University of Hohenheim
Faculty of Agricultural Sciences
Institute of Plant Breeding, Seed Science and Population Genetics
Crop Biodiversity and Breeding Informatics
Prof. Dr. Karl J. Schmid

Speciation and Domestication Genomics of *Amaranthus* spp.

Dissertation
Submitted in fulfillment of the requirements for the degree „Doktor der Agrarwissenschaften“ (Dr. sc. agr. /Ph. D. in Agricultural Sciences) to the Faculty of Agricultural Sciences

Presented by
Markus G. Stetter
from Ehingen (Donau)

Stuttgart-Hohenheim
2016

This thesis was accepted as a doctoral dissertation in fulfillment of the requirements for the degree "Doktor der Agrarwissenschaften" by the Faculty of Agricultural Sciences at the University of Hohenheim, on December 05, 2016.

Examination Committee:

Head of Committee:	Prof. Dr. Andrea Knierim
1^{st} examiner and reviewer:	Prof. Dr. Karl Schmid
2^{nd} examiner and reviewer:	Priv. Doz. Dr. Andreas Börner
3^{rd} examiner:	Prof. Dr. Uwe Ludewig
Day of oral examination:	December 15, 2016

I. Summary

The genus *Amaranthus* consists of 50 to 70 species, including several cultivated and weedy species. The seeds of the three grain amaranth species, *A. caudatus*, *A. cruentus* and *A. hypochondriacus* have a high nutritional value and are gluten free. In this work, three main aspects of amaranth genetics are studied, because previous work was limited to few species and few genetic markers: First, the evolutionary relationship between species in the genus; second, the domestication syndrome of South American grain amaranth; and third, crossing methods and controlled growth conditions for amaranth breeding.

The genus has been taxonomically split into three subgenera, *A. Amaranthus*, *A. Albersia* and *A. Acnida*. Together with their two relatives *A. hybridus* and *A. quitensis*, the three grain amaranths form the Hybridus complex within the *A. Amaranthus* subgenus. We used genotyping by sequencing (GBS) of 94 genebank accessions, representing 35 species to infer the phylogeny of *Amaranthus*. SNPs were called using *de novo* and reference genome based methods and genome sizes of the species were measured using flow cytometry. The analysis of genome size evolution within the genus revealed that with the exception of two lineages polyploidization played a minor role in the history of the genus. A distance-based neighbor joining tree of individual accessions and a species tree based on the multispecies coalescent were constructed. Both phylogenies supported the previous taxonomic classification into three subgenera, but split the *A. Acnida* subgenus into two distant groups. Analyzing the Hybridus complex gave insights into the domestication history of grain amaranth. The complex was well separated from the other species in the *A. Amaranthus* subgenus and included the three grain amaranth species and their wild relatives. Individuals within the Hybridus complex did not cluster by species, but rather by their geographic origin from South and Central America. Geographically separated lineages of *A. hybridus* appeared to be the common ancestor of the three cultivated grain amaranths, while *A. quitensis* was involved in the domestication of *A. caudatus*.

The domestication of grain amaranth remains unclear and seems to be complex, because the domestication syndrome that differentiates crops

from their wild ancestors is only weakly pronounced. Therefore, the domestication syndrome in South American grain amaranth (*A. caudatus*) was studied by characterizing genetic and phenotypic diversity of *A. caudatus* and the two potential wild relatives, *A. hybridus* and *A. quitensis*. To investigate the evolutionary relationship of *A. caudatus* and its potential wild ancestors, 119 amaranth accession from the Andean region were characterized using 9,485 GBS derived SNPs. Additionally, two domestication related phenotypes, seed color and seed size, were analyzed. None of the accessions of wild amaranths had white seeds, while this was the predominant seed color in *A. caudatus*. The seed size did not significantly differ between species, but a genetically distinct group of *A. caudatus* from Bolivia had significantly larger seeds than the other groups. The genetic analysis revealed a strong differentiation of *A. caudatus* from its wild relatives. The two relatives did not cluster according to their species assignment, but rather by their geographic origins from Peru and Ecuador. Surprisingly, *A. caudatus* had a higher genetic diversity than its two close relatives and shared a high proportion of polymorphisms with them, consistent with the absence of strong bottlenecks or high levels of gene flow between them.

