FACULTY OF BIOSCIENCE ENGINEERING

CHARACTERIZING PARENTAL MATERIAL IN THE SWEETPOTATO BREEDING PROGRAM IN WEST AFRICA

word count: ±18000

Cédric Schindfessel

Student ID: 01404478

Supervisor(s): prof. dr. Dirk Reheul, dr. ir. Jolien Swanckaert

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of master in Bioscience Engineering.

Academic year: 2018 - 2019

GHENT UNIVERSITY

Gent, June 4, 2019

The promotor,	The author,
prof. dr. Dirk Reheul, dr. ir. Jolien Swanckaert	Cédric Schindfessel

PREFACE

> *"It's a great thing when you realize you still have the ability to surprise yourself. Makes you wonder what else you can do you've forgotten about."*
> *- Lester Burnham (American Beauty)*

About 12 months ago I discovered the ability to surprise myself. Surprising yourself is a serendipitous act: you have to grasps an opportunity when it arises, sometimes take a decision that is atypical for your personality and without certainty where it might lead you, you have to give it all you have and constantly be on the lookout for things you weren't looking for. If done successfully, you will end up in unique situations that could not have happened if you stayed in your normal routine. Because you did something that is atypical, because you worked hard, because you decided not to wait, because you were open for surprise. I have made a series of important decisions these last 12 months that will influence my personal and professional life. I will not bore you with meaningless details, but I was able to surprise myself sometimes and I'm very pleased. The first in that series of surprising, but benevolent decisions was to leave the safety of our faculty's labs, and to go to Ghana to write my masters' dissertation. The fruits of this decision, you can read on the following pages.

Of course, this dissertation would not exist without the help I received from a few important people. Although I have already thanked most of them, it's good to put it on paper once. First and foremost, I would like to thank Jolien and Mathias for welcoming me in their home during my stay in Ghana, not only for giving me a place to sleep and eat, but also for the many fun and useful talks and discussions. I wish them and their children all the best on their next adventures. I want to thank Jolien, prof. Reheul, Ted and Gerlinde for reviewing the text, concerning both content and spelling. Finally, I would like to thank Gerlinde, Cedric and Giovanni for adding that extra bit of excitement, fun and surprise to this last year, that seem to have made writing this dissertation a far less daunting task for me than it was for many of my peers and predecessors.
Hakuna batata!

Cédric Schindfessel
Marke, May 31, 2019

CONTENTS

SAMENVATTING

Zoete aardappel (*Ipomoea batatas* (L.) Lam.) is een belangrijke voedselbron voor heel wat mensen op deze wereld. Zijn hoge nutritionele waarde, hoge droogte tolerantie en behoorlijke opbrengst, zelfs onder slechte agronomische omstandigheden, maken van dit gewas een zeer goede voedselbron voor de mensen in Sub-Sahara-Afrika. Bijzondere aandacht gaat uit naar de zoete aardappel met oranje vruchtvlees, die, omwille van zijn hoge β-caroteen gehalte, een belangrijke rol kan spelen in het bestrijden van vitamine A deficiënties in dit werelddeel. Zoete aardappel is een hexaploïde kruisbestuiver die vegetatief vermeerderd wordt, deze 3 kenmerken hebben belangrijke gevolgen voor de genetische samenstelling en de veredeling van dit gewas. Het International Potato Center (CIP) leidt het zoete aardappel veredelingsprogramma in West-Afrika, gevestigd in Ghana. Eén van de belangrijkste stappen in de verdeling van zoete aardappel is de selectie van de beste ouders om een verdelingsprogramma mee verder te zetten.

Voor deze dissertatie werden zowel fenotypische data (voornamelijk opbrengst data van 149 kruisingen tussen 22 ouders) als genotypische data (microsatelliet merker data van 48 ouders) van de zoete aardappel ouders in Ghana geanalyseerd, met als hoofddoel de veredelaars in West-Afrika te helpen om een heterosis exploiterend veredelingsschema op te stellen. De fenotypische relatie tussen de ouders en hun nakomelingen werd bestudeerd. Op basis hiervan werd een groot heterosis potentieel vastgesteld, en kwamen een paar experimentele fouten aan het licht, die moeten worden vermeden tijdens toekomstige veldproeven. Uit de genotypische studie werd de grote genetische variabiliteit tussen de ouders duidelijk. Alles tezamen geven de resultaten van deze studie een definitief groen licht aan de veredelaars in West-Afrika om hun oudermateriaal onder te verdelen in verschillende heterotische groepen om zo een heterosis exploiterend veredelingsschema op te starten.

SUMMARY

Sweetpotato (*Ipomoea batatas* (L.) Lam.) is an important food crop for many people across the globe. Its high nutritional value, high drought tolerance and decent yield under poor agronomic conditions make it an excellent food source for people in Sub-Saharan Africa. Orange fleshed sweetpotato in particular can, due to its high *β*-carotene content, play an important role in battling vitamin A deficiencies in this region of the world. Sweetpotato is a cross-fertilizing, hexaploid plant that is vegetatively propagated. These 3 characteristics have important consequences for the genetic constitution and breeding of the crop. The International Potato Center (CIP) leads a sweetpotato breeding initiative in West Africa, hosted in Ghana. One of the most important steps in sweetpotato breeding is the selection of the best parents to continue a breeding program with.

This dissertation analysed both phenotypic data (yield data on 149 cross combination between 22 parents) and genotypic data (SSR marker data on 48 parents) on the parents in the Ghana crossing block, in order to help the breeders in West Africa set up a heterosis exploiting breeding scheme. The phenotypic parent-offspring relationship was studied and a large heterosis potential in this population was identified, as well as some experimental flaws that should be alleviated in future field trials. The genotypic study identified the large genetic variability within the parental population. Taken together, the results of this dissertation give a definite permission for the breeders in West Africa to start dividing their parental population into mutually heterotic groups and start up a heterosis exploiting breeding scheme.

ACRONYMS

AFLP amplification fragment length polymorphism

AMOVA analysis of molecular variance

ANOVA analysis of variance

BIC Bayesian information criterium

CIP ESP: Centro Internacional de la Papa. ENG: International Potato Center

CSIR-SARI Council for Scientific and Industrial Research - Savannah Agricultural Research Institute

CSIR-CRI Council for Scientific and Industrial Research - Crops Research Institute

DAPC Discriminant Analysis of Principal Components

GCA general combining ability

GT4SP genomic tools for sweetpotato

HEBS heterosis exploiting breeding scheme

HI harvest index

ICRISAT International Crops Research Institute for the Semi-Arid Tropics

MH mid-parent heterosis

MVA multivariate analysis

NARS National Agricultural Research System

NCSR number of commercial storage roots

NNCSR number of non-commercial storage roots

OFSP orange fleshed sweetpotato

PCA principal component analysis

PCoA principal coordinate analysis

PCR polymerase chain reaction

PIC polymorphic information content

QTL quantitative trait locus

RAPD random amplification of polymorphic DNA

SASHA Sweetpotato Action for Security and Health in Africa

SCA specific combining ability

SPCSV sweetpotato chlorotic stunt virus

SPFMV sweetpotato feahtery mottle virus

SPMMV sweetpotato mild mottle virus

SPVD sweetpotato virus disease

SSA Sub-Saharan Africa

SSP-WA sweetpotato support platform for West Africa

SSR simple sequence repeat

UPGMA unweighted pair group method with arithmetic mean

WCSR weight of commercial storage roots

WNCSR weight of non-commercial storage roots

CHAPTER 1

INTRODUCTION

The goal of this dissertation was to characterise the parents of the sweetpotato breeding program in Ghana, West Africa. Characterisation in this case means: to gather knowledge on the phenotypic performances of the parents and the parent-offspring relationship, and to identify the genotypic diversity of the parents. The work presented here aims at aiding the breeders in West Africa to set up a heterosis exploiting breeding scheme. All of this is thoroughly explained in the different chapters of this dissertation. This introduction chapter serves only as a guide through those chapters.

Chapter 2 starts by introducing the reader to sweetpotato. This chapter covers the importance of sweetpotato as a crop (Section 2.1), gives a review on the genetic constitution of the plant (Section 2.2) and briefly introduces where and how it is bred (Section 2.3). At the end of this chapter the research questions are defined (Section 2.4).

Chapter 3 reviews the concepts that are needed to answer the research questions of this dissertation. Starting with some broad information on general and specific combining ability (Section 3.1) and heterosis (Section 3.2), and ending with a more thorough background on SSR markers and their analysis in polyploid plants (Section 3.3).

Chapters 4, 5 and 6 cover respectively the materials and methods, the results and discussion and the conclusions of the research part of this dissertation. Throughout these chapters it is explained how the phenotypic and genotypic data on the sweetpotato parents were analysed and how this analysis has served to answer the research questions.

CHAPTER 2

INTRODUCTION TO SWEETPOTATO

2.1 The importance of sweetpotato

Sweetpotato (*Ipomoea batatas* (L.) Lam.) is a dicotyledonous plant belonging to the family of Convolvulaceae. The plant grows as a perennial vine and produces rather large, edible storage roots. Sweetpotato is very diverse, both phenotypically and genetically (see Section 2.2). Most notably are the skin and flesh colours of the storage roots that range from white to orange, red and pink all the way to purple.

The centre of origin of sweetpotato is assumed to be the North-Western part of South America (O'Brien, 1972). Starting from the 16th and 17th century the crop spread through Europe, Asia, the Pacific and Africa and nowadays it is grown on about 9.2 million hectares for an annual storage root production of over 100 million tonnes (compared to about 19.3 million hectares and well over 350 million tonnes for potato (FAOSTAT, 2016)). Figures 2.1 and 2.2 show the trends in sweetpotato yield and harvested area from 1998 to 2016 (FAOSTAT, 2016). It is clear that the key players in sweetpotato production are Asia and Africa, each accounting for roughly half of the world's harvested area. Notice that the harvested area in Africa more than doubled during the last two decades. Nevertheless, it is Asia that accounts for about 80% of the world production expressed in mass equivalents. This is because the yield per hectare in Asia is 4 times higher compared to Africa[1] (20 t/ha versus 5 t/ha). It has been pointed out by other authors already that this gap represents an enormous potential to increase global sweetpotato yield by introducing improved varieties and cultural practices (Oswald *et al.*, 2009; Grüneberg *et al.*, 2015). Figure 2.2 illustrates that in Europe the yield per hectare was able to increase drastically over the course of just a few years, presumably because of better agricultural practices that were introduced due to an increased interest in sweetpotato in this region of the world.

[1] However, Grüneberg *et al.* (2015) mention a probable underestimation of the yield in African countries, due to an overestimation of the harvested area.

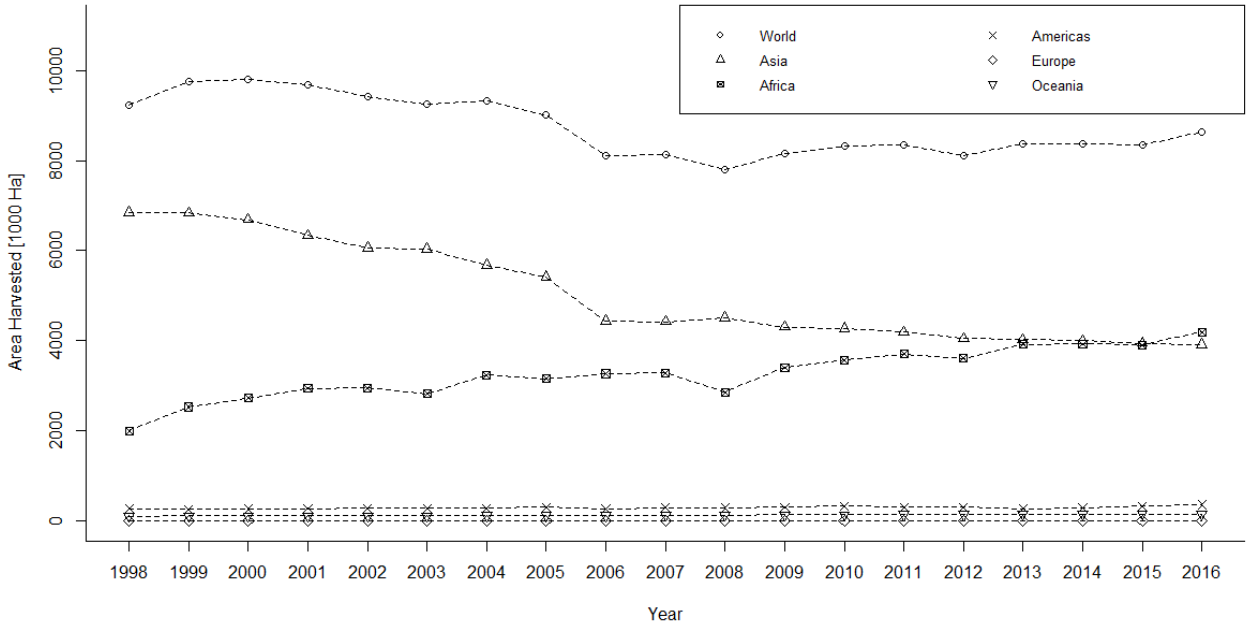


Figure 2.1: The trend in the harvested area of sweetpotato for the entire world, Asia, Africa, the Americas, Europe and Oceania from 1998-2016 (FAOSTAT, 2016).

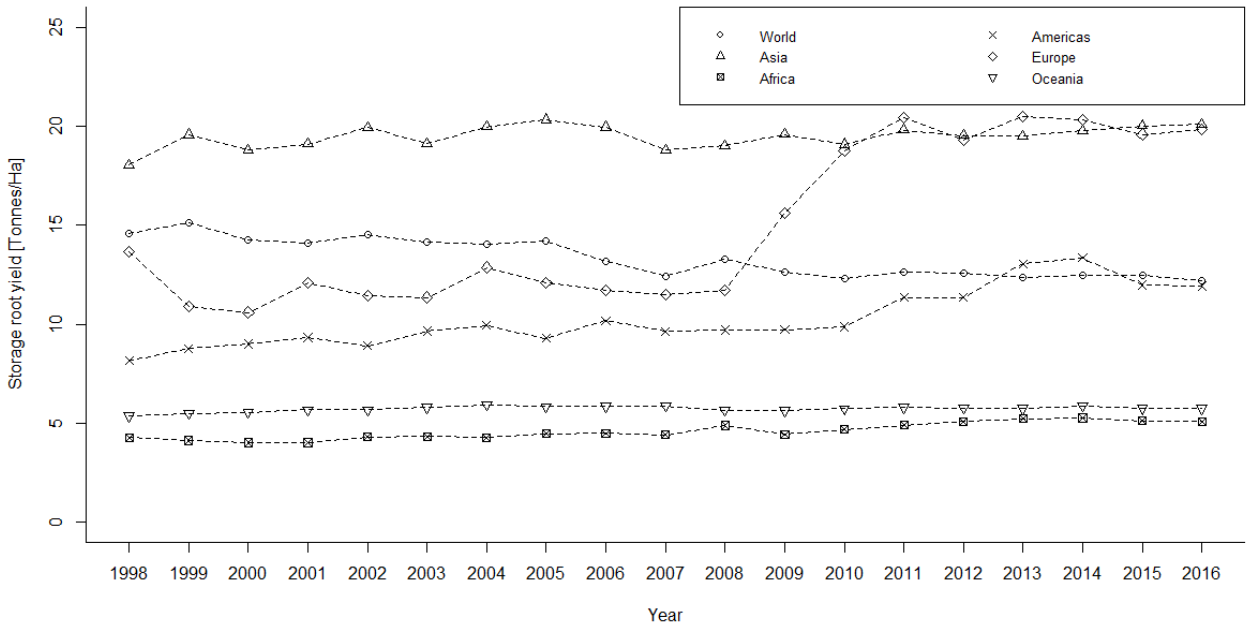


Figure 2.2: The trend in the yield of sweetpotato for the entire world, Asia, Africa, the Americas, Europe and Oceania from 1998-2016 (FAOSTAT, 2016).

Sweetpotato is often acclaimed for its high nutritional value, relatively high drought tolerance once established and decent yield under poor agronomic conditions (Low *et al.*, 2007; Grüneberg *et al.*, 2015). It's a good source of several minerals, including K, P, Mg, Zn and Fe, but also vitamins C, K, E and B (Bovel-Benjamin, 2007; Grüneberg *et al.*, 2015; Low *et al.*, 2017). Orange fleshed sweetpotato (OFSP) on top of that contains *β*-carotene which the body can convert into vitamin A, and thus can play an important role in battling vitamin A deficiencies in Sub-Saharan Africa (SSA) (Low *et al.*, 2001, 2007). An excellent summary of the rise of sweetpotato as a biofortified food security crop in SSA and what discoveries and initiatives made it all possible was given by Low *et al.* (2017). Next to being an important food crop as such, sweetpotato is also processed into many food products and it is used as animal feed (Woolfe, 1992). More recently, sweetpotato is also considered for the use in bio-ethanol production and colourant production (Grüneberg *et al.*, 2015).

2.2 The genetics of sweetpotato and its consequences

Sweetpotato is a hexaploid (Section 2.2.1), cross-fertilizing (Section 2.2.2) plant that can be easily vegetatively propagated (Section 2.2.3). These 3 characteristics have important consequences for the genetic constitution and the breeding of sweetpotato.

2.2.1 Hexaploidy

The genus *Ipomoea* contains species of varying ploidy levels. Sweetpotato is a hexaploid with base chromosome level of 15 (2n = 6x = 90). There has been a lot of debate on the origin of the polyploidy of sweetpotato. Two contrasting hypotheses were formulated in the 1980's: Kobayashi (1984) postulated that polyploidisation within *I. trifida* lay at the basis of hexaploid sweetpotato (autopolyploid), Austin (1988) suggested that it was the interspecific cross between *I. trifida* and *I. triloba* (allopolyploid). Recent molecular work (Roullier *et al.*, 2013; Munoz-Rodriguez *et al.*, 2018; Feng *et al.*, 2018) has shown that allopolyploidy can be ruled out for sweetpotato and that *I. trifida* is the most probable progenitor of hexaploid sweetpotato. The mode of inheritance, auto- or allopolyploid, is of major importance for anyone studying the population genetics of sweetpotato. Indeed, in a strict allopolyploid only bivalents can be formed between homologous chromosomes of the same genome copy during meiosis, leading to disomic inheritance. In an autopolyploid however, multivalents can be formed between all homologous chromosomes, leading to specific phenomena such

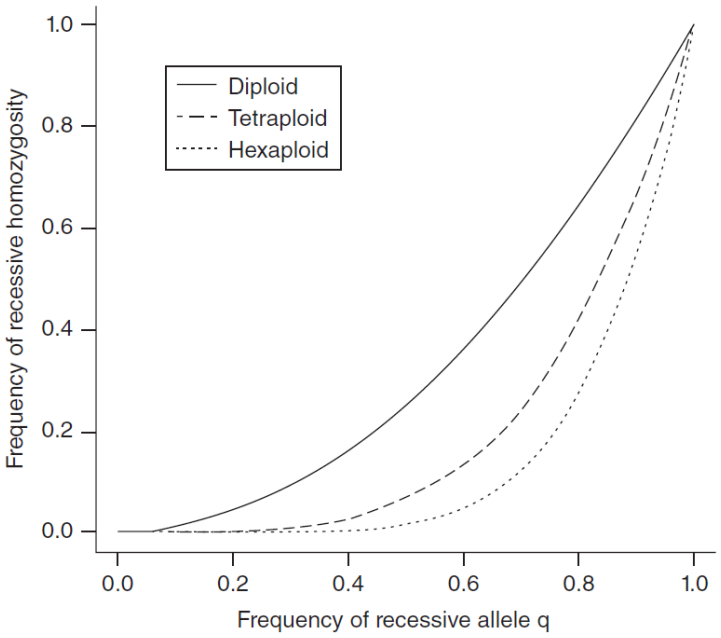


Figure 2.3: Effect of the ploidy level on the frequency of recessive homozygous genotypes in a random mating biallelic population at equilibrium as a function of q, the frequency of the recessive allele. Copied from Grüneberg *et al.* (2015).

as double reduction (the phenomenon whereby sister chromatids end up in the same gamete, thus increasing the homozygosity of a population. Excellent illustration in Parisod *et al.* (2010)) that change the segregation pattern compared to disomic inheritance. This needs to be accounted for when studying and analysing autopolyploid populations (Wu *et al.*, 2001; Parisod *et al.*, 2010; Hardy, 2016). Linkage map studies have shown that sweetpotato follows the autopolyploid route of inheritance (i.e. polysomic inheritance), but some preferential pairing during meiosis occurs (Kriegner *et al.*, 2003; Cervantes-Flores *et al.*, 2008).

The polyploidy of sweetpotato has an important influence on the expression of recessively inherited attributes in a population. Figure 2.3 illustrates that, even when the frequency of a recessive allele is high, the frequency of homozygous genotypes is rather low in polyploids. As a consequence, it is difficult to observe recessive traits, and to select for (or against) them in the field. Important recessive attributes in sweetpotato include virus resistance and non-sweetness after boiling (see Section 2.3.1).

2.2.2 Allogamy

Sweetpotato is a monoecious plant and sexual reproduction happens mostly through cross-fertilisation. Self-fertilisation is hampered by a strong, sporophytic self-incompatibility system (Gurmu *et al.*, 2013). True-seed set through cross-fertilisation happens easily and about 1-3 seeds can be obtained per flower from a successful pollination.

2.2.3 Vegetative propagation

Sweetpotato is very easily vegetatively propagated, predominantly through vine cuttings. Grüneberg *et al.* (2009b) mention a propagation coefficient (multiplication factor) of 30-90 for sweetpotato, depending on the propagation method. Independent of the exact number, this propagation coefficient is very high and it should be clear that a sweetpotato plant can be easily multiplied (cloned) for both agronomical trials and variety release.

When two sweetpotato parents are crossed, the segregation in their offspring is large due to the autopolyploid and generally very heterozygous nature of sweetpotato. In this context the easy clonal propagation is a blessing, since none of this diversity should theoretically be lost. Indeed, every seed from the cross can grow into a plant with a unique genotype that in turn can be propagated infinitely many times and can thus be considered a unique variety (more on variety development in Section 2.3.2). The downsides of vegetative propagation however, should not be forgotten. Barker *et al.* (2009), when reviewing the challenges in the distribution of sweetpotato planting material in SSA, mention two problems that are unavoidably linked to clonal propagation of the crop. First is the inadequate supply of planting material at the onset of the rains. Indeed in regions of the world with an extended dry season (such as West Africa) vegetative parts of the plant have to survive these periods without rain. If not, simply no (or late) planting material is available for the next growing season. This is of course not an issue for crops distributed through seed. Second is the quality assurance of the planting material. It is a fact that diseases spread easily and unavoidably through vegetative propagation. For sweetpotato this is predominantly important for viruses (see Section 2.3.1). Proper screening and *in vitro* cleaning and propagation of planting material are needed to minimise this detrimental effect of clonal propagation.

2.2.4 Genetic mapping

Genetic mapping studies and sequencing of sweetpotato are challenging. This is because, due to its autopolyploid nature, alleles in a segregating population can combine in a large number of combinations and, due to the lack of good self-compatibility and its hexaploid nature, no pure homozygous inbred lines can be made in a reasonable amount of time. On top of that there is the large chromosome number and genome size that add to the complexity of working with sweetpotato genomic data (Kriegner *et al.*, 2003; Cervantes-Flores *et al.*, 2008; Yada *et al.*, 2017b; Yang *et al.*, 2017). Considerable effort has been put in genetic mapping studies for sweetpotato in order to link important traits with genetic markers. This can aid the breeding of sweetpotato by unravelling the genetic correlations among important traits such as β-carotene levels, sugar content and dry matter content, but also help discovering resistances to important diseases (see Section 2.3.1). Once unambiguous markers have been established, these could facilitate the breeding of sweetpotato through marker assisted breeding (Chang *et al.*, 2009; Cervantes-Flores *et al.*, 2011). Notable efforts are the identification of QTL markers for dry matter, starch and β-carotene content and yield traits using both AFLP and SSR markers (Chang *et al.*, 2009; Cervantes-Flores *et al.*, 2011; Yada *et al.*, 2017b) and the identification of viral disease resistances through AFLP and RAPD (Mwanga *et al.*, 2002a; Yada *et al.*, 2017a). Important sequencing efforts include that of the diploid relative of sweetpotato, *I. trifida* (Hirakawa *et al.*, 2015) and recently the resolution of the sweetpotato haplotype (Yang *et al.*, 2017).

2.3 Breeding sweetpotato

2.3.1 The sweetpotato support platform for West Africa

This dissertation was written in collaboration with the sweetpotato support platform for West Africa (SSP-WA). The SSP-WA fits into the larger SASHA (Sweetpotato Action for Security and Health in Africa) initiative that is led by the International Potato Center (CIP) and funded by the Bill & Melinda Gates Foundation. SASHA aims at improving the lives of poor families in SSA by exploiting the so called 'untapped potential of sweetpotato' (Low, 2012). This comprises mainly the alleviation of poverty and undernutrition in SSA.

