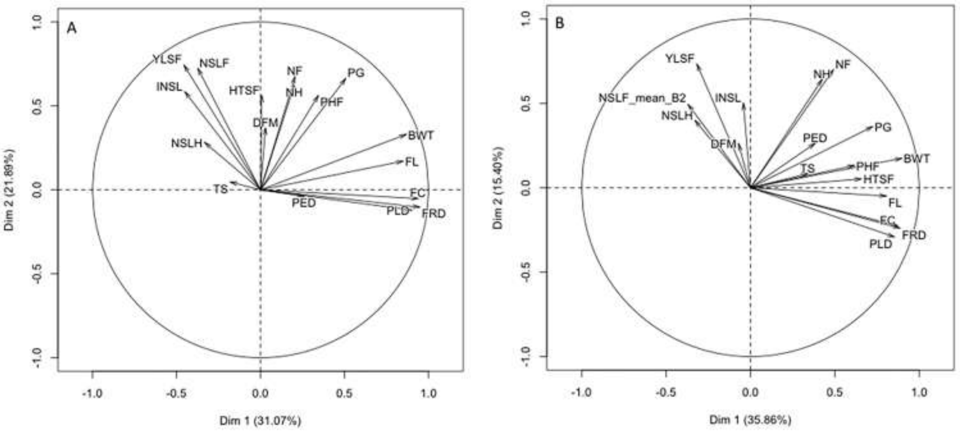


Fig 2. Principal component analysis plots generated in R using package FactoMineR for the traits scored in a banana genomic selection training population. (A) shows the distribution of traits under low input field management (GS1) and (B) shows the distribution of traits under high input field management (GS2) on the first two components.

https://doi.org/10.1371/journal.pone.0178734.g002

NSLF, YLSH, INSL, and NSLH and in contrast they had the lowest average scores on BWT, FL, FC, FRD, and PLD and the reverse was true for those projected on Dim 1.

Analysis of variance

Visual inspection of boxplots for various traits indicated a cycle effect on data distribution of some traits while others were not affected by cycle. For example, Plant height increased at cycle 2 while index of non-spotted leaves did not increase (Fig 5a and 5b) and this was confirmed by ANOVA results. Fruit traits such as FC, FRD and PLD showed a bimodal distribution with the histogram having two peaks. Based on these parameters, the population could be separated into two main groups, poor fruit filling genotypes with FRD < 2.0 cm and FC < 8.0 cm, and good fruit filling genotypes with FRD ≥ 2.0 cm and FC ≥ 8.0 cm (S1A–S1D Fig).

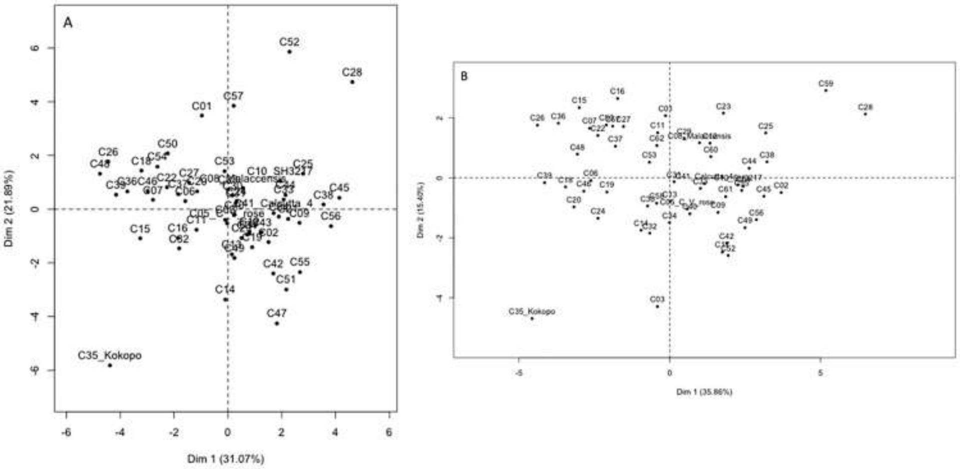


Fig 3. Principal component analysis plots generated in R using package FactoMineR for the cross combinations in a banana genomic selection training population. (A) shows the distribution of cross combinations under low input field management (GS1) and (B) shows the distribution of cross combinations under high input field management (GS2) on the first two components.

https://doi.org/10.1371/journal.pone.0178734.g003

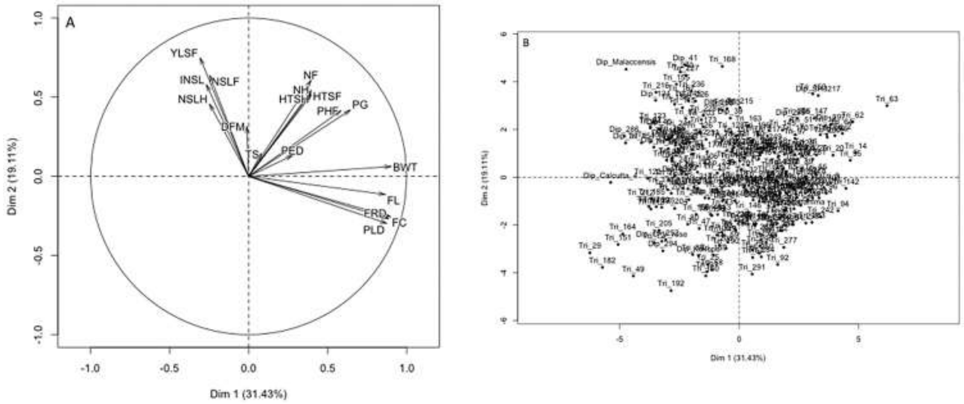


Fig 4. Principal component analysis plots generated in R using package FactoMineR for the traits scored in a banana genomic selection training population. (A) shows the distribution of traits for individual genotypes and (B) shows the distribution of individual genotypes on the first two components based on mean of combined data from the two fields.

https://doi.org/10.1371/journal.pone.0178734.g004

Coefficients of determination showed that under low input, cycle had less effect on NSLF, YLSF, INSL, TS, HTSF and PED across genotype cross combinations (Table 3). The Student's t-test revealed that both PED and HTSF were the most stable traits across cycles at 95% confidence level with P = 0.515 and P = 0.108, respectively. Under high input, cycle accounted for less than 50% of the variation in NSLF, YLSF, INSL, TS, HTSF, DFM, NSLH, NH, NF and PED between cross combinations. Just as in the first field, PED and HTSF were the least affected with P = 0.216 and P = 0.108, respectively. Under high input field management, trait variation due to cycle was more homogenous as compared to low input field management. However, in both cases the effects were statistically significant ($P < 0.001$) indicating that cycle is a source of variation in genotype performance.

When generating ANOVA models, genotype (clone) was assumed to be the main source of variation. In addition to genotype the effect of cycle, field and their interaction with genotype

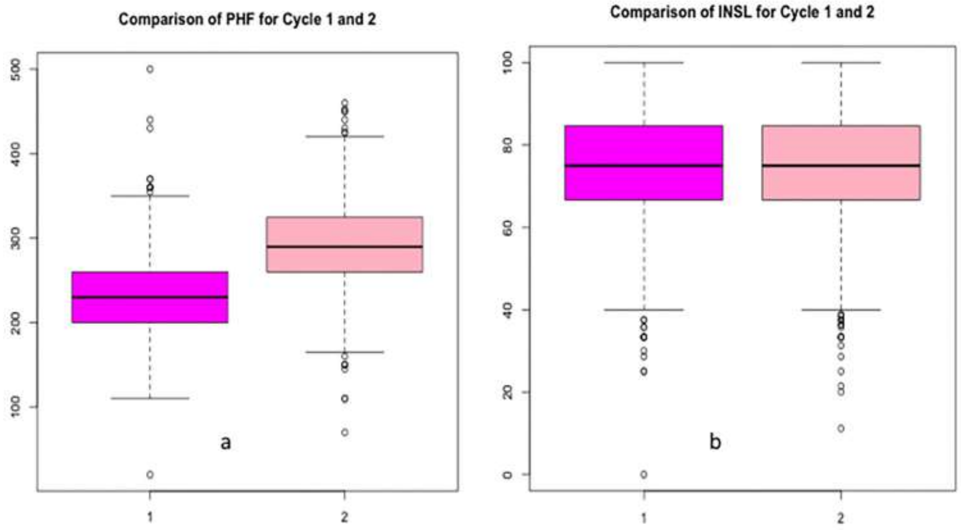


Fig 5. Effect of cycle on trait variation in bananas, where (a) shows an increase in plant height at flowering at cycle 2 while (b) shows no increase in index of non-spotted leaves at cycle 2.

https://doi.org/10.1371/journal.pone.0178734.g005

Table 3. Coefficient of determination and Student's t-test P-values showing the effect of cycle on cross combinations.

	GS1			GS2		
Traits	**df**	**R^2**	**P-value**	**df**	**R^2**	**P-value**
NH	60	0.87	<0.0001	56	**0.44**	<0.0001
PLD	57	0.78	<0.0001	56	0.65	<0.0001
FRD	59	0.77	<0.0001	56	0.68	<0.0001
PED	58	**0.06**	**0.5150**	56	**0.03**	**0.2161**
BWT	60	0.79	<0.0001	56	0.74	<0.0001
NF	60	0.54	<0.0001	56	**0.37**	<0.0001
FL	59	0.77	<0.0001	56	0.64	<0.0001
FC	58	0.79	<0.0001	56	0.73	<0.0001
DFM	59	0.54	<0.0001	56	**0.25**	<0.0001
NSLH	60	0.63	<0.0001	56	**0.38**	<0.0001
PHF	66	0.65	<0.0001	63	0.73	<0.0001
PG	66	0.65	<0.0001	63	0.73	<0.0001
NSLF	66	**0.25**	<0.0001	63	**0.28**	<0.0001
YLSF	66	**0.47**	<0.0001	63	**0.26**	<0.0001
INSL	66	**0.14**	0.0015	63	**0.21**	0.0001
TS	68	**0.12**	0.0032	68	**0.12**	0.0032
HTSF	68	**0.04**	**0.1084**	68	**0.04**	**0.1084**

Df = degrees of freedom, GS1 = low input field, GS2 = high input field and R^2 = coefficient of determination

https://doi.org/10.1371/journal.pone.0178734.t003

was investigated. In all models for all traits, genotype had significant effect on trait variation with $P < 0.001$ (Table 4, S3 Table). Traits that were not affected by the interaction between genotype and field management practice include PHF and PG whereas traits not affected by interaction between genotype and cycle include NSLF, YLSF, INSL, YLSH, FL, FRD and PED ($P > 0.05$). Weak interaction between genotype and cycle was observed on NSLH and HTSH with P = 0.0417 and 0.0408, respectively. In some cases, although there were significant interactions between genotype and field or cycle, either field or cycle did not show significant effect on the trait when interaction was included in the model.

Whereas there were significant interactions between genotype and field management, there was no significant main effect of field on NSLF, YLSF, HTSF, INSL, TS, NSLH, YLSH, HTSH, NH, NF and PED. Similarly, in the presence of significant interaction between genotype and cycle, there was no main effect of cycle on INSL, HTSF, HTSH, FC, PLD and PED (Table 4, S3 Table). When the interactions were removed from the models, all the factors had significant effect on the traits except INSL and PED, for which cycle had no effect. Analysis was repeated on these two traits using type I and type II ANOVA and both produced similar results as that observed with type III SS.

Performance of cross combinations (parental pairs)

The GS training population consisted of 77 different cross combinations representing about two decades of banana breeding activities by IITA and NARO Uganda. Some of these cross combinations gave rise to the tetraploids and improved diploids that are part of the core breeding lines in the program. Tetraploids and triploids were predominantly used as female parents while the diploids provided the pollen source but in some instances 2x by 2x or 4x by 4x crosses were made. The majority of the cross combinations were excluded for this analysis in

Table 4. Effect of genotype (clone), field management, cycle and their interaction on trait variation.

Dep. variable	Indep. variable	Sum Sq	Df	F value	Pr(>F)
PHF	Clone	2222889.11	306	3.77	<0.0001
	Clone:Field	432297.46	284	0.79	0.9947
	Clone:Cycle	332846.71	299	1.05	0.2662
PG	Clone	73176.82	306	4.30	<0.0001
	Clone:Field	12061.30	284	0.76	0.9981
	Clone:Cycle	13057.24	299	1.51	<0.0001
INSL	Clone	116602.02	306	2.44	<0.0001
	Clone:Field	58583.77	284	1.32	0.0005
	Clone:Cycle	51026.49	299	0.95	0.6947
TS^{sqrt}	Clone	240.28	305	3.21	<0.0001
	Clone:Field	100.88	282	1.46	<0.0001
BWT^{sqrt}	Clone	1213.89	303	12.55	<0.0001
	Clone:Field	126.77	269	1.48	<0.0001
	Clone:Cycle	108.68	276	1.49	<0.0001
FC	Clone	9506.06	300	16.11	0.0000
	Clone:Field	733.66	269	1.39	0.0001
	Clone:Cycle	751.00	272	1.29	0.0021
PLD	Clone	865.42	299	17.60	0.0000
	Clone:Field	68.27	269	1.54	<0.0001
	Clone:Cycle	60.55	271	1.29	0.0022
PED	Clone	20.96	299	11.41	<0.0001
	Clone:Field	16.61	269	10.05	<0.0001
	Clone:Cycle	3.15	271	0.80	0.9913

^{sqrt} Original data transformed by square root

https://doi.org/10.1371/journal.pone.0178734.t004

this work because they had less than five hybrids in the population. However, crosses between different EAHB with Calcutta 4 were treated as one cross because the EAHB represent a clone set with very low genetic diversity [9]. In total sixteen cross combinations were compared and they included one 2x by 2x, one 3x by 2x and fourteen 4x by 2x crosses (Table 5 and S2 Table).

The best cross in terms of yield and fruit size was C10 (1201K-1xSH3217). Many hybrids from this cross had the highest bunch weight (R = 3.8) characterized by longer fruit fingers, big fruit circumference and the highest pulp content. However, the plants were very tall with big girth. Their maturity period was shorter (about 4.5 months on average) and comparable to hybrids from EAHBxCalcutta 4. Generally, crosses involving SH3217, SH3362 and 9128–3 as male parents produced hybrids that had good fruit filling characteristics although they varied in Black Sigatoka resistance and suckering behavior. For example, crosses involving 9128–3 generated hybrids that had the lowest INSL.

Hybrids from a cross between 5610S-1 and 2180K-6 produced the highest number of leaves scored at flowering (R = 2.1). They had the highest resistance to Black Sigatoka as reflected by INSL (R = 7.2%) despite the parents being susceptible. They were the shortest (R = -62.3 cm) with smaller plant girth. Their average maturity period was almost two months more than the average of all parental lines (R = 54.6 days) and the longest of all other hybrids. Due to long maturity period the number of standing leaves at harvest was very low because of normal leaf senescence. Despite producing many fruit fingers and slightly more hands per bunch, their average yield and size of fruits were lower than those of the parents. However, some exceptional lines such as 25031S-7 (diploid) had sizable bunch with relatively big fruits.

Table 5. Comparison of mean performance of parental cross combinations (S) and hybrids from those crosses (R) against the mean of all parents.

CROSS	C04	C05	C08	C10	C11	C12	C13	C16	C22	C27	C31	C33	C34	C37	C61	MxC4
S (NSLF)	-0.5	-0.2	1.2	0.4	0.2	-0.2	-0.3	1.2	0.7	1.9	0.1	-0.3	-0.1	1.4	0.4	-1.1
R (NSLF)	-0.4	0.5	0.8	0.0	1.4	0.9	0.1	**1.8**	2.1	**1.8**	0.6	0.3	-0.2	1.4	0.8	0.1
S (YLSF)	-0.7	-0.4	1.4	0.3	0.0	-0.2	-0.3	1.7	0.2	1.7	-0.8	-1.1	-0.7	1.1	-0.3	-1.5
R (YLSF)	-0.7	0.3	0.7	0.0	0.8	0.4	0.1	1.8	**2.2**	1.7	0.4	-0.1	-0.1	1.2	0.7	-0.1
S (PHF)	24.1	-33.5	6.6	35.2	35.8	17.2	-37.5	3.8	-21.8	-1.5	-14.0	-11.4	-58.2	-22.5	0.2	25.9
R (PHF)	14.8	-23.8	10.1	**33.6**	-23.4	-7.4	-39.4	-6.6	-62.3	2.5	0.5	7.9	-31.0	-17.6	-9.5	7.6
S (PG)	9.6	-2.9	5.0	11.1	**11.7**	2.8	-7.5	-0.1	-5.3	1.3	0.9	1.4	-8.5	-1.6	3.6	3.3
R (PG)	3.6	-3.2	1.2	6.0	-1.4	2.3	-6.8	-1.4	-5.7	-0.6	2.2	4.9	-5.4	-2.0	3.3	2.0
S (HTSF)	11.4	-8.5	30.1	24.2	31.7	-5.5	-18.1	20.5	-46.3	23.0	-27.5	-25.0	-33.7	0.3	-4.7	23.0
R (HTSF)	15.0	-10.3	6.3	**23.3**	-21.1	-7.3	-26.8	14.3	-32.5	13.4	0.8	-2.5	-14.4	-4.0	-6.4	4.2
S (INSL)	-1.8	-1.0	4.9	0.9	0.3	0.2	0.7	7.2	-1.5	3.9	-6.4	-6.5	-4.2	1.1	-4.3	-7.0
R (INSL)	-2.9	0.6	1.4	1.2	-1.9	-0.7	1.1	5.7	**7.2**	4.1	0.1	-1.8	1.8	2.7	2.2	-0.9
S (TS)	-1.6	2.8	0.7	-1.0	1.1	-1.1	3.0	1.2	-1.7	0.1	-3.3	-2.9	1.3	-0.7	-0.4	0.0
R (TS)	-0.3	**1.9**	0.6	0.8	-1.0	-1.2	0.8	1.0	-0.4	-0.8	0.3	-1.9	1.2	0.0	0.7	-1.2
S (DFM)	2.4	2.7	15.9	10.0	-1.3	4.9	6.5	31.4	14.2	32.9	8.9	10.9	8.8	28.0	8.2	-21.3
R (DFM)	7.8	6.3	21.1	7.3	-1.9	19.9	1.6	8.3	**54.6**	32.6	23.9	11.2	13.5	22.3	20.7	7.2
S (NSLH)	-0.7	-0.9	0.3	-0.4	-0.5	-0.1	-0.7	1.3	0.5	1.5	0.4	0.1	-0.3	1.4	0.6	-0.7
R (NSLH)	-0.9	0.0	0.8	-0.4	1.5	0.6	0.1	**2.3**	0.3	1.4	0.3	0.1	-0.1	1.1	0.1	0.2
S (YLSH)	-0.4	-0.4	0.3	-0.1	-0.2	0.0	-0.3	1.0	0.1	1.1	-0.1	-0.2	-0.3	0.6	0.0	-0.4
R (YLSH)	-0.5	0.1	0.5	-0.2	**0.9**	0.1	0.0	0.8	0.1	0.6	0.2	0.2	0.1	0.8	0.1	0.1
S (HTSH)	27.6	-0.1	25.2	34.0	26.8	5.9	-21.3	10.8	-21.7	28.9	-2.6	4.6	-21.7	1.6	-2.2	7.7
R (HTSH)	23.4	-0.3	**45.0**	24.0	-18.4	18.4	-23.1	17.3	-15.9	19.1	23.6	9.9	-11.6	15.0	2.9	31.0
S (BWT)	5.6	2.3	4.2	7.2	7.0	1.5	-2.3	-1.5	-0.6	2.1	2.1	1.6	-1.2	-0.2	2.6	-0.7
R (BWT)	3.4	0.7	1.0	**3.8**	-0.9	1.0	0.4	-4.0	-2.3	-2.3	0.7	2.5	-0.1	-2.8	3.4	1.4
S (NH)	0.7	0.1	0.2	2.6	0.5	0.7	-0.1	0.0	1.1	0.6	0.2	0.3	-0.3	-0.3	-0.1	-0.8
R (NH)	0.4	0.4	1.0	0.9	**1.2**	1.1	0.3	0.9	**1.2**	0.4	0.8	**1.2**	-0.4	0.5	0.7	-0.3
S (NF)	22.1	-1.8	19.7	37.2	17.5	7.0	-19.7	7.7	7.5	15.9	8.8	12.2	-13.4	9.2	3.6	-16.0
R (NF)	15.9	9.0	**35.8**	12.8	19.9	13.9	1.5	21.7	27.4	10.7	19.6	25.6	-3.1	16.3	13.5	2.0
S (FL)	1.6	-0.2	0.8	2.8	1.9	0.7	-1.1	-1.4	-1.5	0.4	0.5	1.0	-0.9	-0.2	1.2	0.2
R (FL)	**2.8**	0.3	-0.8	2.5	-1.3	-0.2	-0.2	-3.9	-2.0	-2.6	-0.5	1.6	1.3	-2.6	2.3	0.3
S (FC)	2.2	0.7	2.2	2.1	3.1	1.2	-1.2	-1.1	0.4	0.9	1.2	0.6	-0.7	0.3	1.4	0.9
R (FC)	0.8	0.0	-0.6	**1.2**	-1.8	-0.7	-0.4	-3.4	-2.8	-2.5	-0.8	0.1	-0.4	-3.0	0.6	0.8
S (FRD)	0.6	0.2	0.6	0.6	0.9	0.4	-0.4	0.0	0.2	0.6	0.5	0.3	-0.2	0.2	0.6	0.1
R (FRD)	0.2	0.0	-0.3	0.3	-0.7	-0.3	-0.2	-1.2	-1.0	-0.8	-0.4	-0.1	-0.2	-1.0	0.1	**0.4**
S (PLD)	0.6	0.2	0.6	0.6	0.9	0.3	-0.3	0.0	0.1	0.6	0.5	0.3	-0.1	0.2	0.6	0.1
R (PLD)	0.2	0.0	-0.3	0.3	-0.7	-0.3	-0.1	-1.2	-1.0	-0.9	-0.4	-0.1	-0.2	-1.0	0.1	**0.4**
S (PED)	0.00	0.00	0.02	0.00	0.02	0.03	-0.02	0.01	0.01	0.02	0.00	-0.03	-0.03	-0.02	-0.01	0.01
R (PED)	0.01	0.01	-0.01	0.01	0.00	0.02	-0.01	-0.01	0.01	0.04	0.02	0.01	0.00	0.00	0.01	0.00

S = Selection differential, R = Response to selection, bold values are the highest observations, C04 = 1201K-1x9128-3, C05 = 1201K-1 x cv. Rose, C08 = 1201K-1 x *malaccensis*, C10 = 1201K-1 x SH3217, C11 = 1201K-1 x SH3362, C12 = 1438K-1 x 5610S-1, C13 = 1438K-1 x cv. Rose, C16 = 1438K-1 x *malaccensis*, C22 = 5610S-1 x 2180K-6, C27 = 660K-1 x *malaccensis*, C31 = 917K-2 x 5610S-1, C33 = 917K-2 x 9128–3, C34 = 917K-2 x cv. Rose, C37 = 917K-2 x *malaccensis*, C61 = 917K-2 x SH3362 and MxC4 = Matooke (EAHB) x Calcutta 4

https://doi.org/10.1371/journal.pone.0178734.t005

Crosses involving *M*. *acuminata* ssp. *malaccensis* 250 as male parent produced hybrids that were tall, slender, with bunches that had many fruit fingers poorly filled with pulp but some individual genotype exceptions were observed. The hybrids were resistant to Black Sigatoka and had the highest number of functional leaves at harvesting. Hybrids from cv.