Efficient crosses are an essential tool for plant research and breeding. Three different crossing methods (open pollination, hot water emasculation and hand emasculation) were evaluated for their efficiency and validated with low cost genetic markers. We identified controlled growth conditions for amaranth that allow short generation times of only six weeks instead of six months. All three crossing methods successfully produced offspring, but with different success rates. Open pollination had the lowest (10%) and hand emasculation the highest success rate (74%). Hot water emasculation showed an intermediate success rate (26%), but high maximum of 94%. Additionally, this method is easy to perform and suitable for large-scale hybrid production. Eleven PCR-based SNP markers were found to be sufficient for intra- and interspecific hybrid identification. Despite the very small flowers, amaranth crosses can be carried out efficiently and verified with inexpensive SNP markers with short generation times under suitable conditions.

The phylogeny and population genetic analysis suggest that amaranth

domestication was incomplete. Gene flow from *A. quitensis* into *A. caudatus* might have prevented the fixation of domestication related alleles. The genus phylogeny and the over 200 genotyped accessions in this work provide the largest genomic resource for amaranth so far. The cultivation and crossing methods presented, together with the genome wide markers could be introduced into breeding programs to increase amaranth yield and quality.

II. Zusammenfassung

Die Gattung *Amaranthus* besteht aus 50 bis 70 Arten und beinhaltet mehrerer Kultur- und Unkrautarten. Die drei Körneramarantarten, *A. caudatus*, *A. cruentus* und *A. hypochondriacus*, haben einen hohen Proteingehalt und sind glutenfrei. In dieser Arbeit wurden drei Hauptthemen der Amarantgenetik untersucht: Erstens, die evolutionäre Verwandtschaft aller Amarantarten, zweitens, das Domestikationssyndrom von südamerikanischem Körneramarant und drittens, Kreuzungsmethoden und kontrollierte Wachstumsbedingungen für die Amarantzüchtung.

Die Gattung wurde taxonomisch in drei Untergattungen, *A. Amaranthus*, *A. Albersia* und *A. Acnida*, unterteilt. Gemeinsam mit ihren beiden wilden Verwandten, *A. hybridus* und *A. quitensis*, bilden die drei Körneramarantarten den Hybridus-Artenkomplex. Um die Phylogenie der Gattung *Amaranthus* zu bestimmen und die wilden Vorfahren von Körneramarant zu finden, wurden 94 Genbankakzessionen, die 35 Arten repräsentieren, mit Genotyping-by-Sequncing (GBS) genotypisiert. Die SNPs wurden mit Hilfe von *de novo* und referenzbasierten Methoden gecalled. Weiterhin wurde die Genomgrößen der verschiedenen Arten durch Durchflusszytometrie bestimmt. Die Analyse der Genomgrößenevolution von *Amaranthus* ergab, dass mit Ausnahme von zwei Abstammungslinien Polyploidisierung nur eine untergeordnete Rolle in der Geschichte der Gattung spielte. Ein distanzbasierter phylogenetischer Baum der einzelnen Akzessionen und ein Artenbaum auf Basis der Multispezieskoaleszenz wurden konstruiert. Beide Phylogenien unterstützen die vorherige Einteilung der Untergattungen, jedoch ist *A. Acnida* in zwei voneinander entfernte Gruppen unterteilt. Die Analyse des Hybriduskomplexes konnte Einblick in die Domestikationsgeschichte von Körneramarant geben. Der Komplex war gut von den anderen Arten der *A. Amaranthus* Untergattung unterscheidbar, gruppierten die Individuen innerhalb des Komplexes aber stärker nach ihrer geographischen Herkunft als nach ihrer Artenzugehörigkeit. Es sieht danach aus, dass die drei Körneramarantarten von geographisch distanzierten Abstammungslinien von *A. hybridus* abstammen und *A. quitensis* direkt oder indirekt in die Domestikation von *A. caudatus* involviert war.