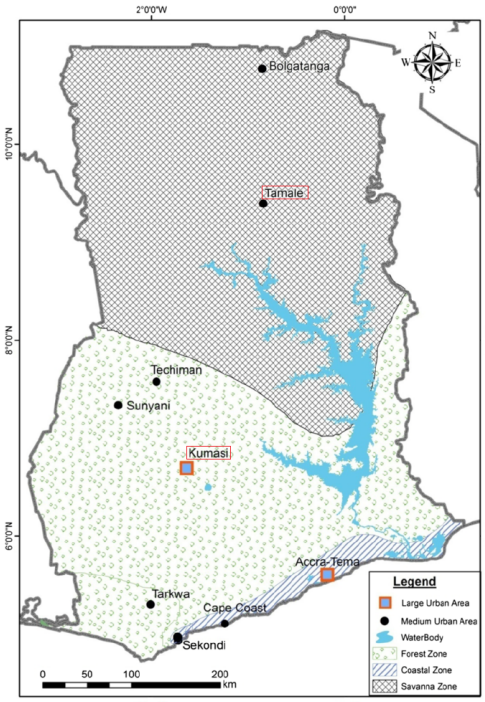


Figure 2.4: Ghana map showing the three important ecological zones and the locations of the CSIR-CRI (Kumasi) and the CSIR-SARI (Tamale). Forest zone = Humid tropics. Adapted from Kudom *et al.* (2015).

The SSP-WA breeding platform is based in Ghana at the CSIR-Crops Research Institute (CSIR-CRI), Kumasi and at the CSIR-Savannah Agricultural Research Institute (CSIR-SARI), Tamale. These locations represent the two main agro-ecological zones in West Africa: Kumasi is located in the humid tropics, Tamale represents the drought-prone savannah (Figure 2.4). Main responsibilities of the platform are the introduction of new germplasm and the early stages of the breeding process after which the national programs take over for variety release and distribution (Carey, 2013).

As mentioned above, sweetpotato has diverse uses and thus breeding sweetpotato can serve many purposes. Important breeding objectives for West Africa include, but are not limited to:

1. Increasing storage root yield. A straightforward increase in yield is presumably the goal of most breeding initiatives. It was already illustrated in Section 2.1 that there is a large potential for storage root yield increase in Africa. Main ways of achieving this goal are the release of new varieties and the introduction of better agricultural practices.

2. Improving quality. This includes objective characteristics such as nutritional value (e.g. the amount of *β*-carotene, micro-nutrients, vitamins) but also subjective characteristics such as taste and mouthfeel. For West Africa the ideal sweetpotato should have a dry mouthfeel, high dry matter and low sweetness (Grüneberg *et al.*, 2009a, 2015; Low *et al.*, 2017). This was initially problematic for the introduction of OFSP in SSA since these in general had a moist mouthfeel, relatively low dry matter and sweet taste (Grüneberg *et al.*, 2015). Unfortunately, a negative genetic correlation was found between *β*-carotene levels and both dry matter and starch content based on field trials and linkage map studies (Grüneberg *et al.*, 2009a; Cervantes-Flores *et al.*, 2011). Despite these difficulties an evaluation of the SSA germplasm has shown increased dry matter and *β*-carotene levels in OFSP varieties (Tumwegamire *et al.*, 2011) and in 2016, 42 OFSP adapted to SSA consumers preferences had already been released (Low *et al.*, 2017).

3. Increasing resistance to sweetpotato virus disease (SPVD). SPVD causes severe losses in sweetpotato yield all over the world (Carey *et al.*, 1999) and the disease is very prevalent in SSA (Mwanga, 2001; Mwanga *et al.*, 2002b). SPVD is caused by the synergistic effect of the infection with 2 viruses: sweetpotato feathery mottle virus (SPFMV) and sweetpotato chlorotic stunt virus (SPCSV). Sometimes a third virus is mentioned: sweetpotato mild mottle virus (SPMMV). For in-depth information on the causes of and yield losses by SPVD I refer to Carey *et al.* (1999); Mwanga (2001); Gibson *et al.* (2004); Tairo *et al.* (2005). Genetic components of resistance to SPVD have been identified in sweetpotato germplasm (Mwanga *et al.*, 2002a; Yada *et al.*, 2017a). It is suggested that 2 separate recessive genes govern resistance to SPCSV and SPFMV separately, but also quantitative gene effects seem to play a role.

2.3.2 Variety and population development

The next two sections summarise a selection of the extensive information provided by reviews of Grüneberg *et al.* (2009b) and Grüneberg *et al.* (2015) on the breeding of vegetatively propagated crops, more specifically sweetpotato.

Its high ploidy, lack of self-compatibility and generally very heterozygous nature make sweetpotato a very diverse crop, both phenotypically and genetically. Crossing two of these heterozygous plants as parents creates a wide range of offspring genotypes/phenotypes, each being a potential variety since every genotype can be cloned indefinitely. In this sense, sweetpotato is referred to as a clone hybrid.

The general breeding scheme of clone hybrids is represented in Figure 2.5. In a first step sexual reproduction breaks the normal vegetative reproduction and variation in genotypes is created. This can happen through open pollination in a polycross field or through controlled crossing of the parents. In subsequent steps the best plants are selected and cloned to be able to test these potential varieties the following growing seasons, each time on larger plots and/or more locations until variety release. If we do not consider somatic mutations, the released variety is genetically identical to the initial true seed plant after the sexual reproduction step. During the early selection stages, emphasis is on qualitative traits, such as storage root shape and colour and nutritional quality. Later on, when plot size and the number of locations increase, selection also happens on quantitative traits such as storage root yield and vine yield (important for the multiplication through vine cuttings). Needless to say the exact execution of this general scheme depends on the breeding goals and resources. Figure 2.6 represents the situation in Ghana for the SSP-WA. The crossing block is located in Kumasi: here true seed formation takes place through open pollination and controlled crosses. Seedling nursery, observational trials, preliminary trials, advanced trials and varietal trials are the names of subsequent selections steps, each time with fewer genotypes that pass the selection, but plot size and the number of locations increase. These trials are separated for the two agro-ecological zones, humid tropics and drought-prone savannah, that are represented by the southern and northern regions of Ghana respectively. Although crossover of material does happen, this separation is necessary because the breeding goals differ slightly between these regions. Drought is for example a more serious threat in the savannah region, whilst viral infection is worse in the humid region (Carey, 2013).

The process described above is referred to as variety development: developing a new variety from the crosses within a parent population. It is *in se* a straightforward process of selecting the best genotype by testing it in many locations at several occasions. Less straightforward is to select which and how many parents should be used to make crosses with. Indeed, after every sexual reproduction cycle, new potential parents are created, but of course not all of these can or should be used as parents the next year. Parents are often selected based on their own performance, but in a highly heterozygous polyploid with strong self-incompatibility, this is not necessarily a good indicator of their breeding value. However, it is exactly the offspring performance that should be improved year by year, through selection of good parents. Improvement of the parent population is difficult but crucial to find improved sweetpotato varieties. The process of parent improvement is referred to as population development.

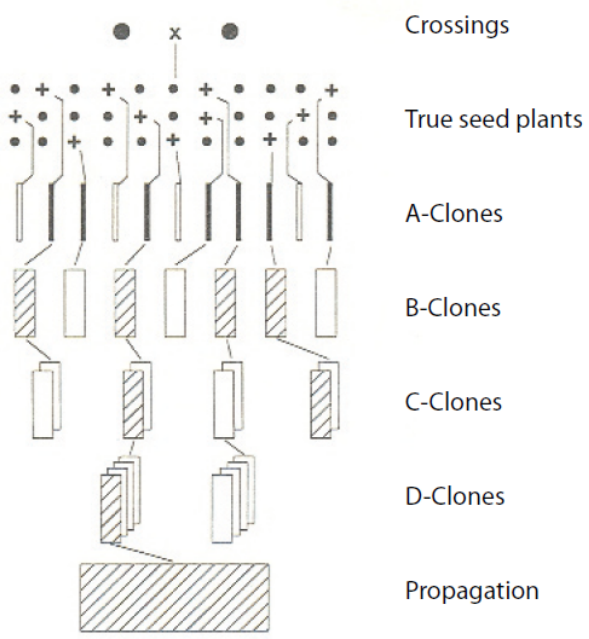


Figure 2.5: Schematical representation of the general breeding scheme for clonally propagated crops. Copied from Grüneberg *et al.* (2009b), who adapted it from Becker (1993).

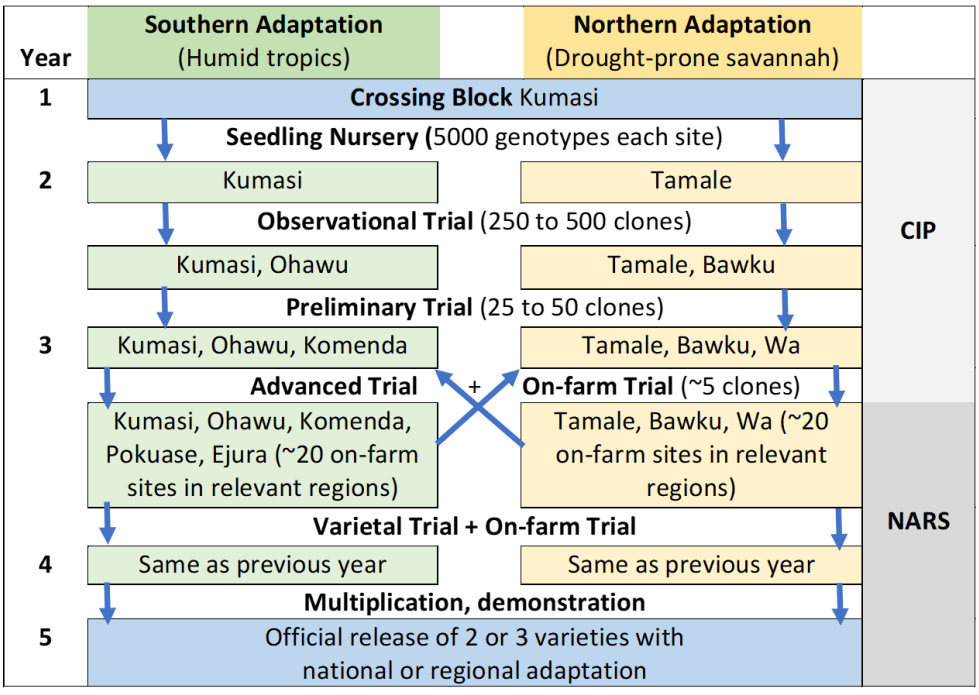


Figure 2.6: Schematical representation of the breeding scheme of the SSP-WA in Ghana. Kumasi, Ohawu, Komenda, Pokuase and Ejura are located in the Humid tropics. Tamale, Bawku and Wa are located in the Savannah region. On the right is indicated which organisation takes responsability for the trials: CIP or the National Agricultural Research System (NARS).

2.3.3 Heterosis exploiting breeding scheme

In recent years there has been increased attention to adapting the breeding scheme of sweetpotato for population improvement based on the exploitation of heterosis. Heterosis is the observed phenomenon that the offspring of a cross can outperform the parents significantly. Heterosis presumably plays a very important role for quantitative traits in sweetpotato. An in-depth review on heterosis is postponed to Chapter 3, but for now it is sufficient to point out that the heterosis effect is only observed between populations at a certain genetic distance from each other; so called heterotic groups. Only crosses between heterotic groups show heterotic effects. In a heterosis exploiting breeding scheme (HEBS) a population of parents is divided into two mutually heterotic groups, A and B. Population development happens separately in both of these groups, i.e. for population development no crosses are made between A and B. This infers a certain degree of inbreeding within each group. Selection of parents within group A or B is then based on reciprocal recurrent selection between the two groups, so based on combining ability and not on parent performance *per se*. Variety development then happens through the general scheme as presented in Section 2.3.2, the sexual reproduction step being a cross between group A and B.

A HEBS has two important characteristics. (1) Maximum exploitation of the heterosis effect should result in rapid improvement for quantitative traits. (2) Due to a certain degree of inbreeding within each heterotic group, the breeding efficiency for recessively inherited traits should improve (see Section 2.2.1).

2.4 Research questions

The main goal of this dissertation is to help the sweetpotato support platform in Ghana set up a HEBS by subdividing the available parents into mutually heterotic groups. The primary questions come down to:

1. What is the phenotypic parent-offspring relation in the starting population?
2. How genetically diverse are the parents in the starting population?
3. How can we subdivide the parents into (preferably 2) groups based on both genotypic and phenotypic data on these parents?

CHAPTER 3

CONCEPTS AND ANALYSES FOR INFERRING HETEROTIC GROUPS

In the following sections the most important concepts, needed to answer the research questions of this dissertation, are described. Starting with some general knowledge on general and specific combining ability and the concept of heterosis (Sections 3.1 and 3.2), and ending with the concepts that form the basis of answering the research questions (Sections 3.2.3 and 3.3).

3.1 General and specific combining ability

This section forms a brief introduction to the concepts of general combining ability (GCA) and specific combining ability (SCA) in quantitative genetics and (plant) breeding. Fully reviewing these concepts is beyond the scope of this dissertation, but it is necessary to highlight a few ideas that are important in the light of heterosis (Section 3.2) and population division (Section 3.2.3).

GCA and SCA are generally used to assess the breeding value of lines (parents) in cross combinations. The terms were first defined by Sprague and Tatum (1942) when evaluating inbred lines of corn used in hybrid production: "The term ’general combining ability’ is used to designate the average performance of a line in hybrid combinations (...) The term ’specific combining ability’ is used to designate those cases in which certain combinations do relatively better or worse than would be expected on the basis of the average performance of the lines involved.".

To clarify and elaborate (based on basic quantitative genetics handbooks of Acquaah (2012) and Falconer and MacKay (1996)): the evaluation of these combining abilities is based on crossing each parent with several other parents and evaluating their offspring. The mean performance of a parent over all of its crosses, assessed by the mean performance of all of its offspring and expressed as a deviation from the overall mean performance of all crosses between all parents, is called the general combin-

ing ability. It gives a *general* assessment of the performance of a parent in a cross combination and is directly linked to the *additive* effect of this parent. Indeed, an expected value of the performance of the offspring of a cross can be estimated by the sum of the general combining abilities of both parents. A deviation of the offspring performance of a *specific* parent combination from this expected performance can be attributed to *non-additive* effects, and is assessed by the specific combining ability. This leads to a very elegant and intuitive statistical interpretation of GCA and SCA: GCA can be considered to be the main effect of each parent and SCA the interaction effect of two parents in a cross combination. This can be described by a linear equation:

$$X_{AB} = \bar{X} + GCA_A + GCA_B + SCA_{AB} \tag{3.1}$$

where subscripts A and B denote two parents, X_{AB} is the measured offspring performance of these two parents and $\bar{X}$ is the overall mean of all offspring of all parents. Note that the values for GCA and SCA are only applicable to the population for which they were calculated, since they are relative to that population's offspring performance. Equation 3.1 forms the most basic model, but several variations exist that include reciprocal and maternal/paternal effects, epistasis and dominance effects (Henderson, 1948; Griffing, 1956).

Several statistical methods exist to estimate GCA and/or SCA for a population based on different crossing schemes (Acquaah, 2012), but the most noteworthy for this dissertation are the methods of Griffing (1956). These methods are extensively used and rely on partial or full diallel crossing schemes with or without including data on the parents. Note that these crossing schemes imply a certain balance in the design (each parent is present in the same number of crosses with each other parent) and involve plenty of cross combinations within the population of parents. An imbalanced diallel or missing data, e.g. because of cross incompatibility or practical difficulties, complicate the analysis. A solution for the analysis of these incomplete diallels comes in the form of mixed linear models and Bayesian statistics (Balzarini, 2002; Lenarcic *et al.*, 2012). Mixed linear models are extensively used in animal breeding and their applications in this field were spear-headed by the work of Charles R. Henderson (Robinson, 1991). They have now found their way into plant breeding as well (Piepho *et al.*, 2008).

Estimations of GCA and SCA are also used in sweetpotato breeding and population studies. A few examples include the studying of inheritance of root dry matter (Shumbusha *et al.*, 2014), drought tolerance, yield and maternal effects (Rukundo *et al.*, 2017) and SPVD resistance (Mwanga *et al.*, 2002b).

3.2 Heterosis

Heterosis, or hybrid vigour, is the observed phenomenon of increased performance of the heterozygous progeny of parental inbred lines, compared to these inbred lines. This increased performance mostly comprises higher biomass, fertility and perseverance under different stress conditions (Hochholdinger and Hoecker, 2007; Veitia and Vaiman, 2011).

Any review on heterosis inevitably points to Darwin (1876) for the first scientific observations on the fact that inbreeding has deleterious effects and cross-fertilization is beneficial. It was not until the early twentieth century, with the influential work of Shull (1908) and East (1908) in maize, that the term heterosis was first used and the phenomenon thoroughly studied and exploited in plant breeding (Hochholdinger and Hoecker (2007); Schnable and Springer (2013) and Crow (1998) for a historical perspective).

The magnitude of heterosis in the F1 generation of the cross of two parental inbred lines is often described by the relative phenotypic values of mid-parent or better-parent heterosis (Figure 3.1). Mid-parent heterosis expresses the performance of the F1 relative to the average of the two parents. Indeed, if we consider phenotypes to be additive, we expect the heterozygous F1 to perform at the average level of its two homozygous parents; any deviation from this mean has to come from other effects, summarised under the term heterosis. Better-parent heterosis expresses the performance of the F1 relative to the performance of the best performing inbred line. With the eye on crop improvement, better-parent heterosis is agronomically the most relevant measure of heterosis (Schnable and Springer, 2013).

3.2.1 The molecular basis of heterosis

Several genetic hypotheses on the causes of heterosis were formulated and extensively discussed over the past century. Three of them that have remained standing and have supporting scientific evidence are summarized below, based on reviews of Hochholdinger and Hoecker (2007); Birchler *et al.* (2010); Veitia and Vaiman (2011); Schnable and Springer (2013).

- The dominance hypothesis, based on the idea of Jones (1917), postulates that heterosis is the result of the complementation of unfavourable recessive alleles present in the inbred parents, by superior dominant alleles coming from the other parent, when brought together in the hybrid. This coming together of superior

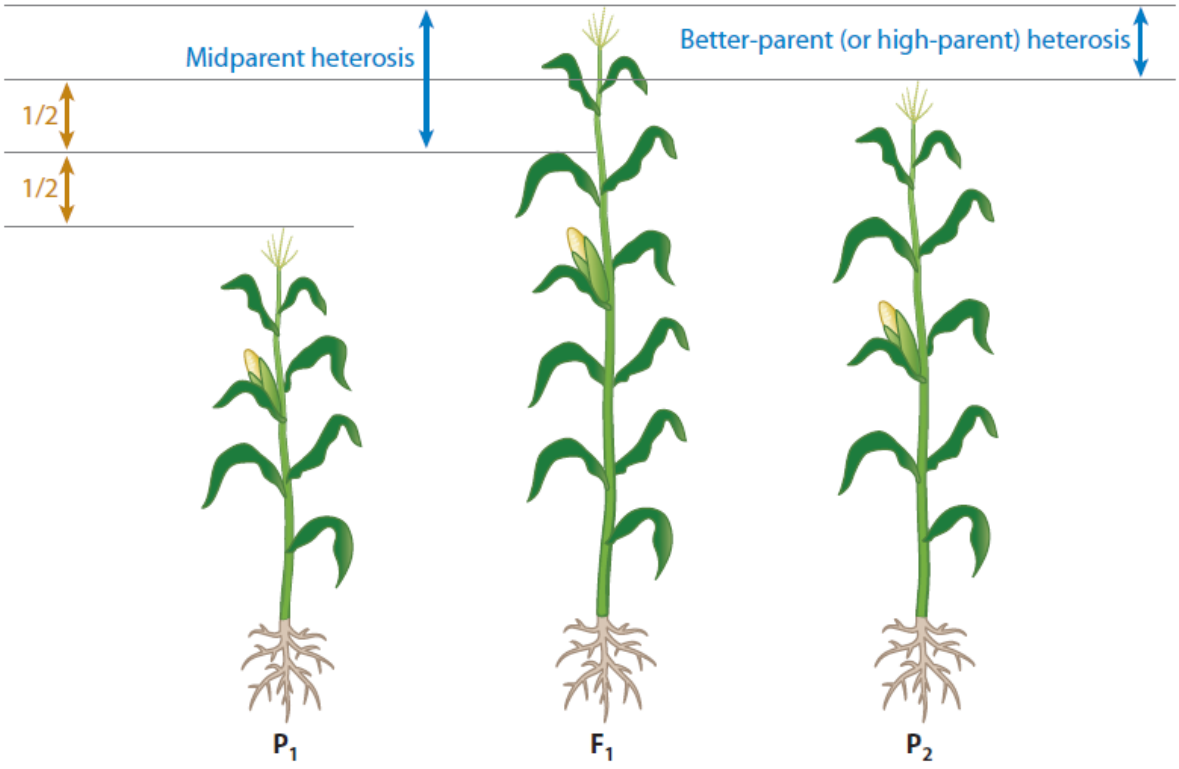


Figure 3.1: An illustration of the concepts of mid-parent and better-parent heterosis. Adapted from Schnable and Springer (2013).

dominant alleles happens over multiple independent loci. As a consequence of this idea, it should be possible to obtain an inbred line containing all of these dominant alleles across all loci that has equal performance to the hybrid.

- The overdominance hypothesis traces back to the original ideas of Shull (1908) and East (1908) that heterozygotes have an intrinsic advantage. It postulates that interaction between the heterozygous alleles at several loci in the hybrid causes the superior phenotype compared to the parental inbred lines. As a consequence it should be impossible to create an inbred line with a performance equal to that of the heterozygote.

- The basis for the epistasis hypothesis was postulated by Powers (1944) as an addition to the work of Jones (1917). It hypothesises that non-allelic interactions between loci cause the superior phenotypes of hybrids.

It was aptly reviewed by more specialised literature (e.g. Birchler *et al.* (2010); Acquaah (2012); Schnable and Springer (2013)) that there is evidence that supports as well as contradicts all of these hypotheses. A consensus, however, is that genetic variation is needed to invoke heterosis.

In a more recent effort to find a unifying theory of the molecular basis of heterosis Birchler and Veitia (2007) built on the gene balance hypothesis. The gene balance hypothesis suggests that regulatory and/or metabolic genes exhibit a stoichiometric balance and thus gene dosage plays an essential role in plant phenotypes. This concept is not new and provides a generally accepted framework to describe aneuploidy phenotypes, sex determination in some species and also links to quantitative traits (Birchler and Veitia, 2007). The gene balance hypothesis, when applied to heterosis, does not exclude dominance, overdominance or epistasis but rather emphasizes the interplay between all inter- and intra-locus interactions as well as allelic dosage in causing heterotic phenotypes (Birchler *et al.*, 2010; Veitia and Vaiman, 2011). Furthermore, it helps to explain peculiar observations concerning heterosis, such as the effect of single genes or genomic segments and progressive heterosis in polyploids (see Section 3.2.2) (Birchler *et al.*, 2010).

To summarize: the molecular basis of heterosis remains elusive, even today. However, one important consensus can be drawn: heterosis is linked to heterozygosity, genetic variation and diversity and is generally determined by a large number of loci (East, 1936; Hochholdinger and Hoecker, 2007; Veitia and Vaiman, 2011; Schnable and Springer, 2013).

3.2.2 Heterosis in polyploids

The conclusion of Section 3.2.1 brings us seamlessly to heterosis in polyploids. It is recognised that heterosis plays an important role in determining the performance of polyploid plants (Grüneberg *et al.*, 2009b; Chen, 2010; Sattler *et al.*, 2016). Considering the fact that allelic interplay plays an important role in heterosis, it is straightforward to see that polyploids have a lot of heterotic potential. Indeed, polyploids have more alleles per locus than diploids, hence many more combinations of alleles can be present in a single plant. On top of that they are intrinsically 'more heterozygous' (see Section 2.2.1).

Polyploids exhibit progressive heterosis. This entails that maximizing the diversity of genomes in a polyploid, maximizes the heterosis; or put differently: it appears that heterosis improves with the greater number of distinct genomes present in the polyploid (Birchler *et al.*, 2003, 2010). Exemplified for a tetraploid: ABCD is most of the times superior to AABB and CCDD. This is in perfect accordance with and serves as evidence for the idea that heterosis increases as the genetic distance between the parental lines increases (East, 1936; Charcosset *et al.*, 1991; Melchinger and Gumber, 1998; Chen, 2010; Sattler *et al.*, 2016).