Rose were slender and shorter and were the highest in sucker production while other traits varied considerably.

Hybrids from different cross combinations had longer maturity period (128–185 days) than EAHB. On average EAHB mature within 90 days after flowering while the average maturity period for all parental lines was 130 days.

Genetic diversity of GS training population

Out of the nineteen SSR markers, eighteen were used to delineate the structure of the study population, because marker mMaCIR164 produced ambiguous allele profiles across samples. From 18 loci, 195 alleles were scored and the number of alleles per locus ranged between 4 and 18 with an average of 10.8. Polymorphism information content (PIC) of the markers was high with an average of 0.87 (0.53–0.95) while the major allele frequency was on average 0.22 (0.1–0.45).

Despite the complex pedigree background of the GS population, SSR markers were informative enough to delineate the structure of the population (Fig 6). Hierarchical clustering based on Ward's criterion revealed ten groups indicating that the genetic diversity of population was high. The triploid East African highland bananas clearly separated from other triploids. They had the lowest within group genetic diversity. The tetraploids that resulted from crossing EAHB by cv. 'Calcutta 4' and *M*. *acuminata* ssp. *malaccensis* 250 formed their own cluster but were closely linked to that of EAHB, thus supporting the hypothesis that the tetraploids were formed after fusion of unreduced gametes from triploid EAHB and haploid gametes from diploid cv. 'Calcutta 4' and *M*. *acuminata* ssp. *malaccensis* 250. The within cluster dispersion was rather homogenous and not highly diverse for the tetraploid hybrids probably due high allele dosage from EAHB. SSR data suggested that the tetraploid presumed to be hybrids of cv. Enzirabahima by *M*. *a malaccensis* 250 (29275S-1, 29275S-4 and 29275S-5), were in fact admixtures from pollination of EAHB with cv. 'Calcutta 4'. These tetraploid inherited 17 alleles specific for cv. 'Calcutta 4' and none of ssp. *malaccensis* 250 specific alleles across the 18 SSR markers used.

Hierarchical clustering of hybrids was much influenced by male parents used in the cross. The biggest percentage of hybrids was produced from crosses involving tetraploids derived from EAHB and cv. 'Calcutta 4'. Hybrids from ssp. *malaccensis* 250 were more distinct from the rest of the population and formed their own cluster. Four hybrids (26998S-1, 27074S-1, 28506S-1 and 27960s-1) presumed to be progeny of 2180K-6, cv. 'Calcutta 4' and cv. 'Rose' as male parents clustered together with ssp. *malaccensis* 250 hybrids. SSR genotype profiles suggested that those four hybrids were misidentified because they had ssp. *malaccensis* 250 specific alleles. The highest genetic diversity was observed in the diploid parents and between families. Diploids that were linked by pedigree clustered together but the within cluster differences were high compared to EAHB and tetraploids. Diploids such as cv. 'Calcutta 4', 861S-1, 5610S-1, 2180K-1, Kokopo, and cv. 'Rose' clustered with their hybrids. Hybrids derived from 5610S-1 x 2180K-1 were all diploids and closely related to cv. 'Calcutta 4' and 861S-1 and formed a separate cluster. Although the pedigree of 2180K-1 could not be traced, there is a possibility that one of its parents was cv. 'Calcutta 4'. Hybrids from cv. 'Long Tavoy' and cv. 'Calcutta 4' were not easily delineated because of the close resemblance of these genotypes. One cluster (J) comprising of triploid hybrids showed high within cluster diversity. Majority of advanced hybrids especially NARITA hybrids comprising of potential candidate varieties are found in this cluster. The ssp. *zebrina* accessions included in the analysis clustered within the main clusters suggesting their genetic relatedness with other *acuminata* genotypes. In the population, some genotypes were duplicates. The duplicates identified included 28465S-2 (A&B), 26337S-11

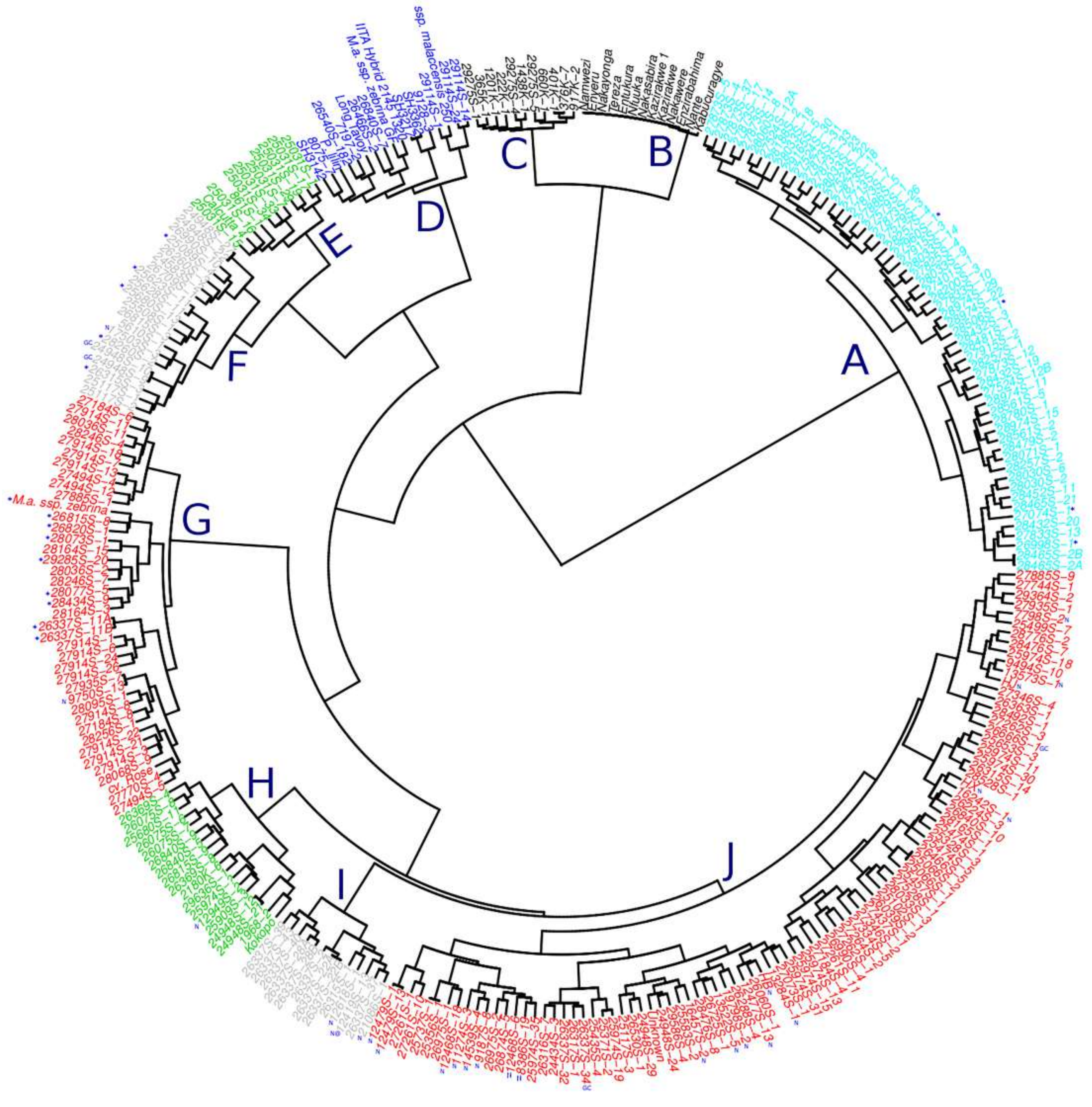


Fig 6. Dendrogram showing the genetic diversity of the genomic selection training population based on 19 informative SSR markers. The squared Euclidean distances were used to generate the hierarchical clusters based on ward.D criterion. Where cluster A = tetraploids (4x) by *M. a. spp. malaccensis* 250, * share only female parent, cluster B = matooke (EAHB), cluster C = tetraploids from EAHB (3x) by Calcutta 4 a wild diploid (2x), cluster D = wild and improved diploids, cluster E = Black Sigatoka resistant diploid hybrids, cluster F = hybrids of 5610S-1 as a male parent, * share grandparent Calcutta 4, GC = good for cooking and N = NARITA hybrid, cluster G = cv. Rose was the main male parent, * share genetic background, cluster H = Long Tavoy and Calcutta 4 are the grandparents, cluster I = mostly hybrids of SH3217 as male parent, N = NARITA, @ = released variety as NARITA 7/M9/ Kiwangazi and cluster J = triploid hybrids with complex pedigree, most advanced hybrids such as NARITAs (N) are found in this cluster of which some are promising variety candidates and GC = good for cooking.

https://doi.org/10.1371/journal.pone.0178734.g006

(A&B) and 26337S-22 (A&B) while 27524S-12 (A&B) that were assumed to be duplicates were clarified to be genetically different although both were progeny of ssp. *malaccensis* 250. Other supposed unique genotypes were identified as likely clonal pairs, such as 24948S-9 and 24948S-10, 24948S-22 and 24948S-27, 25623S-11 and 25628S-11, 24948S-12 and 24948S-21, 12479S-1 and 12479S-13, 25737S-1 and 25356S-1, and 25066S-1 and 25066S-2.

Discussion

Trait evaluation

Bananas express many traits that are used to evaluate hybrids in breeding programs. These traits can be broadly classified as vegetative/agronomic (growth) traits, or yield and consumer appeal (fruit) traits. Growth and yield related traits are used to assess the level of introgression of resistance genes and this is done in the early evaluation trial. The index of non-spotted leaves (INSL) is a measure of resistance to Black Sigatoka, a fungal disease that causes rapid drying of leaves hence reducing the photosynthetic area [7]. Results from ANOVA obtained in this work showed that INSL was not significantly affected by cycle. However, the effect of level of input in field management on INSL depended on genotype. This suggests that resistance to Black Sigatoka might be under strong genetic control and less influenced by cycle.

Correlation analysis showed a positive correlation between INSL, NSLF and YLSF. However, these three had low but significant negative correlations with yield-related traits under low input field management conditions. These results suggest that whereas some Black Sigatoka resistant genotypes give good yield, others produce inferior fruits. Reduction in functional leaves and photosynthetic area has been shown to impact banana yield potential [7]. Tushemereirwe [36] indicated that Black Sigatoka reduced yield of EAHB by more than 30%. Our results show that under high input field management conditions, the impact of the disease on yield traits was less severe (Tables 1 and 2). This result is in agreement with Mobambo et al. [37] who reported that soil fertility had an effect on host plant response to Black Sigatoka and yield in plantains. The symptoms of Black Sigatoka often increase after flowering probably because at that time the ability of a plant to withstand the fungal attack is lowered as it commits most of the energy and resources to the developing inflorescence. Some genotypes had no functional leaves at harvest, indicating that they were very susceptible to Black Sigatoka after flowering. Selection of hybrids based on the number of functional leaves at harvest as a measure of resistance to Black Sigatoka should be done with caution because of the negative association between foliar symptoms to Black Sigatoka and fruit filling.

The present study shows that based on yield and growth traits, four groups of bananas existed in the training population that is, genotypes with high INSL and good fruit filling, high INSL with poor fruit filling, low INSL with good fruit filling and low INSL with poor fruit filling representing the four planes of the two components. However, PCA could not resolve the population structure into clear-cut clusters due to complex pedigrees, although Osuji et al. [38] used this approach to distinguish between different *Musa* triploids. This phenomenon could be attributed to differences in carbon source to sink capacities.

Plant physiological studies have shown that the balance between source and sink translocation of photosynthetic assimilates is key to plant productivity [39]. In bananas, Dens et al. [40] demonstrated the effect of manipulating the carbon source (C-source) and carbon sink (C-sink) of mother plant on ratoon crops in cv. 'Williams' and cv. 'Grand Nain' at a mat level. Their results showed genotype and environmental effect on flowering time, plant height and bunch size for the first ratoon crop. They concluded that the bunch was a larger C-sink than the ratoon crop. At individual plant level, it is likely that difference in C-source to C-sink capacity exists in bananas because our results showed that poor fruit filling genotypes were not significantly affected by cycle and field inputs. It can be postulated that when plants have a strong C-sink capacity they tend to have high yield with increased leaf senescence, while those with low C-sink capacity maintain many leaves with low yield at harvest. More physiological studies in banana are required to shed light on this aspect. It has been reported that at the time of flowering, the fruits and seeds became major sinks and any factor that reduces translocation of photosynthetic assimilates to fruits reduces the harvest index [41].

The training population consisted of poor and good fruit filling genotypes based on FL, FC, FRD and PLD. This characteristic was consistent across cycles and field management, with two overlapping peaks in a binary pattern (S1A Fig). However, given the consistence of the traits under different field conditions, there is likelihood that fruit filling is under control of one or few major-effect quantitative trait loci (QTL). Given that the training population was not a classical bi-parental mapping population this argument may not hold, but investigations using genome wide association studies while accounting for pedigree effect [42] may help to unravel the underlying genetic mechanisms using genome-wide markers such as SNPs.

This study did not find sufficient evidence to show that absolute apical dominance existed in our training population. Different levels of sucker regulation (1–25 suckers) were observed in different cross combinations. This result is in agreement with the observation made by Ortiz and Vuylsteke [43] that non-apical dominance genes were fixed in AA genotypes of *Musa*.

GxE interaction

The effects of cycle and field input management on the genotype and how the genotype interacted with these two aspects of the environment were evaluated. The effect of cross combination was also assessed. Based on coefficients of determination and analysis of variance, genotype, cycle, field and their interactions had different levels of effect on trait variation among cross combinations and individual genotypes. While PHF and PG significantly increased at cycle 2, field management did not have a significant effect on these traits. This could be attributed to the fact that the suckers used were at different physiological maturity. Yield traits were also affected by cycle but the bi-modal distribution was maintained. When bananas are planted in the field they first undergo an establishment phase and build reserves that can accelerate growth of the daughter plants. Therefore, cycle 2 is best to compare genotypes especially with regard to yield traits. Tushemereirwe et al. [16] reported a cycle effect on traits when they analyzed some advanced hybrids, but it was not fully known whether this behavior occurred in different banana genotypes. The effect of cycle alone varied across traits depending on field management except for PED, HTSF and INSL that were most stable. It should however be noted that under optimum field management the cycle explains a small proportion of trait variation in genotypes because most traits had coefficient of determination values below 0.4 in GS2.

The present results show that different banana traits may have different genetic architecture with some traits influenced by GxE. In marker assisted selection this can hamper deployment of classical marker technologies that rely on identifying QTLs. Approaches such as genomic selection that utilize genome-wide markers in complex populations such as in this study provide an opportunity to dissect such traits and could be exploited by banana breeders to increase genetic gain per unit time. Genotype by environment interaction has been shown to affect the accuracy of genomic selection models [24, 44]. Therefore, understanding genotype trait variation across different environments is paramount.

Many hybrids generated from crossbreeding usually have inferior fruit size irrespective of the ploidy level. Such inferiority has been attributed to linkage drag from wild diploids [45].

Bananas have a long selection cycle, they are labor intensive, costly and require large land area for evaluation. Any technology that can discriminate the inferior genotypes from the good ones at a nursery stage could save a lot of resources and time for the breeders thus increasing the breeding efficiency. With the availability of the *Musa* reference genome [46, 47] and decreasing costs of next generation sequencing technologies, high density marker technologies such as genotyping by sequencing are available for many plant species [48]. This provides an opportunity to investigate the application of genomic selection in banana breeding.

Performance of cross combinations

The true breeding value of a genotype is determined by the quality of hybrids produced when it is involved in a cross. By comparing the responses to selection (R) and selection differentials (S) of sixteen cross combinations it was concluded that no single cross combination presented all the good qualities targeted by the breeders in hybrids. This further explains the complex trait variation observed within study population. No attempt was made to determine heritability of the traits because of unbalanced design and the possibility of confounding from heterosis [31]. Some hybrids that had many active leaves at harvest showed variation in fruit filling. Performance of the hybrids was greatly influenced by the male parent involved in the cross. Although both diploids and tetraploids had 50% segregation opportunity, the tetraploids were genetically very similar, whereas the diploids were more genetically diverse with the exception of SH3217 and SH3362 that were closely related. Crosses involving diploid SH3217, SH3362 and 9128–3 produced hybrids which were superior in yield compared to other crosses. These diploids are parthenocarpic, with big fruits and many hands (clusters) per bunch. The best cross combination was C10 (120K-1 x SH3217) that produced hybrids that were fairly resistant to Black Sigatoka, high yielding and quick maturing. Despite the susceptibility of 1201K-1 parent to Black Sigatoka, segregation was observed and some hybrids that had some acceptable levels of resistance were produced.

Tenkouano et al. [49] reported a 4-fold contribution of male parents toward yield traits while Rowe and Rosales [50] highlighted that breeding for improved diploids with pest and disease resistance, parthenocarpy and good yield was the best strategy in banana improvement. Gene pyramiding has also been suggested so that multiple introgressions of good traits are possible [51]. Most of the improved varieties produced by crossbreeding are triploid and all assumed to be completely sterile but no research has been conducted to evaluate their fertility. Further improvement of these triploids is necessary given that no single hybrid has all traits desired by farmers and consumers. The 2x by 2x hybrids were all diploid and some had sizable bunches compared to other diploids in the core breeding set, i.e. could be interesting as improved 2x parents. Further evaluation of these diploids for pollen viability and parthenocarpy will be necessary before they are incorporated in the core breeding set despite their long maturity period. Hybrids that take four months to mature may be considered quick maturing, given that the majority take more than four months.

Genetic diversity of GS training population

Whereas principal component analysis on cross combinations and individual genotypes showed that high genetic diversity existed in the training population, its power to resolve the structure of the population into clear-cut clusters that make biological sense was limited. This was attributed to complex pedigrees in the population with 77 cross combinations represented. The half-sib families were closely related to one another with which they shared a common parent. The population was interconnected due to shared parents in their pedigree. Use of SSR markers proved valuable in delineating the population structure that could be easily interpreted. The set of markers used was reported to be informative and has been used on genotyping the banana collection from the International Transit Center [32]. The polymorphism information content (PIC) of 0.87 was high enough to resolve even the closest genotypes. Up to ten unique clusters were resolved and results showed that clustering was mostly influenced by the genetic diversity in diploid parents.

Triploid EAHB and tetraploids derived from them by crossing with cv. 'Calcutta 4' formed two distinct but closely related clusters, supporting the hypothesis of production of unreduced 3n and reduced n gametes during meiotic events in the tetraploid progenitors [52]. Despite the high PIC of the markers, the EAHB showed a very low genetic diversity consistent with the hypothesis that this group of bananas is an ancient clone set [9]. Even with a high number of polymorphic SSR markers Kitavi et al. and Karamura et al. [9, 53] failed to separate this group into the corresponding phenotype-based clone sets of Karamura [1]. However, some genetic differences were observed between some individual genotypes that could be attributed to mutations within this ancient clone set. The population was predominated with genetic introgression from cv. 'Calcutta 4'. Hybrids from *M*. *acuminata* ssp. *malaccensis* 250 formed a distinct cluster. Three tetraploids presumed to be arising from a cross of EAHB with ssp. *malaccensis* 250 grouped together with those derived from EAHB by cv. 'Calcutta 4'. The presence of Calcutta 4-specific alleles in these tetraploids and the absence of ssp. *malaccensis* 250 specific alleles suggest that these hybrids are progeny of EAHB by cv. 'Calcutta 4' hence the high genetic relationship with the rest of the tetraploids. Nevertheless, these tetraploids should be tested as parents to determine their breeding values so that the breeding genetic pool is expanded.

The SSR markers proved useful in identifying duplicates and closely related genotypes based on pedigree background. A combination of highly polymorphic SSR markers and the power of Ward's clustering method that minimizes the within-group dispersion [34] in the Euclidean space helped to resolve the structure of the population that was highly interlinked by pedigree background. The high level of genetic complexity observed in this population represents different recombination events that make it suitable as a training population for genomic selection.