Die Domestikationsgeschichte von Körneramarant ist noch immer un-

klar und scheint komplex, da das Domestikationssyndrom, welches Kulturpflanzen von ihren wilden Vorfahren unterscheidet, bei Amarant nur schwach ausgeprägt ist. Das Domestikationssyndrom von südamerikanischem Körneramarant (*A. caudatus*) wurde durch die Charakterisierung der genetischen und phänotypischen Diversität in *A. caudatus* und den beiden potentiellen Vorfahren, *A. hybridus* und *A. quitensis*, untersucht. Um das evolutionäre Verhältnis der drei Arten zu untersuchen, wurden 119 Akzessionen aus der Andenregion mit 9.485 SNPs aus GBS analysiert. Darüber hinaus wurden zwei Domestikationsmerkmal, Samengröße und Samenfarbe, verglichen. Keine der Akzessionen der wilden Arten hatte weiße Samen, während die meisten *A. caudatus* Individuen weiße Samen hatten. Es gab jedoch auch *A. caudatus* Individuen mit dunklen Samen. Die Samengröße unterschied sich nicht signifikant zwischen den Arten, obwohl eine genetisch entfernte Gruppe bolivianischer *A. caudatus* Individuen signifikant größere Samen hatte als die anderen Gruppen. Die genetische Analyse zeigte, dass *A. caudatus* stark von seinen wilden Verwandten differenziert war. Die beiden wilden Arten gruppierten sich allerdings stärker nach ihrer Herkunft aus Peru und Ecuador als nach ihrer Artenzugehörigkeit. Erstaunlicherweise war die genetische Diversität bei *A. caudatus* höher als bei seinen beiden nahen Verwandten, was auf einen fehlenden genetischen Flaschenhals oder ein hohes Maß an Genfluss hinweist.

Effiziente Kreuzungen sind essentiell für die Pflanzenforschung und –züchtung. Deshalb wurden drei unterschiedliche Kreuzugsmethoden (Offene Bestäubung, Heißwasseremaskulierung und Handemaskulierung) auf ihre Effizienz getestet und die Nachkommen mit genetischen Markern validiert. Zusätzlich entwickelten wir kontrollierte Wachstumsbedingungen für Amarant die, die Generationszeit von sechs auf zwei Monate heruntersetzt. Alle drei Kreuzungsmethoden führten zu erfolgreichen Kreuzungen, aber mit unterschiedlichen Erfolgsraten. Offene Bestäubung hatte mit durchschnittlich 10% die geringste Erfolgsrate, während Handemaskulierung mit 74% die höchste Erfolgsrate hatte. Heißwasseremaskulierung hatte zwar eine mittlere durchschnittliche Erfolgsrate (26%), jedoch eine maximale Erfolgsrate von 94%. Außerdem ist diese Methode einfach durchzuführen und deshalb für die Anwendung zur Hybriderzeugung im

Feld geeignet. Die elf PCR-basierten SNP Marker waren ausreichend, um sowohl intra- als auch interspezifische Hybride zu identifizieren. Trotz der kleinen Blüten können Amarantkreuzungen effizient erzeugt werden und mit kostengünstigen genetischen Markern identifiziert werden.

Die Phylogenie und die populationsgenetische Analyse weisen darauf hin, dass die Domestikation von Amarant unvollständig war. Genfluss aus *A. quitensis* in *A. caudatus* könnte die Fixierung von Domesikationsallelen verhindert haben. Die Gattungsphylogenie von *Amaranthus* und die mehr als 200 genotypisierten Akzessionen dieser Arbeit stellen die bisher größte genomische Ressource für Amarant dar. Die Wachstumsbedingungen und Kreuzungsmethoden können in Zuchtprogramme integriert werden, um den Ertrag und die Qualität von Amarant zu steigern.

Contents

List of Figures

List of Tables

Abbreviations

ANOVA	Analysis of Variance
CMS	Cytoplasmic Male Sterility
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DAPC	Discriminant Analysis of Principal Components
GBS	Genotyping by Sequencing
GWAS	Genome Wide Association Study
HSW	Hundred Seed Weight
KASP	Kompetitive Allele Specific PCR
MAF	Minor Allele Frequency
MAS	Marker Assisted Selection
PCR	Polymerase Chain Reaction
QTL	Quantitative Trait Locus
sgRNA	single guide RNA
SNP	Single Nucleotide Polymorphism

1. General Introduction

Over the last 10,000 to 12,000 years, 2,500 crops have been domesticated. Only three crops, maize, rice and wheat, provide over one third of calories for human consumption (Gaut *et al*, 2015). The history of crops is tightly linked to human history, because their domestication converted human lifestyle from hunter-gatherer to settled societies. Understanding the process of crop domestication can thus shed light on early civilization and explain, why some crops spread more successfully than others. Understanding this early adaptation process of crops to human needs is also important to improve modern crops and adapt them to changing environments. Several crop species are only partially domesticated, but the potential constraints preventing domestication are not fully understood (Lenser & Theißen, 2013). Further investigation of the domestication process of such crops will help understanding domestication in general.

The genus *Amaranthus* comprises several crop species that are potentially only partially domesticated (Brenner *et al*, 2010). Additionally, grain amaranth seeds have a high nutritional value and the crop has high potential to contribute to human nutrition in the future (Rastogi & Shukla, 2013). Together, these characteristics make the genus, and in particular grain amaranth, interesting models to study domestication, genome evolution and breeding methods.