3.2.3 Heterotic groups

The concept of heterotic groups is very well defined by Melchinger and Gumber (1998): "... a heterotic group denotes a group of related or unrelated genotypes from the same or different populations, which display similar combining ability and heterotic response when crossed with genotypes from other genetically distinct germplasm groups." In other words, if we take two heterotic groups, A and B, inter-group crosses should show greater heterotic effects than intra-group crosses. The division of a population into mutually heterotic groups is also referred to as 'establishing a heterotic pattern'. Once heterotic groups are established, their complementarity can be further improved by reciprocal recurrent selection (see Section 2.3.3 and Falconer and MacKay (1996)). A well-known example of such heterotic groups in maize are the Flint and Dent lines, established in Europe and North America respectively, of which the hybrids Flint x Dent prove to be high yielding (Melchinger and Gumber, 1998; Rincent *et al.*, 2014).

To divide a population into heterotic groups two main strategies can be used (Melchinger, 1999; Fan *et al.*, 2009): the first and oldest method is based on agronomic, geographic and pedigree data, the second, newer method is based on genetic diversity. Both are explained in the paragraphs below.

Heterotic groups based on agronomic data

The division of a population into mutually heterotic groups based on agronomic data relies on the assessment of the SCA between the members of the population, sometimes combined with pedigree information and hybrid heterosis data (Fan *et al.*, 2009; Acquaah, 2012; Tian *et al.*, 2015). Indeed, since heterosis is linked with non-additive phenotypic effects, the use of the SCA makes perfect sense (Section 3.1). The division happens on the basic idea that the SCA between heterotic groups should be large and positive and the SCA within a group should be smaller or negative (Vasal *et al.*, 1992; Menkir *et al.*, 2004; Librando and Magulama, 2008). Caution should be taken on how the SCA values are calculated; it has for example been shown that including reciprocal crosses or not in the diallel can influence the estimation of the SCA values and the subsequently implied heterotic pattern (Fan *et al.*, 2013).

Instead of using the SCA values for every cross combination in the population, often testers are identified. Testers can be identified based on established heterotic patterns (such as the Flint x Dent groups in maize) or can be identified for a new population. The testers for categorising a genotype in either group A or B should have a very large and positive SCA between them. A new genotype that needs to be

categorized, showing a large SCA with the tester for group A and a small SCA with the tester for group B should be classified under group B, and vice versa. For genotypes showing positive SCA with both testers no easy decision can be made (Menkir *et al.*, 2004; Librando and Magulama, 2008; Tian *et al.*, 2015). Identifying testers for a population and classifying new genotypes based on the testcross performance is useful for large populations where the estimation of all SCAs between all crosses becomes cumbersome (Melchinger and Gumber, 1998). Several variations exist on this methodology, by for example also including GCA in the decision making (Fan *et al.*, 2009; Badu-Apraku *et al.*, 2013; Tian *et al.*, 2015).

Heterotic groups based on molecular marker data

Using molecular data to establish a heterotic pattern is based entirely on the idea that heterosis increases as the genetic diversity (distance) between the heterotic groups increases (Section 3.2.1). Hence the problem shifts from population division based on the agronomic assessment of cross performance, to a population structure problem: how can we subdivide a population of individuals (genotypes) based on genetic diversity?

The potential of molecular markers, such as AFLP and SSR, to assess genetic diversity has long been recognized, also in the light of heterosis (Melchinger, 1999). Numerous examples can be cited of their use and development in all crops, also in sweetpotato (Zhang *et al.*, 1999; Yada *et al.*, 2010; Ngailo *et al.*, 2016; David *et al.*, 2018; Meng *et al.*, 2018).

Since the analysis of molecular marker data to infer population structure and diversity is quite a specialized subject that can also be used outside the concept of heterosis, it is reviewed in a separate Section 3.3 together with SSR markers and their analysis in polyploid organisms.

3.3 Molecular markers and their analysis in polyploids

The following sections discuss molecular markers, SSR markers in particular, and their analysis aimed at inferring population structure based on genetic diversity, with an emphasis on particular difficulties in polyploid organisms.

3.3.1 SSR markers

Molecular (DNA) markers reveal sites of variation in DNA at the sequence level (Jones *et al.*, 1997; Collard *et al.*, 2005). Jones *et al.* (1997) describe them as 'neutral', because the variations in molecular markers do not necessarily show themselves in the phenotype, as would be the case with morphological markers. Indeed, markers may indicate variation within a gene, but also in non-coding regions of the DNA. In some cases, molecular markers are tightly linked to a gene, but this is not a prerequisite.

Molecular markers are typically associated with size differences of a particular piece of DNA, differences that arise from mutations and/or rearrangements of the DNA that happen spontaneously over time. Molecular markers are able to reveal genetic differences between individuals that can be visualised with, for example, a PCR reaction and common electrophoresis techniques (Collard *et al.*, 2005). The different sizes of a piece of DNA that can be detected with a particular marker are called marker 'alleles' and similarly one talks about 'loci' to denote the particular DNA positions where these alleles occur.

A particular class of genetic markers are microsatellites, also known as short tandem repeats or simple sequence repeats (SSR). They are non-coding DNA regions made up out of small motifs of 1 to 6 nucleotides that are repeated several times (Vieira *et al.*, 2016). Variation in SSR length (polymorphism) is mainly attributed to the addition or deletion of entire motifs through strand-slippage during DNA replication or repair, and SSR markers show a very high mutation rate compared to other parts of the genome (Oliveira *et al.*, 2006; Sehn, 2015). Assuming this simple mutation model (known as the stepwise mutation model), one can state that two alleles differing by only one motif length are more closely related than if the difference is two motifs or more (Slatkin, 1995; Oliveira *et al.*, 2006).

SSR markers are co-dominant markers, this means that all different alleles can be detected, contrary to dominant markers. Co-dominant markers allow the easy distinction of heterozygotes versus homozygotes (at least in diploids) as is illustrated in Figure 3.2.

3.3.2 SSR marker difficulties

Because of their high mutation rate, high occurrence in the genome of many organisms, co-dominant nature and ability to link related alleles by a mutation model based on allele size, SSR markers are often acclaimed for their use in studies of population structure and genetic diversity (Sunnucks, 2000; Oliveira *et al.*, 2006; Guichoux *et al.*,

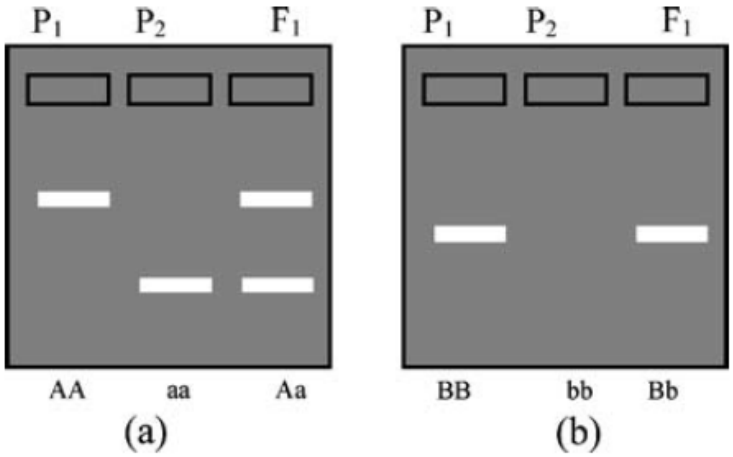


Figure 3.2: (a) Co-dominant markers show a band for every allele and can thus clearly distinguish homozygotes from heterozygotes. (b) For a dominant marker this is not possible. Copied from Collard *et al.* (2005).

2011; Dufresne *et al.*, 2014; Vieira *et al.*, 2016). However, a few important problems exist for the analysis of SSR marker data, especially in the case of polyploids:

- The first and foremost problem when analysing SSR data of polyploids is the fact that allelic dosage cannot be accurately determined (Dufresne *et al.*, 2014; Meirmans *et al.*, 2018). Take a tetraploid organism for which three SSR bands were detected on a gel after electrophoresis. Because of the co-dominant nature of SSR markers we can say this individual has three distinct alleles (A, B and C), with a particular size, at this particular locus. It is however impossible to say whether the allelic constitution of this individual at this locus is either AABC, ABBC or ABCC. An immediate consequence is the fact that it is impossible to estimate allele frequencies for a population, a critical parameter for many population genetics studies. Note that this problem does not arise for diploid organisms. There are a few solutions that, for example, rely on estimating the dosage from the peak ratios of the fluorescent intensities of electrophoresis bands (Esselink *et al.*, 2004) or that use iterative procedures for estimating allele frequencies based on the observed 'allelic phenotypes' (e.g. the method of De Silva *et al.* (2005)). However, none of these solutions are flawless (Dufresne *et al.*, 2014).
- When any evolutionary or phylogenetic interpretation is to be given to the SSR data, the problem of homoplasy plays an important role. Homoplasy entails the fact that, since SSR alleles are based on fragment size, it is impossible to determine whether two bands of the same size (identical in state), are also identical in descent (Chen *et al.*, 2002). Indeed, two bands of the same size can arise from different sources of mutations (insertions, deletions, substitutions) that can only be revealed at the sequence level. Two bands of the same size in two different

genotypes are therefore not necessarily derived from a common ancestor. This is a grave problem and asks for particular caution when inferring relationships from SSR data (Doyle *et al.*, 1998). The only real solution to the problem of homoplasy is shifting from marker analysis based on size to analysis based on sequence as suggested by Guichoux *et al.* (2011) and Dufresne *et al.* (2014).

- SSR regions can show perfect or imperfect repeats. A perfect repeat SSR consists of uninterrupted repeats of the same multi-nucleotide motif, whereas in imperfect SSRs the repeats are interrupted by a pair of bases that does not match the motif (Oliveira *et al.*, 2006). Guichoux *et al.* (2011) and Estoup *et al.* (2001) suggested that using as much as possible perfect repeat SSRs is the best guarantee that the SSRs will follow the stepwise mutation model. Not using perfect repeat SSRs may thus complicate the analysis.

- Since SSR data are based on PCR amplification and subsequent electrophoresis, all problems associated with these techniques can complicate the analysis (reviewed by Guichoux *et al.* (2011)). An example of such a problem is that of 'null alleles'. Null alleles arise when mutations occur in the flanking regions of the SSR locus, resulting in poor PCR primer binding and a lack of amplification of this allele. This results in an apparent homozygote when in fact the genotype is heterozygous. The occurrence of null alleles can be corrected for during data analysis (e.g. Chapuis and Estoup (2007)), but it is best to avoid this problem by thoroughly checking all locus candidates for null alleles (Guichoux *et al.*, 2011). Both the corrections during data analysis and checking of SSR marker loci are based on offspring studies and gene frequencies, not a trivial task considering the previous remarks on determining genetic dosage in polyploids.

3.3.3 SSR marker data analysis

Raw SSR marker data, coming from capillary electrophoresis, consist of a collection of electrophoresis peaks for every locus under consideration and for every genotype being studied. The first step in the analysis is to convert these peaks to alleles defined by a specific size; this step is known as 'allele calling'. Several commercial electrophoresis software packages, such as GeneMapper (Applied Biosystems), are capable of at least partially automating this step. Often a manual check is advisable (as reviewed by Guichoux *et al.* (2011)).

The raw allele sizes obtained in this way are often not integer values corresponding to exact base lengths. Measured fragment lengths may differ slightly from the expected base length, and to assign alleles to these fragments 'binning' is required (Amos

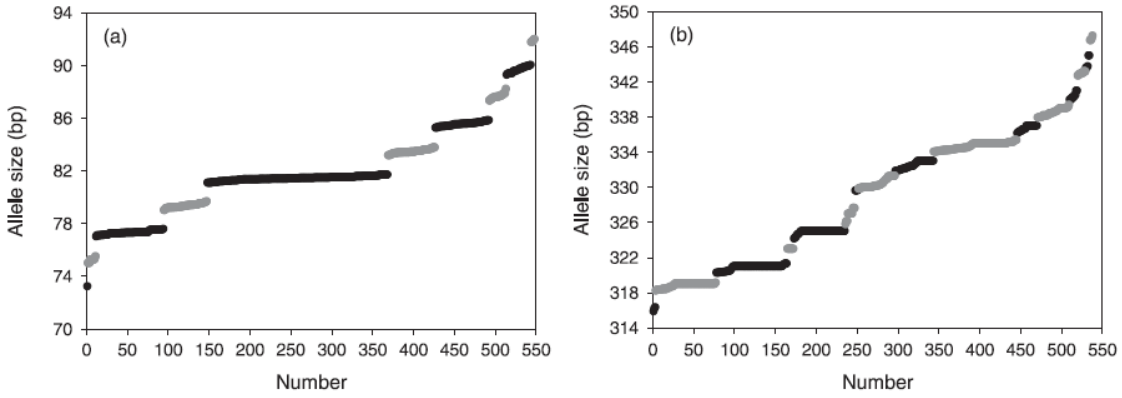


Figure 3.3: Illustration of cumulative allele length distributions for 2 dinucleotide SSR markers. Alternate allele categories (bins) are represented by different colours. (a) This markers shows a very clear spacing between each bin. (b) There is some overlap between the bins, and even manual binning would struggle to categorize these sizes correctly. Adapted from Amos *et al.* (2007).

et al., 2007; Guichoux *et al.*, 2011). Binning categorises measured allele lengths (real number) to a specific allele size (integer). The most straightforward means of binning include rounding to the nearest integer size or fitting bins defined by the repeat motif length (Amos *et al.*, 2007). For example: a dinucleotide marker with measured allele sizes 200.45 and 204.60 would be converted to actual alleles 200 and 204, based on repeat motif size. Although perfect repeat SSR marker loci are expected to differ in exact motif sizes from each other, this is not always the case. Measuring differences due to GC content and single base mutations may sometimes change this pattern and thus complicate the binning procedure (Amos *et al.*, 2007). Free software has been developed, based on least-squares minimization procedures, that performs automated binning and allows for a manual check of the result (e.g. FLEXIBIN (Amos *et al.*, 2007), ALLELOBIN (Idury and Cardon, 1997), see also review by Guichoux *et al.* (2011)). An excellent way of checking the binning procedure is by plotting the cumulative allele size distribution for every locus (Figure 3.3). A clear spacing between every allele bin indicates good allele calling and binning.

Population structure and differentiation

The binned SSR data can now be used for several genetic analyses. For this dissertation it is useful to briefly describe some of the most important methods to infer population structure and differentiation. These methods are based either on (1) distance metrics, (2) multivariate analysis or (3) cluster analysis (excellent reviews by Dufresne *et al.* (2014) and Meirmans *et al.* (2018), also in the light of polyploid analysis).

1. The result of a distance metrics analysis is a matrix with pairwise genetic distances between all individuals. This distance matrix can be visually represented by a dendrogram. To understand distance metrics each individual can be represented by a vector of alleles, take individual A [200,202] and individual B [202,204]. These vectors can be compared by using the mathematical concept of group similarity/dissimilarity, for which there are plenty of measures such as the Jaccard index, the Dice index or the Lynch index (Kosman and Leonard, 2005; Dufresne *et al.*, 2014). These measures are thus not based on any biological assumptions, but simply compare the presence and/or absence of alleles within each individual. As such, allele 200 is never equal to allele 202, but 202 of individual A is equal to 202 of individual B. An alternative distance measure that allows for a more biologically relevant interpretation, and that was specifically developed for polyploid analysis is the Bruvo distance (Bruvo *et al.*, 2004; Clark and Jasieniuk, 2011; Dufresne *et al.*, 2014). The Bruvo distance measure assumes that the stepwise mutation model for SSR loci is valid and it thus allows to account for these mutations. Alleles 200 and 202 are no longer completely unequal (distance = 1), but instead they are said to be only one mutational step apart (for a dinucleotide marker), and thus the distance is lower than 1. Subsequently the distance between alleles 200 and 204 (two mutational steps) is also smaller than 1 but larger than the distance between 200 and 202.

   Distance metrics are commonly used for population genetics, but the result strongly depends on the choice of the metric. No universal and completely correct choice exists amongst the similarity measures for any type of marker or ploidy level (Kosman and Leonard, 2005), and even the result of the Bruvo distance should be interpreted with caution (Meirmans *et al.*, 2018).

2. Multivariate analyses (MVAs), such as principal component analysis (PCA), are common mathematical tools in all scientific fields. Several methods have been adapted for the use with genetic data for inferring population structure (Jombart *et al.*, 2009). These methods usually rely on very few biological assumptions. Discriminant Analysis of Principal Components (DAPC) (Jombart *et al.*, 2010) and AMOVA-based K-means clustering (Meirmans, 2012) are two methods built on well-established mathematical frameworks: PCA, discriminant analysis, K-means clustering and/or AMOVA (Excoffier *et al.*, 1992). These two methods rely on the estimation of allele frequencies and might therefore not be ideal for the analysis of polyploid data where dosage estimation can be troublesome (Dufresne *et al.*, 2014; Meirmans *et al.*, 2018). Other methods, such as a principal coordinate analysis (PCoA), only rely on a pairwise distance matrix, which can always be calculated, but then the choice of the distance metric becomes crucial again.

3. A commonly used clustering method is implemented in Structure (Pritchard *et al.*, 2000). Structure uses a Bayesian approach and the idea is to assign individuals to one or more populations such that deviations from a Hardy-Weinberg equilibrium are minimized (Pritchard *et al.*, 2000; Dufresne *et al.*, 2014; Meirmans *et al.*, 2018). Although Structure can be used for polyploid data, it has one very strong assumption: the populations are assumed to be in Hardy-Weinberg equilibrium. This assumption is very hard to accomplish in polyploids such as sweetpotato[1], thus causing potential problems and complications when using structure. There is a need for simulation studies to assess how grave these problems by assumption violations might be, but to date no such studies were published (Dufresne *et al.*, 2014; Meirmans *et al.*, 2018).

Given the difficulties with allele dosage estimation for polyploids, and the consequent impossibility to accurately estimate allele frequencies for polyploid populations, one should be very cautious when choosing a method to analyse polyploid SSR data that relies on these allele frequencies. Add to this problem the other issues and uncertainties off SSR data (Section 3.3.2) and it becomes clear that analysing such data is always an *ad hoc* procedure of which every step should be thoroughly considered.

The SSR analysis presented in Chapters 4 and 5 will therefor rely mostly on good quality control of the data and avoiding the use of methods that rely on allele frequencies. Of the methods presented in this section, distance metrics seem the most appropriate for this dissertation's goals. Indeed, since these methods calculate pairwise distances between individuals they do not rely on population allele frequencies and the Bruvo distance was even built with the idea of handling missing dosage information at the level of the individual. In some cases the distance matrix can then be used for a MVA, such as a PCoA or AMOVA, to further investigate population structure.

[1] Several assumptions of a Hardy-Weinberg equilibrium are violated in sweetpotato, for example: random mating cannot be assumed due to self-incompatibility and vegetative propagation (Section 2.2). Due to polysomic inheritance, Hardy-Weinberg equilibrium is also reached more gradually than in diploid plants (Parisod *et al.*, 2010).

CHAPTER 4

CHARACTERISING SWEETPOTATO PARENTAL MATERIAL: MATERIALS AND METHODS

The following chapters seek to answer the research questions proposed in Section 2.4, based on phenotypic and genotypic data on the sweetpotato parents in the Ghana crossing block. This chapter describes the materials and methods that were used, Chapter 5 summarises and discusses the results and Chapter 6 ends with the final conclusions.

The materials and methods are divided into 2 parts: Section 4.1 describes the phenotypic analysis of the sweetpotato parents based on agronomic tests of parent and offspring performance; Section 4.2 describes the genotypic (SSR marker) analysis.

4.1 Phenotypic analysis

4.1.1 Plant material

Twenty-two of the sweetpotato parents at the Kumasi crossing block were used for the phenotypic heterosis trial. Controlled, manual crosses between these parents resulted in enough planting material to conduct a field trial with 149 different cross combinations. Parents were used both as males and females, but not necessarily in equal amounts (see Figure 4.1). These offspring, together with the 22 parents, were planted in 5-plant plots of $1.5m^2$ which were replicated 4 times. Each 'genotype' is thus represented by 20 plants, but note that whereas for the parents these are indeed 20 identical clones, for the cross combinations these are actually 20 different siblings.

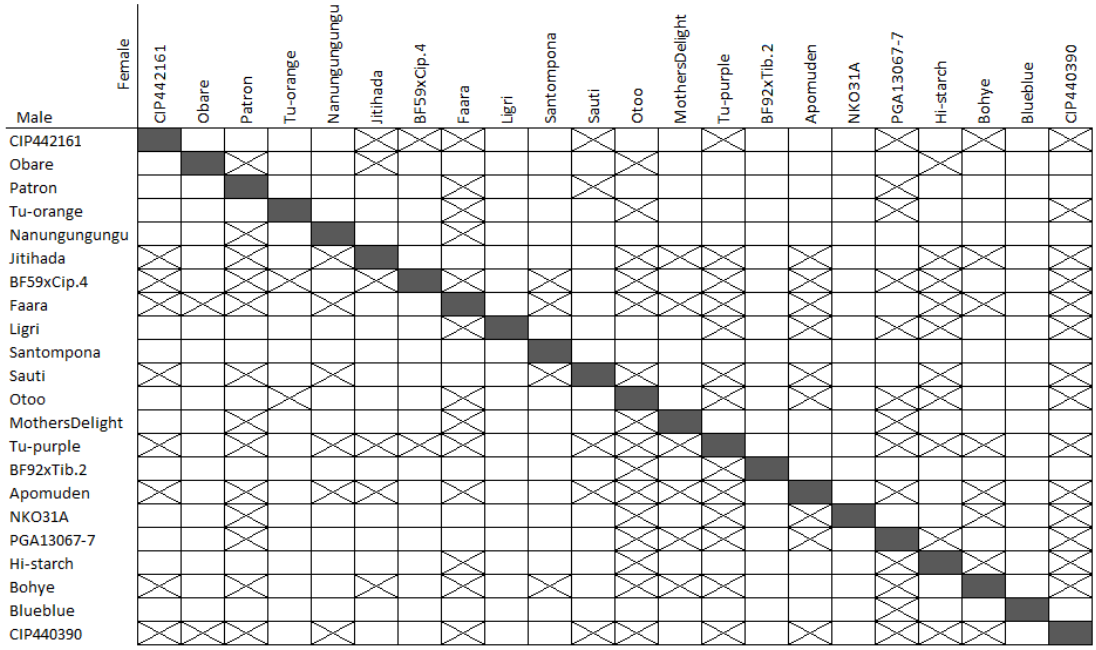


Figure 4.1: The crossing scheme of the 149 cross combinations between the 22 parents that were made and used in the heterosis field trial. Crossed out cells denote the successful crosses.

We are thus assessing family performance of the cross combinations and genotype performance of the parents, but for simplicity I will, from now on, refer to these as 171 (149 + 22) different genotypes.

The plots ($4 \cdot 171 = 684$) were planted according to a Westcott design (Westcott, 1981) with the parents 'Bohye' and 'Ligri' as checks (see Section 4.1.3), at the trial site in Fumesua (Kumasi), Ghana on the 3th of August 2017. The plots were grown under good agronomical practices until harvest and measurements took place on the 1st of December 2017.

4.1.2 Measurements

Before and during harvest several parameters were measured or estimated from the plots, the full dataset is freely available from SweetPotatoBase (https://sweetpotatobase.org/), with trial name: 2017ASPGH_HT-FUMESUA. The measurements are briefly described here:

- The number of commercial ($> 100g$ per root; NCSR) and non-commercial ($\leq 100g$ per root or damaged; NNCSR) storage roots were counted and expressed as number of roots per ha.

- The weight of the commercial (WCSR) and non-commercial storage roots (WNCSR) as well as the weight of the vines were measured separately and expressed as tons per ha (fresh weight). The sum of these weights equals the total biomass.
- The number of plants that successfully established, the number of plants that were harvested and the number of plants that grew storage roots were recorded.
- Vine vigour was estimated with a score from 1 to 9 during harvest. A score of 1 indicates nearly no vines have grown, a score of 9 indicates very strong, thick and long vines with short internodes.
- Six weeks after planting and 1 month before harvest, virus symptoms were visually estimated from the plots (virus symptoms 1 and 2 respectively), with a score from 1 to 9. A score of 1 indicates no virus symptoms were present on any of the plants within a plot, a score of 9 indicates severe virus symptoms on all plants in a plot.
- The harvest index (HI) was calculated as the weight of the commercial storage roots divided by the total biomass weight (based on fresh weight).