Apart from obtaining important data on the banana GS training population, important lessons were learned during the course of this work. Dedicated efforts are required to understand the genome organization of bananas through cytological approaches. Ploidy analysis should be routinely employed in breeding programs to differentiate ploidy levels so that different selection criteria are used to select hybrids intended for the breeding pipeline from those eligible for variety release. Despite a majority of the improved hybrids being triploids, their fertility should be tested so that further improvements can be made on them as a way to achieve gene pyramiding while minimizing inbreeding.

Conclusion

The response of genotype trait expression to cycle and field management practices varied greatly. The largest proportion of genetic variation was due to the greater genetic diversity of male parents used in crosses since the tetraploids used in the majority of crosses as female parents were genetically related. Yield traits accounted for 31–35% of the total principal component variation observed in the population and were loaded on the first component while vegetative traits contributed to the second component with 15–22%. A high level of correlation within vegetative- and yield-related traits was observed but correlation between vegetative and yield traits was low and depended on the interaction with field management practices. Therefore, genomic selection models could be developed for traits that are easy to measure. It is likely that the predictive ability of traits that are difficult to phenotype will be similar to traits easily measured but highly correlated. The study population was observed to be genetically diverse with complex pedigree structure. Yield-related traits showed a bi-modal distribution, which was not influenced by cycle or field management. Resistance to Black Sigatoka was also stable across cycles but varied under different field management depending on the genotype. Principal component analysis could not delineate this complex population structure but the application of SSR markers in combination with Ward's hierarchical clustering proved powerful and resolved the structure into biologically meaningful groups.

Supporting information

S1 Fig. Variation in fruit characteristics. (A) is a histogram showing the bimodal distribution of fruit circumference (FC), (B) cross sections of poor filling fruits, (C) good filling fruits with fruit diameter (FRD) and pulp diameter (PLD) values in cm, and (D) poor filling and good filling banana fruits.
(TIF)

S1 Table. List of genotypes in a banana genomic selection training population.
(DOCX)

S2 Table. Data used to calculate selection differential and response to selection for the sixteen cross combinations.
(XLSX)

S3 Table. Summary of all trait variations in response to cycle and field management.
(DOCX)

Acknowledgments

We thank the field staff of IITA-Uganda banana-breeding program for helping in data collection, Mrs. Marie Seifertová from the Institute of Experimental Botany, Olomouc, Czech Republic for the technical service in SSR fragment analysis and NARO, Uganda for providing some of the genotypes used in this study.

Author Contributions

Conceptualization: JL MN JD.

Data curation: MN MB PC.

Formal analysis: MN PC.

Funding acquisition: JL RS BU JD.

Investigation: MN MB PC.

Methodology: MN MB PC.

Project administration: JD RS.

Resources: JD EH.

Software: PC.

Supervision: JL JD RS.

Validation: JD BU RS EH.

Writing – original draft: MN.

Writing – review & editing: BU JD AB PC RS MB EH JL.

References

1. Karamura D. Numerical taxonomic studies of the East African highland bananas (Musa AAA- East Africa) in Uganda. Doctor of Philosophy Thesis, The University of Reading. 1998;1–192.
2. Karamura DA, Karamura E, Tinzaara W. editors. Banana cultivar names, synonyms and their usage in East Africa. Bioversity International, Uganda; 2012.
3. Ortiz R, Swennen R. Review: From crossbreeding to biotechnology-facilitated improvement of banana and plantain. Biotechnol Adv. 2014; 32: 158–169. https://doi.org/10.1016/j.biotechadv.2013.09.010 PMID: 24091289
4. Perrier X, De Langhe E, Donohue M, Lentfer C, Vrydaghs L, Bakry F, et al. Multi-disciplinary perspectives on banana (*Musa* spp.) domestication. Proc Natl Acad Sci USA. 2011; 108(28): 11311–11318. https://doi.org/10.1073/pnas.1102001108 PMID: 21730145
5. Van Asten P, Fermont AM, Taulya G. Drought is a major yield loss factor for rainfed East African highland banana. Elsevier, Agric Water Manag. 2011; 98: 541–552.
6. Lorenzen J, Tenkouano A, Bandyopadhyay R, Vroh B, Coyne D, Tripathi L. Overview of Banana and Plantain (*Musa* spp.) Improvement in Africa: Past and Future. Acta Hortic. 2010; 879: 595–603.
7. Barekye A. Breeding investigations for black Sigatoka resistance and associated traits in diploids, tetraploids and the triploid progenies of bananas in Uganda. Thesis: Doctor of Philosophy (PhD) in Plant Breeding, University of KwaZulu-Natal, Republic of South Africa. 2009:1–230.
8. Gold CS, Kagezi GH, Night G, Ragama P. The effects of banana weevil, *Cosmopolites sordidus*, damage on highland banana growth, yield and stand duration in Uganda. Ann Appl Biol. 2004; 145: 263–269.
9. Kitavi M, Downing T, Lorenzen J, Karamura D, Onyango M, Nyine M, et al. The triploid East African Highland Banana (EAHB) genepool is genetically uniform arising from a single ancestral clone that underwent population expansion by vegetative propagation. Theor Appl Genet. 2016. https://doi.org/10.1007/s00122-015-2647-1 PMID: 26743524
10. Pillay M, Ogundiwin E, Nwakanma D, Ude G, Tenkouano A. Analysis of genetic diversity and relationships in East African banana germplasm. Theor Appl Genet. 2001; 102(6–7): 965–970.
11. Simmonds NW. Bananas, Musa cvs. In: Simmonds NW, editor. Breeding for durable resistance in perennial crops. FAO Technical Papers 70. Food and Agriculture Organization, Rome; 1986; pp. 17–24.
12. Rowe PR. Breeding bananas and plantains for resistance to fusarial wilt: the track record. In: Ploetz RC, editor. Fusarium wilt of bananas. APS, St. Paul, MN. 1990; pp. 115–119
13. Pillay M, Tenkouano A. Genome, cytogenetics, and flow cytometry of *Musa*. In: Pillay M, Tenkouano A, editors. Banana breeding: progress and challenges. CRC Press, Boca Raton. 2011; pp. 53–70.
14. Ssebuliba R, Talengera D, Makumbi D, Namanya P, Tenkouano A, Tushemereirwe W, et al. Reproductive efficiency and breeding potential of East African highland (*Musa* AAA-EA) bananas. Field Crops Res. 2006; 95: 250–255.
15. Vuylsteke D, Swennen R, Ortiz R. Development and performance of black sigatoka resistant tetraploid hybrids of plantain (*Musa* spp., AAB group). Euphytica. 1993; 65: 33–42.
16. Tushemereirwe W, Batte M, Nyine M, Tumuhimbise R, Barekye A, Ssali T, et al. Performance of NARITA hybrids in the preliminary yield trial for three cycles in Uganda. IITA, NARO, Uganda; 2015.
17. Desta ZA, Ortiz R. Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 2014; 19(9): 592–601. https://doi.org/10.1016/j.tplants.2014.05.006 PMID: 24970707
18. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157: 1819–1829. PMID: 11290733
19. Pérez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 2014; 198: 483–495. https://doi.org/10.1534/genetics.114.164442 PMID: 25009151
20. Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, et al. Genomic selection in plant breeding: knowledge and prospects. Adv Agron. 2011; 110: 77–123.
21. Nakaya A, Isobe SN. Will genomic selection be a practical method for plant breeding? Ann Bot. 2012; 110: 1303–1316. https://doi.org/10.1093/aob/mcs109 PMID: 22645117
22. Togashi K, Lin CY, Yamazaki T. The efficiency of genome-wide selection for genetic improvement of net merit. J Anim Sci. 2011; 89: 2972–2980. https://doi.org/10.2527/jas.2009-2606 PMID: 21512116

23. Goddard ME, Hayes BJ. Review article: Genomic selection. J Anim Breed Genet. 2007; 124: 323–330. https://doi.org/10.1111/j.1439-0388.2007.00702.x PMID: 18076469

24. Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J. et al. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity. 2014; 112: 48–60. https://doi.org/10.1038/hdy.2013.16 PMID: 23572121

25. Beaulieu J, Doerksen T, Clément S, MacKay J, Bousquet J. Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. Heredity. 2014; 113: 343–352. https://doi.org/10.1038/hdy.2014.36 PMID: 24781808

26. Onogi A, Watanabe M, Mochizuki T, Hayashi T, Nakagawa H, Hasegawa T, et al. Toward integration of genomic selection with crop modelling: the development of an integrated approach to predicting rice heading dates. Theor Appl Genet. 2016; 129: 805–817. https://doi.org/10.1007/s00122-016-2667-5 PMID: 26791836

27. De Oliveira EJ, de Resende MDV, Santos VS, Ferreira CF, Oliveira GAF, da Silva MS, et al. Genome-wide selection in cassava. Euphytica 2012.

28. Doležel J, Lysák MA, Van den Houwe I, Doleželová M, Roux N. Use of flow cytometry for rapid ploidy determination in *Musa* species. New Methods, INFOMUSA. 1997; 6(1): 6–9.

29. Doležel J. Flow cytometric analysis of nuclear DNA content in higher plants. Phytochem Anal. 1991; 2: 143–154.

30. Craenen K. Black Sigatoka disease of banana and plantain: A reference manual, International Institute of Tropical Agriculture, Nigeria, Ibadan. Host plant response to black sigatoka. 1998; pp. 41–45.

31. Piepho H-P, Möhring J. Computing Heritability and Selection Response from Unbalanced Plant Breeding Trials. Genetics. 2007; 177: 1881–1888. https://doi.org/10.1534/genetics.107.074229 PMID: 18039886

32. Christelová P, Valárik M, Hřibová, Van den houwe I, Channelière S, Roux N, et al. A platform for efficient genotyping in *Musa* using microsatellite markers. AoB Plants. 2011. https://doi.org/10.1093/aobpla/plr024 PMID: 22476494

33. Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963; 58: 236–244.

34. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J Classif. 2014; 31: 274–295.

35. Liu K, Muse SV. PowerMarker: integrated analysis environment for genetic marker data. Bioinformatics. 2005; 21(9): 2128–2129. https://doi.org/10.1093/bioinformatics/bti282 PMID: 15705655

36. Tushemereirwe WK. Factors influencing the expression of leaf spot diseases of highland bananas in Uganda. PhD thesis. University of Reading, U.K; 1996.

37. Mobambo K, Zoufa K, Gauhl F, Adeniji M, Pasberg-Gauhl C. Effect of soil fertility on host response to black leaf streak of plantain (*Musa* spp., AAB group) under traditional farming systems in southeastern Nigeria. Int J Pest Manag. 1994; 40: 75–80

38. Osuji JO, Okol BE, Vuylsteke D, Ortiz R. Multivariate pattern of quantitative trait variation in triploid banana and plantain cultivars. Sci Hortic. 1997; 71: 197–202.

39. Lemoine R, La Camera S, Atanassova R, Dédaldéchamp F, Allario T, Pourtau N, et al. Source-to-sink transport of sugar and regulation by environmental factors. Front Plant Sci. 2013; 4: 1–21.

40. Dens KR, Romero RA, Swennen R, Turner DW. Removal of bunch, leaves, or pseudostem alone, or in combination, influences growth and bunch weight of ratoon crops in two banana cultivars. J Hortic Sci Biotechnol. 2008; 83(1): 113–119.

41. Wardlaw IF. Tansley Review No. 27—The control of carbon partitioning in plants. New Phytol. 1990; 116: 341–381.

42. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013; 9: 29. https://doi.org/10.1186/1746-4811-9-29 PMID: 23876160

43. Ortiz R, Vuylsteke DR. Genetics of Apical Dominance in Plantain (*Musa* spp., AAB Group) and Improvement of Suckering Behavior. J Am Soc Hortic Sci. 1994; 119(5): 1050–1053.

44. Burgueño J, de los Campos G, Weigel K, Crossa J. Genomic prediction of breeding values when modeling genotype x environment interaction using pedigree and dense molecular markers. Crop Sci. 2012; 52: 707–719.

45. Lorenzen J, Tenkouano A, Bandyopadhyay R, Vroh I, Coyne D, Tripathi L. The role of crop improvement in pest and disease management. Acta Hortic. 2009; 828: 305–314.

46. Martin G, Baurens FC, Droc G, Rouard M, Cenci A, Kilian A. et al. Improvement of the banana "*Musa acuminata*" reference sequence using NGS data and semi-automated bioinformatics methods. BMC Genomics. 2016; 17: 243. https://doi.org/10.1186/s12864-016-2579-4 PMID: 26984673

47. D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, et al. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature. 2012; 488: 213–219. https://doi.org/10.1038/nature11241 PMID: 22801500

48. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. PLoS ONE. 2012; 6(5): e19379. https://doi.org/10.1371/journal.pone.0019379 PMID: 21573248

49. Tenkouano A, Ortiz R, Vuylsteke D. Combining ability for yield and plant phenology in plantain derived populations. Euphytica. 1998; 104: 151–158.

50. Rowe P, Rosales F. Diploid breeding at FHIA and the development of Goldfinger. InfoMusa. 1993; 2: 9–11.

51. Pillay M, Ude G, Kole C, editors. Genetics, Genomics, and Breeding of Bananas. Science Publishers; 2012. pp. 116–123.

52. Ssali RT, Nawankunda K, Erima RB, Batte M, Tushemereirwe WK. On-Farm Participatory Evaluation of East African Highland Banana 'Matooke' Hybrids (*Musa* spp.). Acta Hortic. 2010; 879: 585–591.

53. Karamura D, Kitavi M, Nyine M, Ochola D, Muhangi S, Talengera D, et al. Genotyping the local banana landrace groups of East Africa. Acta Hortic. 2016; 1114: 67–73.

S1 Table. List of genotypes in the genomic selection training population

S/No	Cross	Genotype name	Female parent	Male parent	Ploidy	Description
1		Enzirabahima			3x	Parent
2		Kabucuragye			3x	Parent
3		Tereza			3x	Parent
4		Enyeru			3x	Parent
5		Nakayonga			3x	Parent
6		Namwezi			3x	Parent
7		Entukura			3x	Parent
8		Nakasabira			3x	Parent
9		Nakawere			3x	Parent
10		Nante			3x	Parent
11		Kazirakwe			3x	Parent
12		Nfuuka			3x	Parent
13		Calcutta 4			2x	Parent
14	C45	1201K-1	Nakawere	Calcutta 4	4x	Parent
15	C41	917K-2	Enzirabahima	Calcutta 4	4x	Parent
16	C41	660K-1	Enzirabahima	Calcutta 4	4x	Parent
17	C40	1438K-1	Entukura	Calcutta 4	4x	Parent
18	C51	222K-1	Nfuuka	Calcutta 4	4x	Parent
19	C49	376K-7	Nante	Calcutta 4	4x	Parent
20	C67	365K-1	Kabucuragye	Calcutta 4	4x	Parent
21	C40	401K-1	Entukura	Calcutta 4	4x	Parent
22	C66	2180K-6			2x	Parent
23	C53	8075-7	SH3362	Calcutta 4	2x	Parent
24	C54	7197-2	SH3362	Long Tavoy	2x	Parent
25	C63	SH3142	SH1734	Pisang Jari Buaya	2x	Parent
26	C64	SH3362	SH3217	SH3142	2x	Parent
27	C52	SH3217	SH2095	SH2766	2x	Parent
28	C43	5610S-1	Kabucuragye	7197-2	2x	Parent
29	C65	9128-3	Tjau lagada	Pisang lilin	2x	Parent
30	C57	1968-2	Who-gu	Calcutta 4	3x	Parent
31	C48	861S-1	Namwezi	Calcutta 4	2x	Parent
32		cv. Rose			2x	Parent
33		Pisang Lilin			2x	Parent
34		Kokopo			2x	Parent
35		Long Tavoy			2x	Parent
36		*malaccensis* 250			2x	Parent
37	C01	28165S-1	1201K-1	1968-2	3x	Hybrid
38	C02	25583S-2	1201K-1	5610S-1	3x	Hybrid
39	C02	26660S-1	1201K-1	5610S-1	3x	Hybrid
40	C02	28434S-9	1201K-1	5610S-1	3x	Hybrid
41	C68	17503S-3	1201K-1	7197-2	3x	Hybrid
42	C03	16242S-1	1201K-1	8075-7	3x	Hybrid
43	C04	12479S-1	1201K-1	9128-3	3x	Hybrid

44	C04	12479S-13	1201K-1	9128-3	3x	Hybrid
45	C04	26317S-1	1201K-1	9128-3	3x	Hybrid
46	C04	27262S-1	1201K-1	9128-3	3x	Hybrid
47	C04	27262S-3	1201K-1	9128-3	3x	Hybrid
48	C05	27770S-20	1201K-1	cv. Rose	3x	Hybrid
49	C05	27770S-4	1201K-1	cv. Rose	3x	Hybrid
50	C05	27935S-1	1201K-1	cv. Rose	3x	Hybrid
51	C05	27960S-1	1201K-1	cv. Rose	3x	Hybrid
52	C05	28036S-11	1201K-1	cv. Rose	3x	Hybrid
53	C05	28036S-2	1201K-1	cv. Rose	3x	Hybrid
54	C05	28164S-3	1201K-1	cv. Rose	3x	Hybrid
55	C05	28246S-4	1201K-1	cv. Rose	3x	Hybrid
56	C05	28246S-7	1201K-1	cv. Rose	3x	Hybrid
57	C05	27935S-7	1201K-1	cv. Rose	3x	Hybrid
58	C06	26363S-1	1201K-1	Kokopo	3x	Hybrid
59	C07	26075S-6	1201K-1	Long Tavoy	3x	Hybrid
60	C07	26075S-7	1201K-1	Long Tavoy	3x	Hybrid
61	C07	26075S-8	1201K-1	Long Tavoy	3x	Hybrid
62	C08	27346S-2	1201K-1	*malaccensis* 250	3x	Hybrid
63	C08	27346S-4	1201K-1	*malaccensis* 250	3x	Hybrid
64	C08	27437S-1	1201K-1	*malaccensis* 250	3x	Hybrid
65	C08	27579S-1	1201K-1	*malaccensis* 250	3x	Hybrid
66	C08	27579S-3	1201K-1	*malaccensis* 250	3x	Hybrid
67	C08	28030S-2	1201K-1	*malaccensis* 250	3x	Hybrid
68	C08	28030S-6	1201K-1	*malaccensis* 250	3x	Hybrid
69	C08	28071S-1	1201K-1	*malaccensis* 250	3x	Hybrid
70	C08	28465S-2	1201K-1	*malaccensis* 250	3x	Hybrid
71	C08	28465S-21	1201K-1	*malaccensis* 250	3x	Hybrid
72	C08	28479S-2	1201K-1	*malaccensis* 250	3x	Hybrid
73	C10	26337S-22A	1201K-1	SH3217	3x	Hybrid
74	C10	26337S-40	1201K-1	SH3217	3x	Hybrid
75	C11	26840S-7	1201K-1	SH3362	2x	Hybrid
76	C09	26315S-1	1201K-1	SH3142	3x	Hybrid
77	C10	12419S-13	1201K-1	SH3217	3x	Hybrid
78	C10	26337S-11A	1201K-1	SH3217	3x	Hybrid
79	C10	26337S-2	1201K-1	SH3217	3x	Hybrid
80	C10	26337S-34	1201K-1	SH3217	3x	Hybrid
81	C10	26337S-37	1201K-1	SH3217	3x	Hybrid
82	C10	26337S-39	1201K-1	SH3217	3x	Hybrid
83	C10	26337S-43	1201K-1	SH3217	3x	Hybrid
84	C10	28263S-2	1201K-1	SH3217	3x	Hybrid
85	C11	12618S-1	1201K-1	SH3362	3x	Hybrid
86	C11	26316S-7	1201K-1	SH3362	3x	Hybrid
87	C11	26840S-10	1201K-1	SH3362	3x	Hybrid
88	C58	25328S-3	1438K-1	1537K-1	3x	Hybrid

89	C12	24948S-10	1438K-1	5610S-1	3x	Hybrid
90	C12	24948S-13	1438K-1	5610S-1	3x	Hybrid
91	C12	24948S-24	1438K-1	5610S-1	3x	Hybrid
92	C12	24948S-9	1438K-1	5610S-1	3x	Hybrid
93	C69	26060S-1	1438K-1	9128-3	3x	Hybrid
94	C70	13573S-1	1438K-1	9719-7	3x	Hybrid
95	C13	27914S-1	1438K-1	cv. Rose	3x	Hybrid
96	C13	27914S-13	1438K-1	cv. Rose	3x	Hybrid
97	C13	28095S-1	1438K-1	cv. Rose	3x	Hybrid
98	C13	27264S-2	1438K-1	cv. Rose	2x	Hybrid
99	C13	27914S-24	1438K-1	cv. Rose	3x	Hybrid
100	C13	27914S-26	1438K-1	cv. Rose	3x	Hybrid
101	C13	27914S-3	1438K-1	cv. Rose	3x	Hybrid
102	C14	25474S-1	1438K-1	Kokopo	3x	Hybrid
103	C15	26369S-4	1438K-1	Long Tavoy	3x	Hybrid
104	C16	28481S-1	1438K-1	*malaccensis* 250	3x	Hybrid
105	C16	28561S-2	1438K-1	*malaccensis* 250	3x	Hybrid
106	C19	26725S-1	1438K-1	SH3362	3x	Hybrid
107	C17	25499S-7	1438K-1	SH3142	3x	Hybrid
108	C18	26039S-2	1438K-1	SH3217	3x	Hybrid
109	C20	26466S-2	1977K-1	5610S-1	3x	Hybrid
110	C20	26466S-5	1977K-1	5610S-1	3x	Hybrid
111	C71	22598S-2	365K-1	1201K-1	3x	Hybrid
112	C59	14539S-4	365K-1	660K-1	3x	Hybrid
113	C21	9750S-13	401K-1	9128-3	3x	Hybrid
114	C22	25031S-1	5610S-1	2180K-6	2x	Hybrid
115	C22	25031S-15	5610S-1	2180K-6	2x	Hybrid
116	C22	25031S-16	5610S-1	2180K-6	2x	Hybrid
117	C22	25031S-17	5610S-1	2180K-6	2x	Hybrid
118	C22	25031S-19	5610S-1	2180K-6	2x	Hybrid
119	C22	25031S-27	5610S-1	2180K-6	2x	Hybrid
120	C22	25031S-33	5610S-1	2180K-6	2x	Hybrid
121	C22	25031S-34	5610S-1	2180K-6	2x	Hybrid
122	C22	25031S-7	5610S-1	2180K-6	2x	Hybrid
123	C24	24583S-2	660K-1	5610S-1	3x	Hybrid
124	C24	26260S-3	660K-1	5610S-1	3x	Hybrid
125	C25	13284S-1	660K-1	9128-3	3x	Hybrid
126	C25	25371S-2	660K-1	9128-3	3x	Hybrid
127	C25	9187S-8	660K-1	9128-3	3x	Hybrid
128	C26	26709S-1	660K-1	Calcutta 4	3x	Hybrid
129	C27	27713S-1	660K-1	*malaccensis* 250	3x	Hybrid
130	C27	27825S-4	660K-1	*malaccensis* 250	3x	Hybrid
131	C27	27873S-18	660K-1	*malaccensis* 250	3x	Hybrid
132	C27	27873S-38	660K-1	*malaccensis* 250	3x	Hybrid
133	C27	27873S-4	660K-1	*malaccensis* 250	3x	Hybrid