1. The Genus *Amaranthus*

The genus *Amaranthus* is part of the *Amarantaceae* family. Within the genus, there are 50 to 70 species with a world wide distribution (Costea & DeMason, 2001; Sauer, 1967). The taxonomy of the genus is complex and there is no clear key for species identification in the *Amaranthus*

genus. Furthermore, duplicated naming for several species complicates the identification and study of the genus (Brenner *et al*, 2010; Das, 2014). Nevertheless, there have been attempts to organize the *Amaranthus* genus. Based on morphological features, the genus has been split into three subgenera, namely *A. Amaranthus*, *A. Acnida* and *A. Albersia* (Mosyakin & Robertson, 1996).

The genus *Amaranthus* went through an ancient allotetraploidization, but most of its species now inherit their chromosomes as diploids (Clouse *et al*, 2016). However, other species underwent further genome duplication and are tetraploids (Behera & Patnaik, 1982). Systematic studies on the genome size evolution in the *Amaranthus* genus are missing.

Within the genus there are several species with a high economical and historical importance. Weedy *Amaranthus* species lead to substantial yield losses in field crops. The well-known palmer amaranth (*A. palmeri*) leads to yield losses in soybean of 30 to 70 % (Bensch *et al*, 2003; Davis *et al*, 2015). One of the main reasons palmer amaranth causes such strong damage is its resistance to the herbicide glyphosate. Other members of the *Amaranthus* genus also reduce yields of grain crops, namely *A. retroflexus*, *A. rudis* and *A. tuberculatus* (Steckel & Sprague, 2004; Bensch *et al*, 2003).

Amaranthus species are not only weeds, but are also used by humans for multiple purposes. In Africa and Asia amaranth is used as leafy vegetable, while in the Americas and in Europe it is mainly used as pseudo-cereal. The main species consumed as leafy vegetable is *A. tricolor*, but other species (*A. cruentus*, *A. hybridus* and *A. dubius*) are grown as vegetable crop, too. Species of the *Amaranthus* genus are also used as ornamental plants because of there intensive inflorescence color (Brenner *et al*, 2010).

There are three main grain amaranth species, *A. caudatus*, *A. cruentus* and *A. hypochondriacus*. Together with the two potential ancestors *A. hybridus* and *A. quitensis* these species build the Hybridus complex (Costea & DeMason, 2001, Figure 1.1).

2. Grain Amaranth

The word "amaranth" itself derives from Greek and means "everlasting" (Kauffman & Weber, 1990). Indeed, grain amaranth has a long history of cultivation and consumption. The oldest finding of collected amaranth seeds in a North Argentinian cave dates back to the mid-Holocene (Arreguez *et al*, 2013). In the Aztec Empire, amaranth was of high importance, and tributes payed in amaranth grain were nearly as high as those payed in maize (Sauer, 1967). Since the arrival of the Spanish in South America the cultivation area has strongly declined. Different hypotheses are trying to explain the decline in amaranth production. One of these states that the cultivation of amaranth was forbidden by the Spanish, because of its religious importance to the indigenous population. However, no clear evidence supports this. The fact that the production of maize, which was also of ceremonial importance, did not decline, but was distributed world wide, contradicts this hypothesis. The small seed size of amaranth might have led to a decline of production and replacement by maize. Indeed, maize has larger seeds that make handling simpler than amaranth cultivation (Kauffman & Weber, 1990; Sauer, 1967).

Amaranth cultivation has been preserved mainly in rural places of Mexico and the Andean highlands (Kauffman & Weber, 1990). Nowadays, amaranth is gaining recognition in North America and Europe for its high nutritional value. The protein content of amaranth seeds is with 14.5% higher than that of cereals (maize 9%; rice 7% and rye 13%). Their amino acid composition is favorable for human consumption and animal feed. The grains are high in essential amino acids, e.g. 0.85% lysine (Rastogi & Shukla, 2013). Additionally, amaranth is gluten-free and provides a suitable source of carbohydrates for coeliac and gluten intolerant people (Brenner *et al*, 2010).