4.1.3 Westcott design and adjustment

The plots were planted according to a Westcott design (Westcott, 1981) that allows for an adjustment of the data for spatial heterogeneity. Figure 4.2 illustrates such a design. In a Westcott design all test plots (t) are positioned completely at random, and every n columns, a column of check plots (A and B) is planted. The checks alternate in both columns and rows and form a grid that spans the entire field. Since these checks are the same (vegetative clones) across the entire field, we can observe field heterogeneity by observing the variation in the performances of the checks, and adjust the measurements of the test plots for this heterogeneity. A basic adjustment would proceed as follows: the mean (e.g. for commercial root

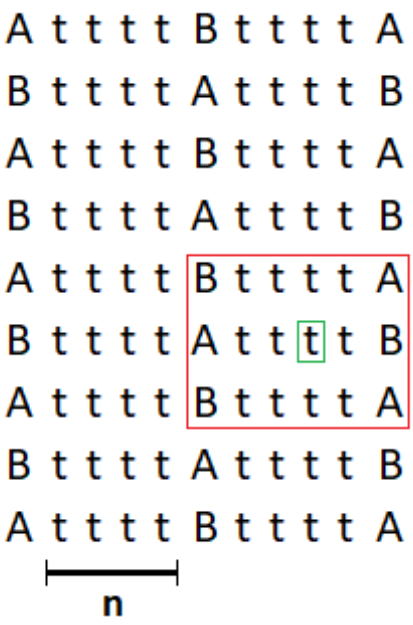


Figure 4.2: An illustration of a Westcott design with testplots (t), checks (A and B) and $n = 4$. A subregion of 3 rows for the testplot in the green rectangle, is indicated in red.

weight) of the checks across the entire field will be calculated, say 100 t/ha. Next, the mean of the checks for a small subregion around a test plot is calculated; the subregion consists typically of the 3 or 5 rows around the test plot and the mean is calculated based on these 6 or 10 check performances (see Figure 4.2). Take the mean of this subregion for the checks to be 110 t/ha and the measurement for the test plot to be 90 t/ha. Since the performance of the checks in this subregion is higher than the average performance of the checks across the entire field, the measurement of the test plot should be adjusted downwards with a factor 110/100. The adjusted plot value is then 90/1.1 = 81.8 t/ha.

For the actual adjustment a few extra parameters are available:

- The means of the checks for both the entire plot and the subregion can be calculated for each genotype (A and B) separately or the values for both genotypes can be combined in a single check mean. Consequently the adjustment of the test plots can happen based on the difference of the subregional mean(s) from the separate means or the combined mean.
- The adjustment can be weighted by the distance of the test plot to the checks, the closest checks would receive more weight in this case.
- The number of check rows to take into consideration can be adjusted, typical values are 3 or 5 check rows.
- In all cases a weighing factor can be added to the adjustment factor, ranging from 1 (full adjustment) to 0 (no adjustment).

To select the best possible adjustment for field heterogeneity from these different options, a simple simulation can be run. For this simulation we will use the fact that the parents in this field trial are each replicated 4 times as identical clones. These parents are distributed completely at random across the field and in a (hypothetical) homogeneous field we expect the replicates to perform equally. To select the best parameters for the Westcott adjustment, we can simulate the adjustment of the most important parameter (WCSR in our case) with all possible parameter combinations (combined or separate means, distance weighted or not, 3 or 5 check rows and global adjustment ranging from 0 to 1 in steps of 0.05). The best adjustment parameters are then those that minimize the variation in WCSR for all parents together. This is assessed by minimizing the mean squared error (MSE) of a simple ANOVA model:

$$WCSR_p = G_p + \epsilon \tag{4.1}$$

where G_p is the mean of the p^{th} parent and subscript p implies that the model is only fitted for the parents; ϵ is a normally distributed error term.

The adjustment parameter combination with the lowest MSE gives us the optimal parameter combination. The same optimal adjustment parameters are then used to adjust the other quantitative measurements (other than the WCSR for which the adjustment was optimized).

The Westcott design for the heterosis trial was a 23 by 34 grid of plant plots throughout which the 4 replicates of the 171 genotypes were distributed completely at random. For the checks the parents 'Bohye' and 'Ligri' were used; the 1^{st}, 12^{th}, 23^{th} and last column were check columns (for a total of 92 check plots). The last 6 plot positions of the last row were not used, indeed $23 \cdot 34 - 6 = 4 \cdot 171 + 92 = 776$, which is the total number of plots that were planted and measured.

4.1.4 Analysis

The full analysis of the phenotypic field trial data (including the Westcott adjustment) was performed using R software version 3.3.1 (R Core Team, 2016). Most of the analysis makes use of basic mathematics and statistics. To test the assumptions of a statistical test, Levene's test was used (to test for homoscedasticity), and visual assessment (for normality). If all assumptions were fulfilled, parametric statistics were always used (ANOVA, t-test), otherwise less powerful non-parametric statistics were used (Kruskal-Wallis test, Wilcoxson rank-sum test); exceptions to this rule are explicitly mentioned. All hypothesis tests were conducted at the 5% significance level. An overview of the performed analysis is presented here:

1. *Erroneous data and data adjustment*. The raw data were checked for errors and inconsistencies, such as improbable outliers (using the inter-quartile range method from Tukey (1977) combined with a case-by-case assessment) and negative or clearly incorrect data input. Any such data point was set to NA. Next, the data were adjusted with the method of Westcott (1981) which is implemented in the st4gi R package (Eyzaguirre, 2019).

2. *Data exploration*. A basic data exploration was performed to get a general view on the data. Emphasis was put on comparing the parent population with the offspring population.

3. *Virus effect*. From the data exploration of virus symptoms 1 and 2 a remarkable difference between the parents and the offspring was observed. With basic hypotheses testing statistics it was tested whether there was a difference in viral

infection between parents and offspring and if this affected the yield measurements.

4. *Reciprocal effect*. Some of the parents were used in a sufficient number of reciprocal cross combinations to be able to test whether there was an effect on the WCSR of these parents' offspring, depending of the direction of the cross. To do this a simple ANOVA model was fitted for each parent separately and only for the reciprocal crosses that were available:

$$WCSR_{pij} = F_i + N_j + F_i \times N_j + \epsilon \tag{4.2}$$

whereby F represents the family (e.g. if we are investigating the reciprocal effect of 'Apomuden', the crosses 'Apomuden×Sauti' and 'Sauti×Apomuden' both belong to the same family), N is a binary that denotes the direction of the cross (e.g. 'Apomuden×Sauti' has direction 0; 'Sauti×Apomuden' has direction 1), and $F \times N$ is the interaction effect. The subscript p refers to the fact that the model is fitted for 1 parent at the time, subscipts i and j refer to the family of the cross and the direction respectively. A parental effect was considered to be significant if the interaction term ($F \times N$) was not significant (indeed, a significant interaction between the family and the direction of the cross would mean that the effect of the direction changes depending on the other parent, and thus is not consistent for the parent under investigation), and if the direction term (N) was significant (after removing the non-significant interaction term from the model).

5. *Mid-parent heterosis*. For all of the offspring a value for the mid-parent heterosis of the WCSR and total biomass was determined as:

$$MH = (\frac{\bar{Y}}{\bar{Y_{ab}}} - 1) \cdot 100 \tag{4.3}$$

Where $\bar{Y}$ is the mean of the offspring of parent combination $a \times b$, $\bar{Y_{ab}}$ is the mid-parent of parents a and b and MH is the value for the mid-parent heterosis expressed as a percentage difference from the mid-parent. The significance of every MH value was tested with a t-test as derived by Soehendi and Srinives (2005). Non-significant MH values were put to 0.

6. *Heterotic grouping*. Based on the parent-offspring agronomic data of this field trial it was considered how to determine heterotic groups for a certain characteristics using Griffing's method (Griffing, 1956) or mixed linear models.

4.2 Genotypic analysis

4.2.1 Plant material

A total of 51 sweetpotato accessions were used for the genotyping study. Forty-eight of them are parents at the Ghana crossing block. These parents are either introductions from all over the world, or the result of a successful cross in Ghana (Table A.1). Three accessions were added from the CIP genebank as checks: 'CIP199062.1', 'Cemsa' and 'Tanzania'. Although propagated separately for many years, these accessions were once identical to (and should thus still show great resemblance to) the Ghanaian parents 'Bohye', 'Ligri' and 'Sauti', respectively.

4.2.2 SSR marker amplification and allele calling

This part of the analysis was performed by dr. Mercy Kitavi, a postdoctoral scientist working for CIP in Nairobi, Kenya.

DNA from every accession was extracted from 1g of leaf tissue, using a modified method of Dellaporta *et al.* (1983) and Mace *et al.* (2003) (fully described in Section A.1.1). The DNA concentration was estimated using spectrophotometry (NanoDrop, Thermo Scientific) and consequently diluted to a working concentration of 20 ng/μL.

Thirty-six previously developed SSR loci were amplified using PCR (fully described in Section A.1.2), using the correct primers and annealing temperatures (Table 4.1). Forward primers were fluorescently labelled and all loci were amplified individually. Four PCR products were then multiplexed, based on dye colour and expected fragment size, for the consecutive separation using capillary electrophoresis (3730xl Genetic Analyzer, Applied Biosystems), sizing (500 LIZ internal standard, GeneScan) and allele calling (using GeneMapper software v5.0 (Applied Biosystems) with manual verification).

4.2.3 Binning and marker selection

Binning was performed using allelobin, a program written by Prasanth *et al.* (2006), based on the least-squares minimisation algorithm by Idury and Cardon (1997). The allelobin software was kindly provided by dr. Abhishek Rathore of ICRISAT upon request. In order to run the 32-bit program on a 64-bit personal computer the MS-DOS emulator DOSBox (https://www.dosbox.com/download.php?main=1) was used. Allelobin was chosen over other binning software because of its simple workflow that

Table 4.1: Description of the 36 SSR markers that were analysed. This table includes marker names, forward and reverse primer sequences, repeat motifs, anealing temperatures and literature references.

Marker	Forward primer	Reverse primer	Ta [°C]	Repeat motif	Reference
IB297	gcaatttcacacacaaacacg	cccttcttccaccactttca	58	$(ct)_{13}$	Buteler *et al.* (1999)
IBC3	caagaaaaagaagtgaaacaaagg	ctgctgctgttgctgttcatt	56	$(agaag)_5$	Huamani *et al.* (unp.)
IBJ559	ctcactctttcctttctctctctg	acagcatgatctcgccggaacc	55	$(tc)_7(ta)_7$	Huamani *et al.* (unp.)
IBJ664E	cacatgccatggacgctccaa	gattcttctccttccagctcct	55	$(ctt)_6$	Huamani *et al.* (unp.)
IBN24	taatgaggtgtgatgatgggtacta	agtgaagttgaggtcaggaaaatc	60	$(ta)_5ga(ta)_3$	Huamani *et al.* (unp.)
IBN35	cgggactaagaccttcctctctat	agagcatctgcgtagtacttatcgt	62	$(ta)_8$	Huamani *et al.* (unp.)
IBN36	tttatactctcggaaccctaccc	cgggtgatagagagactgttgtt	62	$(tc)_8$	Huamani *et al.* (unp.)
IBN37	catgatggagctcataaatctcg	gtcactgtgtcctccagtttttc	55	$(ta)_7t$	Huamani *et al.* (unp.)
IBS137	tcaacagacgtcttcacttacc	tcgatagtatgatgtgaatcgc	60	$(ctt)_8$	Schafleitner *et al.* (2010)
IBS141	gaagcagtagttgtgttgcttt	ctctatctttatctcttccggc	60	$(cttt)_6$	Schafleitner *et al.* (2010)
IBS144	tcgaacgctttctacactctt	ctgtgtttatagtctctggcga	60	$(ttc)_9$	Schafleitner *et al.* (2010)
IBS146	gcaaacctcaaaaagcgtaa	tagaggaattgtagggagtggt	60	$(gtct)_5$	Schafleitner *et al.* (2010)
IBS147	tgtgtacatgagtttggttgtg	gaagtgcaactaggaaacatga	55	$(gca)_8$	Schafleitner *et al.* (2010)
IBS149	ccacctccttaggtatcagact	actactagcgctgcaaccttat	60	$(aga)_8$	Schafleitner *et al.* (2010)
IBS150	agtcccttgaatgtgtacctct	agctgcaatcatacagtcaatc	60	$(ct)_{13}$	Schafleitner *et al.* (2010)
IBS169	cgtactatgtttcccccattac	aatgcatctaccctccttacac	53	$(ttg)_8$	Schafleitner *et al.* (2010)
IBS174	agagaacaaaatcgggaagaac	cgaaatagagattgtaatgggg	60	$(aga)_7$	Schafleitner *et al.* (2010)
IBS186	cagaaacaagcaaagatctcac	ctgttgcttctcttctccttct	60	$(aag)_8$	Schafleitner *et al.* (2010)
IBS199	taactaggttgcagtggtttgt	ataggtccatatacaatgccag	60	$(aca)_7$	Schafleitner *et al.* (2010)
IBY40	agtgttgggactcataaagattctg	gaatgaaatacagtgacccgagag	60	$(gcg)_7gc$	Huamani *et al.* (unp.)
IBY41	gacgaggattcaaaggagaaatatg	gatatcttcatgagattaggcttcc	62	$(gaa)_6ga$	Huamani *et al.* (unp.)
IBY43	tcctagtattctacaccgttccttg	cggccaccggttatcgtcctcgt	62	$(gaa)_6g$	Huamani *et al.* (unp.)
IBY44	caagaagagcataagcgtgagat	gcgatctgagaaggtgataattg	52	$(aga)_6$	Huamani *et al.* (unp.)
IBY45	gtggctatcggttttcatctcc	cgatcatcaaagggtactgaaac	55	$(tca)_6$	Huamani *et al.* (unp.)
IBY46	tagtaacaccattacttattaactttg	tgtaatctcatggattgctcgtag	55	$(atc)_5at$	Huamani *et al.* (unp.)
IBY47	cttacagttcagtaccccgcaccat	tctggtaccgcctccgagagt	53	$(cag)_5ca$	Huamani *et al.* (unp.)
IBY48	cacccctatttctttctctccagt	taagtccggacctctttcctaatca	60	$(ccg)_5cc$	Huamani *et al.* (unp.)
IBY50	ctctcttcttatgagaaagccctgt	ttgatacattgtagtcctgctgct	55	$(aac)_5a$	Huamani *et al.* (unp.)
IBY51	gatgtcgtttagcggactgag	gtatcgtcacattcagcagcag	55	$(gcg)_5g$	Huamani *et al.* (unp.)
IBY52	aaacagatagcagagacgagatgag	cagatagtgtcaccaacactgaaga	55	$(gcg)_5g$	Huamani *et al.* (unp.)
IBY53	ccacgatctcggaaaccgccat	ggggcaaaaggtcttattcatat	55	$(gga)_5g$	Huamani *et al.* (unp.)
IBY54	gtccaagagaaagaaactgaagatg	aactattctgcacaactacatgctc	57	$(tgt)_5t$	Huamani *et al.* (unp.)
IBY56	caccatggatttcaaaccactactt	agggggagttgtcttgactggt	52	$(cct)_5$	Huamani *et al.* (unp.)
IBY58	acgacatggctctctctttctc	agtttcctttctcgacgcttct	55	$(gcg)_5$	Huamani *et al.* (unp.)
IBY59	gattaagcaggtgaaaagggaagt	gaagataccttcacttcagaacag	62	$(ggc)_5$	Huamani *et al.* (unp.)
IBY60	tctctctgttatgttatggtgatgg	gcgttttacaagattcagaaaccac	62	$(tat)_5$	Huamani *et al.* (unp.)

allows it to be usable for polyploid data with minimal manipulation of the data and allows for easy export to other programs. The program also provides the user with a useful quality parameter to assess the quality of the binning (see below).

Since there are many difficulties when analysing polyploid SSR data (see Section 3.3.2) and a large number of markers was available for this study, it seemed appropriate to make a selection of the best (most trustworthy) markers to continue the analysis with. This selection was based on several criteria:

- *Occurrence of a seventh allele*. Since sweetpotato is a hexaploid plant, we expect to see a maximum of six different allele calls per genotype per locus for a full heterozygote. However, for some genotypes on some loci, more than six alleles were called. Whether this is the result of a calling mistake, a mistake during SSR amplification or if this is biological in nature (e.g. aneuploidy), is uncertain. Because of a lack of access to the raw electrophoresis peak data, it was impossible to properly check the cause of this 'seventh allele' problem. From a precautionary point of view, it therefor seemed best to omit all loci where more than six alleles were called for at least one genotype.

- *Perfect repeat markers*. As mentioned in Section 3.3.2, the use of perfect repeat SSRs is preferred if one wants to make use of the stepwise mutation model to analyse SSR data. Since calculation of the Bruvo distance indeed implies this mutation model, as much as possible perfect repeat loci were selected for the analysis.

- *Allelobin measure of quality*. Allelobin calculates a measure of quality of the binning, as described by Idury and Cardon (1997). In essence, this quality measure is a measure of the variation of the allele length distributions within all bins for a given locus. A large variation within each bin indicates possibly poorly defined boundaries of the bins. Indeed, if we look at the illustrations of some cumulative allele length distributions in Figure 3.3, we see that appropriate binning coincides with a small variance within each bin's size distribution; a large variance within each bin makes it difficult for the program to unambiguously assign an allele length to the appropriate bin and leads to poor binning. A large value for the allelobin measure of quality indicates that a visual inspection of the binning is required (see below).

- *Visual binning control*. As a final control, cumulative allele length distribution plots were made for every locus, and the quality of the binning was visually checked. If the allelobin measure of quality had a very high value and visually the binning looked poor, the locus was removed from the analysis. In some cases the binning could be improved by manually re-assigning some alleles to a new bin.

After this quality check of the markers, the markers were also checked for their usefulness. Simple measures, such as the allele diversity and the polymorphic information content (PIC)[1], give an indication of the variation that is present in the population at this locus. A locus with high variation is more appropriate (more useful) to differentiate the different genotypes than a locus with low variation (loci with a PIC higher than 0.50 are considered highly informative (Botstein *et al.*, 1980)). Note that calculating the PIC relies on the estimation of the allele frequencies (which is inherently wrong for polyploid data with missing dosage information), nevertheless the PIC value remains indicative of the variation that is present. Estimation of allele frequencies and calculation of allele diversity and PIC were done using the R package polysat (Clark and Jasieniuk, 2011).

[1] Allele diversity measures the number of alleles present at a certain locus. The PIC is defined by Botstein *et al.* (1980) as: $PIC = 1 - (\sum_{i=1}^{n} p_i^2) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2p_i^2 p_j^2$, where p_i and p_j are allele frequencies of alleles i and j respectively, and n is the number of alleles. A value of 0 means there is no allelic variation, a value of 1 means all alleles are unique in the population.

4.2.4 SSR analysis

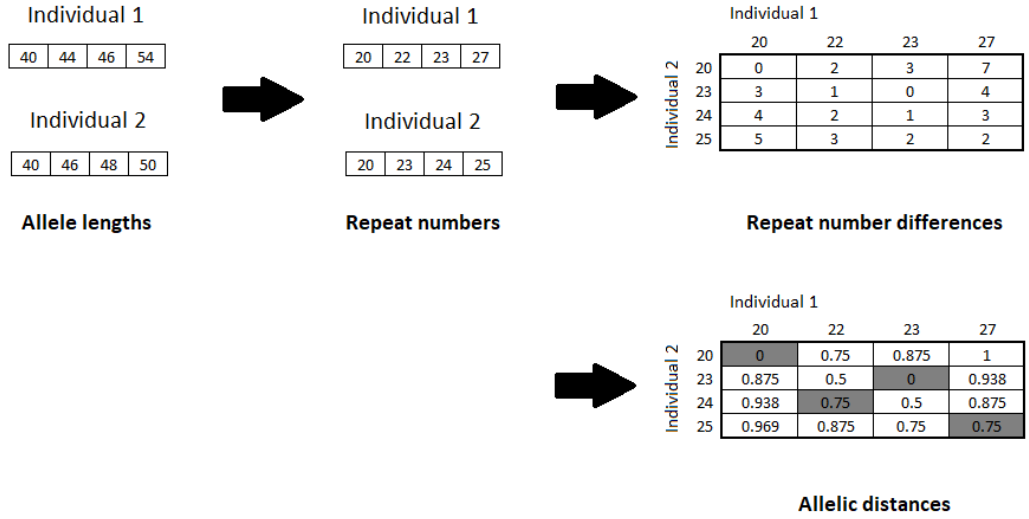
Three distance metrics were used as a first step to analyse the SSR data: (1) the Lynch distance (2) the Bruvo distance and (3) the FullBruvo distance. Because none of these distance metrics is necessarily perfectly suitable for this analysis (see Section 3.3.3), using 3 different distance metrics allows to compare their results and look for a consensus that can be presumed to be correct. All of these metrics were implemented in or adapted from the polysat package.

1. The Lynch distance (Lynch, 1990) is based on band similarity and does not take into account any mutation model. Lynch similarity is defined as 2 times the number of alleles that 2 individuals have in common, divided by the total number of alleles the 2 individuals have. The Lynch distance equals 1 minus the similarity. This distance is first calculated between all pairs of individuals for every locus separately, and then averaged over all loci to get a single value for the genetic distance between each pair of individuals.

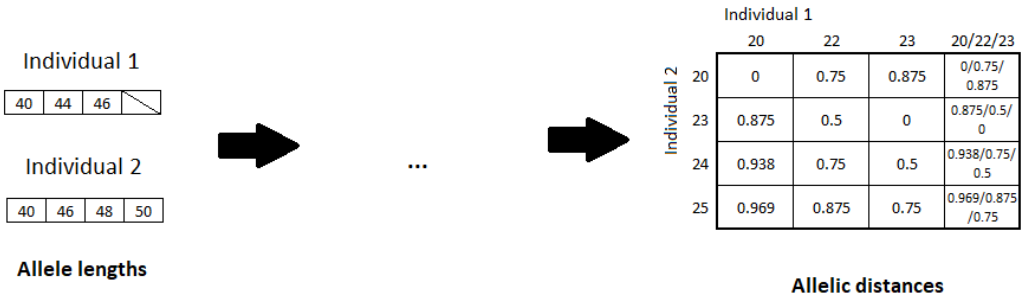
2. The Bruvo distance (Bruvo *et al.*, 2004) is based on the stepwise mutation model. Its calculation, as used in this dissertation, is illustrated in Figure 4.3a, for a single dinucleotide repeat locus, between 2 tetraploid individuals. First, all allele lengths are converted to their respective repeat number (e.g. for a dinucleotide marker, an allele length of 40bp becomes a repeat number of 20). Next, a matrix is constructed with every cell representing a repeat number difference between the alleles of both individuals. These repeat number differences are converted into genetic distances as:

$$d_a = 1 - 2^{-|x|} \tag{4.4}$$

where x represents the repeat number difference and d_a is the distance between 2 alleles (allelic distance). Next, the genetic distance between individuals (d_l) for this locus is calculated by finding the minimal sum of 4 numbers from the allelic distances matrix, whereby each time taking only one value from each row and one value from each column. In this way, every allele from individual 1 is compared to only one allele of individual 2 and the minimal distance between the 2 individuals is calculated (maximum parsimony principle). For the example in Figure 4.3a these 4 numbers are indicated by the grey boxes. Note that there are $4! = 24$ combinations that need to be calculated and compared for a tetraploid organism (6! for a hexaploid). The resulting minimal sum is normalized by dividing by the ploidy level. For the example in Figure 4.3a: $d_l = \frac{0+0.75+0+0.75}{4} = 0.375$. Averaging this d_l value over all loci (in our example



(a) The calculation in the case of full dosage information for both individuals



(b) The calculation in the case of missing dosage information for individual 1

Figure 4.3: Examples of the Bruvo distance calculation, for a single dinucleotide repeat locus, between 2 tetraploid individuals (a) with full dosage information or (b) with missing dosage information for individual 1; as explained in the text. Allelic distances are calculated using Equation 4.4, with x comming from the repeat number differences matrix.

only 1 locus is considered) gives us a single value for the distance between a pair of individuals.

Figure 4.3b illustrates the Bruvo distance calculation in the case of missing dosage information for individual 1. Calculations proceed in the same way as in Figure 4.3a but the missing allele of individual 1 is now replaced by each of the other 3 alleles present in this individual. As such, the abovementioned calculations happen 3 times, once for every possible genotype, and are then averaged. For the example in Figure 4.3b, we get 3 different minimal sums (1.375, 1.625 and 1.716) by each time replacing the missing allele in individual 1 by one of the 3 other possible alleles; these values are then averaged to get a single d_l value: $d_l = \frac{1.375+1.625+1.716}{3 \cdot 4} = 0.393$. In this way, the Bruvo distance is able to account for missing dosage information by simply assuming all alleles that are present

in a certain genotype have an equal probability of occurring more than once. For clarity: a tetraploid with only 3 called alleles [A,B,C] can have 3 genotypes, [A,B,C,A],[A,B,C,B] or [A,B,C,C], and in this case the Bruvo distance calculation assumes each of the genotypes is equally possible.

One can imagine that for an organism of high ploidy and a lot of missing dosage information, a large number of calculations have to be made, and calculating the Bruvo distance becomes very computationally intensive. Polysat therefor implements a simplified version of the Bruvo distance [2]: when calculating the distance between two tetraploid individuals with missing dosage information [A,B,C] and [C,D], they will be treated as triploids. In this way, not all possible genotypes have to be compared and this decreases the computational time. This method will from now on be referred to as the Bruvo distance, as compared to the Full-Bruvo distance (see below)

3. The FullBruvo distance calculates the complete Bruvo distance, with no compromise on computational time. This means that when dosage information is missing, all possible genotypes are considered. The code for the FullBruvo distance calculation was adapted from the polysat Bruvo2.distance function, and can be found in Section A.2. The algorithm was also made more efficient by parallelising some of the calculations, thus decreasing computational time.