134	C27	27873S-5	660K-1	*malaccensis* 250	3x	Hybrid
135	C27	28188S-2	660K-1	*malaccensis* 250	3x	Hybrid
136	C28	25623S-11	8817S-1	917K-2	3x	Hybrid
137	C29	28492S-1	917K-2	1968-2	3x	Hybrid
138	C30	26998S-1	917K-2	2180K-6	3x	Hybrid
139	C30	27074S-1	917K-2	2180K-6	3x	Hybrid
140	C31	25117S-1	917K-2	5610S-1	3x	Hybrid
141	C31	25117S-2	917K-2	5610S-1	3x	Hybrid
142	C31	25117S-3	917K-2	5610S-1	3x	Hybrid
143	C31	25508S-1	917K-2	5610S-1	3x	Hybrid
144	C31	25628S-11	917K-2	5610S-1	3x	Hybrid
145	C31	26815S-3	917K-2	5610S-1	3x	Hybrid
146	C31	26815S-8	917K-2	5610S-1	3x	Hybrid
147	C31	26815S-9	917K-2	5610S-1	3x	Hybrid
148	C31	26990S-10	917K-2	5610S-1	3x	Hybrid
149	C31	26990S-11	917K-2	5610S-1	3x	Hybrid
150	C31	26990S-4	917K-2	5610S-1	3x	Hybrid
151	C31	27073S-1	917K-2	5610S-1	3x	Hybrid
152	C31	27744S-1	917K-2	5610S-1	3x	Hybrid
153	C60	12949S-2	917K-2	7197-2	3x	Hybrid
154	C60	25909S-3	917K-2	7197-2	3x	Hybrid
155	C32	25089S-4	917K-2	861S-1	3x	Hybrid
156	C33	19798S-2	917K-2	9128-3	3x	Hybrid
157	C33	24434S-3	917K-2	9128-3	3x	Hybrid
158	C33	25435S-11	917K-2	9128-3	3x	Hybrid
159	C33	25435S-4	917K-2	9128-3	3x	Hybrid
160	C33	25737S-1	917K-2	9128-3	3x	Hybrid
161	C33	26288S-4	917K-2	9128-3	3x	Hybrid
162	C33	26975S-1	917K-2	9128-3	3x	Hybrid
163	C33	26975S-2	917K-2	9128-3	3x	Hybrid
164	C33	7798S-2	917K-2	9128-3	3x	Hybrid
165	C34	27184S-4	917K-2	cv. Rose	3x	Hybrid
166	C34	27885S-9	917K-2	cv. Rose	3x	Hybrid
167	C34	27184S-8	917K-2	cv. Rose	3x	Hybrid
168	C34	27494S-12	917K-2	cv. Rose	3x	Hybrid
169	C34	27494S-4	917K-2	cv. Rose	3x	Hybrid
170	C34	27494S-5	917K-2	cv. Rose	3x	Hybrid
171	C34	28068S-9	917K-2	cv. Rose	3x	Hybrid
172	C34	27184S-6	917K-2	cv. Rose	3x	Hybrid
173	C34	27885S-1	917K-2	cv. Rose	3x	Hybrid
174	C35	24410S-2	917K-2	Kokopo	3x	Hybrid
175	C36	25680S-11	917K-2	Long Tavoy	3x	Hybrid
176	C36	25680S-13	917K-2	Long Tavoy	3x	Hybrid
177	C37	27261S-1	917K-2	*malaccensis* 250	3x	Hybrid
178	C37	27261S-10	917K-2	*malaccensis* 250	3x	Hybrid

179	C37	27261S-11	917K-2	*malaccensis* 250	3x	Hybrid
180	C37	27334S-5	917K-2	*malaccensis* 250	3x	Hybrid
181	C37	27401S-1	917K-2	*malaccensis* 250	3x	Hybrid
182	C37	27524S-12A	917K-2	*malaccensis* 250	3x	Hybrid
183	C37	27524S-12B	917K-2	*malaccensis* 250	3x	Hybrid
184	C37	27524S-22	917K-2	*malaccensis* 250	3x	Hybrid
185	C37	27524S-30	917K-2	*malaccensis* 250	3x	Hybrid
186	C37	27833S-10	917K-2	*malaccensis* 250	3x	Hybrid
187	C37	27833S-13	917K-2	*malaccensis* 250	3x	Hybrid
188	C37	27886S-5	917K-2	*malaccensis* 250	3x	Hybrid
189	C37	28033S-14	917K-2	*malaccensis* 250	3x	Hybrid
190	C37	28033S-15	917K-2	*malaccensis* 250	3x	Hybrid
191	C37	28033S-18	917K-2	*malaccensis* 250	3x	Hybrid
192	C37	28033S-23	917K-2	*malaccensis* 250	3x	Hybrid
193	C37	28033S-3	917K-2	*malaccensis* 250	3x	Hybrid
194	C37	28060S-8	917K-2	*malaccensis* 250	3x	Hybrid
195	C37	28200S-3	917K-2	*malaccensis* 250	3x	Hybrid
196	C37	28257S-1	917K-2	*malaccensis* 250	3x	Hybrid
197	C37	28257S-2	917K-2	*malaccensis* 250	3x	Hybrid
198	C37	28257S-4	917K-2	*malaccensis* 250	3x	Hybrid
199	C37	28432S-19	917K-2	*malaccensis* 250	3x	Hybrid
200	C37	28432S-20	917K-2	*malaccensis* 250	3x	Hybrid
201	C37	28432S-3	917K-2	*malaccensis* 250	3x	Hybrid
202	C37	28780S-1	917K-2	*malaccensis* 250	3x	Hybrid
203	C61	26874S-5	917K-2	SH3362	3x	Hybrid
204	C38	12468S-18	917K-2	SH3217	3x	Hybrid
205	C38	12477S-13	917K-2	SH3217	3x	Hybrid
206	C38	8386S-19	917K-2	SH3217	3x	Hybrid
207	C61	13522S-5	917K-2	SH3362	3x	Hybrid
208	C61	25974S-?	917K-2	SH3362	3x	Hybrid
209	C61	25974S-19	917K-2	SH3362	3x	Hybrid
210	C61	25974S-21	917K-2	SH3362	3x	Hybrid
211	C61	25974S-30	917K-2	SH3362	3x	Hybrid
212	C61	25974S-35	917K-2	SH3362	3x	Hybrid
213	C61	26666S-1	917K-2	SH3362	3x	Hybrid
214	C61	28476S-7	917K-2	SH3362	3x	Hybrid
215	C61	9494S-10	917K-2	SH3362	3x	Hybrid
216	C62	16457S-2	Entukura	365K-1	3x	Hybrid
217	C39	26540S-182	Entukura	8075-7	2x	Hybrid
218	C41	28260S-2	Enzirabahima	Calcutta 4	3x	Hybrid
219	C72	21086S-1	Kazirakwe	7197-2	3x	Hybrid
220	C46	28073S-1	Namwezi	7197-2	3x	Hybrid
221	C55	25356S-1	Tereza	7197-2	3x	Hybrid
222	C75	HB	unknown	unknown	3x	Hybrid
223	C76	HJ	unknown	unknown	3x	Hybrid

224	C77	HX	unknown	unknown	3x	Hybrid
225	C10	26337S-11B	1201K-1	SH3217	3x	Hybrid
226	C73	16285S-13	Calcutta 4	660K-1	2x	Hybrid
227	C10	26337S-22B	1201K-1	SH3217	3x	Hybrid
228	C73	16285S-3	Calcutta 4	660K-1	2x	Hybrid
229	C10	26337S-28	1201K-1	SH3217	3x	Hybrid
230	C14	25066S-1	1438K-1	Kokopo	3x	Hybrid
231	C73	16285S-6	Calcutta 4	660K-1	2x	Hybrid
232	C14	25066S-2	1438K-1	Kokopo	3x	Hybrid
233	C73	16285S-8	Calcutta 4	660K-1	2x	Hybrid
234	C61	25974S-11	917K-2	SH3362	3x	Hybrid
235	C61	25974S-15	917K-2	SH3362	3x	Hybrid
236	C14	25457S-1	1438K-1	Kokopo	3x	Hybrid
237	C74	16191S-6	Calcutta 4	917K-2	2x	Hybrid
238	C35	24797S-7	917K-2	Kokopo	3x	Hybrid
239	C35	25102S-1	917K-2	Kokopo	3x	Hybrid
240	C44	28452S-11	Nakasabira	Calcutta 4	3x	Hybrid
241	C37	28033S-9	917K-2	*malaccensis* 250	3x	Hybrid
242	C61	25974S-13	917K-2	SH3362	3x	Hybrid
243	C34	28256S-1	917K-2	cv. Rose	3x	Hybrid
244	C61	25974S-17	917K-2	SH3362	4x	Hybrid
245	C38	12468S-6	917K-2	SH3217	3x	Hybrid
246	C13	27914S-11	1438K-1	cv. Rose	3x	Hybrid
247	C13	27914S-18	1438K-1	cv. Rose	3x	Hybrid
248	C13	27914S-21	1438K-1	cv. Rose	3x	Hybrid
249	C13	27914S-22	1438K-1	cv. Rose	3x	Hybrid
250	C13	27914S-6	1438K-1	cv. Rose	3x	Hybrid
251	C13	27914S-7	1438K-1	cv. Rose	3x	Hybrid
252	C13	27914S-8	1438K-1	cv. Rose	3x	Hybrid
253	C27	27873S-12	660K-1	*malaccensis* 250	3x	Hybrid
254	C27	27873S-14	660K-1	*malaccensis* 250	3x	Hybrid
255	C27	27873S-17	660K-1	*malaccensis* 250	3x	Hybrid
256	C27	27873S-33	660K-1	*malaccensis* 250	3x	Hybrid
257	C27	27873S-37	660K-1	*malaccensis* 250	3x	Hybrid
258	C27	27873S-7	660K-1	*malaccensis* 250	3x	Hybrid
259	C11	26224S-3	1201K-1	SH3362	3x	Hybrid
260	C11	26840S-9	1201K-1	SH3362	3x	Hybrid
261	C11	26316S-14	1201K-1	SH3362	3x	Hybrid
262	C11	26224S-2	1201K-1	SH3362	3x	Hybrid
263	C11	26840S-5	1201K-1	SH3362	3x	Hybrid
264	C09	25653S-3	1201K-1	SH3142	3x	Hybrid
265	C09	26315S-3	1201K-1	SH3142	3x	Hybrid
266	C06	28528S-1	1201K-1	Kokopo	3x	Hybrid
267	C15	26369S-8	1438K-1	Long Tavoy	3x	Hybrid
268	C19	26530S-1	1438K-1	SH3362	3x	Hybrid

269	C16	27528S-1	1438K-1	*malaccensis* 250	3x	Hybrid
270	C16	27915S-3	1438K-1	*malaccensis* 250	3x	Hybrid
271	C16	28561S-5	1438K-1	*malaccensis* 250	3x	Hybrid
272	C16	27915S-2	1438K-1	*malaccensis* 250	3x	Hybrid
273	C16	28974S-11	1438K-1	*malaccensis* 250	3x	Hybrid
274	C16	28974S-15	1438K-1	*malaccensis* 250	3x	Hybrid
275	C16	28974S-22	1438K-1	*malaccensis* 250	3x	Hybrid
276	C16	28974S-29	1438K-1	*malaccensis* 250	3x	Hybrid
277	C23	29114S-14A	5610S-1	*malaccensis* 250	2x	Hybrid
278	C23	29114S-14B	5610S-1	*malaccensis* 250	3x	Hybrid
279	C23	29114S-19	5610S-1	*malaccensis* 250	3x	Hybrid
280	C23	29114S-24	5610S-1	*malaccensis* 250	3x	Hybrid
281	C27	27873S-26	660K-1	*malaccensis* 250	3x	Hybrid
282	C27	27873S-31	660K-1	*malaccensis* 250	3x	Hybrid
283	C27	29165S-5	660K-1	*malaccensis* 250	3x	Hybrid
284	C40	28506S-1	Entukura	Calcutta 4	3x	Hybrid
285	C47	29364S-2	Namwezi	cv. Rose	4x	Hybrid
286	C50	28077S-5	Nfuuka	8075-7	3x	Hybrid
287	C05	28164S-15	1201K-1	cv. Rose	3x	Hybrid
288	C05	29285S-20	1201K-1	cv. Rose	3x	Hybrid
289	C10	26337S-32	1201K-1	SH3217	3x	Hybrid
290	C11	27684S-5	1201K-1	SH3362	3x	Hybrid
291	C12	24948S-12	1438K-1	5610S-1	3x	Hybrid
292	C12	24948S-21	1438K-1	5610S-1	3x	Hybrid
293	C12	24948S-27	1438K-1	5610S-1	3x	Hybrid
294	C12	29586S-4	1438K-1	5610S-1	3x	Hybrid
295	C12	24948S-22	1438K-1	5610S-1	3x	Hybrid
296	C12	24948S-2	1438K-1	5610S-1	3x	Hybrid
297	C12	24948S-29	1438K-1	5610S-1	3x	Hybrid
298	C29	26820S-1	917K-2	1968-2	3x	Hybrid
299	C32	25474S-5	917K-2	861S-1	3x	Hybrid
300	C61	25974S-18	917K-2	SH3362	3x	Hybrid
301	C61	28476S-8	917K-2	SH3362	3x	Hybrid
302	C61	25974S-31	917K-2	SH3362	3x	Hybrid
303	C42	29275S-1	Enzirabahima	*malaccensis* 250	4x	Hybrid
304	C42	29275S-4	Enzirabahima	*malaccensis* 250	4x	Hybrid
305	C42	29275S-5	Enzirabahima	*malaccensis* 250	4x	Hybrid
306	C55	29636S-1	Tereza	7197-2	4x	Hybrid
307	C56	28776S-2	Tereza	8075-7	3x	Hybrid

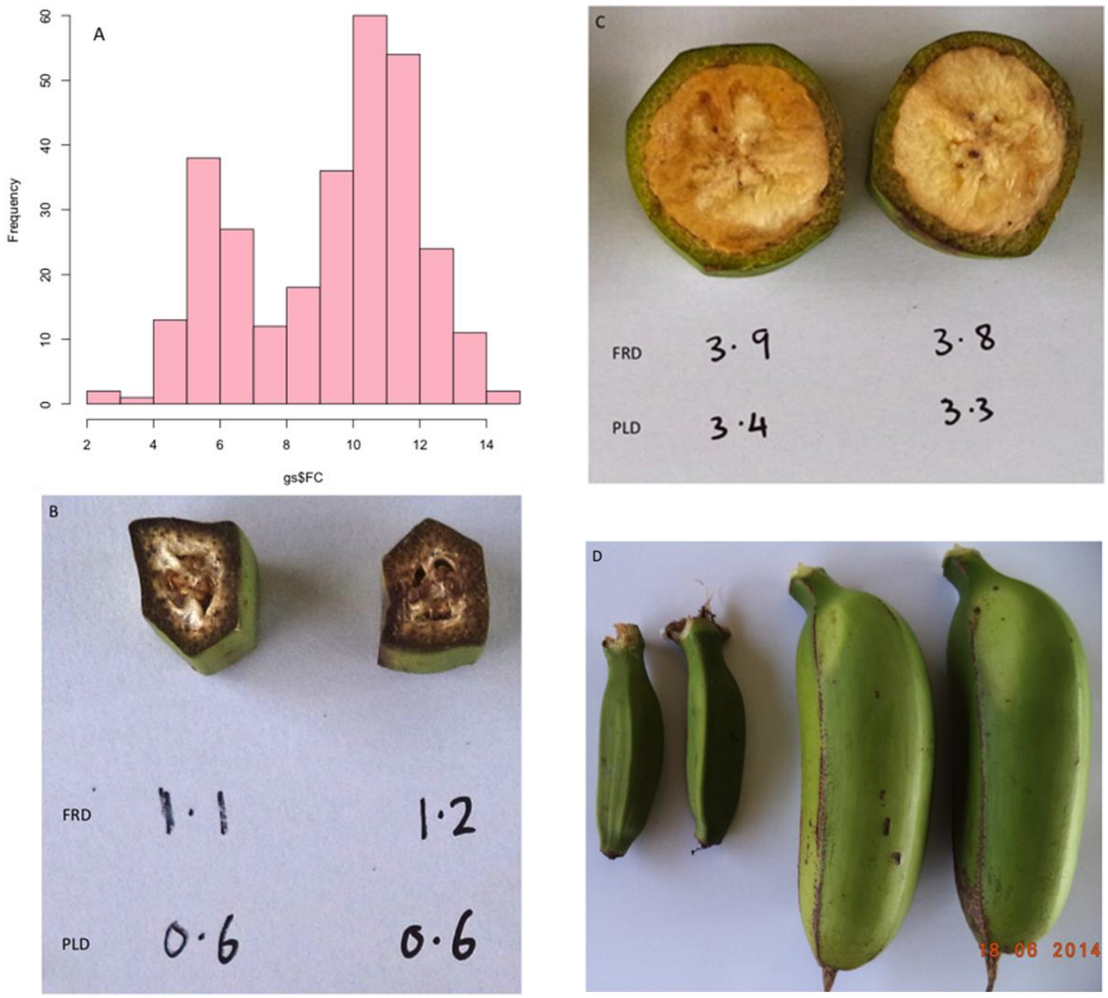
2x = diploid, 3x = triploid and 4x = tetraploid

S3 Table: Summary of all trait variations in response to cycle and field management.

Dep. variable	Indep. variable	Sum Sq	Df	F value	Pr(>F)
NSLF	Clone	3901.23	306	3.63	<0.0001
	Field	4.67	1	1.33	0.2492
	Clone:Field	1360.46	284	1.36	0.0001
	Cycle	2.63	1	0.69	0.4052
	Clone:Cycle	1174.64	299	1.04	0.3283
YLSF	Clone	4790.39	306	4.50	<0.0001
	Field	2.63	1	0.75	0.3852
	Clone:Field	1483.14	284	1.50	<0.0001
	Cycle	0.00	1	0.00	1.0000
	Clone:Cycle	1102.33	299	0.85	0.9669
PHF	Clone	2222889.11	306	3.77	<0.0001
	Field	1126.34	1	0.58	0.4449
	Clone:Field	432297.46	284	0.79	0.9947
	Cycle	8714.88	1	8.25	0.0041
	Clone:Cycle	332846.71	299	1.05	0.2662
PG	Clone	73176.82	306	4.30	<0.0001
	Field	1.52	1	0.03	0.8686
	Clone:Field	12061.30	284	0.76	0.9981
	Cycle	351.48	1	12.11	0.0005
	Clone:Cycle	13057.24	299	1.51	<0.0001
HTSF	Clone	2151815.75	306	2.96	<0.0001
	Field	1075.15	1	0.45	0.5014
	Clone:Field	895154.77	284	1.33	0.0005
	Cycle	59.52	1	0.02	0.8836
	Clone:Cycle	976295.15	299	1.18	0.0276
INSL	Clone	116602.02	306	2.44	<0.0001
	Field	4.96	1	0.03	0.8584
	Clone:Field	58583.77	284	1.32	0.0005
	Cycle	141.37	1	0.79	0.3740
	Clone:Cycle	51026.49	299	0.95	0.6947
TS^{sqrt}	Clone	240.28	305	3.21	<0.0001
	Field	0.24	1	0.99	0.3204
	Clone:Field	100.88	282	1.46	<0.0001
NSLH	Clone	4746.65	303	5.14	<0.0001
	Field	7.50	1	2.46	0.1170
	Clone:Field	958.14	269	1.17	0.0417
	Cycle	20.74	1	6.78	0.0093
	Clone:Cycle	1154.94	276	1.37	0.0002
YLSH	Clone	2261.86	303	4.18	<0.0001
	Field	3.33	1	1.87	0.1719
	Clone:Field	649.25	269	1.35	0.0003
	Cycle	4.14	1	2.01	0.1562
	Clone:Cycle	579.70	276	1.02	0.4063

HTSH	Clone	2714448.28	303	4.55	<0.0001
	Field	7053.33	1	3.58	0.0587
	Clone:Field	1190067.21	269	2.25	<0.0001
	Cycle	1920.12	1	0.65	0.4196
	Clone:Cycle	949051.52	276	1.17	0.0408
BWT^{sqrt}	Clone	1213.89	303	12.55	<0.0001
	Field	1.4	1	4.38	0.0365
	Clone:Field	126.77	269	1.48	<0.0001
	Cycle	4.04	1	15.24	<0.0001
	Clone:Cycle	108.68	276	1.49	<0.0001
NH	Clone	3334.02	303	8.67	<0.0001
	Field	0.03	1	0.03	0.8713
	Clone:Field	569.58	269	1.67	<0.0001
	Cycle	7.43	1	6.01	0.0143
	Clone:Cycle	429.09	276	1.26	0.0048
NF	Clone	1380508.67	303	5.46	<0.0001
	Field	112.13	1	0.13	0.7139
	Clone:Field	333080.59	269	1.49	<0.0001
	Cycle	4742.88	1	6.13	0.0134
	Clone:Cycle	262980.73	276	1.23	0.0092
FL	Clone	16284.98	300	13.49	<0.0001
	Field	33.92	1	8.43	0.0037
	Clone:Field	1982.34	269	1.83	<0.0001
	Cycle	5.95	1	1.10	0.2944
	Clone:Cycle	1328.62	273	0.90	0.8661
FC	Clone	9506.06	300	16.11	0.0000
	Field	17.79	1	9.04	0.0027
	Clone:Field	733.66	269	1.39	0.0001
	Cycle	2.78	1	1.30	0.2548
	Clone:Cycle	751.00	272	1.29	0.0021
FRD	Clone	1003.46	299	17.55	0.0000
	Field	2.52	1	13.19	0.0003
	Clone:Field	139.73	269	2.72	<0.0001
	Cycle	0.44	1	1.75	0.1866
	Clone:Cycle	70.74	271	1.04	0.3331
PLD	Clone	865.42	299	17.60	0.0000
	Field	2.70	1	16.42	<0.0001
	Clone:Field	68.27	269	1.54	<0.0001
	Cycle	0.52	1	3.03	0.0820
	Clone:Cycle	60.55	271	1.29	0.0022
PED	Clone	20.96	299	11.41	<0.0001
	Field	0.00	1	0.08	0.7799
	Clone:Field	16.61	269	10.05	<0.0001
	Cycle	0.00	1	0.13	0.7192
	Clone:Cycle	3.15	271	0.80	0.9913

^{sqrt} Original data transformed by square root,



S1 Fig. Variation in fruit characteristics. (A) is a histogram showing the bimodal distribution of fruit circumference (FC), (B) cross sections of poor filling fruits, (C) good filling fruits with fruit diameter (FRD) and pulp diameter (PLD) values in cm, and (D) poor filling and good filling banana fruits.