Figure 1.1.: Inflorescence of grain amaranth and *A. hybridus*. **A** *A. caudatus*, **B** *A. cruentus*, **C** *A. hypochondriacus* and **D** *A. hybridus*

3. Characterization of Species in the Hybridus Complex

3.1. *A. caudatus*

Amaranthus caudatus is mainly grown in highland areas of South America and Asia (Brenner *et al*, 2010). The seeds of cultivated *A. caudatus* are mostly white, but accessions with dark seeds also exist (Sauer, 1967). One of the main characteristics of *A. caudatus* is its dropping inflorescence, which led to its use as ornamental plant (Figure 1.1A).

3.2. *A. cruentus*

Amaranthus cruentus originated in Southern Mexico or Guatemala and is the most photoperiod insensitive of all three grain amaranths (Sauer, 1967). Historically, white seeded *A. cruentus* landraces were preferred by ancient farmers, but genotypes with dark seeds also exist and might have been generated by hybridization with weedy relatives (Sauer, 1967). The inflorescence of *A. cruentus* is erect and dense (Figure 1.1B). The only commercial amaranth variety in Germany is "Bärnkrafft", a selection from

A. cruentus landraces from Central America.

3.3. *A. hypochondriacus*

Amaranthus hypochondriacus is the predominant grain amaranth cultivated in North America (Brenner *et al*, 2010). Similar to the other two grain amaranths *A. hypochondriacus* genotypes have white seeds, but there are several exceptions with dark seeds. One particular case is a genotype from Southern Mexico that was selected for dark seeds for a special traditional dish (Sauer, 1967).

3.4. *A. hybridus*

Amaranthus hybridus is not cultivated as grain amaranth, but is of high importance for understanding the history of the three grain amaranths, as this species is the potential ancestor of the three grain species. The wide range of distribution and geographic overlap with the grain amaranth species suggest the ancestry (Sauer, 1967; Brenner *et al*, 2010). Unlike the three grain species, *Amaranthus hybridus* has purely dark seeds, displays branching and has smaller and less dense flowers (Figure 1.1D).

3.5. *A. quitensis*

Amaranthus quitensis is native to South America, where it has been growing in close proximity to cultivated *A. caudatus*. It was tolerated and collected by farmers in fields and home gardens, but not cultivated as field crop itself. Its intensive color was used for food coloring (Sauer, 1967). *A. quitensis* has been suggested to be a subspecies of *A. hybridus*, as their growth habit and appearance are very similar (Coons, 1978, 1982). The seeds of *A. quitensis* are dark brown, red or black.

4. Crop Domestication

Over 10,000 years ago humans started to domesticate plants and wild animals, leading to a major change in lifestyle for human communities. The so-called Neolithic Revolution was the transition from hunter-gatherer communities to an agricultural society. Domestication took place in different regions of the world at similar times and gave rise to some of today's most important crops (maize, rice and wheat). Archaeological findings in the Fertile Crescent, Central Asia, Sub-Saharan Africa and Meso and South America reveal these regions as early centers of domestication (Figure 1.2).

While field crops were likely domesticated on large fields cleared by burning or floods of rivers, other crops were likely leftovers from gathered fruits and vegetables. Cleared open areas favor the growth of grasses from seeds that were collected and brought to the camps. Hunter-gatherers followed seasonal routes and returned to similar locations in annual rhythms, finding regrown seeds from the previous years (Doebley *et al*, 2006; Gaut *et al*, 2015). This led to the selection of superior plants and eventually no further collection of wild fruits and seeds was necessary, because desirable fruits were collected and selected each season.

The process of crop domestication can be described as the genetic modification of wild plants that adapts them to human needs (Doebley *et al*, 2006). There is a particular set of morphological and physiological changes - the domestication syndrome - that can be observed in crops with similar uses. These changes include the loss of seed shattering, reduced seed dormancy, determined growth, larger seeds and color variation (Abbo *et al*, 2014; Hake & Ross-Ibarra, 2015, Figure 1.3). The type and extent of domestication syndrome for each crop depends on their life history and use (Meyer *et al*, 2012), although crops from distantly related plant families frequently show similar domestication phenotypes. In maize (*Zea mays*), domestication drastically changed physiological and morphological characteristics of teosinte (*Zea mays* ssp. *parviglumis*) leaving little similarity between the crop and its wild ancestor. The differences between wild species and the cultivated crop species are less pronounced in other cases, e.g. in lettuce and carrot (Iorizzo *et al*, 2013).