Using these 3 distance metrics, bootstrapped (10000 replications) dendrograms were made with the unweighted pair group method with arithmetic mean (UPGMA). Polysat does not have a built-in bootstrapping function, but a bootstrapping function was written based on the resampling of loci using the R package ape (Paradis and Schliep, 2018). Bootstrap values are added at each bifurcation in the dendrogram and represent the number of times a specific cluster appeared over the 10000 bootstrap replications, divided by the number of replications and rounded to 2 decimal numbers.

In order to confirm the results of the distance metrics, a DAPC (Jombart *et al.*, 2010) was performed. Therefore, the data were converted into binary, presence/absence data for input into the adegenet R package (Jombart, 2008). A very interesting feature of a DAPC is its ability to determine the ideal number of groups (clusters) in a population, based on K-means clustering and the minimisation of the Bayesian information criterium (BIC). Note that a DAPC relies on allele frequency estimations, these results thus have an inherent error in the case of missing dosage information. Nevertheless, they can serve as a good coherency check with other results.

[2] Polysat actually has different functions for calculating the Bruvo distance, each suitable for different applications. Under consideration here is the Bruvo2.distance function, using the genome addition model (Bruvo *et al.*, 2004; Clark and Jasieniuk, 2011).

CHAPTER 5

CHARACTERISING SWEETPOTATO PARENTAL MATERIAL: RESULTS AND DISCUSSION

5.1 Phenotypic analysis

5.1.1 Westcott adjustment

The optimal parameter set for the Westcott adjustment of the heterosis field trial data, as determined by simulating the adjustment for the WCSR measurements, was to use separate means for each check, weigh the adjustment according to the distance of the checks, use a subregion of 5 rows and use a global weighing factor of 0.5.

Figure 5.1 shows the WCSR measurements across the field before and after the Westcott adjustment. Before the adjustment we see clearly lower yield values in the lower left corner and higher values in the upper left corner of Figure 5.1, because of field heterogeneity. After adjustment this pattern is removed, which indicates the data adjustment worked properly.

5.1.2 Data exploration

Figure 5.2 compares the parent and offspring population on several measurements, Table 5.1 summarises important results on the parents. We can observe the following:

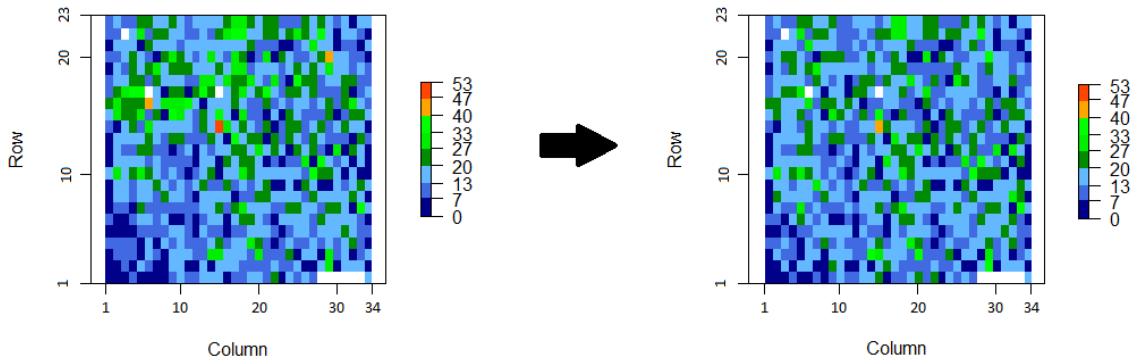


Figure 5.1: Raster plot of the heterosis trial field for the weight of commercial storage roots (WCSR), before (left) and after the Westcott adjustment (right). Scale: t/ha.

- *Virus symptoms 1 and 2*. The scoring ranges from 1 to 6 and 2 to 7 for virus symptoms 1 and 2 respectively. This means that viral infection was definitely present in the field, but never the most severe forms of infection were observed.

  The most striking observation is that virus symptoms were significantly higher for the parent population compared to the offspring population (chi-square test with simulated p-values < 0.0001 for both virus symptoms). This is especially striking since SPVD is considered to be mostly governed by 2 recessive genes (Mwanga *et al.*, 2002a). We suspect that this infection difference comes from the infection carried in the vine cuttings of the parents, which was thus already present before the planting of the trial. Indeed, the parents in this trial were planted after propagation through vine cuttings whilst the offspring were grown from seed prior to the trial. Material grown from seed can be considered 'more clean' than material propagated through clonal propagation, unless the material was first properly cleaned. This degeneration due to viruses, as well as the effects of virus-cleaning the planting material is reviewed by Gibson and Kreuze (2015).

- *WCSR*. WCSR for both the parents and their offspring ranges from 0 to over 45 t/ha. WCSR yield for the offspring population is significantly higher than for the parents (Wilcoxon rank sum test p-value $< 1{\cdot}10^{-15}$), average values are 10 and 16.5 t/ha for the parent and offspring population respectively. These values are extremely high if we consider the average yields in Africa (Section 2.1), but it has to be considered these results come from experimental fields with only $1.5m^2$ per plot, not farmlands. In fact, the results presented here coincide well with the results of an evaluation of 1174 CIP germplasm clones conducted in 5 different environments in Peru (Grüneberg *et al.*, 2015). These authors report an average storage root yield (WCSR + WNCSR) of 19 t/ha across all environments with individual yields ranging from 0 to 55.5 t/ha. A smaller field trial conducted

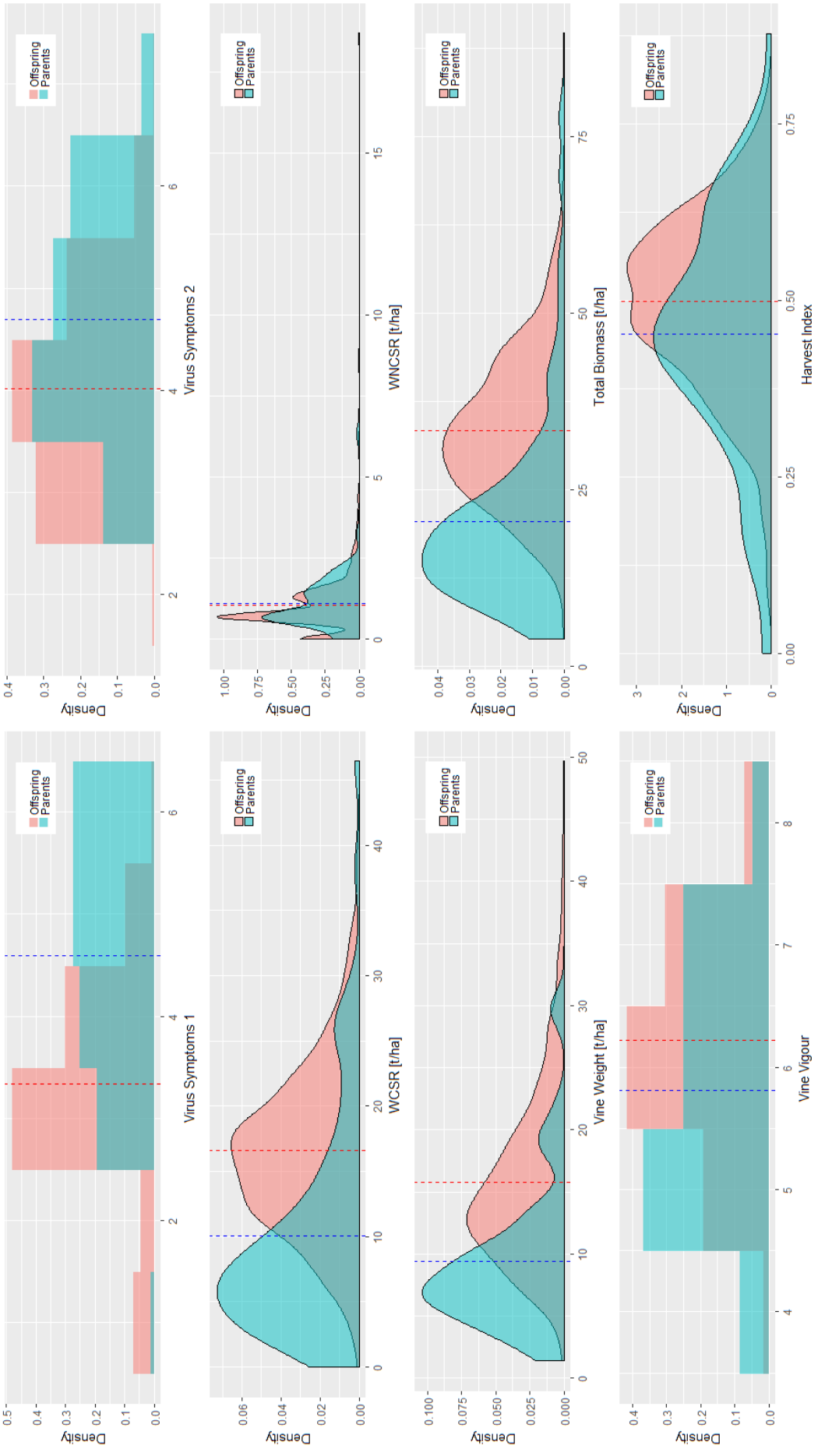


Figure 5.2: Histograms for the categorical estimations and density plots for the continuous measurements of the heterosis field trial. The parent and offspring population are represented separately and with a different colour. Dashed vertical lines indicate the mean for each population (red = offspring; blue = parents).

with 25 genotypes across 6 environments in Ethiopia reports fresh root yields (WCSR + WNCSR) ranging from 0.77 to well over 50 t/ha, with an average of 19.45 t/ha across all environments (Gurmu *et al.*, 2017).

Although generally the offspring seem to perform better than the parents, the highest measurement comes from the parent 'CIP442162'. Some other parents, such as 'Nanungungungu', 'Obare', 'Ligri' and 'PGA13067-7', also perform well above average, but 'CIP442162' outperforms them by more than 10 t/ha (Table 5.1).

- *WNCSR*. The WNCSR averages at about 1 t/ha for both the parents and the offspring, no significant difference was detected (Wilcoxon rank sum test p-value = 0.4). Although most measurements fall within the range of 0 to 2.5 t/ha we observe a few remarkable outliers. These are the result of damaged storage roots (e.g. weevil damage) that are big enough to be considered commercial storage roots, but could never be sold due to this damage (these apparent outliers are thus not removed from the data).

- *Vine weight*. The weight of the vines ranges from about 1.5 to almost 50 t/ha. Average values are 8 and 14.5 t/ha for the parent and offspring population respectively (significant difference, Wilcoxon rank sum test p-value $< 1 \cdot 10^{-15}$). These results are somewhat on the low side compared to the large study of Grüneberg *et al.* (2015) who report average foliage yields of 22.6 t/ha and individual measurements ranging from 0 to 67.8 t/ha. This difference is not worrying since the trial presented here uses only small plots ($1.5m^2$) and few replicates, yield results are thus not necessarily very accurate.

- *Total biomass*. Total biomass, as calculated by the sum of WCSR, WNCSR and vine weight, ranges from about 4 t/ha to almost 90 t/ha. Unsurprisingly we again observe a significant (Wilcoxon rank sum test p-value $< 1 \cdot 10^{-15}$) difference between parents and offspring, with average values of 20.4 and 33.4 t/ha for the parent and offspring population respectively. This extreme range of biomass yields was also observed by Grüneberg *et al.* (2015) who report biomass yields ranging from 2 to almost 100 t/ha. 'CIP442162' again pops out as a remarkably well performing parent.

  The heterosis trial was set up with the intention to assess the heterosis potential of the Ghana parents (especially for WCSR). From the previous observations it seems obvious this heterosis potential is very big. Indeed, the offspring generally seems to outperform the parents when it comes to WCSR, vine weight and total biomass. If only additive gene effects were present, we would not expect this great increase in the average performance for the offspring (we would expect the offspring to perform around the same average as the parents. See Section

3.2), and thus we might conclude that heterosis is very prevalent. However, we have also noticed a remarkable and unexpected difference in virus symptoms between the parents and the offspring. And on top of that we can show there is a significant influence of both virus symptoms 1 and 2 on the total biomass yield (Kruskal-Wallis test p-values are both $< 1 \cdot 10^{-15}$). We can thus state that the observed difference in the yield performance between parents and offspring is biased by a difference in the severity of viral infection in the 2 populations, and the difference in viral infection is probably not due to genotype but due to the method of propagation (from seed vs. through vine cuttings). Since there is no way of correcting the yield data for this bias (proper reference plots would be needed for that) we can do nothing more than take it into consideration when trying to assess the heterotic increments (Section 5.1.3).

- *Vine vigour*. The vine vigour score ranges from 4 to 8. We notice again a significant (chi-square test with simulated p-value < 0.0001) difference between parents and offspring. However, since stunted growth is an important symptom of viral infection (Mwanga *et al.*, 2002b), this observation might again be biased by the difference in viral infection between the parent and offspring populations.
- *Harvest Index*. The HI ranges from 0 to 0.88. A small but significant difference was detected (Wilcoxon rank sum test p-value = 0.006) between parents and offspring, average values are 0.45 and 0.50 for the parent and offspring population respectively. These results again coincide well with the results of the extensive sweetpotato study presented by Grüneberg *et al.* (2015), who report a HI average of 0.48 and a range of 0 to 1.

5.1.3 Heterosis and heterotic groups

Mid-parent heterosis increments (MH-values) were calculated for all 149 cross combinations for both WCSR and total biomass. These results are presented in Figures A.1 and A.2 respectively (note that more than a third of the values in these figures equal 0, this means the heterosis increment was not significantly different from 0).

MH-values for WCSR range from -48% to 400% with an average of 80%. MH-values for total biomass range from -40% to 252% with an average of 65%. Most of the MH-values, for both WCSR and total biomass, are highly positive, an observation that was also made by Gurmu *et al.* (2018), who studied a 7x7 half-diallel of sweetpotato parents in Ethiopia. This indicates once more the potential of exploiting heterosis in sweetpotato (see Section 2.3.3). In fact, the only parent exhibiting negative MH-values for WCSR and total biomass is 'CIP442162'.

Table 5.1: The frequency (f) with which every parent was used as a male or female and performances of the parents and offspring. Columns (4) and (5): the average weight of commercial storage roots (WCSR) [t/ha] and total biomass [t/ha] for every parent; Columns (6) and (7): the average WCSR [t/ha] and total biomass [t/ha] over all offspring of the parents; Columns (8) and (9): the average mid-parent heterosis values (MH) for WCSR [%] and total biomass [%] over all offspring of the parents. For every column, the top ranked parents are indicated in green and the bottom ranked parents in red. The table is sorted on the average WCSR for the offspring.

Name	f Fem	f Male	Parents		Offspring		MH	
(1)	(2)	(3)	WCSR (4)	Biomass (5)	WCSR (6)	Biomass (7)	WCSR (8)	Biomass (9)
PGA13067-7	13	7	13.55	21.15	19.38	34.55	60.49	68.88
Ligri	0	6	17.65	26.59	19.03	35.75	34.86	42.68
MothersDelight	6	4	4.00	12.15	18.99	33.80	159.70	111.04
BF92xTib.2	0	2	1.32	9.85	17.95	37.84	335.62	195.10
BF59xCip.4	2	12	7.03	15.44	17.77	35.40	117.66	93.84
Apomuden	9	12	11.63	18.06	17.67	31.23	62.92	50.24
Obare	2	4	21.66	31.72	17.42	38.78	8.49	44.79
Tu-orange	2	4	2.28	9.48	17.14	33.41	231.66	141.73
Bohye	8	10	7.21	21.56	17.03	35.19	90.82	53.83
Blueblue	0	1	3.68	9.78	16.72	35.65	0.00	130.58
Otoo	15	7	7.65	16.40	16.53	34.36	108.89	98.06
Tu-purple	13	13	6.31	15.27	16.51	33.72	100.59	77.64
CIP442162	8	8	33.41	59.82	16.46	33.45	-5.89	-2.47
CIP440390	14	12	7.30	14.67	16.26	31.80	88.02	68.93
Hi-starch	10	5	6.54	14.36	15.84	32.84	78.11	71.61
Jitihada	6	10	12.81	29.03	15.81	35.33	21.04	12.99
Nanungungungu	6	2	21.73	31.57	15.78	30.04	2.54	12.55
Faara	13	12	6.94	17.14	15.68	32.91	81.93	79.24
Patron	13	3	9.44	19.83	15.22	32.37	56.94	47.79
Sauti	5	9	2.81	17.62	14.43	30.78	127.05	66.79
Santompona	4	0	3.84	7.46	13.01	32.76	146.49	160.01
NKO31A	0	6	10.35	28.05	12.96	29.74	16.11	0.00

It should be noted that the heterosis values presented here are extremely high when compared to other heterosis studies on sweetpotato storage root yield. Grüneberg *et al.* (2009b) reported MH-values of 84% among 48 cross combinations for their sweetpotato experiments (although little further information was provided by the authors), and Grüneberg *et al.* (2015) present a small trial with 16 parents (12 female, 4 male) with heterosis values ranging up to 59%. To my knowledge the most extreme values reported for mid-parent heterosis for storage root yield reach up to 120%, 126% and 170% (Gurmu *et al.*, 2018). It can be assumed that the values presented in the present study are biased upwards by the difference in viral infection between the parents and their offspring. Emphasis should thus not be put on the exact numerical values, but nevertheless it is clear that the potential for the exploitation of heterosis at the Ghana sweetpotato support platform is great.

Table 5.1 presents us with some interesting summary statistics on the parents. The first thing that can be observed is that the average WCSR yield for the parents coincides well with the total biomass, this is of course a consequence of the high HI we observed previously. Secondly, high yield for the parents does not seem to correlate necessarily with high average yield for their offspring (also noticed by Grüneberg *et al.* (2009b)): highly ranked parents, on average, do not necessarily produce better offspring than poorly preforming parents. It should be noted here that there is an inherent difficulty in comparing these results between the parents, since not all par-

ents were used in the same cross combinations with all other parents. Nevertheless, we can observe one final thing in the light of heterosis: almost all parents have high positive average values for mid-parent heterosis (notable exception is 'CIP442162') and this is even more true for the poorest performing parents. This is a clear indication that non-additive effects (heterosis) play a crucial role in this population and that parents can and should not be judged on their own performance but more so on the performance of their offspring. Of course, this observation will be biased by the low amount of crosses with some parents and perhaps some lucky or unlucky crosses with others, but without fixating on the exact numerical values, this statement remains valid.

Concerning the formation of heterotic groups based on these data I can be very brief: the data are too sparse to use any of the conventional methods. A half diallel crossing scheme would require 231 cross combinations ($(22 \cdot 21)/2$), but only 149 cross combinations were made. The methods of Griffing (1956) rely at least on a half diallel crossing scheme to be able to estimate all GCA and SCA parameters and the data were even too sparse to use mixed linear models (Raul Eyzaguirre, CIP statistician - personal communications).

The lack of crossing data is of course also the reason why such *ad hoc* methods were used to estimate the reciprocal effect (Section 4.1.4) or to analyse the performance of the parents (Table 5.1; this would normally be done by fitting a linear model. See Section 3.1). This is not the first time sweetpotato research is troubled by a lack of crosses, a recent example is the study of Gurmu *et al.* (2018), who admit they had to cut down their initially planned 10x10 crossing scheme to a 7x7 scheme due to a lack of sufficient flowering of some parents. Other possible problems might be cross incompatibility, cross success rates and germination rates; subjects recently studied by Rukundo *et al.* (2017).

5.1.4 Reciprocal effects

For 15 out of the 22 parents, reciprocal crosses were available. For only 9 parents a sufficient number (5 or more) of reciprocal crosses was made to properly assess a reciprocal effect. These data are summarised in Figure 5.3. For the parents 'Sauti', 'Faara' and 'Apomuden' a significant reciprocal effect for WCSR was found, according to the methodology described in Section 4.1.4. The offspring of 'Apomuden' perform better when the parent is used as a male, the offspring of 'Sauti' and 'Faara' perform better when these parents are used as female. It is not useful to fully quantify these results since the data on this subject are very preliminary, with only a limited number

of crosses, and not necessarily with the same partners, for all the parents under investigation.

It is difficult to maintain a reciprocal effect in sexually reproducing plants, but in a vegetatively propagated crop the effect can be fixed due to clonal propagation. The breeding of sweetpotato can thus benefit from the identification of these reciprocal effects (Falconer and MacKay, 1996; Rukundo *et al.*, 2017). An example of the study of reciprocal effects in sweetpotato is the paper of Rukundo *et al.* (2017) who studied, amongst others, the effect on vine yield and total biomass.

The results presented here seem to indicate that a reciprocal effect is present for at least some of the parents at the Ghana crossing block and it might be of interest to the breeders to make this a subject for further investigation.

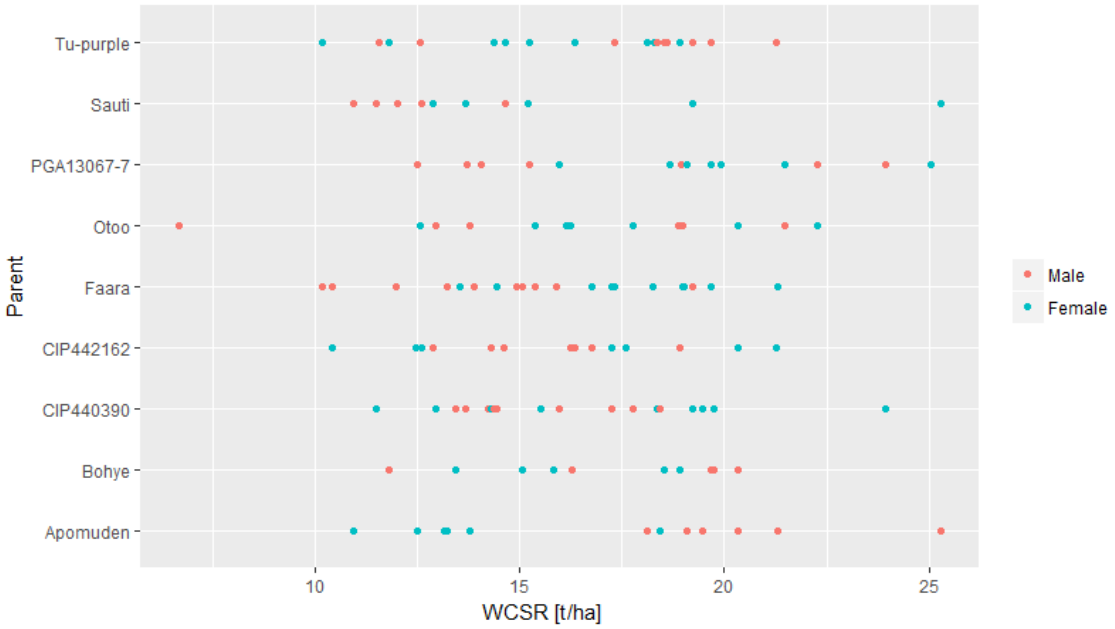


Figure 5.3: Comparison between the average weight of commercial storage roots (WCSR) of the offspring of different parents when these parents were either used as male or female.

5.2 Genotypic analysis

5.2.1 Marker selection

Before commencing the SSR analysis, the 36 available markers were evaluated on their quality, in order to minimize errors and difficulties during the analysis (Table 5.2):

- Twelve markers were immediately discarded because of the 'seventh allele' problem. Aneuploidy of somatic cells has been reported in sweetpotato when investigated with chromosome staining techniques (Oracion, 1995) and more recently through whole genome sequencing (Wu *et al.*, 2018). The high occurrence of this 'seventh allele', however, is striking and might also indicate experimental errors. Unfortunately the data of this study are not sufficient to further investigate the issue, and from a precautionary point of view, these markers were omitted.
- Another 5 markers were discarded because the allelobin quality measure had a value higher than 0.4 and a subsequent visual check of the cumulative allele distribution plot confirmed the binning problem (examples in Section A.5). On rare occasions, an allele was manually moved to a different bin to improve the binning.

Table 5.2: Summary of the marker selection procedure. This table indicates whether a seventh allele was present. If not, the table also includes the allelobin binning quality index, whether visually the binning was good or poor and if a manual binning correction was necessary. The final verdict is either to discard the marker or to use it for further analysis.