Appendix III

Poster presentation: Trait Variation in a Banana Training Population for Genomic Selection. At: Annual banana meeting, April, 2017, Kampala, Uganda.

Poster presentation: Trait Variation in a Banana Training Population for Genomic Selection. At: P4D and R4D meeting, November, 2016 at IITA, Ibadan, Nigeria.

Poster presentation: Towards marker assisted breeding in banana. At: R4D meeting, November, 2015 at IITA, Ibadan, Nigeria.

Poster Presentation: Genomic selection to accelerate banana breeding. At: Roots, Tubers and Bananas (RTB) project evaluation, February, 2015 at IITA, Sendusu, Uganda.

IITA
Transforming African Agriculture
50
1967 - 2017

Breeding Better Bananas

Trait Variation in a Banana Training Population for Genomic Selection

Nyine M[1,3], Uwimana B[1], Swennen R[2], Batte M[1], Brown A[2], Christelová P[3], Hřibová E[3], Lorenzen J[1*], Doležel J[3]

[1] International Institute of Tropical Agriculture (IITA), Kampala, Uganda

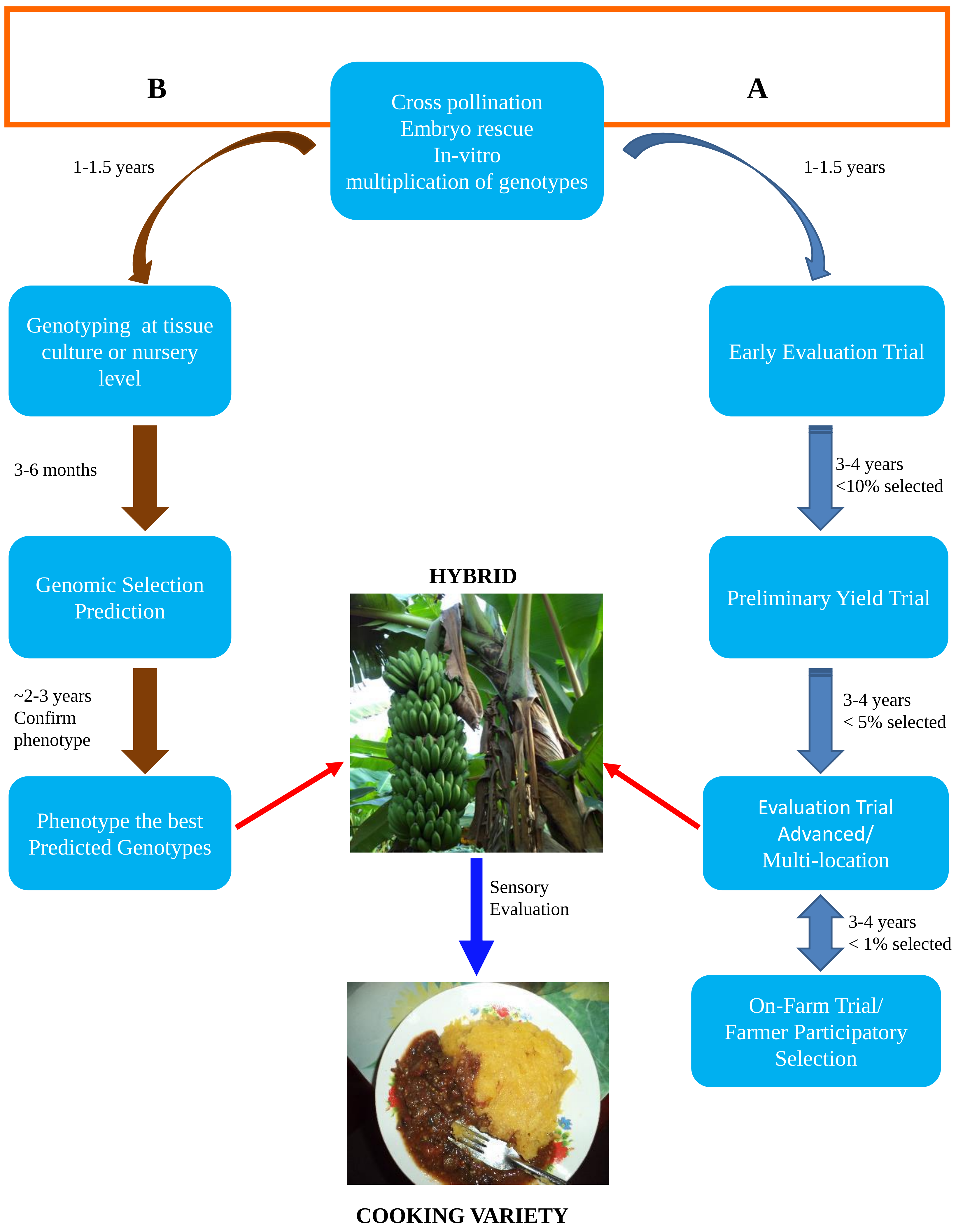
[2] International Institute of Tropical Agriculture (IITA), Arusha, Tanzania

[3] Institute of Experimental Botany & Palacký University, Czech Republic

* Currently with the Bill and Melinda Gates Foundation

Introduction

Conventional crossbreeding is the main approach used in banana improvement. However, the method requires up to two decades of crossing and field evaluation to develop a new hybrid. This is because selection is carried out at different levels (Fig 1). At every level, plants are evaluated after three crop cycles, each taking about a year. Yield traits can only be scored at harvest while organoleptic traits are recorded after harvesting, making the selection process slow, expensive and labour intensive. Molecular tools with the potential to improve banana breeding efficiency are being investigated. These include genomic selection (GS), which will benefit breeding through increased genetic gain per unit time (Meuwissen et al. 2001; Nakaya and Isobe 2009). Understanding trait variation and the correlation among economically important traits is an essential first step in the development of GS models. In this study we tested the hypothesis that trait variations in bananas are not affected by cross combination, cycle, field management and their interaction with genotype.



Cross pollination
Embryo rescue
In-vitro
multiplication of genotypes

Genotyping at tissue culture or nursery level

Early Evaluation Trial

Genomic Selection Prediction

Preliminary Yield Trial

Phenotype the best Predicted Genotypes

Evaluation Trial
Advanced/
Multi-location

On-Farm Trial/
Farmer Participatory
Selection

Fig 1: Approaches to hybrid selection in banana breeding program. (A) the classical phenotypic selection of banana hybrids and (B) integrated genomic selection and phenotypic selection approach being investigated.

Materials and Methods

The training population consists of 307 genotypes that include parents and the resulting hybrids. The population was phenotyped under low (no mulch and NPK fertilizer) and high (mulch + NPK) field input management at Namulonge research station. Data collected on two crop cycles were analysed using R statistical software. The correlations and significance of correlations were determined using R package Hmisc. Analysis of variance was performed to understand the effect of genotype and the interaction between genotype and cycle, and genotype and field management on trait variation.

Results and Discussion

A high level of correlation among vegetative and yield related traits was observed (Table 1). This could mean that the predictive ability of traits that are difficult to phenotype will be similar to less difficult traits they are highly correlated with. Therefore, genomic selection models could be developed for traits that are easily measured. Table 2 summarizes the genotypic effects and the interaction between genotype and cycle and genotype and field management on the traits. Black Sigatoka-related traits were not affected by crop cycle. These could be measured in the first cycle thus reducing on phenotyping burden. Growth traits such as plant height and girth were the least affected by field input management. Conversely, yield-related traits such as bunch weight, number of hands and number of fingers were significantly affected by both crop cycle and field input management. The variation in traits observed suggest that different genomic selection models should be tested. For traits affected by cycle and field management, models that account for non-additive genetic effect are likely to have better predictive ability on them. Integration of genomic selection in crossbreeding allows simultaneous prediction and selection of best hybrids. This is likely to reduce the selection cycle and increase genetic gain per unit time.

Table 1: Pearson's correlation coefficients of traits under high input field management

	Pant height	Plant girth	Index of non-spotted leaf	Bunch weight	Number of hands	Number of fruits	Fruit length	Fruit circumference	Fruit diameter
Plant girth	0.77*								
Index of non-spotted leaf	0.21	0.27							
Bunch weight	0.37*	0.62*	−0.13						
Number of hands	0.22	0.42*	0.10	0.52*					
Number of fruits	0.37*	0.58*	0.19	0.57*	0.84*				
Fruit length	0.20	0.44*	−0.15	0.83*	0.28*	0.27*			
Fruit circumference	0.33*	0.45*	−0.15	0.81*	0.15	0.15	0.85*		
Fruit diameter	0.39*	0.48*	−0.16	0.79*	0.16	0.18	0.80*	0.97*	
Pulp diameter	0.39*	0.45*	−0.16	0.74*	0.11	0.13	0.76*	0.94*	0.99*

* Significant correlation with P-value < 0.05

Table 2: Effect of genotype and genotype interaction with cycle and field management on the traits

Trait	Indep. variable	Sum Sq	Df	F value	*P* value
Plant height	Genotype	2222889	306	3.77	<0.0001
	Genotype x Field	432297	284	0.79	0.995
	Genotype x Cycle	332846	299	1.05	0.266
Plant girth	Genotype	73176	306	4.30	<0.0001
	Genotype x Field	12061	284	0.76	0.998
	Genotype x Cycle	13057	299	1.51	<0.0001
Index of non-spotted leaf	Genotype	116602	306	2.44	<0.0001
	Genotype x Field	58584	284	1.32	0.0005
	Genotype x Cycle	51026	299	0.95	0.695
Bunch weight*	Genotype	1214	303	12.55	<0.0001
	Genotype x Field	127	269	1.48	<0.0001
	Genotype x Cycle	109	276	1.49	<0.0001
Number of hands	Genotype	3334	303	8.67	<0.0001
	Genotype x Field	570	269	1.67	<0.0001
	Genotype x Cycle	429	276	1.26	0.005
Number of fruits	Genotype	1380509	303	5.46	<0.0001
	Genotype x Field	333081	269	1.49	<0.0001
	Genotype x Cycle	262981	276	1.23	0.009

* Original data was square root transformed

Conclusion

Genomic selection as a form of marker assisted selection is a non-stand alone approach but if integrated into conventional crossbreeding it has the potential to accelerate the breeding process. The effectiveness of genomic selection in banana will greatly depend on the prediction accuracy of the genomic selection models.

References

1.Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157: 1819-1829.

2.Nakaya A, Isobe SN. Will genomic selection be a practical method for plant breeding? Ann Bot. 2012;110: 1303-1316.

www.iita.org I www.cgiar.org

Trait Variation in a Banana Training Population for Genomic Selection

Nyine, Moses[1,3]
Uwimana, Brigitte[1]
Swennen, Rony[2]
Batte, Michael[1]
Brown, Allan[2]
Christelová, Pavla[3]
Hřibová , Eva[3]
Lorenzen, Jim[1*]
Doležel, Jaroslav[3]

[1] International Institute of Tropical Agriculture (IITA), Kampala, Uganda

[2] International Institute of Tropical Agriculture (IITA), Arusha, Tanzania

[3] Institute of Experimental Botany, Palacký University, Czech Republic

* Currently with the Bill and Melinda Gates Foundation

M.Nyine@cgiar.org

B.Uwimana@cgiar.org

Introduction

Conventional crossbreeding is the main approach used in banana improvement. However, the method requires up to two decades of crossing and field evaluation to develop a new hybrid. This is because selection is carried out at different levels (Fig 1). At every level, plants are evaluated after three crop cycles, each taking about a year. Yield traits can only be scored at harvest while organoleptic traits are recorded after harvesting, making the selection process slow, expensive and labour intensive. Molecular tools with the potential to improve banana breeding efficiency are being investigated. These include genomic selection (GS), which will benefit breeding through increased genetic gain per unit time (Meuwissen et al. 2001; Nakaya and Isobe 2009). Understanding trait variation and the correlation among economically important traits is an essential first step in the development of GS models. In this study we tested the hypothesis that trait variations in bananas are not affected by cross combination, cycle, field management and their interaction with genotype.

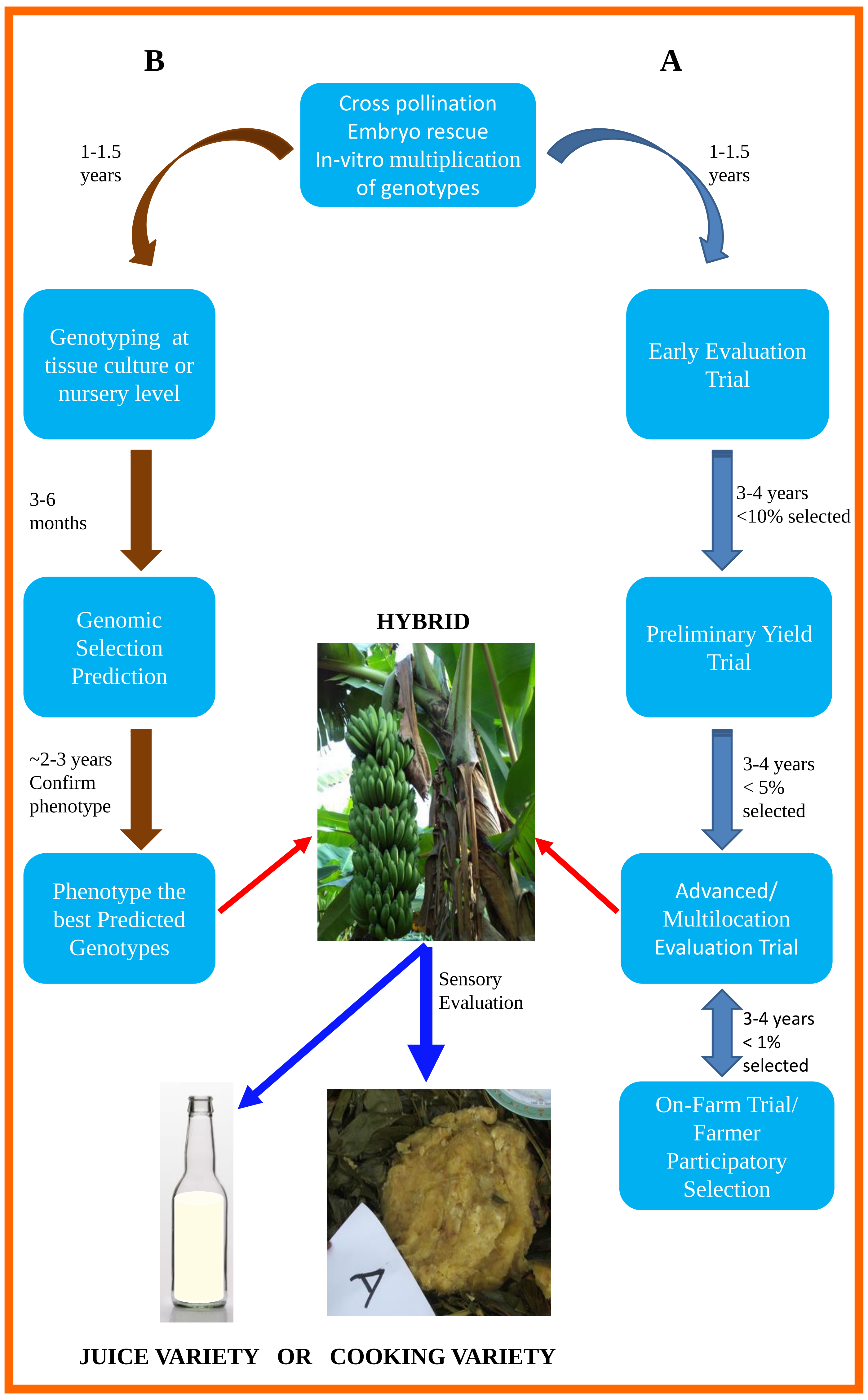

B
A
Cross pollination
Embryo rescue
In-vitro multiplication of genotypes
1-1.5 years
1-1.5 years
Genotyping at tissue culture or nursery level
Early Evaluation Trial
3-6 months
3-4 years
<10% selected
Genomic Selection Prediction
HYBRID
Preliminary Yield Trial
~2-3 years
Confirm phenotype
3-4 years
< 5% selected
Phenotype the best Predicted Genotypes
Advanced/ Multilocation Evaluation Trial
Sensory Evaluation
3-4 years
< 1% selected
On-Farm Trial/ Farmer Participatory Selection
JUICE VARIETY OR COOKING VARIETY

Fig 1: Approaches to hybrid selection in banana breeding program. (A) the classical phenotypic selection of banana hybrids and (B) integrated genomic selection and phenotypic selection approach being investigated.

Materials and Methods

The training population consists of 307 genotypes that include parents and the resulting hybrids. The population was phenotyped under low (no mulch and NPK fertilizer) and high (mulch + NPK) field input management at Namulonge research station. Data collected on two crop cycles were analysed using R statistical software. The correlations and significance of correlations were determined using R package Hmisc. Analysis of variance was performed to understand the effect of genotype and the interaction between genotype and cycle, and genotype and field management on trait variation.

Results and Discussion

A high level of correlation among vegetative and yield related traits was observed (Table 1). This could mean that the predictive ability of traits that are difficult to phenotype will be similar to less difficult traits they are highly correlated with. Therefore, genomic selection models could be developed for traits that are easily measured. Table 2 summarizes the genotypic effects and the interaction between genotype and cycle and genotype and field management on the traits. Black Sigatoka-related traits were not affected by crop cycle. These could be measured in the first cycle thus reducing on phenotyping burden. Growth traits such as plant height and girth were the least affected by field input management. Conversely, yield-related traits such as bunch weight, number of hands and number of fingers were significantly affected by both crop cycle and field input management. The variation in traits observed suggest that different genomic selection models should be tested. For traits affected by cycle and field management, models that account for non-additive genetic effect are likely to have better predictive ability on them. Integration of genomic selection in crossbreeding allows simultaneous prediction and selection of best hybrids. This is likely to reduce the selection cycle and increase genetic gain per unit time.

Table 1: Pearson's correlation coefficients of traits under high input field management

	Pant height	Plant girth	Index of non-spotted leaf	Bunch weight	Number of hands	Number of fruits	Fruit length	Fruit circumference	Fruit diameter
Plant girth	0.77*								
Index of non-spotted leaf	0.21	0.27							
Bunch weight	0.37*	0.62*	−0.13						
Number of hands	0.22	0.42*	0.10	0.52*					
Number of fingers	0.37*	0.58*	0.19	0.57*	0.84*				
Fruit length	0.20	0.44*	−0.15	0.83*	0.28*	0.27*			
Fruit circumference	0.33*	0.45*	−0.15	0.81*	0.15	0.15	0.85*		
Fruit diameter	0.39*	0.48*	−0.16	0.79*	0.16	0.18	0.80*	0.97*	
Pulp diameter	0.39*	0.45*	−0.16	0.74*	0.11	0.13	0.76*	0.94*	0.99*

* Significant correlation with P-value < 0.05

Table 2: Effect of genotype and genotype interaction with cycle and field management on the traits

Trait	Indep. variable	Sum Sq	Df	F value	*P* value
Plant height	Genotype	2222889	306	3.77	<0.0001
	Genotype x Field	432297	284	0.79	0.995
	Genotype x Cycle	332846	299	1.05	0.266
Plant girth	Genotype	73176	306	4.30	<0.0001
	Genotype x Field	12061	284	0.76	0.998
	Genotype x Cycle	13057	299	1.51	<0.0001
Index of non-spotted leaf	Genotype	116602	306	2.44	<0.0001
	Genotype x Field	58584	284	1.32	0.0005
	Genotype x Cycle	51026	299	0.95	0.695
Bunch weight*	Genotype	1214	303	12.55	<0.0001
	Genotype x Field	127	269	1.48	<0.0001
	Genotype x Cycle	109	276	1.49	<0.0001
Number of hands	Genotype	3334	303	8.67	<0.0001
	Genotype x Field	570	269	1.67	<0.0001
	Genotype x Cycle	429	276	1.26	0.005
Number of fruits	Genotype	1380509	303	5.46	<0.0001
	Genotype x Field	333081	269	1.49	<0.0001
	Genotype x Cycle	262981	276	1.23	0.009

* Original data square root transformed

Conclusions and Recommendations

Genomic selection as a form of marker assisted selection is a non-stand alone approach but if integrated into conventional crossbreeding it has the potential to accelerate the breeding process. The effectiveness of genomic selection in banana will greatly depend on the prediction accuracy of the genomic selection models.