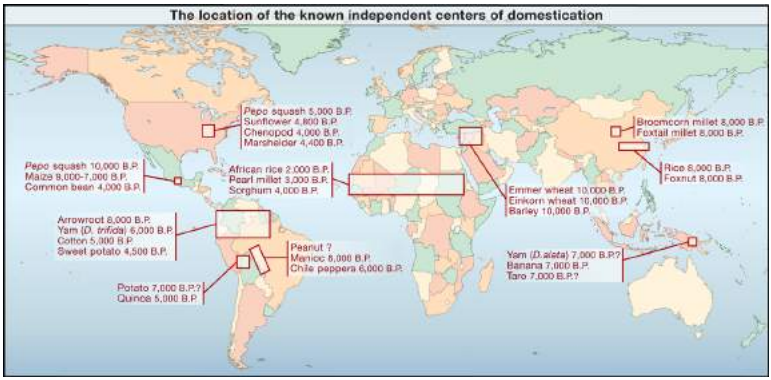

Figure 1.2.: Known centers of domestication around the world (see Doebley *et al* (2006))

The history of crop domestication is tightly linked to human history. Understanding the driving forces and genetic background of crop domestication ca provide precious information on early human societies.

5. Domestication Genomics

Besides the phenotypic changes, the particular history of crop domestication shapes a genome wide signature (Meyer *et al*, 2012). The signature is driven by strong artificial selection, genetic drift and population bottlenecks that populations undergo during the domestication process. An overall decrease of genetic diversity and particular regions with low diversity and high genetic differentiation between wild and domesticated populations are expected for classical domestication scenarios (Doebley *et al*, 2006; Olsen & Wendel, 2013; Sang & Li, 2013; Nabholz *et al*, 2014).

The genome wide pattern of decreased genetic diversity can be modified, because humans distributed crops over long distances. This ge-

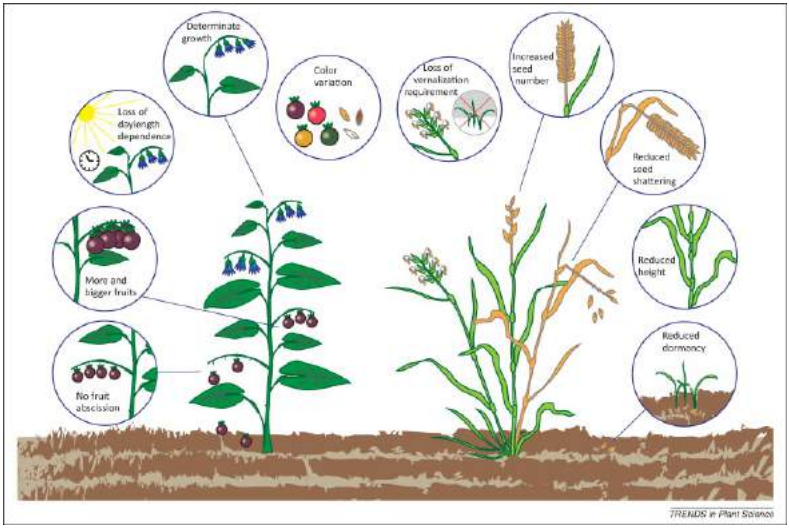

Figure 1.3.: Domestication syndromes of different crops depending on their use (see Lenser & Theißen (2013))

ographic expansion of domesticated crops provided the opportunity for gene flow with new crop wild relatives, which further contributed to genetic differentiation from wild ancestors and increased the genetic diversity in crops (Hake & Ross-Ibarra, 2015). To elucidate the different processes of plant domestication, it is essential to develop and test potential model scenarios of domestication (Beissinger *et al*, 2016). These scenarios are expected to differ between crops, but taken together they can provide a global understanding of the evolutionary processes driving domestication (Gaut *et al*, 2015; Poets *et al*, 2015; Lin *et al*, 2014).

For several crops, individual genes under selection during the domestication process have been identified. The first domestication gene characterized was *teosinte branched1* (*tb1*) which controls the difference in plant architecture between maize and its wild ancestor (Meyer & Purugganan,

2013). Since this first discovery many other domestication genes in several crops have been identified (Meyer *et al*, 2012; Lenser & Theißen, 2013). The identification of the underlying genes of the domestication syndrome is important to adapt crops and use their wild relatives in breeding programs.,