Marker	7th Allele	Quality Index	Visual Binning Control	Final Verdict	Marker	7th Allele	Quality Index	Visual Binning Control	Final Verdict
IB297	No	0.46	Poor	Discard	**IBS199**	Yes	-	-	Discard
IBC3	Yes	-	-	Discard	**IBY40**	No	0.48	Poor	Discard
IBJ559	Yes	-	-	Discard	**IBY41**	No	0.32	Good	Use
IBJ664E	Yes	-	-	Discard	**IBY43**	No	0.24	Good	Use
IBN24	Yes	-	-	Discard	**IBY44**	No	0.29	Good	Use
IBN35	Yes	-	-	Discard	**IBY45**	No	0.31	Good	Use
IBN36	No	0.32	Manual	Use	**IBY46**	No	0.37	Manual	Use
IBN37	Yes	-	-	Discard	**IBY47**	No	0.35	Good	Use
IBS137	No	0.43	Manual	Use	**IBY48**	No	0.38	Good	Use
IBS141	No	0.34	Good	Use	**IBY50**	No	0.32	Good	Use
IBS144	Yes	0.30	-	Discard	**IBY51**	Yes	-	-	Discard
IBS146	No	0.25	Good	Use	**IBY52**	No	0.36	Manual	Use
IBS147	Yes	-	-	Discard	**IBY53**	No	0.24	Good	Use
IBS149	No	0.10	Good	Use	**IBY54**	No	0.38	Good	Use
IBS150	No	0.25	Good	Use	**IBY56**	No	0.47	Poor	Discard
IBS169	No	0.47	Poor	Discard	**IBY58**	Yes	-	-	Discard
IBS174	No	0.44	Manual	Use	**IBY59**	No	0.40	Manual	Use
IBS186	Yes	-	-	Discard	**IBY60**	No	0.41	Poor	Discard

The remaining 19 markers are all perfect repeat markers with di-, tri- or tetranucleotide repeats. This will allow an interpretation through the stepwise mutation model and the use of the Bruvo distance to analyse these markers.

Allele diversity of these 19 markers ranged from 3 to 15 with an average of 7.6 alleles per locus, and the PIC ranged from 0.41 to 0.83 with an average value of 0.67 (Table 5.3). These results are in accordance with other SSR studies that assessed sweetpotato diversity or characterised sweetpotato germplasm in different parts of the world: Koussao *et al.* (2014) report numbers of alleles per locus between 1 and 12 and an average PIC of 0.73 based on 30 markers. Ngailo *et al.* (2016) observed between 4 and 17 alleles per locus and an average PIC of 0.78 based on 9 markers. David *et al.* (2018) report an average number of alleles of 7.13 and an average PIC of 0.75 based on 31 markers. Yada *et al.* (2010) used 10 SSR markers with an average PIC of only 0.62 and a number of alleles ranging from 2 to 6.

Table 5.3: Allele diversity and PIC for the 19 selected SSR markers.

Marker	**Allele diversity**	**PIC**
IBN36	3	0.55
IBS137	7	0.75
IBS141	7	0.76
IBS146	8	0.68
IBS149	14	0.83
IBS150	12	0.64
IBS174	15	0.79
IBY41	6	0.58
IBY43	6	0.72
IBY44	9	0.83
IBY45	6	0.60
IBY46	7	0.78
IBY47	5	0.54
IBY48	4	0.41
IBY50	5	0.62
IBY52	7	0.76
IBY53	5	0.58
IBY54	9	0.76
IBY59	10	0.65
Average	7.63	0.67

It is thus safe to say that the 19 SSR markers selected in this study will suffice, in both quantity and quality, to assess the diversity of the sweetpotato parents under investigation.

5.2.2 SSR analysis

Three different genetic distance metrics were used to calculate 3 pairwise distance matrices between all 51 accessions. The Lynch distance is based only on band similarity and does not take into account any mutational model, its calculation for 19 loci across 51 individuals took mere seconds. The Bruvo distance takes into account a stepwise mutational model and allows for a more biologically relevant interpretation, its calculation took 70 seconds. The FullBruvo distance is similar to the Bruvo distance, but without simplifications to cut on computational time, its calculation took almost 6 hours, using 3 out of 4 cores on a 5 year old Acer Aspire E1-771G laptop.

The pairwise distance matrices are represented as bootstrapped, UPGMA dendrograms (Figures 5.4, 5.6 and 5.8) and histograms of the distance metrics (Figures 5.5, 5.7 and 5.9). We can observe the following:

- There is a larger range of different distances and the mean distance is higher for the Lynch distance compared to the Bruvo distance (Figures 5.5, 5.7 and 5.9). It is of course not appropriate to compare different distance measures this way (because they are calculated in a completely different manner), but it makes sense that the method that takes mutational steps into account gives rise to generally smaller distances between individuals. The mean distance rises again for the FullBruvo distance, this is because the pairwise distances rise automatically when dosage information is missing (Bruvo *et al.*, 2004).

- The check pairs ('Cemsa'+'Ligri'; 'CIP199062.1'+'Bohye'; 'Tanzania'+'Sauti') are grouped close together with reasonable confidence by the Lynch distance. 'Tanzania' and 'Sauti' are still clustered together by the Bruvo and FullBruvo distances, but 'Ligri' and 'Cemsa', and especially 'Bohye' and 'CIP199062.1' are more scattered by the Bruvo and FullBruvo distances. The check pairs are indicated in different colours in Figures 5.4, 5.6 and 5.8. Genetic variation within the same cultivar due to clonal propagation in sweetpotato has been reported decades ago (He *et al.* (1995) and references therein). It was thus not unlikely that the check pairs would have differentiated from each other because of many years of separate propagation.

  Both variants of the Bruvo distance seem to have the most difficulty to group checks together. It was already hinted by Meirmans *et al.* (2018) with a simple simulation study that, despite being developed for the analysis of polyploid SSR data (with missing dosage information), the Bruvo distance has a certain bias, even when comparing genotypes with the same ploidy level.

- Bootstrapping values, especially at the higher hierarchical levels, are extremely low and grouping at higher levels is inconsistent across the different distance metrics. This means that the groupings presented in Figures 5.4, 5.6 and 5.8 are nothing more than artefacts of the clustering algorithm and the chosen SSR loci. There is no statistical confidence to divide the accessions of the Ghana crossing block into clearly distinct groups based on these data.

  Many studies have grouped sweetpotato varieties based on SSR markers. In some cases a clear link between grouping and geographical region of origin was detected (e.g. Roullier *et al.* (2011) and David *et al.* (2018)); this is not the case for the accessions in this study, most probably because the parents in the Ghana crossing block were introduced from many different regions of the world and not

a single region is represented by enough accessions to observe a clear grouping (Table A.1).

It is rare to see any bootstrapping procedure for the correct interpretation of dendrograms in sweetpotato SSR marker studies (e.g. Yada *et al.* (2010); Koussao *et al.* (2014); Ngailo *et al.* (2016); Meng *et al.* (2018) and David *et al.* (2018) all produce dendrograms without bootstrapping). In many cases this is not considered necessary because the grouping from a dendrogram can be checked by other analyses such as AMOVA, DAPC or STRUCTURE. The bootstrapping procedure used in this study is very simple and easy to interpret, but nevertheless provides us with valuable insights on the (in)stability of the presented dendrograms.

- At the lower hierarchical levels (groups of 2-3 accessions) there are 11 groups that appear consistently across the different distance metrics and with reasonable bootstrap values (at least >0.2, preferably >0.5). Bootstrapping values of 0.2-0.5 are still very low, but the fact that these groups appear for all 3 distance metrics boosts confidence in these groupings. These small groups are indicated in different shades of green in Figures 5.4, 5.6 and 5.8.

  'Patron' and 'Otoo' are both accessions originating from Burundi. Due to a great phenotypic resemblance, it was argued before that these accessions might be clones (Jolien Swanckaert, CIP - personal communications). The results presented here indicate a great genetic resemblance as well. It seems unnecessary to keep both accessions at the Ghana crossing block.

  The other small groups are not so easily explained. It might be worthwhile to check, based on phenotypic information, if it is useful for the breeding program to keep every accession of such a group. The most striking example of such a group is 'Obare' and 'BF92XTib.2'. These accessions are genetically closely linked, but phenotypically they are not alike.

The most important result from the previous paragraphs is that no appropriate grouping of the accessions at the Ghana crossing block seems possible. Presumably because the accessions are very diverse in nature. Indeed, sweetpotato is known for its great phenotypic and genotypic diversity (Section 2.2) and on top of that, accessions that were introduced in Ghana came from all over the world.

To confirm this result, the first steps of a DAPC were performed. Figure 5.10 shows the BIC as a function of the amount of inferred clusters (all principal components were retained for this step). The ideal number of groups to divide the population into, is reached when the BIC reaches its minimum. Contrary to what one hopes to see in a DAPC (a clear minimum at a certain number of groups), Figure 5.10 has no distinct

minimum. However, 2 minima are observed at the edges of the plot: at 1 and 50 clusters. Put differently: the ideal way of grouping the Ghana accessions according to the DAPC is by either making a single group (= no grouping) or by putting every accession in its own separate group (= no grouping). We do not need to continue the DAPC as this result already confirms the result of the distance metrics.

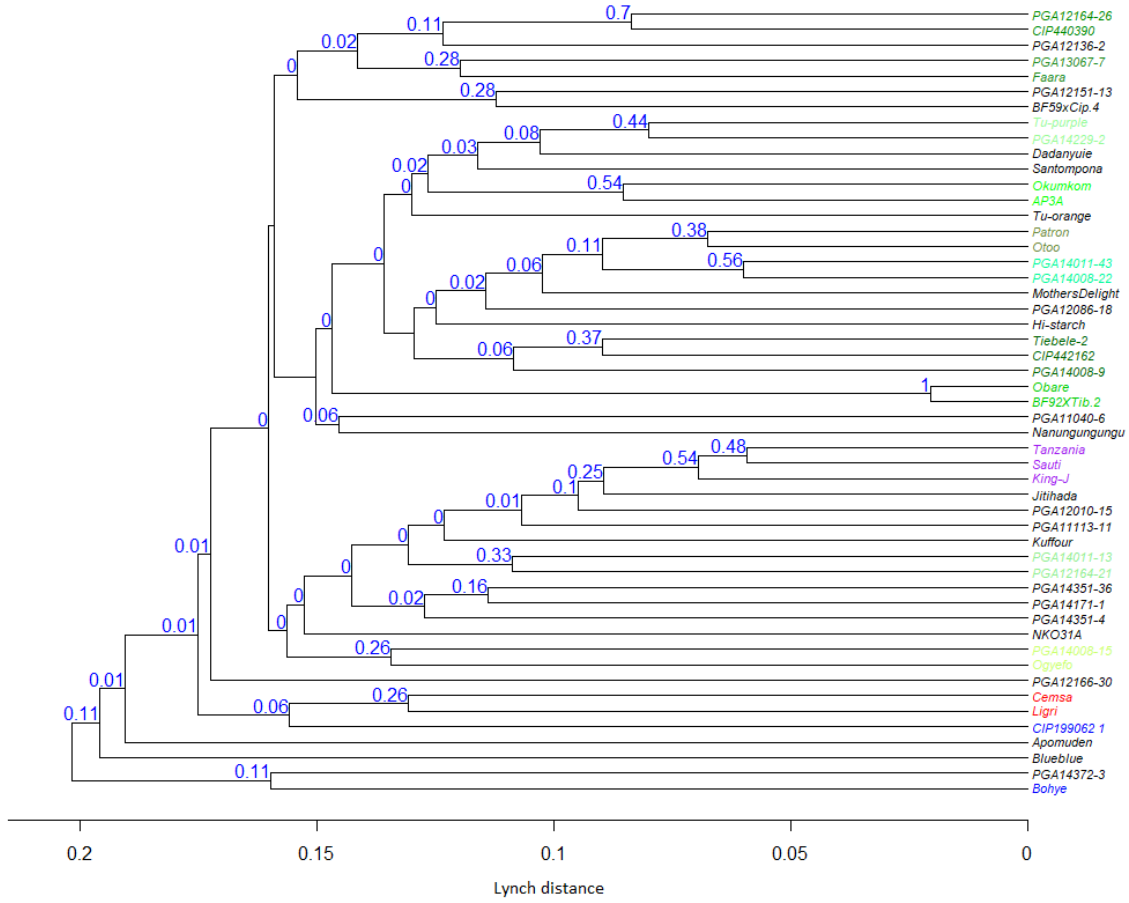


Figure 5.4: The dendrogram constructed from the **Lynch** pairwise distance matrix. Bootstrap values are indicated in blue for every cluster. Check pairs are indicated in red, blue and purple respectively. Different small groups of 2-3 accessions, that are shared amongst the 3 dendrograms (Figures 5.4, 5.6 and 5.8) with reasonable bootstrap values are indicated in 10 different shades of green (except for the group 'Sauti'+'Tanzania'+'King-J').

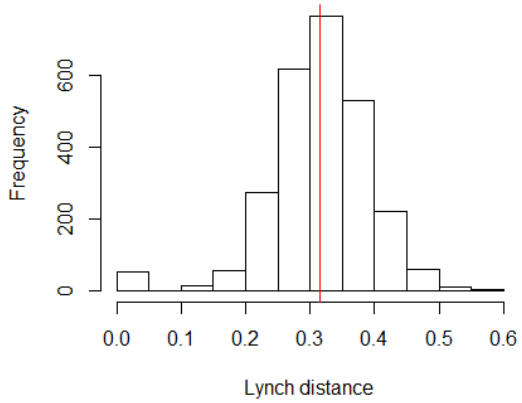


Figure 5.5: Histogram of the **Lynch** pairwise distances. The mean is indicated by the red vertical line.

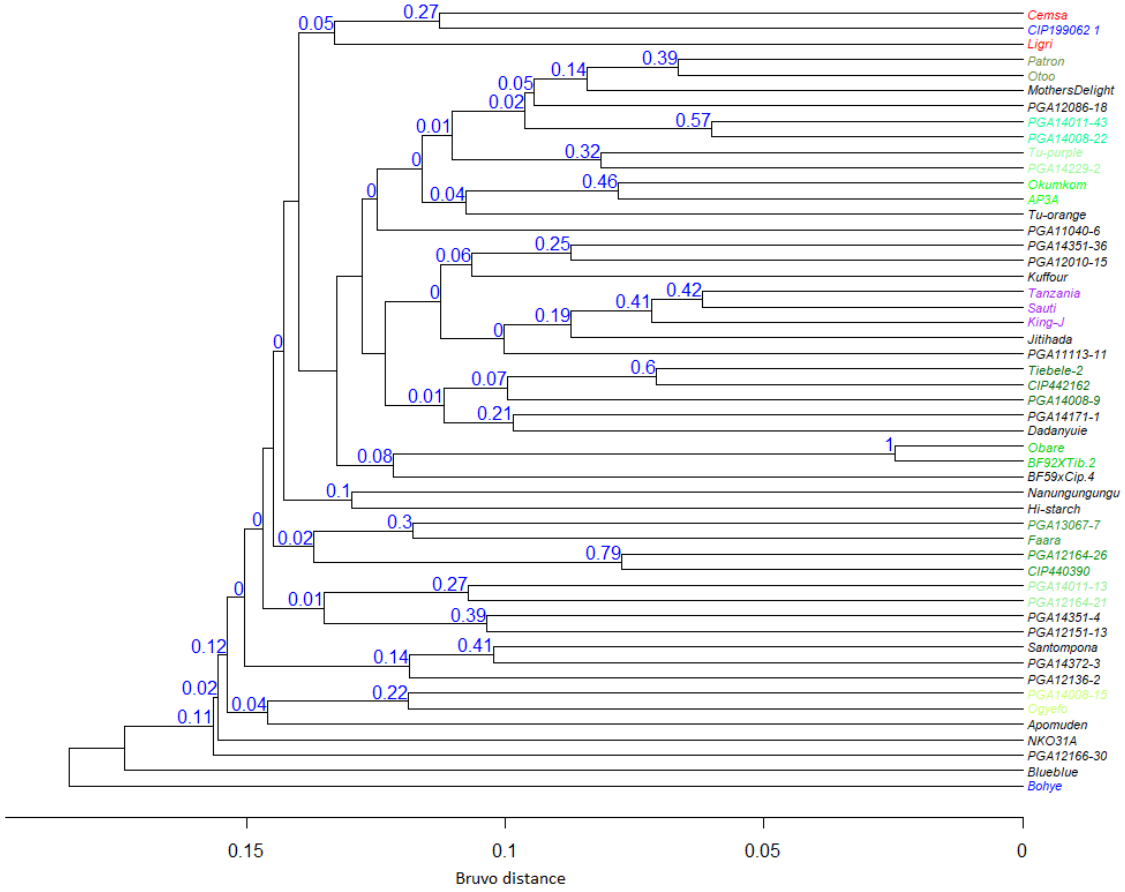


Figure 5.6: The dendrogram constructed from the **Bruvo** pairwise distance matrix. Bootstrap values are indicated in blue for every cluster. Check pairs are indicated in red, blue and purple respectively. Different small groups of 2-3 accessions, that are shared amongst the 3 dendrograms (Figures 5.4, 5.6 and 5.8) with reasonable bootstrap values are indicated in 10 different shades of green (except for the group 'Sauti'+'Tanzania'+'King-J').

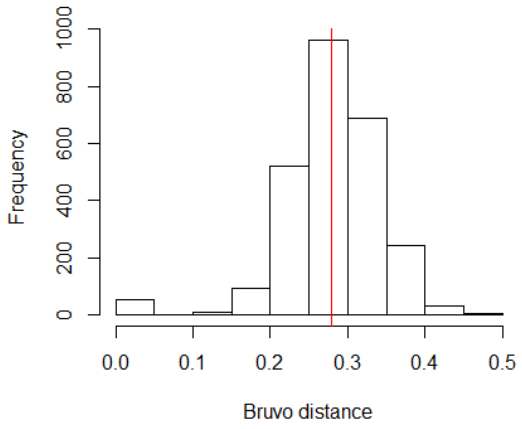


Figure 5.7: Histogram of the **Bruvo** pairwise distances. The mean is indicated by the red vertical line.

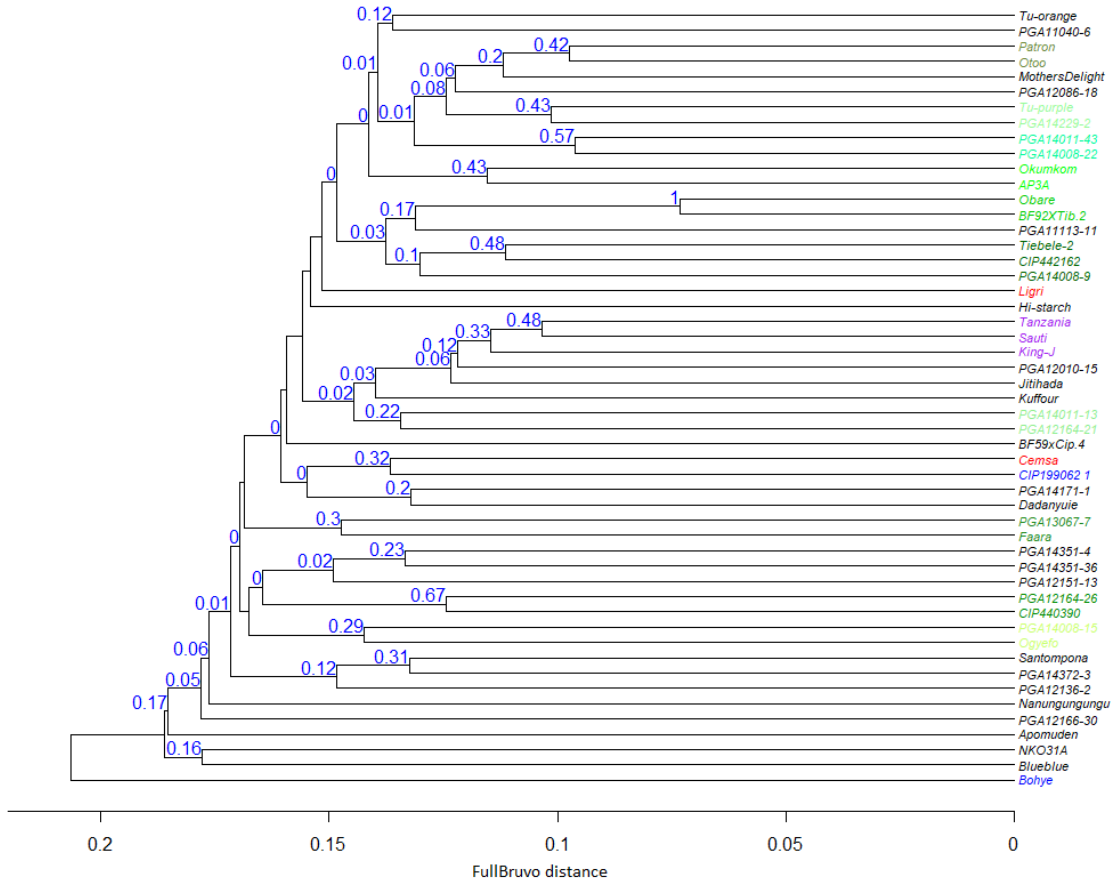


Figure 5.8: The dendrogram constructed from the **FullBruvo** pairwise distance matrix. Bootstrap values are indicated in blue for every cluster. Check pairs are indicated in red, blue and purple respectively. Different small groups of 2-3 accessions, that are shared amongst the 3 dendrograms (Figures 5.4, 5.6 and 5.8) with reasonable bootstrap values are indicated in 10 different shades of green (except for the group 'Sauti'+'Tanzania'+'King-J').

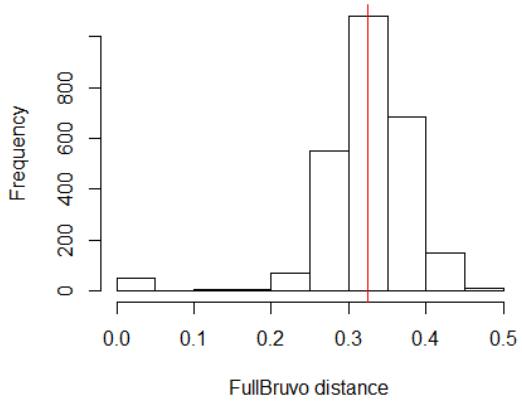


Figure 5.9: Histogram of the **FullBruvo** pairwise distances. The mean is indicated by the red vertical line.

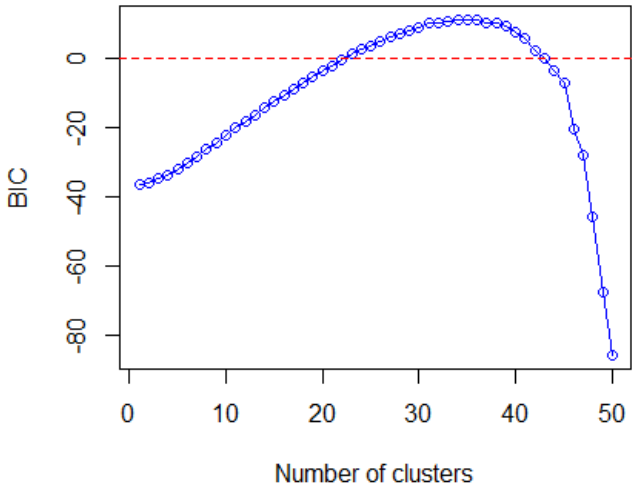


Figure 5.10: Scatter plot of the BIC versus the inferred number of cluster from the DAPC.

CHAPTER 6

CHARACTERISING SWEETPOTATO PARENTAL MATERIAL: CONCLUSIONS

Selection of the best parents for further breeding is a crucial step for population and variety development in sweetpotato (Section 2.3.2). Based on phenotypic and genotypic data on the parents in the crossing block of the Ghana SSP-WA, this dissertation tried to answer 3 questions concerning the characterisation of this parental material (Section 2.4):

1. *What is the phenotypic parent-offspring relation in the starting population?*

   Twenty-two parents were evaluated by studying 149 of their cross combinations in a field trial in Kumasi, Ghana. At first glance the offspring seemed to perform better than the parents on every level, having higher yields for both above and below ground biomass, higher harvest index and being less prone to viral infection. All of this cumulated in an apparent very high heterosis potential between most of the parents and clear indications of non-additive gene effects that play an important role in the population. However, an experimental flaw was uncovered: the parents were propagated through vine cuttings, whereas their offspring was grown from seed prior to the field trial. This is probably the reason for the lower viral infection on the offspring, which must account (at least partially) for the higher yields of the offspring as well. The offspring performance is thus overestimated compared to the parents. To avoid this bias, it is suggested to use virus cleaned material for the parents and the offspring in future experiments. It is also useful not to use seedling material of the offspring for this kind of quantitative measurements, but to propagate them through vine cuttings prior to an experiment. Indeed, there have been earlier indications that the performance of a variety changes between the initial plant from seed and the later cuttings (Jolien Swanckaert, CIP - personal communications). Since it are the cuttings

that are to be used in the farmers' fields, it makes sense to focus quantitative evaluation on vine cuttings rather than plants from seed.