References

1.Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157: 1819-1829.

2.Nakaya A, Isobe SN. Will genomic selection be a practical method for plant breeding? Ann Bot. 2012;110: 1303-1316.

IITA is a member of the CGIAR Consortium

www.iita.org

Towards Marker Assisted Breeding in Banana

Nyine, Moses[1]
Uwimana, Brigitte[1]
Ssali, Tendo Reuben[2]
Kubiriba, Jerome[2]
Amorim, Edson[3]
Othman, Yasmin[4]
Swennen, Rony[1]
Batte, Michael[1]
Hribova, Eva[5]
Dolezel, Jaroslav[5]

[1] International Institute of Tropical Agriculture (IITA)

[2] National Agricultural Research Organisation (NARO)

[3] Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA)

[4] Centre for Research in Biotechnology for Agriculture (CEBAR), University of Malaya

[5] Palacky University

B.Uwimana@cgiar.org

Introduction

Breeding offers the most sustainable solution to most of the crop yield-limiting factors such as pests, diseases and abiotic stress. Crossbreeding is the main approach used in banana improvement but the method requires up to 20 years to release a variety. Integration of molecular tools into crossbreeding speeds up variety development through marker-assisted selection (MAS) and genomic selection (GS). With these approaches, banana breeders can shorten the breeding cycle to less than a decade. The initial stage towards MAS is to generate segregating populations followed by mapping of quantitative trait loci (QTL) affecting target traits using linkage mapping. Genome-wide association studies (GWAS) are useful in underpinning alleles responsible for phenotypic variation. Given the high cost of phenotyping and the ever-decreasing cost of genotyping, genomic selection (GS) is being considered for routine use in breeding programs and as such predictive genomic selection models are being developed. IITA banana breeding, in collaboration with other partners, is fully committed to developing an integrated approach to banana improvement with the aim of increasing genetic gain per unit time while reducing the selection cycle.

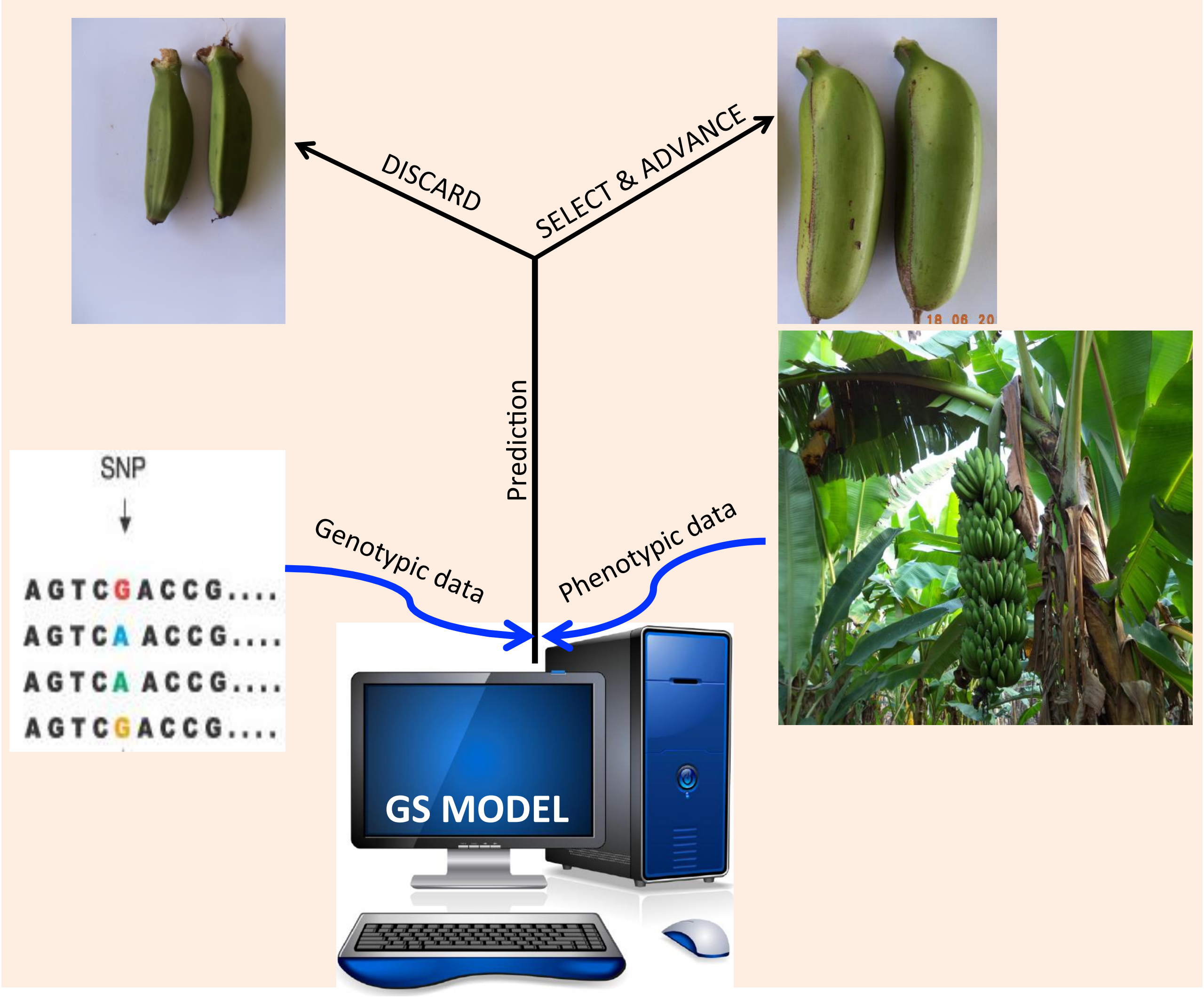
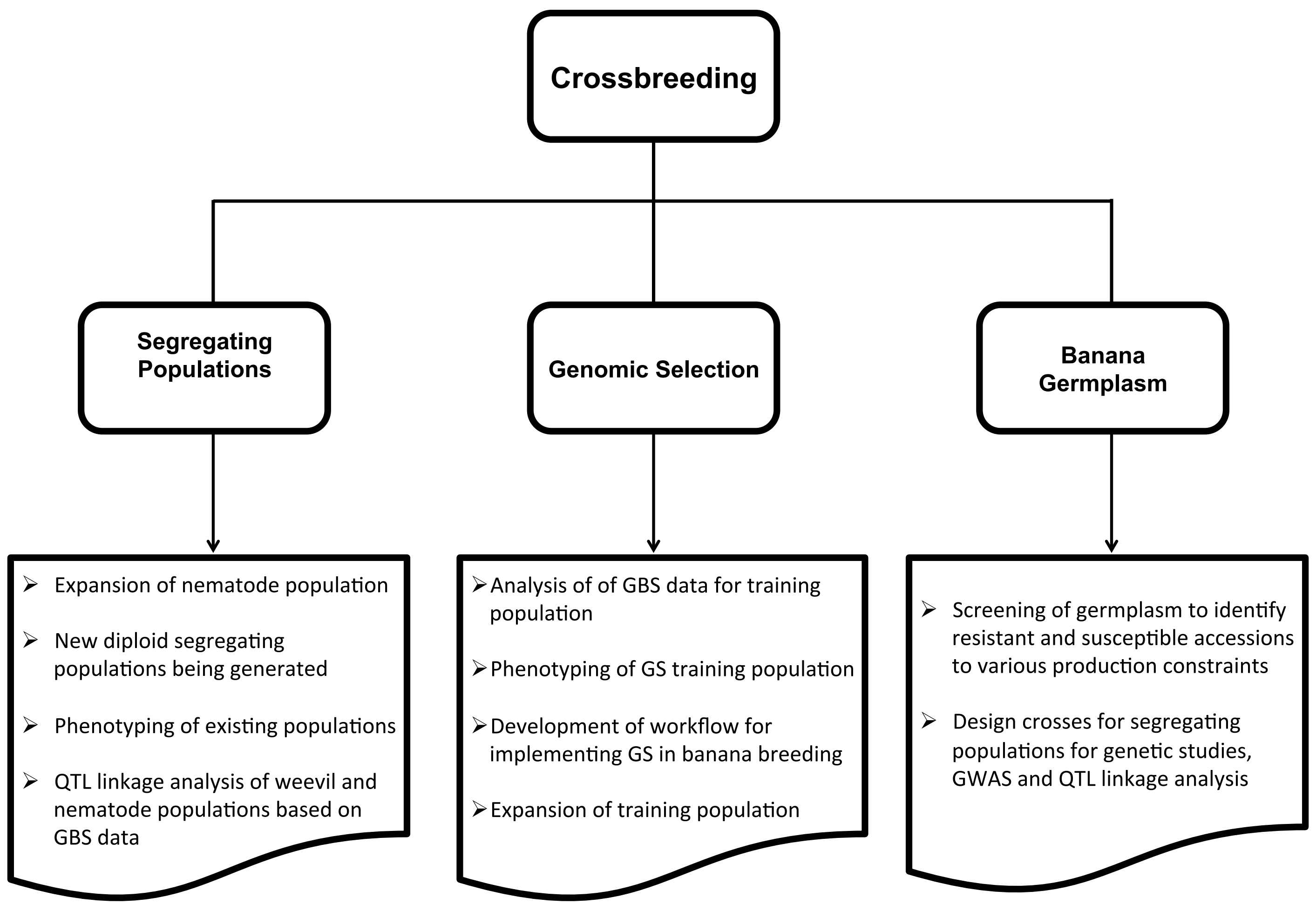


Figure 1. A simplified depiction of genomic selection model development and application in crossbreeding



Crossbreeding

Segregating Populations

- Expansion of nematode population
- New diploid segregating populations being generated
- Phenotyping of existing populations
- QTL linkage analysis of weevil and nematode populations based on GBS data

Genomic Selection

- Analysis of of GBS data for training population
- Phenotyping of GS training population
- Development of workflow for implementing GS in banana breeding
- Expansion of training population

Banana Germplasm

- Screening of germplasm to identify resistant and susceptible accessions to various production constraints
- Design crosses for segregating populations for genetic studies, GWAS and QTL linkage analysis

Figure 2: On-going activities related to molecular breeding of bananas within IITA in collaboration with other institutions such as NARO, Palacky University, EMBRAPA, and the University of Malaya

Materials and Methods

Target traits: Fusarium wilt, weevil, burrowing nematode, yield and agronomic traits.

The nematode segregating population consists of two half-sib populations with one common male parent (Mbanjo et al., 2012). The weevil segregating population was derived from selfing the F1 progeny of Borneo and Kasasika. The Fusarium segregating population was derived from selfing F1 progeny of 8075-7 and sukali ndizi. The training population for GS consists of all breeder's parental stock used by IITA and NARO and advanced hybrids and hybrids from early evaluation trials. DNA from all these populations was extracted and submitted to Cornell for sequencing using the genotyping by sequencing (GBS) approach. On-going activities are summarized in figure 2.

Results and Discussion

Diploid nematode-segregating population was developed by IITA banana breeding (Dochez et al., 2009). The population was used to generate genetic linkage maps by Mbanjo et al., 2012 (figure 3A) based on SSR markers designed from expressed sequence tags (Lorenzen et al., 2011) and diversity array technology (DArT) markers. The population has also been genotyped by sequencing (GBS) to identify SNP markers. Together with the phenotype data, the QTLs responsible for nematode resistance will be identified once the analysis is complete.

Two F_2 populations developed by NARO and segregating for weevil and Fusarium resistance have been genotyped by IITA by GBS and SNP data are being analyzed to generate SNP-based genetic linkage map for identification of QTLs for weevil and Fusarium resistance.

Application of genomic selection is being tested at IITA-Uganda for the first time ever in banana breeding. Over 300 accessions including parental lines and hybrids (training population) have been genotyped by GBS and are being phenotyped in three fields (figure 3B). Disease and pest resistance in plants is controlled by one or few QTLs with major effect on the phenotype. However, yield and many agronomically important traits are controlled by many QTLs with small effects on the phenotype. GS is ideal for such traits as it utilizes genome-wide markers to determine the genomic estimated breeding value (GEBV) of the individual plant (Meuwissen et al., 2001, Nakaya and Isobe 2012). This is a model-based approach which requires the breeder to generate genotypic data which are fed into the model to predict phenotypic performance of the individual plant (figure 1). This approach holds promise to improve the efficiency of crossbreeding by reducing the selection cycle yet increasing genetic gain per unit time.

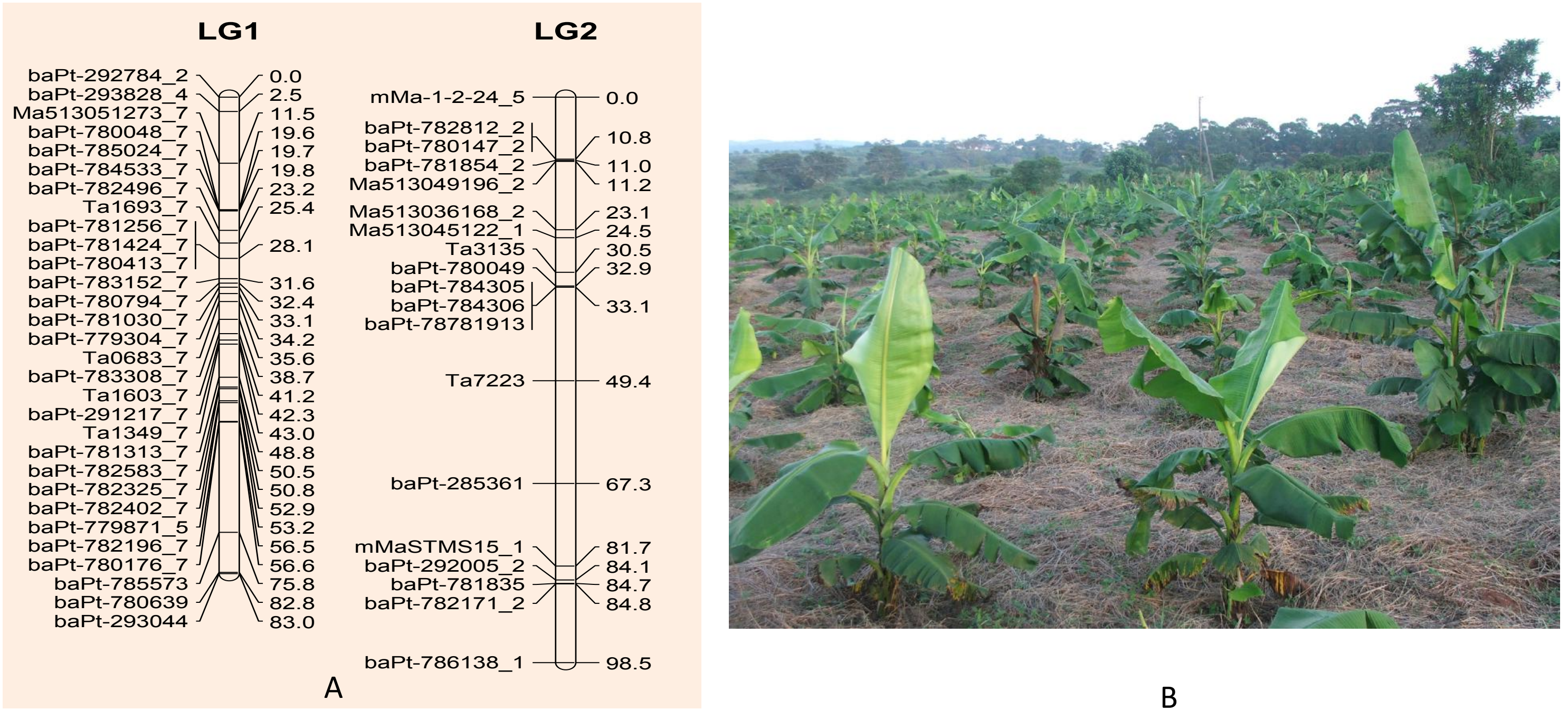


Figure 3. A - Part of linkage maps from nematode segregating population generated by Mbanjo *et al*. 2012, B – One of the genomic selection training population phenotyping field at Namulonge, Uganda.

Conclusions

In marker assisted breeding, no single marker technique is sufficient to address all the breeders' questions. In bananas different traits have different mechanisms of genetic control ranging from single gene with major effect to multiple genes with small additive effects on the phenotype. The interaction of genes by environment makes the interpretation of results even more challenging. However, with genomic selection this can be corrected for in the model development once phenotype data is collected in different environments while QTL mapping could help in selection for pest and disease resistance. Despite the challenges, IITA in collaboration with other partners is committed to the development of platforms for marker assisted breeding in banana as a model polyploid plant. Once a break-through is realized this will set a precedence for other polyploidy breeding programs to embrace marker assisted breeding.

References

1. A. Nakaya and S. Isobe (2012). Review: Will genomic selection be a practical method for plant breeding? Annals of Botany 110:1303-1316.
2. C. Dochez (2004). Breeding for resistance of Rhadopholus similis in East African highland bananas (Musa spp.). Dissertation, Katholieke Universiteit Leuven.
3. E.G.N. Mbanjo, F. Tchoumbougnang, A.S. Mouelle, J.E. Oben, M. Nyine, C. Dochez, M.E. Ferguson and J. Lorenzen (2012). Molecular marker_based genetic linkage map of a diploid banana population (*Musa acuminata* Colla). Euphytica 188:369-386.
4. J. Lorenzen, S. Hearne, G. Mbanjo, M. Nyine and T. Close (2011). Use of Molecular markers in Banana and Plantain Improvement. Acta Hort. 897:231-236.
5. T.H.E. Meuwissen, B.J. Hayes and M.E. Goddard (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

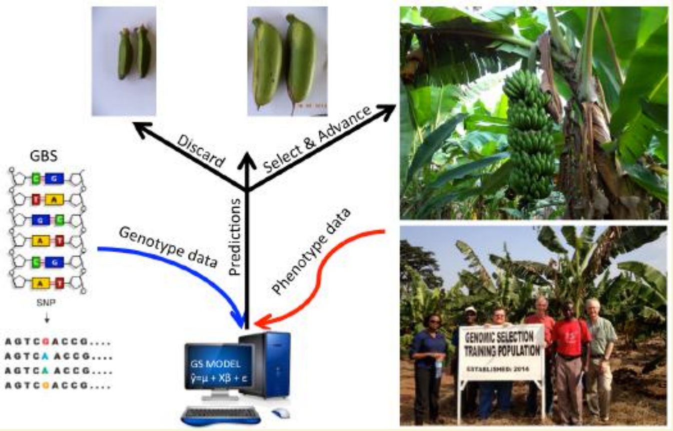
GENOMIC SELECTION TO ACCELERATE BANANA BREEDING

M.Nyine, B.Uwimana, R. Swennen, M. Batte, E. Hribova, J. Lorenzen J. Dolezel

Introduction

✧ Genomic selection (GS) is a form of marker-assisted selection which involves the use of markers across the genome to predict the genetic estimated breeding value (GEBV) of a plant.

✧ Phenotype prediction is based on a genomic selection model



State of the training population

✧ 320 accessions genotyped

✧ Three phenotyping established

✧ Experimental design: CRD with three plants per accession

✧ Target traits: fruit filling, stature and suckering

✧ Data is being collected on 22 additional traits

Why genomic selection?