6. Amaranth Domestication

The long history of amaranth cultivation in Central and South America, where other crops (e.g. maize, bean, potato and cotton) were domesticated, should have led to similar morphological changes to those observed in other field crops (Sang & Li, 2013; Lenser & Theißen, 2013, Figure 1.2). However, in cultivated amaranth species (*A. caudatus*, *A. cruentus* and *A. hypochondricaus*) typical domestication traits, such as seed shattering or seed size, are surprisingly indistinct from their wild relatives (Sauer, 1967; Brenner *et al*, 2010). Consequently, the history of amaranth cultivation and the extent of its domestication are still under discussion. Since the first hypotheses proposed by Sauer in 1967, the scenarios for amaranth domestication have been further developed and studied (Maughan *et al*, 2011; Kietlinski *et al*, 2014). For the three grain amaranth species, distinct centers of domestication (Central and South America) have been proposed and three species have been suggested as potential wild ancestors (*A. hybridus*, *A. quitensis* and *A. powelli*; see chapter 3). However, other species cannot yet be excluded, as a comprehensive phylogeny of the genus is missing.

Several possible reasons could explain why typical domestication traits are not fixed in amaranth, despite its long cultivation history. Genetic constraints, a lack of genetic variation or (agri-)cultural reasons might might have prevented a pronounced domestication syndrome (Lenser & Theißen, 2013). The comparison of population structure and genetic diversity between cultivated and wild amaranth can reveal traces of domestication. The systematic evaluation of domestication related traits can indicate if the domestication syndrome is only weakly pronounced or not present at all.

7. Amaranth Breeding

The strong decline of amaranth cultivation in the Americas after the arrival of the Spanish is one reason for the little breeding effort made in amaranth until recently. One difficulty in amaranth breeding is the inflorescence morphology of the monoecious plant. Amaranth is mainly self-pollinating and has numerous intricate flowers, which make crosses more difficult than in other crops (Sauer, 1967). Effective crosses are an important tool for plant research and essential for crop breeding. Crosses on a large scale with little effort is of central importance for the development and production of hybrid crop varieties.

Hybrid varieties are characterized by strongly increased yields in many crops because of heterosis, the superiority of an offspring over its parents (Duvick, 2001). In amaranth a mid-parent heterosis of up to 88% has been reported (Lehmann *et al*, 1991), but hybrid breeding has not yet been established for the crop. For many orphan crop species, like amaranth no efficient crossing methods are available (Veerappan *et al*, 2014). Such methods have to be developed to produce hybrid varieties and generate new genetic variation and to introgress exotic material into breeding populations. Crosses are also important to study the genetic basis of resistances and other important cultivation traits (Moose & Mumm, 2008; Olsen & Wendel, 2013).

The predominant traits of improvement for grain amaranth are related to consumption, crop management and productivity. Taste is one of the traits that could be subject to improvement and a breeding target. While some people describe the taste of amaranth grain as nutty, others find it rather unpleasant. Therefore, weaker taste would be a desirable improvement. As very small seeds cause difficulties during production and processing, increased seed size would simplify the handling and potentially improve seedling development. Synchronous flowering and ripening are also important traits to facilitate the expansion of amaranth production. Other traits to be selected for are similar to those in cereals, yield, pest resistance and tolerance to abiotic stresses (Kauffman & Weber, 1990; Brenner *et al*, 2010). Amaranth improvement could increase its production and use in North America and Europe and help small farmers in developing

countries.

The identification of genetically diverse material and heterotic groups provide possibilities to accelerate the breeding progress in amaranth. The use of genetic markers for QTL mapping, marker assisted breeding and hybrid identification are further applications of genomics in amaranth breeding.

8. Genotyping by Sequencing

While genome wide genetic markers used to be reserved for model species, recent advances in sequencing technologies and genotyping methods provide new tools for non-model organisms (Andrews *et al*, 2016). Genotyping by Sequencing (GBS) is a method based on restriction enzymes and consecutive short read sequencing of the sequence fragments around restriction sites (Elshire *et al*, 2011). The method allows high multiplexing, making GBS a cost efficient method to genotype hundreds of samples. The ligation of specific DNA barcodes allows multiplexing. The barcodes identify sequences of each individual and allow their split up during data analysis (Figure 1.4). High throughput short read sequencing methods allow multiplexing of up to 384 samples on a single sequencing lane with sufficient coverage to call thousands of SNPs (Andrews *et al*, 2016). Library construction is relatively simple and the hands on time is low for GBS. Therefore, it is a useful tool for non-model organisms such as amaranth, for which resources are limited. The choice and number of the restriction enzymes used can control the total coverage of the genome sequence and the coverage per locus. Using rare cutters leads to less restriction sites and higher coverage per read, but less of the genome is covered. On the other hand, common cutters lead to many restriction sites and more of the genome is being sequenced, but with lower coverage per read (Fu *et al*, 2016).