On top of that, the relatively low number of cross combinations makes it difficult to correctly interpret the phenotypic results. Not all parents were used in the same number of cross combinations or with the same partners. Thus, lucky or unlucky cross pairs may provide another bias in this study. The obvious solution is to only evaluate (at least) a half-diallel crossing scheme, which will also allow for parameter assessment through conventional methods (e.g. Griffing's method). Of course, this is more easily said than done: not all cross combinations may be possible (e.g. because of cross incompatibility) and the working circumstances in Ghana (e.g. periods of severe drought) do not always allow all seedlings to survive.

Despite these difficulties, it is impossible to deny that non-additive gene effects are very prevalent in this population and a HEBS seems an excellent choice to improve sweetpotato cultivars in West Africa. Based on these phenotypic data, however, it is not possible to determine clear heterotic groups.

One parent, 'CIP442162', deviates from the others as being an extremely well performing parent, but having poorly performing offspring. Based on the results of this study, it is recommended to discard this parent from the crossing block since it does not seem to be able to contribute to better variety or population development.

Reciprocal effects seem to be present in the population. Certain parents produce better offspring when used only as male or female in cross combinations. Using these parents in the 'wrong' directions for variety development seems a waste of resources. Although the results of this study are very preliminary, this may thus be an interesting and useful area for further research.

2. *How genetically diverse are the parents in the starting population?*

   Forty-eight parents were genotyped using 36 SSR markers. After a critical assessment of marker quality, 19 markers were retained for further evaluation and analysis, using 3 different genetic distance metrics. Despite being developed for polyploid SSR analysis, the Bruvo and FullBruvo metrics did not seem to perform as well to characterise the parents as did the simple Lynch distance based on band similarity. To my knowledge, this is the first time the Bruvo distance was used to investigate sweetpotato SSR data, but unfortunately it does not seem to be very useful for this kind of investigation. Meirmans *et al.* (2018) already hinted in this direction.

   None of the distance metrics were able to distinctly and consistently group the parents at higher hierarchical levels and further investigation using a DAPC also

indicated a lack of clear grouping based on these genetic data. All of this points into the direction of the presence of great genetic diversity at the Ghana crossing block. This should not come as a surprise since accessions from all over the world, both landraces and products of different breeding programs, were brought together in Ghana by the SSP-WA.

At lower hierarchical levels, some surprising groups were formed. Although some of these are without doubt coincidental, some are very striking and hard to explain. For example, the genetic resemblance between 'Obare' and 'BF92xTib.2' is very large, but phenotypically these accessions are very distinct. Reinvestigating some of these cases would be useful before making the decision to discard any of these parents. 'Patron' and 'Otoo' were identified to be genetically very similar and are also phenotypically alike, it does not seem worthwhile to keep both of these parents at the Ghana crossing block.

There are plenty of difficulties when analysing SSR marker data for a hexaploid plant such as sweetpotato (Section 3.3.2). Especially the problems with missing dosage information introduce a lot of uncertainties into the analysis. Although SSR analysis has been used in sweetpotato a lot and has proven to be useful, it should be considered by the breeding programs to make a switch to next generation sequencing techniques for a more in-depth analysis of sweetpotato genotypes. Of course, full sequencing efforts for sweetpotato are difficult (Section 2.2.4) and come with their own set of problems, but steps are made into the right direction with the development of new marker-based genotyping tools and new sequencing efforts (for example by the genomic tools for sweetpotato (GT4SP) improvement project (Yencho, 2015)).

3. *How can we subdivide the parents into (preferably 2) groups based on both genotypic and phenotypic data on these parents?*

   The division of a population into mutually heterotic groups has been discussed in Section 3.2.3 for both phenotypic and genotypic data. From the phenotypic data we would need SCA values in order to divide the population, unfortunately the data presented here were too sparse to calculate these parameters and so a proper division based on phenotypic data is not possible. A genotypic division into heterotic groups is based on genetic diversity and, as was mentioned in point 2, no appropriate grouping was possible, based on these genotypic SSR marker data.

   Does this mean a division into heterotic groups for the parents at the Ghana crossing block is not feasible? Certainly not. The phenotypic and genotypic data agree perfectly on one point: there seems to be a large genetic diversity between the parents and a clear potential for heterosis. Indeed, as was reviewed

in Section 3.2, the phenotypic observation of heterosis is clearly linked to genetic diversity. The practical conclusion for the breeders at the SSP-WA should be this: any division of the parents is a good division. Any division into 2 groups, A and B, will result in positive heterosis increments for inter group crosses, which is excellent for variety development. Reciprocal recurrent selection will allow the groups to be bred towards each other, and intra group crosses can be made for population development (with all the benefits of a HEBS, Section 2.3.3). The small groups of 2 or 3 genetically similar individuals are best kept together, or a few of these parents discarded. A division into more than 2, smaller groups, each with an emphasis on a specific trait (e.g. orange flesh colour, high dry matter, or SPVD resistance), should also be possible given the fact that sweetpotato is genetically very diverse and even small breeding populations suffer very little from inbreeding. A final and strict division is not presented in this dissertation, since the results give the liberty to the breeders to make the division whichever way they believe is most appropriate, based on their expertise.

BIBLIOGRAPHY

Acquaah, G. (2012). *Principles of Plant Genetics and Breeding*. Wiley-Blackwell, Oxford, United Kingdom.

Amos, W., Hoffman, J. I., Frodsham, A., Zhang, L., Best, S., and Hill, A. V. S. (2007). Automated binning of microsatellite alleles: problems and solutions. *Molecular Ecology Notes*, 7(1):10–14.

Austin, D. (1988). *The Taxonomy, Evolution and Genetic Diversity of Sweet Potato and Related Wild Species*. International Potato Center (CIP), Lima, Peru.

Badu-Apraku, B., Oyekunle, M., Fakorede, M., Vroh, I., Akinwale, R., and Aderounmu, M. (2013). Combining ability, heterotic patterns and genetic diversity of extra-early yellow inbreds under contrasting environments. *Euphytica*, 192:413–433.

Balzarini, M. (2002). Applications of mixed models in plant breeding. In Kang, M., editor, *Quantitative genetics, genomics and plant breeding*, pages 353–364. CABI Publishing, Wallingford, United Kingdom.

Barker, I., Andrade, M., Labarta, R., Mwanga, R. O. M., Kapinga, R., Fuentes, S., and Low, J. (2009). Challenge theme paper 2: sustainable seed systems. In Barker, C., editor, *Unleashing the potential of sweetpotato in Sub-Saharan Africa: current challenges and way forward*, pages 43–72. International Potato Center (CIP), Lima, Peru.

Becker, H. (1993). *pflanzenzüchtung*. Eugen Ulmer, Stuttgart, Germany.

Birchler, J. A., Auger, D. L., and Riddle, N. C. (2003). In search of the molecular basis of heterosis. *Plant Cell*, 15(10):2236–2239.

Birchler, J. A. and Veitia, R. A. (2007). The gene balance hypothesis: From classical genetics to modern genomics. *Plant Cell*, 19(2):395–402.

Birchler, J. A., Yao, H., Chudalayandi, S., Vaiman, D., and Veitia, R. A. (2010). Heterosis. *Plant Cell*, 22(7):2105–2112.

Botstein, M., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32:314–331.

Bovel-Benjamin, A. (2007). Sweet potato: a review of its past, present, and future role in human nutrition. *Advances in Food and Nutrition Research*, 52:1–59.

Bruvo, R., Michiels, N. K., D'Souza, T. G., and Schulenburg, H. (2004). A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. *Molecular Ecology*, 13(7):2101–2106.

Buteler, M., Jarret, R., and LaBonte, D. (1999). Sequence characterization of microsatellites in diploid and polyploid *Ipomoea*. *Theoretical and Applied Genetics*, 99:123–132.

Carey, E. (2013). *Progress and Plans at the Sweetpotato Support Platform for West Africa*. International Potato Center (CIP).

Carey, E. E., Gibson, R. W., Fuentes, S., Machmud, M., Mwanga, R. O. M., Turyamureeba, G., Zhang, L., Ma, D., Abo El-Abbas, F., EL-Bedewy, R., and Salazar, L. (1999). The causes and control of virus diseases of sweetpotato in developing countries. is sweetpotato virus disease the main problem? In *Impact on a Changing World. Program Report 1997-1998.*, pages 241–248. International Potato Center (CIP), Lima.

Cervantes-Flores, J. C., Sosinski, B., Pecota, K. V., Mwanga, R. O. M., Catignani, G. L., Truong, V. D., Watkins, R. H., Ulmer, M. R., and Yencho, G. C. (2011). Identification of quantitative trait loci for dry-matter, starch, and beta-carotene content in sweetpotato. *Molecular Breeding*, 28(2):201–216.

Cervantes-Flores, J. C., Yencho, G. C., Kriegner, A., Pecota, K. V., Faulk, M. A., and Mwanga, R. O. M. (2008). Development of a genetic linkage map and identification of homologous linkage groups in sweetpotato using multiple-dose AFLP markers. *Molecular Breeding*, 21(4):21.

Chang, K. Y., Lo, H. F., Lai, Y. C., Yao, P. J., Lin, K. H., and Hwang, S. Y. (2009). Identification of quantitative trait loci associated with yield-related traits in sweet potato (*Ipomoea batatas*). *Botanical Studies*, 50(1):43–55.

Chapuis, M.-P. and Estoup, A. (2007). Microsatellite null alleles and estimation of population differentiation. *Molecular Biology Evolution*, 24(3):621–631.

Charcosset, A., Lefortbuson, M., and Gallais, A. (1991). Relationship between heterosis and heterozygosity at marker loci - a theoretical computation. *Theoretical and Applied Genetics*, 81(5):571–575.

Chen, X., Cho, Y. G., and McCouch, S. R. (2002). Sequence divergence of rice microsatellites in *Oryza* and other plant species. *Molecular Genetics and Genomics*, 268(3):331–343.

Chen, Z. J. (2010). Molecular mechanisms of polyploidy and hybrid vigor. *Trends in Plant Science*, 15(2):57–71.

Clark, L. V. and Jasieniuk, M. (2011). POLYSAT: an R package for polyploid microsatellite analysis. *Molecular Ecology Resources*, 11(3):562–566.

Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., and Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, 142(1-2):169–196.

Crow, J. F. (1998). 90 years ago: The beginning of hybrid maize. *Genetics*, 148(3):923–928.

Darwin, C. (1876). *The effects of cross- and self-fertilization in the vegetable kingdom*. John Murray, London, United Kingdom.

David, M. C., Diaz, F. C., Mwanga, R. O. M., Tumwegamire, S., Mansilla, R. C., and Grüneberg, W. J. (2018). Gene pool subdivision of East African sweetpotato parental material. *Crop Science*, 58(6):2302–2314.

De Silva, H. N., Hall, A. J., Rikkerink, E., McNeilage, M. A., and Fraser, L. G. (2005). Estimation of allele frequencies in polyploids under certain patterns of inheritance. *Heredity*, 95(4):327–334.

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

Doyle, J. J., Morgante, M., Tingey, S. V., and Powell, W. (1998). Size homoplasy in chloroplast microsatellites of wild perennial relatives of soybean (*Glycine* subgenus *Glycine*). *Molecular Biology and Evolution*, 15(2):215–218.

Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology*, 23(1):40–69.

East, E. (1908). Inbreeding in corn. *Conneticut Agricultural Experiment Station Report 1907*, pages 419–428.

East, E. (1936). Heterosis. *Genetics*, pages 375–397.

Esselink, G. D., Nybom, H., and Vosman, B. (2004). Assignment of allelic configuration in polyploids using the MAC-PR (microsatellite DNA allele counting-peak ratios) method. *Theoretical and Applied Genetics*, 109(2):402–408.

Estoup, A., Wilson, I., Sullivan, C., J.-M., C., and Moritz, C. (2001). Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics*, 159:1671–1687.

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes - application to human mitochondrial-DNA restriction data. *Genetics*, 131(2):479–491.

Eyzaguirre, R. (2019). st4gi: Statistical tools for genetic improvement. R package version 2.2.8. `https://github.com/reyzaguirre/st4gi`.

Falconer, D. and MacKay, T. (1996). *Introduction To Quantitative Genetics*. Pearson Education Limited, London, United Kingdom, 4th edition.

Fan, X., Zhang, Y., Yao, W., Bi, Y., Liu, L., Chen, H., and Kang, M. (2013). Reciprocal diallel crosses impact combining ability, variance estimation, and heterotic group classification. *Crop Science*, 54:89–97.

Fan, X. M., Zhang, Y. M., Yao, W. H., Chen, H. M., Tan, J., Xu, C. X., Han, X. L., Luo, L. M., and Kang, M. S. (2009). Classifying maize inbred lines into heterotic groups using a factorial mating design. *Agronomy Journal*, 101(1):106–112.

FAOSTAT (2016). `http://www.fao.org/faostat/en/#data/QC`. [Online; accessed 14-November-2018].

Feng, J. Y., Li, M., Zhao, S., Zhang, C., Yang, S. T., Qiao, S., Tan, W. F., Qu, H. J., Wang, D. Y., and Pu, Z. G. (2018). Analysis of evolution and genetic diversity of sweetpotato and its related different polyploidy wild species *I. trifida* using RAD-seq. *BMC Plant Biology*, 18:12.

Gibson, R. and Kreuze, J. (2015). Degeneration in sweetpotato due to viruses, virus-cleaned planting material and reversion: a review. *Plant Pathology*, 64:1–15.

Gibson, R. W., Aritua, V., Byamukama, E., Mpembe, I., and Kayongo, J. (2004). Control strategies for sweet potato virus disease in Africa. *Virus Research*, 100(1):115–122.

Griffing, B. (1956). Concept of general and specific combining ability in relation to diallel crossing systems. *Australian Journal of Biological Sciences*, 9:463–493.

Grüneberg, W. J., Ma, D., Mwanga, R. O. M., Carey, E. E., Huamani, K., Diaz, F., Eyzaguirre, R., Guaf, E., Jusuf, M., Karuniawan, A., Tjintokohadi, K., Song, Y. S., Anil, S. R., Hossain, M., Rahaman, E., Attaluri, S. I., Some, K., Afuape, S. O., Adofo, K., Lukonge, E., Karanja, L., Ndirigwe, J., Ssemakula, G., Agili, S., Randrianaivoarivony, J. M., Chiona, M., Chipungu, F., Laurie, S. M., Ricardo, J., Andrade, M., Fernandes, F. R., Mello, A. S., Khan, M. A., Labonte, D. R., and Yencho, G. C. (2015). Advances in sweetpotato breeding from 1992 to 2012. In Low, J., Nyongesa, M., Quinn, S., and Parker, M., editors, *Potato and Sweetpotato in Africa: Transforming the Value Chains for Food and Nutrition Security*, pages 3–68. CAB International, Wallingford, United Kingdom.

Grüneberg, W. J., Mwanga, R. O. M., Andrade, M., and Dapaah, H. (2009a). Challenge theme paper 1: sweetpotato breeding. In Barker, C., editor, *Unleashing the potential of sweetpotato in Sub-Saharan Africa: current challenges and way forward*, pages 1–42. International Potato Center (CIP), Lima, Peru.

Grüneberg, W. J., Mwanga, R. O. M., Andrade, M., and Espinoza, J. (2009b). Selection methods part 5: breeding clonally propagated crops. In Ceccarelli, S., Guimaraes, E., and Weltzien, E., editors, *Plant Breeding and Farmer Participation*, pages 275–322. Food and Agriculture Organization of the United Nations (FAO), Rome.

Guichoux, E., Lagache, L., Wagner, S., Chaumeil, P., Leger, P., Lepais, O., Lepoittevin, C., Malausa, T., Revardel, E., Salin, F., and Petit, R. J. (2011). Current trends in microsatellite genotyping. *Molecular Ecology Resources*, 11(4):591–611.

Gurmu, F., Hussein, S., and Lain, M. (2013). Self- and cross-incompatibilities in sweetpotato and their implications on breeding. *Australian Journal of Crop Science*, 7(13).

Gurmu, F., S., H., and Laing, M. (2017). Genotype-by-environment interaction and stability of sweetpotato genotypes for root dry matter, β-carotene and fresh root yield. *Open Agriculture*, 2:473–485.

Gurmu, F., S., H., and Laing, M. (2018). Combining ability, heterosis, and heritability of storage root dry matter, beta-carotene, and yield-related traits in sweetpotato. *HortScience*, 53(2):167–175.

Hardy, O. J. (2016). Population genetics of autopolyploids under a mixed mating model and the estimation of selfing rate. *Molecular Ecology Resources*, 16(1):103–117.

He, G., Prakash, C., and Jarret, R. (1995). Analysis of genetic diversity in a sweetpotato (*Ipomoea batatas*) germplasm collection using DNA amplification fingerprinting. *Genome*, 38:938–945.

Henderson, C. (1948). *Estimation of general, specific and maternal combining abilities in crosses among inbred lines of swine*. PhD thesis, Iowa State College.

Hirakawa, H., Okada, Y., Tabuchi, H., Shirasawa, K., Watanabe, A., Tsuruoka, H., Minami, C., Nakayama, S., Sasamoto, S., Kohara, M., Kishida, Y., Fujishiro, T., Kato, M., Nanri, K., Komaki, A., Yoshinaga, M., Takahata, Y., Tanaka, M., Tabata, S., and Isobe, S. N. (2015). Survey of genome sequences in a wild sweet potato, *Ipomoea trifida* (H. B. K.) G. Don. *DNA Research*, 22(2):171–179.

Hochholdinger, F. and Hoecker, N. (2007). Towards the molecular basis of heterosis. *Trends in Plant Science*, 12(9):427–432.

Idury, R. M. and Cardon, L. R. (1997). A simple method for automated allele binning in microsatellite markers. *Genome Research*, 7(11):1104–1109.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24:1403–1405.

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, 11.

Jombart, T., Pontier, D., and Dufour, A. B. (2009). Genetic markers in the playground of multivariate analysis. *Heredity*, 102(4):330–341.

Jones, D. (1917). Dominance of linked factors as a means of accounting for heterosis. *Genetics*, 2:466–475.

Jones, N., Ougham, H., and Thomas, H. (1997). Markers and mapping: we are all geneticists now. *New Phytologist*, 137(1):165–177.

Kobayashi, M. (1984). The *Ipomoea trifida* complex closely related to sweet potato. In Shideler, S. and Rincon, H., editors, *Proceedings of the Sixth Symposium of the International Society of Tropical Root Crops*, pages 561–568. International Potato Center (CIP), Lima, Peru.

Kosman, E. and Leonard, K. (2005). Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Molecular Ecology*, 14:415–424.

Koussao, S., Gracen, V., Asante, I., Danquah, E., Ouedraogo, J., Baptiste, T., Jerome, B., and Vianney, T. (2014). Diversity analysis of sweet potato (*Ipomoea batatas* [L.] Lam) germplasm from Burkina Faso using morphological and simple sequence repeats markers. *African Journal of Biotechnology*, 13(6):729–742.

Kriegner, A., Cervantes, J. C., Burg, K., Mwanga, R. O. M., and Zhang, D. P. (2003). A genetic linkage map of sweetpotato *Ipomoea batatas* (L.) Lam. based on AFLP markers. *Molecular Breeding*, 11(3):169–185.

Kudom, A. A., Mensah, B. A., Froeschl, G., Boakye, D., and Rinder, H. (2015). Preliminary assessment of the potential role of urbanization in the distribution of carbamate and organophosphate resistant populations of *Culex* species in Ghana. *PARASITES & VECTORS*, 8(8).

Lenarcic, A. B., Svenson, K. L., Churchill, G. A., and Valdar, W. (2012). A general Bayesian approach to analyzing diallel crosses of inbred strains. *Genetics*, 190(2):413–U221.

Librando, R. and Magulama, E. (2008). Classifying white inbred lines into heterotic groups using yield combining ability effects. *University of Southern Mindanao Research and Development Journal*, 16(1):99–103.

Low, J. (2012). *Sweetpotato Action for Security and Health in Africa (SASHA)*. International Potato Center (CIP).

Low, J. W., Arimond, M., Osman, N., Cunguara, B., Zano, F., and Tschirley, D. (2007). Food-based approach introducing orange-fleshed sweet potatoes increased vitamin A intake and serum retinol concentrations in young children in rural Mozambique. *Journal of Nutrition*, 137(5):1320–1327.

Low, J. W., Mwanga, R. O. M., Andrade, M., Carey, E., and Ball, A. M. (2017). Tackling vitamin A deficiency with biofortified sweetpotato in sub-Saharan Africa. *Global Food Security-Agriculture Policy Economics and Environment*, 14:23–30.

Low, J. W., Walker, T., and Hijmans, R. (2001). *The potential impact of orange-fleshed sweetpotatoes on vitamin A intake in Sub-Saharan Africa*. International Potato Center (CIP), Lima, Peru.

Lynch, M. (1990). The similarity index and DNA fingerprinting. *Molecular Biology and Evolution*, 7:478–484.

Mace, E., Buhariwalla, H., and J.H., C. (2003). A high-throughput DNA extraction protocol for tropical molecular breeding programs. *Plant Molecular Biology Reporter*, 21(4):459–460.

Meirmans, P. G. (2012). Amova-based clustering of population genetic data. *Journal of Heredity*, 103(5):744–750.

Meirmans, P. G., Liu, S. L., and van Tienderen, P. H. (2018). The analysis of polyploid genetic data. *Journal of Heredity*, 109(3):283–296.

Melchinger, A. E. (1999). Genetic diversity and heterosis. In Coors, J. and Pandy, S., editors, *The Genetics and Exploitation of Heterosis in Crops*, pages 99–108. American Society of Agronomy, Inc and Crop Science Society of America, Inc, Madison, USA.

Melchinger, A. E. and Gumber, R. K. (1998). Overview of heterosis and heterotic groups in agronomic crops. In Lamkey, K. R. and Staub, J. E., editors, *Concepts and Breeding of Heterosis in Crop Plants*, CSSA Special Publications, pages 29–44.

Meng, Y. S., Zhao, N., Li, H., Zhai, H., He, S. Z., and Liu, Q. C. (2018). SSR fingerprinting of 203 sweetpotato (*Ipomoea batatas* (L.) Lam.) varieties. *Journal of Integrative Agriculture*, 17(1):86–93.

Menkir, A., Melake-Berhan, A., The, C., Ingelbrecht, I., and Adepoju, A. (2004). Grouping of tropical mid-altitude maize inbred lines on the basis of yield data and molecular markers. *Theoretical and Applied Genetics*, 108:1582–1590.

Munoz-Rodriguez, P., Carruthers, T., Wood, J. R. I., Williams, B. R. M., Weitemier, K., Kronmiller, B., Ellis, D., Anglin, N. L., Longway, L., Harris, S. A., Rausher, M. D., Kelly, S., Liston, A., and Scotland, R. W. (2018). Reconciling conflicting phylogenies in the origin of sweet potato and dispersal to Polynesia. *Current Biology*, 28(8):1246–+.

Mwanga, R. O. M. (2001). *Nature of Resistance and Response of Sweetpotat to Sweetpotato Virus Disease*. Thesis, North Carolina State University.

Mwanga, R. O. M., Kriegner, A., Cervantes-Flores, J. C., Zhang, D. P., Moyer, J. W., and Yencho, G. C. (2002a). Resistance to sweetpotato chlorotic stunt virus and sweetpotato feathery mottle virus is mediated by two separate recessive genes in sweetpotato. *Journal of the American Society for Horticultural Science*, 127(5):798–806.

Mwanga, R. O. M., Yencho, C. G. C., and Moyer, J. W. (2002b). Diallel analysis of sweetpotatoes for resistance to sweetpotato virus disease. *Euphytica*, 128(2):237–248.

Ngailo, S., Shimelis, H., Sibiya, J., Amelework, B., and Mtunda, K. (2016). Genetic diversity assessment of Tanzanian sweetpotato genotypes using simple sequence repeat markers. *South African Journal of Botany*, 102:40–45.

O'Brien, P. J. (1972). The sweet potato: Its origin and dispersal. *American Anthropologist*, 74(3):342–365.

Oliveira, E. J., Padua, J. G., Zucchi, M. I., Vencovsky, R., and Vieira, M. L. C. (2006). Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology*, 29(2):294–307.

Oracion, M. (1995). Determination and analysis of somatic chromosome numbers in sweetpotato cultivars related *Ipomoea* species, and interspecific hybrids. *Annals of Tropical Research*, 17:11–23.

Oswald, A., Kapinga, R., Lemaga, B., Ortiz, O., Krochel, J., and Lynam, J. (2009). Challenge theme paper 5: integrated crop management. In Barker, C., editor, *Unleashing the potential of sweetpotato in Sub-Saharan Africa: current challenges and way forward*, pages 130–153. International Potato Center (CIP), Lima, Peru.

Paradis, E. and Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528.

Parisod, C., Holderegger, R., and Brochmann, C. (2010). Evolutionary consequences of autopolyploidy. *New Phytologist*, 186(1):5–17.

Piepho, H. P., Mohring, J., Melchinger, A. E., and Buchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161(1-2):209–228.

Powers, L. (1944). An expansion of Jones's theory for the explanantion of heterosis. *The American Naturalist*, 78(776):275–280.