✧ Conventional banana breeding is much slower

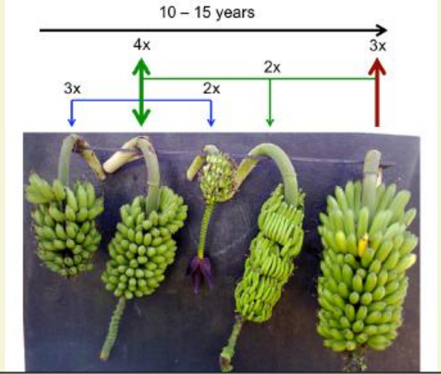
✧ To increase genetic gain per unit time

✧ Genotyping is becoming much cheaper than phenotyping

✧ To improve both variety and parental lines development pipelines

✧ Selection is possible at nursery stage

Time scale for conventional breeding



Conclusion

Genomic selection coupled with increased hybrids from cross-breeding should increase efficiency of banana improvement thus, faster variety release

Acknowledgements:

IEB-Czech Repulic, IGD-Cornell and Dr. Rabbi Ismail

This work is supported by RTB consortium and BMGF through IITA

Appendix IV

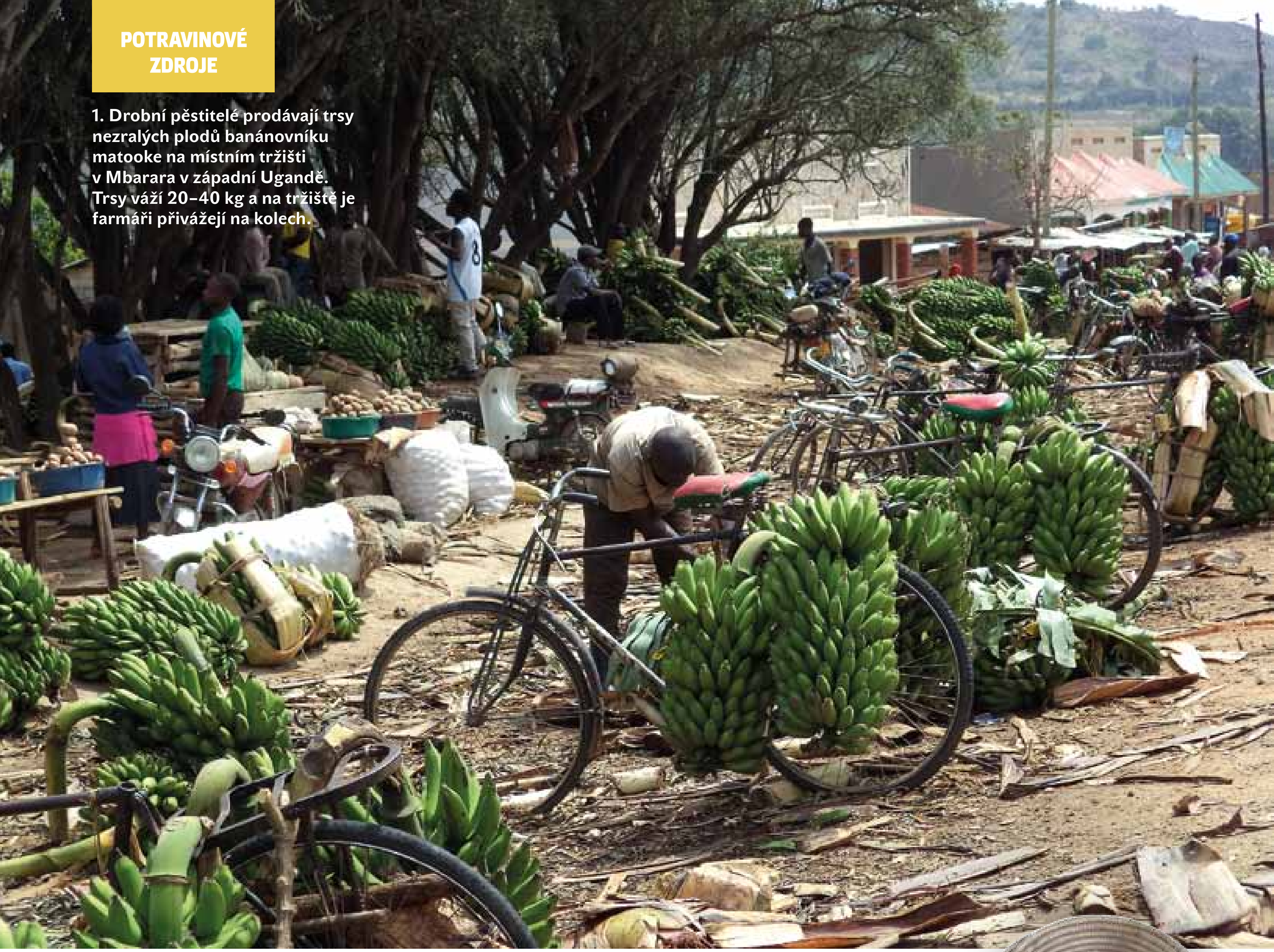
Articles in a popular magazine, Vesmír

1. Banánovník z východoafrické vysočiny: Základní potravina pro miliony

2. Banánovník z východoafrické vysočiny: Záhada původu a pěstování

3. Banánovník z východoafrické vysočiny: Šlechtění banánovníku

POTRAVINOVÉ ZDROJE



1. Drobní pěstitelé prodávají trsy nezralých plodů banánovníku matooke na místním tržišti v Mbarara v západní Ugandě. Trsy váží 20–40 kg a na tržiště je farmáři přivážejí na kolech.

MOSES NYINE
JAROSLAV DOLEŽEL

Banánovník z východoafrické vysočiny

1. Základní potravina pro miliony

Plody banánovníku jsou pro miliony lidí základní potravinou. Je tomu tak zejména v Ugandě a dalších zemích africké oblasti Velkých jezer. Banánovník má však mnohem širší využití. Může být například stavebním materiálem, dokáže nahradit talíře a afričtí kluci si z něj umějí udělat kopací míč.

Banánovníky jsou jednoděložné vytrvalé byliny vlhkých tropů a subtropů (box). Jejich kulturní formy lze rozdělit podle plodů. Některé jsou po dozrání sladké a jedí se syrové, jiné se konzumují nezralé po tepelném zpracování. Existují také banánovníky s plody vhodnými pro přípravu nápojů, zejména banánového piva, banánovníky vhodné pro získávání vláken i okrasné banánovníky. Na mezinárodním trhu převládají banány sladké (ovocné) a z nich zejména plody odrůdy Cavendish. Méně se na zahraniční trhy vyvážejí banány s vysokým obsahem škrobu používané na pečení (plantains). Plody pro vaření a přípravu nápojů jsou konzumovány téměř výhradně v oblastech, kde se pěstují. Okrasné banánovníky nemají jedlé plody a zdobí je květy a panašované listy. Z nepravého stonku další skupiny banánovníků (především *Musa textilis*) se získávají vlákna bohatá na celulózu. Po vyčištění se označují jako manilské konopí a používají se např. na výrobu speciálního papíru pro filtry a bankovky.

Banánovník ve východní Africe

Pro obyvatele východní Afriky a zejména Ugandy má banánovník velký kulturní, společenský a ekonomický význam. V této oblasti se pěstují hlavně odrůdy s plody vhodnými pro vaření (matooke) a v menší míře s plody používanými na přípravu džusu a banánového piva (mbidde). V obou případech se sklízí nezralé zelené plody. Dužina plodů odrůd typu mbidde obsahuje taniny a má svíravou chuť. Proto se nechávají dozrát, získají žlutou barvu a jejich dužina sladkou chuť. Teprve poté se z nich připravují nápoje. Dužina nezralých plodů skupiny odrůd matooke nemá svíravou chuť a na vaření se používají nezralé plody. Tento typ banánů představuje hlavní složku výživy obyvatelstva a jejich spotřeba dosahuje 400–600 kg na osobu za rok, což je nejvíce na světě.

V Ugandě se banánovník pěstuje na ploše asi 1,5 milionů hektarů a sklidí se více než 10 milionu tun banánů, z nichž se 80 % spotřebuje v místě produkce (obr. 1). Pěstování banánů je hlavním zdrojem příjmů mnoha farmářů, zejména v centrální a západní Ugandě. Banány vhodné pro vaření mají široké využití. Oloupané se vaří v páře a jedí se jako kaše s různými omáčkami. Jejich dužina má krémově bílou nebo světle žlutou barvu a vařením v páře se stává zlatožlutou (obr. 3). Jídlo známé jako katogo (obr. 3 vpravo) se připravuje vařením oloupaných banánů společně s fazolemi nebo s pastou z burských oříšků, masem, rybou nebo vnitřnostmi. Obvykle se podává k snídani a zahřívá tělo v době ranního chladu.

Ženy po porodu dostávají k jídlu katogo připravené s vnitřnostmi, protože se věří, že zahřátí břišní dutiny pomůže odstranit zbytek krve z dělohy a stimuluje produkci mateřského mléka. Banánová kaše je používána při přechodu kojenců k normální stravě. Vařené banány jsou považovány za ideální stravu nemocných lidí, kteří ztratili chuť k jídlu. Pokud pacient nejí ani je, naznačuje to vážný stav a blízkost smrti.

Z varných typů banánů se vyrábí mouka, která je vhodná pro přípravu kaše, instantního banánového pokrmu (instantní tooke), pečiva a sladkostí.

Z banánů se vyrábějí také různé typy lupínků a čipsů. Navíc balením potravin do listů banánovníku před vařením v páře (obr. 4) dostává jídlo unikátní chuť. Ještě nerozvinuté listy (tzv. cigar leaves) se udí, balí se do nich maso s kořením a připravuje se tak tradiční pokrm luwombo (obr. 2). Ten se podává jen významným hostům a v restauracích je dražší než ostatní jídla. Z plodů banánovníků mbidde, druhého nejčastějšího typu ve východní Africe, se připravuje banánové pivo známé jako „tonto“, dále džusy, víno a džin.

2. Příprava tradičního pokrmu luwombo vařením masa baleného v uzených listech banánovníku.

Moses Nyine, MSc., (*1978) vystudoval molekulární biologii na Makerere University v Ugandě a v Mezinárodním ústavu tropického zemědělství v Ugandě se věnuje genetice a šlechtění banánovníku. V současné době je doktorandem Univerzity Palackého v Olomouci a na olomouckém pracovišti Ústavu experimentální botaniky AV ČR se zabývá genetickou diverzitou rodu banánovník a vývojem genomických metod šlechtění banánovníku.

Prof. Ing. Jaroslav Doležel, DrSc., (*1954) vystudoval Agronomickou fakultu na Vysoké škole zemědělské v Brně. Zabývá se strukturou a evolucí genomu rostlin, vede Centrum strukturní a funkční genomiky Ústavu experimentální botaniky AV ČR a přednáší na Přírodovědecké fakultě Univerzity Palackého v Olomouci. Od roku 2004 je členem Učené společnosti ČR, v roce 2012 mu předseda AV ČR udělil prestižní Akademickou prémii – Praemium Academiae – a v roce 2014 obdržel cenu ministra školství, mládeže a tělovýchovy za mimořádné výsledky výzkumu, experimentálního vývoje a inovací.

3. Vlevo jídlo připravené z kaše banánů matooke a omáčky z ryby, lilku a jiné zeleniny. Vpravo jídlo zvané katogo připravené z banánů vařených společně s vnitřnostmi.

Ani nejedlé části banánovníku nepřijdou nazmar. Slupky a nepravé stonky se používají na krmení hospodářských zvířat. V některých rodinách je zase zvykem podávat jídlo místo na talířích na banánových listech. V chudších oblastech si lidé z listů banánovníku staví paravány pro dočasné venkovní koupelny. U příležitosti různých slavností se z banánových listů zhotovují kostýmy pro tradiční tance. V některých oblastech se vlákna z nepravého stonku používají na stavbu střech chýší. Děti z chudých rodin, které si nemohou dovolit drahý kožený míč, si hrají s míči vyrobenými z banánovníku (obr. 5). Sušená hlavní žilka listu se navíc používá pro pletení košíků (viz malý obrázek nad nadpisem).

Botanická klasifikace banánovníku

Banánovníky patří do řádu zázvorníkotvaré (Zingiberales), čeledi banánovníkovité (Musaceae), rodu banánovník (*Musa*). Druhy banánovníku, kterých je asi sedmdesát, se na základě novějších molekulárních analýz člení do dvou sekcí: *Musa* s diploidním počtem chromozomů rovným 22 a *Callimusa* s diploidním počtem chromozomů rovným 20 nebo 18. Na evoluci kulturních forem se podílela jak vnitrodruhová, tak mezidruhová hybridizace a vedle diploidních klonů se dvěma sadami chromozomů existují klony triploidní se třemi sadami chromozomů a tetraploidní se čtyřmi sadami chromozomů. Prostřednictvím mezidruhové hybridizace se na evoluci pěstovaných forem podílely zejména diploidní druhy *Musa acuminata* s genomem A, *M. balbisiana* s genomem B a jen ve velmi malé míře pak *M. textilis* s genomem T a *M. schizocarpa* s genomem S. Složení genomů pěstovaných forem je tedy velmi pestré a zahrnuje diploidy (AA, BB, AB, AS), triploidy (AAA, AAB, ABB, AAT) a tetraploidy (AAAA, AABB, ABBB, ABBT).

Banánovníky východoafrické vysočiny patří do sekce *Musa* (podskupina Lujugira-Mutika) a jsou to triploidní kultivary s genomem AAA. Předpokládá se, že vznikly vnitrodruhovou hybridizací mezi diploidními poddruhy *M. acuminata* s genomy AA. Jsou bezsemenné, obvykle sterilní a množí se výhradně vegetativně. Tyto banánovníky dobře rostou ve vyšších nadmořských výškách (1400–2000 m n. m.) a pro optimální růst a vývoj vyžadují průměrné roční srážky okolo 1300 mm. Pěstují se v oblasti velkých jezer a zejména v oblasti Viktoriina jezera a na vysočinách východoafrických zemí. Odtud také jejich název. Na základě morfologie je celkem 84 kultivarů pěstovaných v Ugandě klasifikováno do čtyř skupin: Nfuuka, Nakitembe, Nakabululu a Musakala. I když jsou mezi jednotlivými skupinami dobře patrné morfologické rozdíly, molekulární analýzy naznačují velkou podobnost jejich dědičných informací.

Příště o původu banánovníku a jeho pěstování.

4. Nahoře: Vaření banánů balených v listech banánovníku.
5. Uprostřed: Listy banánovníku a jiné části rostliny jsou používány při zhotovování uměleckých předmětů, levných míčů a užitných předmětů.

Morfologie banánovníku

Banánovníky jsou vytrvalé jednoděložné byliny a pěstované druhy patří k nejstatnějším bylinám vůbec. Zkrácený podzemní stonek (oddenek) nese kořeny, které rostou jen do hloubky 30–45 cm, a proto čerpají živiny z povrchových půdních vrstev. Apikální meristém oddenku se nachází pod povrchem půdy nebo na jeho úrovni. Postupně z něj ve šroubovici vyrůstají nové listy a odstředivě směrem ven vytlačují listy starší. Nové listy jsou stočené, a proto se jím říká „cigar leaf“. Listové pochvy jsou tuhé (vytrvávají i po odumření čepelí), dlouhé, vzájemně těsně shloučené a vytvářejí tak nepravý stonek, který zdánlivě vypadá jako kmen. Před rozkvětem přestanou růst nové listy a z apikálního meristému vyroste květenství, jehož dlouhá stopka prorůstá vnitřkem nepravého stonku. Vlastní květenství se pak objeví na bázi shluku listových čepelí v horní části rostliny, často bývá převislé. V jeho spodní části se nejdříve ve shlucích vytvářejí samičí květy, které se vyvíjejí v plody (bobule), uprostřed jsou jalové květy a na konci květenství pak shluky samčích květů. Shluky květů jsou podepřené nápadně zbarvenými toulcovitými listeny. Pěstované odrůdy jsou bezsemenné a množí se vegetativně pomocí odnoží, které vyrůstají z postranních pupenů na zkráceném stonku. Růst odnoží reguluje apikální meristém a mezi druhy a odrůdami banánovníku existují velké rozdíly v počtu rychlosti jejich růstu.

Banánovník z východoafrické vysočiny

2. ZÁHADA PŮVODU A PĚSTOVÁNÍ

Banánovník pochází z jihovýchodní Asie, kde byly některé jeho typy asi před deseti tisíci lety domestikovány a kde se také nachází primární centrum jeho diverzity. V tomto teritoriu se vyvinuly banánovníky typické zvlášť pro indomalajskou a australasijskou oblast (viz rámeček na s. 38).

text **MOSES NYINE, JAROSLAV DOLEŽEL**

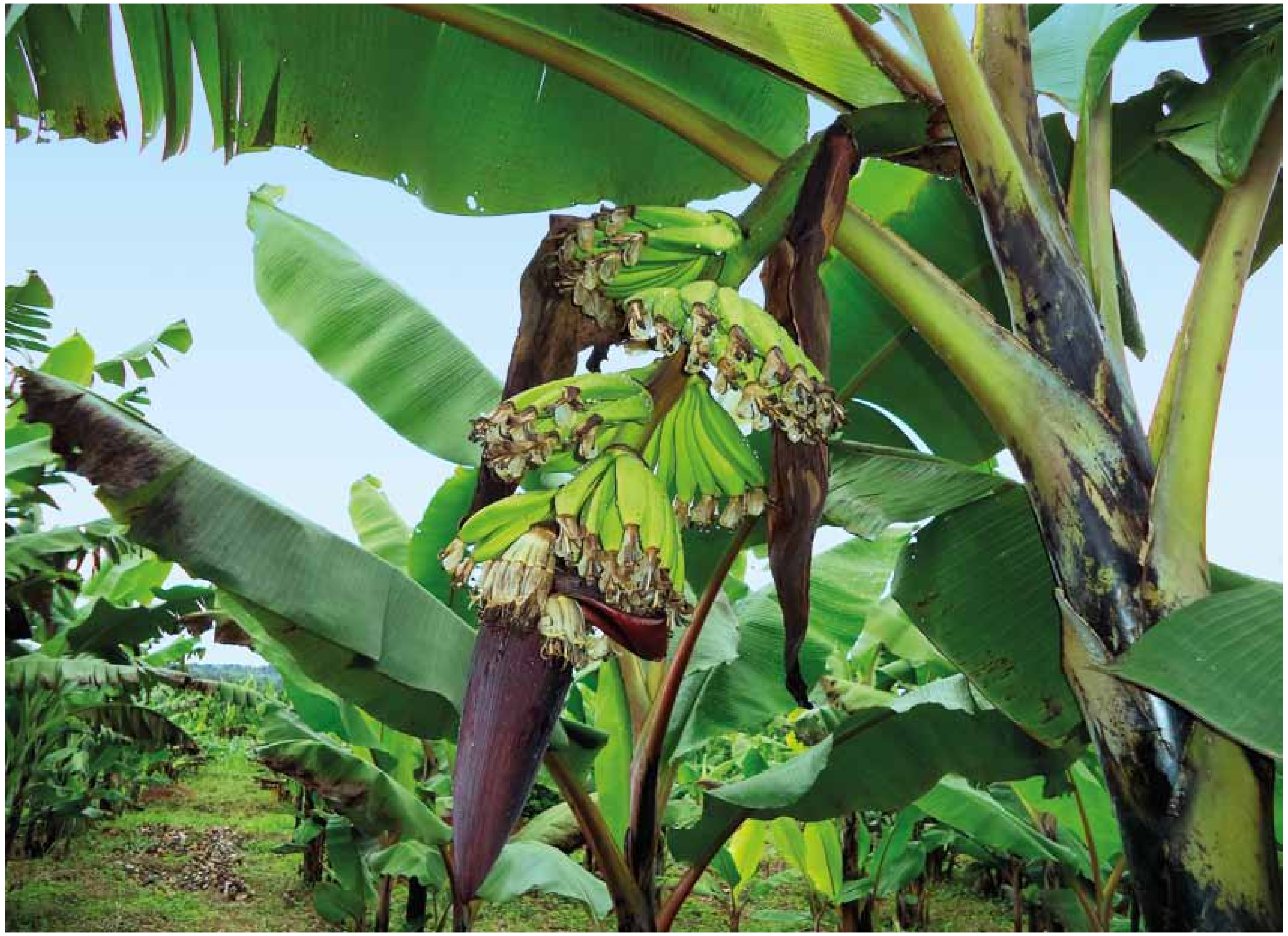
VÝCHODNÍ AFRIKA je sekundárním centrem diverzity s asi 120 klonově množenými odrůdami. Původ východoafrických banánovníků je však nejasný a vysvětlit se jej snaží několik hypotéz.

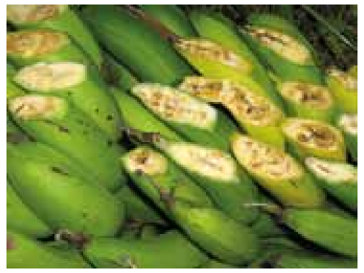
První z nich předpokládá, že se tyto odrůdy do Afriky dostaly prostřednictvím obchodníků, kteří se plavili Indickým oceánem mezi jihovýchodní Asií a východní Afrikou. Ti mohli v období 100–600 n. l. do východní Afriky přivézt odnože jedlých odrůd. Největším problémem této hypotézy je absence stejných forem banánovníku v jihovýchodní Asii.

Druhá hypotéza vysvětluje původ východoafrických banánovníků křížením mezi diploidními druhy, které se do východní Afriky dostaly z jihovýchodní Asie. Dosud se však nepodařilo nalézt žádné diploidní druhy nebo klony, jejichž dědičná informace by se podobala východoafrickým banánovníkům.

Třetí hypotéza předpokládá změnu dědičné informace odrůd přivezených z Asie následkem spontánních mutací, což mohlo mít za následek vznik kultivarů s odlišným fenotypem. Tzv. *somaklonální variabilita* spočívá v mutacích somatických (tělních) buněk, které zahrnují změny počtu a struktury chromozomů. Somaklonální variabilita byla popsána u rostlin regenerovaných z buněk kultivovaných *in vitro*. Protože jsou banánovníky množeny vegetativně, změna dědičné informace v jejich somatických buňkách může být přenesena do další generace. Pro existenci takové variability u rostlin pěstovaných na poli však neexistují žádné důkazy, a tak ani tato hypotéza nebyla dosud prokázána.

V poslední době se výzkumné týmy věnují možnému podílu epigenetických změn na morfologické variabilitě východoafrických banánovníků. Epigenetické změny jsou dědičné a mohou mít za následek změnu fenotypu, a to aniž by došlo ke změně sekvencí DNA. Podstatou těchto změn, které ovlivňují funkci genů, jsou metylace DNA a modifikace histonů. Je známo, že mohou být vyvolány vnějším prostředím. Tyto změny lze jen obtížně identifikovat pomocí molekulárních markerů. Nicméně pokrok v oblasti molekulární biologie a genomiky dává naději, že bude v brzké době možné ověřit případný podíl epigenetických změn na vzniku východoafrických banánovníků.

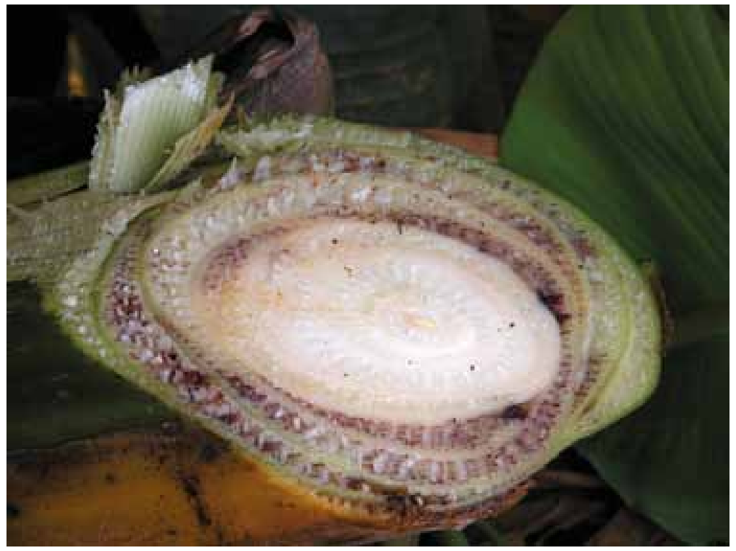




2. **PLODY BANÁNOVNÍKU znehodnocené infekcí bakterií *Xanthomonas campestris* pv. *musacearum*, která způsobuje chorobu zvanou bakteriální vadnutí.**

Snímky na s. 36–38 © Moses Nyine.