Prasanth, V., Chandra, S., Jayashree, B., and Hoisington, D. (2006). AlleloBin - A program for allele binning of microsatellite markers based on the the algorithm of Idury and Cardon (1997). International Crops Research Institute for the Semi-Arid Tropics (ICRISAT).

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rincent, R., Nicolas, S., Bouchet, S., Altmann, T., Brunel, D., Revilla, P., Malvar, R. A., Moreno-Gonzalez, J., Campo, L., Melchinger, A. E., Schipprack, W., Bauer, E., Schoen, C. C., Meyer, N., Ouzunova, M., Dubreuil, P., Giauffret, C., Madur, D., Combes, V., Dumas, F., Bauland, C., Jamin, P., Laborde, J., Flament, P., Moreau, L., and Charcosset, A. (2014). Dent and flint maize diversity panels reveal important genetic potential for increasing biomass production. *Theoretical and Applied Genetics*, 127(11):2313–2331.

Robinson, G. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32.

Roullier, C., Duputie, A., Wennekes, P., Benoit, L., Bringas, V. M. F., Rossel, G., Tay, D., McKey, D., and Lebot, V. (2013). Disentangling the origins of cultivated sweet potato (*Ipomoea batatas* (L.) Lam.). *Plos One*, 8(5):12.

Roullier, C., Rossel, G., Tay, D., McKey, D., and Lebot, V. (2011). Combining chloroplast and nuclear microsatellites to investigate origin and dispersal of new world sweet potato landraces. *Molecular Ecology*, 20:3963–3977.

Rukundo, P., Shimelis, H., Laing, M., and Gahakwa, D. (2017). Combining ability, maternal effects, and heritability of drought tolerance, yield and yield components in sweetpotato. *Frontiers in Plant Science*, 7.

Sattler, M. C., Carvalho, C. R., and Clarindo, W. R. (2016). The polyploidy and its key role in plant breeding. *Planta*, 243(2):281–296.

Schafleitner, R., Tincopa, L., Palomino, O., Rossel, G., Robles, R., Alagon, R., Rivera, C., Quispe, C., Rojas, L., Pacheco, J., Solis, J., Cerna, D., Kim, J., Hou, J., and Simon, R. (2010). A sweetpotato gene index established by *de novo* assembly of pyrosequencing and Sanger sequences and mining for gene-based microsatellite markers. *BMC Genomics*, 11:604.

Schnable, P. S. and Springer, N. M. (2013). Progress toward understanding heterosis in crop plants. In Merchant, S. S., editor, *Annual Review of Plant Biology, Vol 64*, pages 71–88.

Sehn, J. (2015). Chapter 9 - insertions and deletions (indels). In Kulkarni, S. and Pfeifer, J., editors, *Clinical Genomics*, pages 129–150. Academic Press, Missouri, USA.

Shull, G. (1908). The composistion of a field of maize. *Journal of Heredity*, 4(1):296–301.

Shumbusha, D., Tusiime, G., Edema, R., Gibson, P., Adipala, E., and Mwanga, R. (2014). Inheritance of root dry matter content in sweetpotato. *African Crop Science Journal*, 22(1):69–78.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139(1):457–462.

Soehendi, R. and Srinives, P. (2005). Significance of heterosis and heterobeltiosis in an F1 hybrid of mungbean (*Vigna radiata* (L.) Wilczek) for hybrid seed production. *SABRAO Journal of Breeding and Genetics*, 37(2):97–105.

Sprague, G. and Tatum, L. (1942). General vs. specific combining ability in single crosses of corn. *Journal of the American Society of Agronomy*, 34:923–932.

Sunnucks, P. (2000). Efficient genetic markers for population biology. *Trends in Ecology and Evolution*, 15(5):199–203.

Tairo, F., Mukasa, S. B., Jones, R. C., Kullaya, A., Rubaihayo, P. R., and Valkonen, J. P. T. (2005). Unravelling the genetic diversity of the three main viruses involved in sweet potato virus disease (SPVD), and its practical implications. *Molecular Plant Pathology*, 6(2):199–211.

Tian, H., S.A., C., and Hu, S. (2015). Heterotic grouping and the heterotic pattern among chinese rapeseed (*Brassica napus* L.) accessions. *Agronomy Journal*, 107(4):1321–1330.

Tukey, J. (1977). *Exploratory data analysis*. Addison-Wesley, Reading Massachusetts, USA.

Tumwegamire, S., Kapinga, R., Rubaihayo, P. R., LaBonte, D. R., Grüneberg, W. J., Burgos, G., zum Felde, T., Carpio, R., Pawelzik, E., and Mwanga, R. O. M. (2011). Evaluation of dry matter, protein, starch, sucrose, beta-carotene, iron, zinc, calcium, and magnesium in East African sweetpotato *Ipomoea batatas* (L.) Lam germplasm. *Hortscience*, 46(3):348–357.

Vasal, S., Srinivasan, G., Han, G., and F.C., G. (1992). Heterotic patterns of eighty-eight white subtropical CIMMYT maize lines. *Maydica*, 37:319–327.

Veitia, R. A. and Vaiman, D. (2011). Exploring the mechanistic bases of heterosis from the perspective of macromolecular complexes. *Faseb Journal*, 25(2):476–482.

Vieira, M., Santini, L., Diniz, A., and Munhoz, C. (2016). Microsatellite markers: what they mean and why they are so useful. *Genetics and Molecular Biology*, 39(3):312–328.

Westcott, B. (1981). Two methods for early generation yield assessment in winter wheat. In *Proc. of the 4th meeting of the Biometrics in Plant Breeding Section of Eucarpia*, pages 91–95. INRA, Poitier, France.

Woolfe, J. (1992). *Sweet potato: an untapped food resource*. Cambridge University Press, Cambridge.

Wu, R. L., Gallo-Meagher, M., Littell, R. C., and Zeng, Z. B. (2001). A general polyploid model for analyzing gene segregation in outcrossing tetraploid species. *Genetics*, 159(2):869–882.

Wu, S., Lau, K., Cao, Q., Hamilton, J., Sun, H., Zhou, C., Eserman, L., Gemenet, D., Olukolu, B., Wang, H., Crisovan, E., Godden, G., Jiao, C., Wang, X., Kitavi, M., Manrique-Carpintero, N., Vaillancourt, B., Wiegert-Rininger, K., Yang, X., Bao, K., Schaff, J., Kreuze, J., Grüneberg, W., Khan, A., Ghislain, M., Ma, D., Jiang, J., Mwanga, R., Leebens-Mack, J., Coin, L., Yencho, G., Buell, C., and Fei, Z. (2018). Genome sequences of two diploid wild relatives of cultivated sweetpotato reveal targets for genetic improvement. *Nature Communications*, 9(1):4580.

Yada, B., Alajo, A., Ssemakula, G. N., Mwanga, R. O. M., Brown-Guedira, G., and Yencho, G. C. (2017a). Selection of simple sequence repeat markers associated with inheritance of sweetpotato virus disease resistance in sweetpotato. *Crop Science*, 57(3):1421–1430.

Yada, B., Brown-Guedira, G., Alajo, A., Ssemakula, G. N., Owusu-Mensah, E., Carey, E. E., Mwanga, R. O. M., and Yencho, G. C. (2017b). Genetic analysis and association of simple sequence repeat markers with storage root yield, dry matter, starch and beta-carotene content in sweetpotato. *Breeding Science*, 67(2):140–150.

Yada, B., Tukamuhabwa, P., Wanjala, B., Kim, D. J., Skilton, R. A., Alajo, A., and Mwanga, R. O. M. (2010). Characterization of Ugandan sweetpotato germplasm using fluorescent labeled simple sequence repeat markers. *Hortscience*, 45(2):225–230.

Yang, J., Moeinzadeh, M. H., Kuhl, H., Helmuth, J., Xiao, P., Haas, S., Liu, G. L., Zheng, J. L., Sun, Z., Fan, W. J., Deng, G. F., Wang, H. X., Hu, F. H., Zhao, S. S., Fernie, A. R., Boerno, S., Timmermann, B., Zhang, P., and Vingron, M. (2017). Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nature Plants*, 3(9):696–703.

Yencho, C. (2015). *The genomic tools for sweetpotato improvement project - GT4SP*. North Carolina State University, USA.

Zhang, D., Carbajulca, D., Ojeda, L., Rossel, G., Milla, S., Herrera, C., and Ghislain, M. (1999). Microsatellite analysis of genetic diversity in sweetpotato varieties from Latin America. *CIP program report, 2000*, pages 295–301.

APPENDIX A

SUPPLEMENTARY MATERIAL

A.1 Protocols of DNA extraction and PCR

A.1.1 DNA extraction, quantification and normalization

Approximately 1g of lyophilized leaf tissues were put in 2mL screw cap microfuge tubes (Sarstedt, Germany) with two washed and autoclaved, 4mm diameter, stainless steel ball bearings (Spex CertiPrep, USA). Leaves were completely and thoroughly ground (20sec or more) until fine powder was obtained using the FastPrep-2 5G tissue homogenizer (MP Biomedicals). DNA extraction followed a modification of combined Dellaporta *et al.* (1983) and Mace *et al.* (2003) protocols optimized for pure, high quality and quantity DNA from plant tissues. In the fume hood chamber 1ml of pre-warmed (65°C) CTAB buffer (200mM Tris-HCl pH 8, 50mM EDTA, 2M NaCl, 2% CTAB and 3% β-mercaptoethanol) was added to the leaf powder and mixed by vortex at 3,000rpm for 30sec. Tubes were heated in a water bath at 65°C for 45 minutes and were inverted after every 10 minutes. Non-soluble debris was removed by adding 500μl chlroform:isoamyl alcohol (24:1), mixed by gently inverting the tubes until the mixture appeared milky. Centrifugation of tubes was done at maximum speed for 15 minutes and supernatant recovered to new Eppendorf safe-lock tubes. This step (chloroform:isoamyl alcohol) was repeated once before transferring the aqueous phase to fresh tubes. DNA was precipitated with 1/5 volume of 5M NaOAC and 2 volumes of cold absolute ethanol. The samples were mixed gently and incubated at -20°C for 60 minutes. Tubes were centrifuged for 25 minutes, the supernatant removed by decanting and DNA pellet washed by adding 500μl of cold 70% (v/v) ethanol. By vortexing, the pellet was dislodged, or by flicking with finger. Maximum centrifugal force was applied for 10 minutes in a microcentrifuge at room temperature and the 70% ethanol supernatant removed, taking care not to disturb the pellet. This step was repeated twice, then the DNA pellet was air dried to remove residual ethanol. The white pellet, a mixture of DNA and RNA was suspended in 100μl low salt TE buffer (1mM Tris-HCl pH 8, 0.1mM EDTA) followed by RNAse treatment (4μL of 10mg/μL of RNase-A stock solution). DNA samples were allowed to dissolve overnight. DNA amount and pu-

rity were pre-estimated in this step by measuring the DNA concentration (ng/μl) with NanoDrop (50 times the value of absorbance at 260nm) and then diluted to a working concentration of 20ng/μl.

A.1.2 PCR

PCR reactions for SSR amplification were done in 10μl reaction volume containing 5μl Taq PCR MasterMix (https://bioneer.com.au/Files/Link/DNA-Amp/manual-Taq-PCR-PreMix.aspx with Taq DNA polymerase, dNTPs, reaction buffer and stabilizer but without tracking dye), 1μl fluorescently labeled Forward primer (5pmol/μl; 4 different dyes incorporated during oligo synthesis: VIC, FAM, PET and NED (Applied Biosystems)), 1μl Reverse primer (5pmol/μl), 2μl DNA template and ddH20. Amplification steps followed (i) initial denaturation at 95°C for 5 minutes; (ii) 40 cycles at 95°C for 0.30 minutes, 1 minute of 50°C to 61°C annealing temperature (primer pair specific), and 72°C for 2 minutes (iii) final extension of 20 minutes at 72°C. All loci were individually amplified, four PCR primer products were multiplexed based on the dye and expected size of the fragment.

A.2 FullBruvo R code

Running the following lines of R code will enable the FullBruvo calculation. Four functions are defined in this code *alcomb*, *getalldist*, and *FullBruvo* are needed for the calculation of the FullBruvo distance with the *FullBruvo2* function. This requires the parallel package in R. This code was written by the author of this dissertation, based on the polysat package by Clark and Jasieniuk (2011).

```
library(parallel)

alcomb <- function(genotype,diff) {
  lg <- length(genotype)
  mat <- matrix(nrow = diff, ncol = lg^diff)
  for (i in 1:diff) {
    mat[i, ] <- rep(genotype, times = lg^(i - 1),
    each = lg^(diff - i))
  }
  if (diff > 1) {
    mat <- apply(mat, 2, sort)
  }
  return(mat)
}

getalldist <- function(dmat, nrow) {
  thesedist <- dmat[, 1]
  if (dim(dmat)[2] > 1) {
    newdist <- numeric(0)
    for (i in 1:nrow) {
      newdist <- c(newdist, thesedist[i] + getalldist(dmat[-i, -1, drop = FALSE],
          nrow - 1))
    }
    thesedist <- newdist
  }
  return(thesedist)
}

FullBruvo <- function (genotype1, genotype2, usatnt = 2, missing = -9,realPloidy =
    6)
{
  if (is.na(usatnt))
  stop("ErrorType1")
  if (genotype1[1] == missing | genotype2[1] == missing) {
    dist <- NA
  }
  else {
    if (identical(genotype1, genotype2)) {
      dist <- 0
    }
```

```
39     else if(length(genotype1) != realPloidy | length(genotype2) != realPloidy){
40       stop("ErrorType2")
41     }
42     else {
43       genotypeL <- genotype1/usatnt
44       genotypeS <- genotype2/usatnt
45       allele.distances <- array(0, c(realPloidy, realPloidy))
46       for (n in 1:realPloidy) {
47         for (m in 1:realPloidy) {
48           allele.distances[n, m] <- genotypeL[n] - genotypeS[m]
49         }
50       }
51       geometric.distances <- 1 - 2^-abs(allele.distances)
52       mindist <- min(getalldist(geometric.distances, realPloidy))
53       dist <- mindist/realPloidy
54     }
55   }
56   return(dist)
57 }
58
59 FullBruvo2 <- function (genotype1, genotype2, usatnt = 2, missing = -9, realPloidy =
       6)
60 {
61 if (length(genotype1) == realPloidy & length(genotype2) == realPloidy || genotype1
      [1] == missing || genotype2[1] == missing) {
62   d <- FullBruvo(genotype1, genotype2, usatnt = usatnt, missing = missing,
       realPloidy = realPloidy)
63 }
64 else {
65   diff1 <- realPloidy - length(genotype1)
66   diff2 <- realPloidy - length(genotype2)
67   alleleadd1 <- alcomb(genotype1,diff1)
68   alleleadd2 <- alcomb(genotype2,diff2)
69   nocores <- detectCores()-1
70   cl <- makeCluster(nocores)
71   clusterExport(cl,list('genotype1','genotype2','usatnt','missing','realPloidy','
       alleleadd1','alleleadd2','FullBruvo','getalldist','alcomb'),envir =
       environment())
72   distaddpar <- parApply(cl,alleleadd1,2,function(x){apply(alleleadd2,2,function(y){
       FullBruvo(c(genotype1,x), c(genotype2,y),usatnt = usatnt, missing = missing,
       realPloidy = realPloidy)})})
73   stopCluster(cl)
74   d <- mean(distaddpar, na.rm = TRUE)
75   }
76 return(d)
77 }
```

A.3 Extra information on the parents

Table A.1: Extra information on the parents in the Ghana crossing block. Whether the parent was used in the phenotypic and/or genotypic analysis, information on their background (country of origin, breeding program) or pedigree and a unique identifier (DOI).

Accession	Phenotyped	Genotyped	Background/Pedigree	DOI
AP3A	No	Yes	Apomuden x OP	10.18730/MYCAD
Apomuden	Yes	Yes	Bangladesh	10.18730/MYC47
BF59xCip.4	Yes	Yes	Burkina Faso	10.18730/MYC58
BF92XTib.2	Yes	Yes	Burkina Faso	10.18730/MYC69
Blueblue	Yes	Yes	IITA Tib 2?	10.18730/MYC7A
Bohye	Yes	Yes	Peru	10.18730/MYC8B
CIP440390	Yes	Yes	IITA-TIS 87/0087	10.18730/M9T00
CIP442162	Yes	Yes	IITA-TIS 82/0270-OP-1-105	10.18730/MYCBE
Dadanyuie	No	Yes	KEMB 37 (KSP20/TIS 2534?)	10.18730/MYCCF
Faara	Yes	Yes	IITA-TIS 3017	10.18730/MYCDG
Hi-starch	Yes	Yes	Japan	10.18730/SC8WC
Jitihada	Yes	Yes	Kenya	10.18730/MYCEH
King-J	No	Yes	Nigeria	10.18730/MYCFJ
Kuffour	No	Yes	Ghana landrace	10.18730/SC8XD
Ligri	Yes	Yes	Cuba	10.18730/MYCGK
MothersDelight	Yes	Yes	Unknown	10.18730/MYCHM
NKO31A	Yes	Yes	New Kawogo x OP	10.18730/MYCJN
Nanungungungu	Yes	Yes	Burkina Faso	10.18730/SC8YE
Obare	Yes	Yes	Ghana landrace	10.18730/SC8ZF
Ogyefo	No	Yes	Rwanda	10.18730/MYCKP
Okumkom	No	Yes	IITA-TIS 8266	10.18730/MYCMQ
Otoo	Yes	Yes	Burundi	10.18730/MYCNR
PGA11040-6	No	Yes	Beauregard x BOT03/036	10.18730/SC90G
PGA11113-11	No	Yes	Sauti x Otoo	10.18730/SC91H
PGA12010-15	No	Yes	CIP 440293 x OP	10.18730/SC92J
PGA12086-18	No	Yes	BM85-42 x 03DM	10.18730/SC93K
PGA12136-2	No	Yes	Santompona x Faara	10.18730/SC94M
PGA12151-13	No	Yes	Apomuden x CRI-PC	10.18730/SC95N
PGA12164-21	No	Yes	Faara x PC	10.18730/SC96P
PGA12164-26	No	Yes	Faara x PC	10.18730/SC97Q
PGA12166-30	No	Yes	Faara x PC	10.18730/SC98R
PGA13067-7	Yes	Yes	Faara x Otoo	10.18730/SC99S
PGA14008-15	No	Yes	Otoo x PC	10.18730/MYCRV
PGA14008-22	No	Yes	Otoo x PC	10.18730/MYCSW
PGA14008-9	No	Yes	Otoo x PC	10.18730/MYCTX
PGA14011-13	No	Yes	Apomuden x PC	10.18730/MYCVY
PGA14011-43	No	Yes	Apomuden x PC	10.18730/MYCWZ
PGA14171-1	No	Yes	CIP199062.1 x Faara	10.18730/MYCX*
PGA14229-2	No	Yes	Faara x 040-6	10.18730/MYCY~
PGA14351-36	No	Yes	PGA12160-72 x PGA12151-75	10.18730/MYCZ$
PGA14351-4	No	Yes	PGA12160-72 x PGA12151-75	10.18730/MYD0=
PGA14372-3	No	Yes	PGA12151-53 x PC	10.18730/MYD1U
Patron	Yes	Yes	Burundi	10.18730/SC9AT
Santompona	Yes	Yes	IITA-TIS 84/0320	10.18730/SC9BV
Sauti	Yes	Yes	Uganda	10.18730/MYCPS
Tiebele-2	No	Yes	Burkina Faso	10.18730/MYCQT
Tu-orange	Yes	Yes	USA	10.18730/SC9CW
Tu-purple	Yes	Yes	USA	10.18730/SC9DX
CIP199062.1	No	Yes	= Bohye	/
Cemsa	No	Yes	= Ligri	10.18730/SC9EY
Tanzania	No	Yes	= Sauti	10.18730/M9T11

A.4 Mid-parent heterosis values

Male \ Female	CIP442162	Obare	Patron	Tu-orange	Nanungungungu	Jitihada	BF59xCip.4	Faara	Ligri	Santompona	Sauti	Otoo	MothersDelight	Tu-purple	BF92xTib.2	Apomuden	NKO31A	PGA13067-7	Hi-starch	Bohye	Blueblue	CIP440390
CIP442162						0	0	0			0			0				0		0		0
Obare			0			0						51							0			
Patron								123			0							63				
Tu-orange								334				225						109				322
Nanungungungu			0					20														
Jitihada	-46		0		0							0	102	91		0			0	0		0
BF59xCip.4	0		128	400		94		115		168		178		172		80		120	0			193
Faara	-48	0	82		0					0		111	191	0		0			106	113		170
Ligri								0						0		0		113	96			0
Santompona																						
Sauti	0		96		0					252		242		221		0			97			128
Otoo				0				160						171		0		103	166			0
MothersDelight			152					0				163						185				
Tu-purple	0		122		0	95	0	161			321	0	227					0	0	174		170
BF92xTib.2												297		374								
Apomuden	0		0		0	0		129			250	111	221	130				0		144		106
NKO31A			0									41		0		56				0		0
PGA13067-7			0									110	116	0		0			0			130
Hi-starch								101				129						98		149		124
BohyeParent	0		76			0		178		166		0	240	75				63				172
Blueblue																		0				
CIP440390	0	0	69		0			0			171	138		111		95		0	106	85		

Figure A.1: The average mid-parent heterosis (MH) values for the weigth of commercial storage roots (**WCSR**) for all cross combinations, expressed as a percentage difference from the mid-parent

Male \ Female	CIP442162	Obare	Patron	Tu-orange	Nanungungungu	Jitihada	BF59xCip.4	Faara	Ligri	Santompona	Sauti	Otoo	MothersDelight	Tu-purple	BF92xTib.2	Apomuden	NKO31A	PGA13067-7	Hi-starch	Bohye	Blueblue	CIP440390
CIP442162						0	0	0			0			0				0		0		0
Obare			46			0						76							0			
Patron								106			0							0				
Tu-orange								178				131						79				210
Nanungungungu			0					47														
Jitihada	0		0		0							0	0	89		0			0	0		0
BF59xCip.4	0		110	252		0		61		213		207		143		68		95	0			163
Faara	-40	72	59		0					130		146	143	0		0			98	60		135
Ligri								0						0		0		137	119			0
Santompona																						
Sauti	0		54		0					166		138		84		0			69			63
Otoo				0				119						127		0		94	189			80
MothersDelight			107					169				90						125				
Tu-purple	0		104		54	119	0	142			159	98	111					115	0	105		136
BF92xTib.2												178		212								
Apomuden	0		63		0	0		128			117	126	156	130				0		91		107
NKO31A			0									0		0		0				0		0
PGA13067-7			0									116	100	90		0			0			152
Hi-starch								106				135						93		74		95
BohyeParent	0		61			0		121		132		0	108	0				50				117
Blueblue																		131				
CIP440390	0	74	54		0			0			85	108		0		68		0	96	50		

Figure A.2: The average mid-parent heterosis (MH) values for **total biomass** for all cross combinations, expressed as a percentage difference from the mid-parent

A.5 Examples of good and poor binning

Figure A.3 shows an example for the cumulative allele length distributions of a well (IBS149; allelobin quality index = 0.10) and a poorly binned marker (IBY51; allelobin quality index = 0.45) from the SSR data. Similar to Figure 3.3. Interpretation explained in Sections 3.3.3 and 4.2.3. Notice the difference in within bin variation.

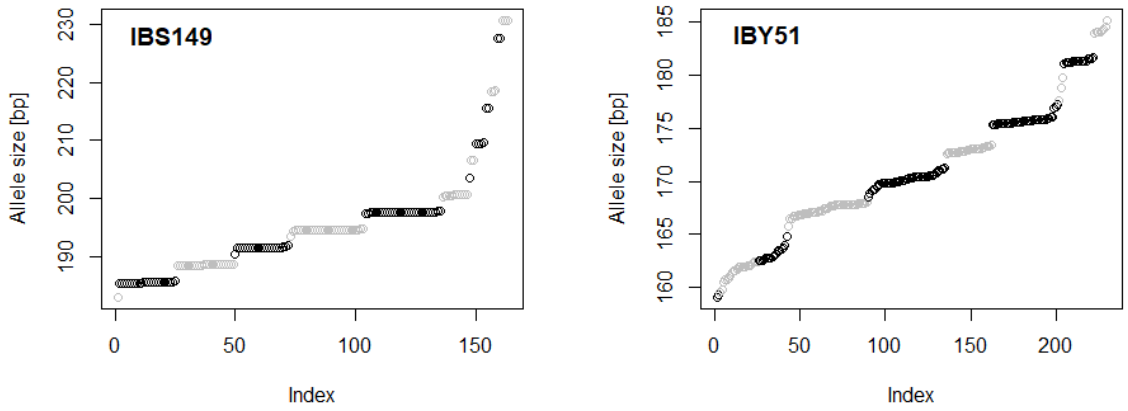


Figure A.3: Example of the cumulative allele length distributions of a well (left) and a poorly binned marker (right). Alternating bins are represented by different colours