3. **KVĚTENSTVÍ východoafrického banánovníku.**



1. **ŘEZ NEPRAVÝM STONKEM BANÁNOVNÍKU napadeného houbou *Fusarium oxysporum* f. sp. *cubense*, která způsobuje fusariové vadnutí.**

MOSES NYINE, MSc., (*1978) vystudoval molekulární biologii na Makerere University v Ugandě a v Mezinárodním ústavu tropického zemědělství v Ugandě se věnuje genetice a šlechtění banánovníku. V současné době je doktorandem Univerzity Palackého v Olomouci a na olomouckém pracovišti Ústavu experimentální botaniky AV ČR se zabývá genetickou diverzitou rodu banánovník a vývojem genomických metod šlechtění banánovníku.

Prof. Ing. JAROSLAV DOLEŽEL, DrSc., (*1954) vystudoval Agronomickou fakultu na Vysoké škole zemědělské v Brně. Zabývá se strukturou a evolucí genomu rostlin, vede Centrum strukturní a funkční genomiky Ústavu experimentální botaniky AV ČR a přednáší na Přírodovědecké fakultě Univerzity Palackého v Olomouci. Od roku 2004 je členem Učené společnosti ČR, v roce 2012 mu předseda AV ČR udělil prestižní Akademickou prémii – Praemium Academiae – a v roce 2014 obdržel Cenu ministra školství, mládeže a tělovýchovy za mimořádné výsledky výzkumu, experimentálního vývoje a inovací.

Posledním uvažovaným zdrojem morfologických odlišností východoafrických banánovníků je epistáze. Tento jev zahrnuje situaci, kdy je jeden fenotypový znak ovlivněn více geny.

BANÁNOVNÍK A OSTATNÍ ORGANISMY

Východoafrické banánovníky ohrožuje řada chorob a napadá je mnoho škůdců. To může mít negativní dopad na výživu místních obyvatel a snížit příjmy malých farmářů, kteří si nemohou dovolit používání drahých pesticidů. Snad nejničivější chorobou východoafrických banánovníků, stejně jako ostatních odrůd, je bakteriální vadnutí způsobené patologickou variantou *musacearum* bakterie *Xanthomonas campestris*.[1] Infekce má za následek úplnou ztrátu úrody. Symptomy zahrnují předčasné dozrávání plodů a změnu barvy jejich dužiny (**obr. 2**), nekrózu samčího pupenu a přítomnost žlutého slizu na řezu nepravým stonkem (**obr. 4**). Dosud se nepodařilo nalézt odolné genotypy. Řešení by mohly přinést metody genetického inženýrství. Jediné, co mohou farmáři v současné době dělat, je omezovat negativní dopady choroby vhodnými agrotechnickými postupy.

Hlavními škůdci pěstovaných banánovníků jsou nosatcovitý brouk *Cosmopolites sordidus* (**obr. 5**) a háďátka *Radopholus similis* (**obr. 6**), *Pratylenchus* spp. a *Helicotylenchus* spp. Nosatec *C. sordidus* páchá největší škody v larválním stadiu, kdy v oddenku vyžírá tunely, poškozuje růstový vrchol a cévní svazky. To způsobuje snížení příjmu vody a živin, zastavení růstu a vývoje. Sklizeň pak bývá ztrátová. Háďátka parazitují na kořenech a takto vzniklá poškození vyvolávají nekrózy. Následkem je redukovaný příjem vody a živin a celková destrukce kořenového systému. V případě silných větrů dochází k vyvrácení rostlin, které nejsou v půdě dostatečně ukotveny.

Z houbových chorob napadajících banánovník je nejzávažnější „Black Sigatoka“, kterou způsobuje houba *Mycospaerella*

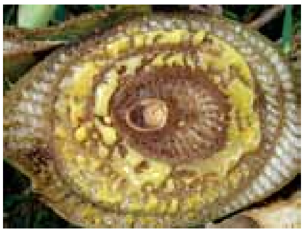
[1] Tato choroba je v anglické literatuře označovaná jako Banana Xanthomonas Wilt (BXW).

fijiensis. Její spory se šíří větrem a houba napadá listy, které zasychají, a to vede ke snížení fotosynteticky aktivní plochy rostliny (**obr. 7**) a ke snížení výnosu. Banánovníky východní Afriky napadají také viry BSV (banana streak virus) a BBTV (banana bunchy top virus), které však produkci zásadním způsobem neohrožují.

PANAMSKÁ CHOROBA

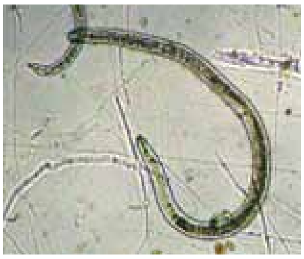
Závažnou chorobou banánovníku je fusariové vadnutí, které způsobuje houba srpovnička *Fusarium oxysporum* f. sp. *cubense*. Do rostliny vniká přes kořeny a ucpává její cévní systém (**obr. 1**). Napadá mnoho kultivarů, zejména ovocné typy nesoucí sladké plody. Naštěstí pro východoafrické farmáře však nenapadá jimi pěstované odrůdy. Proti této chorobě, nazývané též „panamská“, dosud neexistuje účinná ochrana. Přitom to byla právě ona, která zásadním způsobem změnila produkci sladkých banánů pro export. Její tropická rasa 1 (TR1) v letech 1940–1960 postupně zničila plantáže osazené monokulturami odrůdy Gros Michel. Shodou okolností je proti této rase odolná odrůda Cavendish, která nahradila na infikovaných plantážích odrůdu Gros Michel a zachránila tak celé odvětví produkce banánů pro export. V současné době se však začíná šířit tropická rasa 4 (TR4), proti které není odrůda Cavendish odolná, a budoucí produkce banánů pro export začíná být opět ohrožena. ●

v dalším čísle
o šlechtění banánovníku

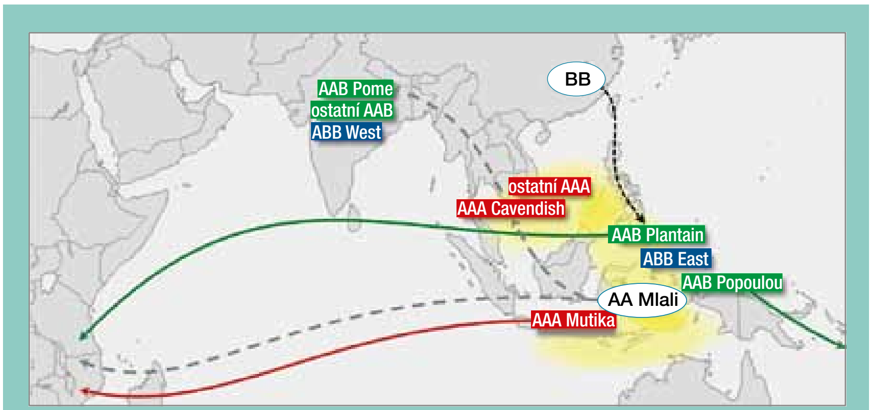


4. **NA ŘEZU NEPRAVÝM STONKEM se bakteriální vadnutí projevuje přítomností žlutého slizu.**

5. **NOSATEC *Cosmopolites sordidus* s čerstvě nakladenými vajíčky. Larvy vyžírají tunely ve stonku a poškozují růstový vrchol a cévní svazky.**



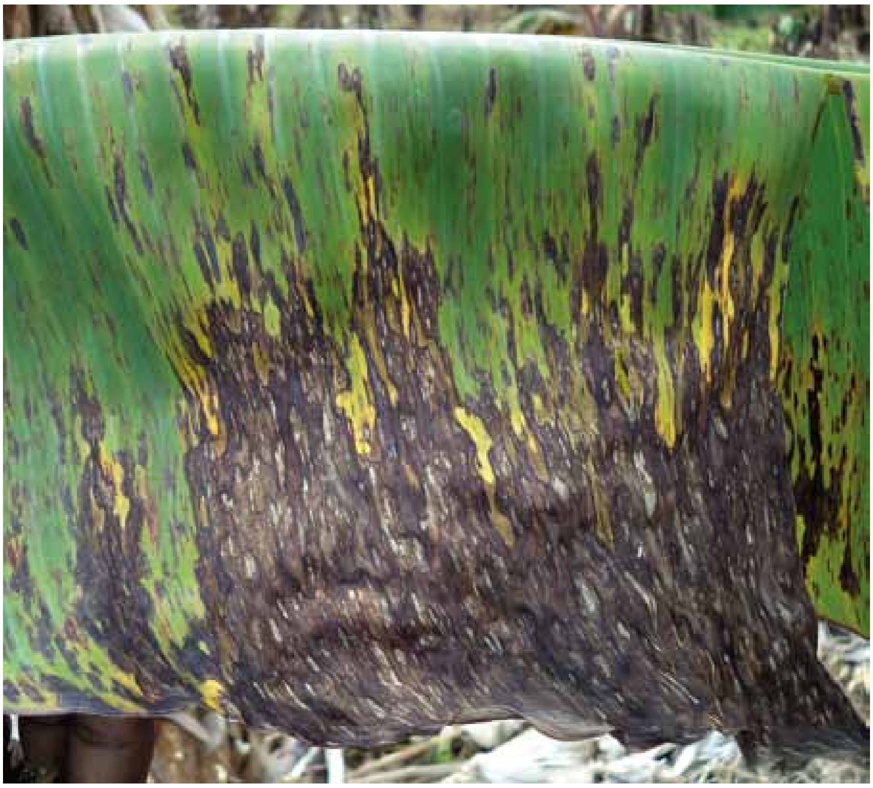
6. **HÁĎÁTKO *Radopholus similis* parazitující na kořenech banánovníku způsobuje nekrózy a ničí kořenový systém.**



Původ banánovníku

Předpokládaný původ a migrace banánovníku. Křížením mezi poddruhy banánovníku *Musa acuminata* (*banksii*, *errans*, *malaccensis*, *microcarpa*, *zebrina*) vnikly v oblasti jihovýchodní Asie bezsemenné diploidní formy s genomem AA. Některé z těchto klonů, které řadíme do podskupiny Mlali (šedé čárkované šipky), migrovaly na asijskou pevninu, kde mimo jiné daly vznik triploidnímu kultivaru Cavendish (s genomem AAA); křížením s druhem *Musa balbisiana* (s genomem BB) na indickém subkontinentu vznikly triploidní kultivary podskupiny Pome s genomem AAB. Zástupci podskupiny Mlali také migrovali na východoafrické pobřeží (čárkované šipky). Dnes se v jihovýchodní Asii nevyskytují a nacházejí se jen na východoafrickém pobřeží a přilehlých ostrovech (Zanzibar, Madagaskar a Komory). V jihovýchodní Asii vznikly rovněž triploidní klony nesoucí škrobové plody, které se však zde v podstatě nepěstují, a plné šipky znázorňují jejich migrace. Banánovníky s genomem AAB vzniky křížením diploidních forem (genom AA) s druhem *M. balbisiana* (genom BB) a migrovaly do západní Afriky (AAB Plantain) a opačným směrem na pacifické ostrovy (AAB Popoulou). Křížením diploidů s *M. balbisiana* jak na indickém subkontinentu, tak v jihovýchodní Asii vznikly rovněž triploidní klony s genomem ABB. Někteří autoři předpokládají migraci triploidních klonů s genomem AAA řazených do podskupiny Mutika na východoafrické pobřeží, kde mohly dát vznik v současnosti pěstovaným kultivarům východoafrické vysočiny.

Zdroj: Perrier et al.: Multidisciplinary perspectives on banana (*Musa* spp.) domestication. PNAS, July 12, 2011, vol. 108, no. 28, 11311–11318.



7. **LIST BANÁNOVNÍKU napadený houbou *Mycospaerella fijiensis*, která způsobuje chorobu Black Sigatoka.**

Banánovník z východoafrické vysočiny

3. ŠLECHTĚNÍ BANÁNOVNÍKU

Klasické způsoby šlechtění banánovníku jsou časově velmi náročné a mohly by být výrazně urychleny pomocí nových genomických metod pro vyhledávání perspektivních kříženců.

text **MOSES NYINE, JAROSLAV DOLEŽEL**

NENÍ POCHYB O TOM, že nejefektivnější obranou proti chorobám a škůdcům, která nemá negativní vliv na prostředí, je pěstování rezistentních odrůd. Kromě choroby Banana Xanthomonas Wilt (Vesmír 95, 36, 2016/1) lze mezi planými druhy banánovníků nalézt zdroje rezistence a je velký zájem využít je ve šlechtění. U banánovníku je však problém se sterilitou – pěstované odrůdy jsou bezsemenné. Přesto má křížení smysl: s velmi malou frekvencí vznikají funkční pohlavní buňky i u sterilních triploidů a lze získat hybridní semena. Úspěšným pionýrem šlechtění byl Phil Rowe (1939–2001) z USA, který v Hondurasu pracoval pro několik společností. Jeho úspěšná strategie je založena na šlechtění diploidních fertilních linií, které slouží jako zdroje pylu pro křížení se sterilními triploidními odrůdami. Kříženci, které takto získal, jsou odolní vůči mnoha chorobám a škůdcům, mají vysoké výnosy a patří mezi první nové odrůdy, které farmáři ochotně přijali. O své práci Rowe řekl: „křížíte rostliny, které netvoří semena, abyste získali lepší rostliny, které nemají semena“.

Na základě poznatků Phila Rowa bylo r. 1994 zahájeno šlechtění východoafrických banánovníků, na němž se podílejí Mezinárodní ústav tropického zemědělství a Národní organizace pro zemědělský výzkum. Cílem je zlepšit odolnost místních odrůd vůči chorobě Black Sigatoka, broukům a háďátkům přenesením rezistence z planých druhů. Postup ukazuje **rámeček** na protější straně. Nejprve byly na základě výsledků křížení s diploidním klonem Calcutta 4 planého druhu *M. acuminata* ssp. *burmanicoides* vybrány místní odrůdy s nejvyšší samičí fertilitou. Pak bylo jedenáct odrůd zařazeno do šlechtitelského programu (**obr. 1**). Jejich křížením s klonem Calcutta 4 (**obr. 2**) se získaly tetraploidní hybridy s jednou sadou chromozomů diploidního klonu a třemi sadami chromozomů triploidních odrůd východoafrických banánovníků. Tetraploidní hybridy (**obr. 4** a velký snímek) se potom křížily s diploidními klony (**obr. 3**), které mají lepší vlastnosti než Calcutta 4. Výsledkem byly triploidní hybridy s jednou sadou chromozomů diploidního rodiče a dvěma sadami chromozomů tetraploidního rodiče (**obr. 5**).

Hybridní semena mají velmi nízkou klíčivost a jejich embrya jsou proto vyjmuta a dopěstovávána v podmínkách in vitro. Po postupném otužení se nové hybridy přesazují na pokusné lokality a hodnotí se jejich odolnost vůči chorobám a škůdcům, kvalita plodů a výnos. Za 20 let trvání šlechtitelského programu se získalo 27 nových odrůd zvaných Narita, z nichž farmáři již jednu pěstují ve velkém. I když je šlechtitelský program úspěšný, je náročný na ruční práci a čas, a je tedy drahý.

ZVYŠOVÁNÍ EFEKTIVITY ŠLECHTĚNÍ

Získání nové odrůdy banánovníku vyžaduje nejméně 10 až 15 let a její přijetí konzervativními farmáři a konzumenty není jisté (jeden cyklus hodnocení výnosu a kvality plodů trvá jeden až jeden a půl roku od vysazení odnože na pole). Rychlé šíření nových chorob a škůdců však vyžaduje rychlejší reakci šlechtitelů. Jednou z možností, jak urychlit hodnocení získaných kříženců, je jejich výběr v raných fázích růstu pomocí markerů DNA. Protože klasické šlechtění spočívá v opakování cyklů křížení a výběru potomstev s požadovanými vlastnostmi, výběr pomocí markerů (marker assisted selection, MAS) může vhodně doplnit klasické postupy šlechtění. Pokud je určitý marker DNA v těsné vazbě na daný znak, může být jeho nositel identifikován v rané fázi růstu na základě analýzy DNA. Šlechtění pomocí markerů se však u banánovníku zatím neuplatňuje, protože důležité znaky jsou komplexní a pěstované odrůdy jsou triploidní.

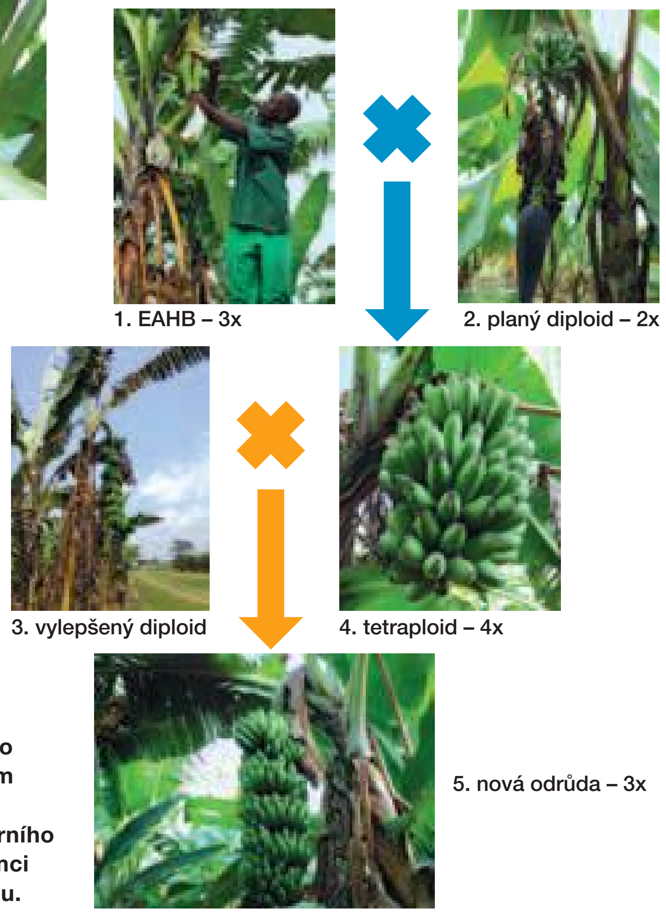
Nadějí tak mohou být genomické metody, jejichž uplatnění se stalo reálné díky pokroku v nových technologiích sekvenování a analýze takto získaných dat. Za relativně nízkou cenu je dnes možné podrobně charakterizovat genomy mnoha jedinců. Charakterizace dědičné informace každého jedince pomocí velmi vysokého počtu markerů DNA (typicky polymorfismy individuálních nukleotidů v sekvenci DNA, tzv. SNP) umožňuje navrhnout modely pro tzv. genomickou selekci. Tyto modely budou používány při výběru rodičovských partnerů, aniž by bylo nutné identifkovat markery vázané na konkrétní znaky. Genomická selekce je variantou výběru pomocí markerů, v níž jsou všechny dostupné markery souhrnně používány pro odhad šlechtitelské hodnoty jedince, a to pomocí matematického modelu. Správnost modelu se ověřuje v tzv. trénovací populaci. V případě správného modelu je pak možné identifikovat hybridy s požadovanými vlastnostmi už v raných fázích jejich růstu. Použití genomické selekce ve šlechtění východoafrických banánovníků se v současné době testuje, a pokud se tento přístup osvědčí, může podstatně zefektivnit šlechtění a získávání nových odrůd s požadovanými vlastnostmi. ●

MOSES NYINE, MSc., (*1978) a **prof. Ing. JAROSLAV DOLEŽEL, DrSc.,** (*1954) viz Vesmír 95, 36, 2016/1.

Snímky na této dvoustraně Moses Nyine

Šlechtění banánovníků východoafrické vysočiny

Východoafrické banánovníky jsou citlivé vůči mnoha chorobám a škůdcům a nesou bezsemenné plody, které se vytvářejí bez oplodnění (jsou partenokarpické). Plané diploidní (2x) formy jsou odolné vůči mnoha chorobám a škůdcům a jsou fertilní s plody plnými semen. Tetraploidní hybrid (4x) se čtyřmi sadami chromozomů vzniká splynutím neredukované gamety odrůdy východoafrického banánovníku, která je triploidní (3x), a redukované gamety diploida, která je haploidní (x). Ve srovnání s odrůdou východoafrického banánovníku je trs plodů tetraploidního hybridu menší. Avšak potomek křížení tohoto hybridu s vybraným diploidem vykazuje heterózní efekt. Takto získaný sekundární triploid (3x) je větší než kterýkoliv z rodičů, má větší trs plodů a je odolný vůči chorobám a škůdcům. Pokud uspěje u farmářů a spotřebitelů, může být uvolněn jako nová odrůda.



1. OPYLOVÁNÍ východoafrického banánovníku.
2. PLODY banánovníku Calcutta 4, jehož pyl se používá při křížení s východoafrickými banánovníky.
3. DIPLOIDNÍ KLON banánovníku vybraný pro křížení s tetraploidními hybridy.
4. PLODY tetraploidního hybridu získaného křížením triploidní odrůdy východoafrického banánovníku s diploidním klonem Calcutta 4.
5. TRS PLODŮ sekundárního triploida získaného v rámci šlechtitelského programu.