An extension of the traditional GBS protocol is the use of two distinct restriction enzymes. This method uses usually a rare cutter and a common cutter enzyme to achieve equal distribution of fragments that are sequenced (Poland *et al*, 2012b). Adapters are designed in a way that only

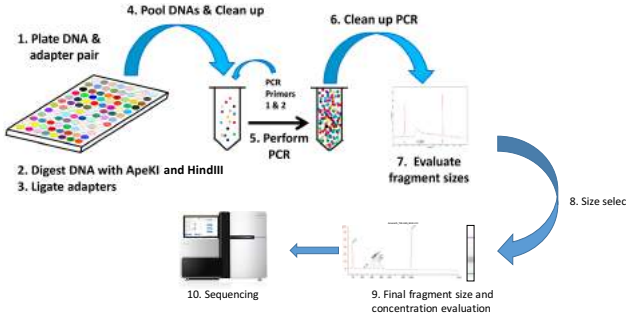

Figure 1.4.: Work flow for laboratory preparation of genotyping by sequencing (adapted from Elshire *et al* (2011))

fragments ending with both restriction sites are amplified and sequenced. An additional size selection step further decreases the number of fragments and therefore improves the sequencing coverage per fragment.

The sequenced reads are aligned to a reference genome, when available, which is often not the case for non-model organisms. Approaches to call SNPs without reference genomes are necessary. The `Stacks` and `UNEAK` pipelines provide methods that construct a reference set of unique reads from the sequenced data. All reads are then aligned to this internal reference, allowing SNP calling without reference genome (Catchen *et al*, 2013; Lu *et al*, 2013).

SNPs resulting from GBS are well-suited for population genetic analysis, because they are distributed over the whole genome. However, GBS data are biased, due to the fact that SNPs in the recognition sequence of the enzyme lead to a loss of the sequence in the concerned individuals (known as 'allele drop-out'). Additionally, there are less restriction sites in the centromere region (Narum *et al*, 2013). Nevertheless, the high number of markers, obtained from GBS, and the genome-wide distribution of them, make GBS a valuable marker system for population genetic studies

in many organisms.

GBS or other high density marker systems have not been used in amaranth. In the present work GBS was used to study the evolutionary history of the genus *Amaranthus* and the domestication syndrome of South American grain amaranth.

9. Objectives

Although the species in the genus *Amaranthus*, and in particular grain amaranth, represent interesting models to study evolutionary processes, little is known about their evolutionary history and the domestication process of grain amaranth. To address these questions, there are three main objectives of my thesis that add basic understanding about the genus and introduce the knowledge into practical amaranth breeding. First, we study the evolutionary history of the whole genus *Amaranthus* and provide a phylogeny of the genus that was missing so far. Second, we investigate the domestication syndrome of South American grain amaranth (*A. caudatus*). Third, we develop controlled growth conditions and evaluate three crossing methods for amaranth that facilitate faster plant regeneration and breeding.

The particular goals of the study were:

- Reconstruct the phylogeny and genome size evolution of the genus *Amaranthus*, using genome wide SNP markers from GBS and a multispecies coalescence method.
- Identify the potential ancestors of grain amaranths
- Investigate the domestication syndrome in South American grain amaranth (*A. caudatus*), by comparing domestication related seed traits (seed size and seed color) of wild amaranth species and cultivated *A. caudatus*
- Reveal the status of genetic diversity and genetic structure of *A. caudatus* and compare it to its potential wild ancestors

- Develop efficient crossing methods for amaranth
- Provide low cost genetic markers for the identification of successful crosses

To achieve these objectives, we used genome wide SNPs derived from GBS and state of the art statistical approaches.

2. Analysis of Phylogenetic Relationships and Genome Size Evolution of the *Amaranthus* Genus Using GBS Indicates the Ancestors of an Ancient Crop

Markus G. Stetter and Karl J. Schmid*
Institute of Plant Breeding, Seed Science and Population Genetics,
University of Hohenheim, Stuttgart, Germany

Accepted for publication in
Molecular Phylogenetics and Evolution
Available as preprint on
BioRxiv
doi: https://doi.org/10.1101/085472

* Corresponding author: karl.schmid@uni-hohenheim.de

1. Abstract

The genus *Amaranthus* consists of 50 to 70 species and harbors several cultivated and weedy species of great economic importance. A small number of suitable traits, phenotypic plasticity, gene flow and hybridization made it difficult to establish the taxonomy and phylogeny of the whole genus despite various studies using molecular markers. We inferred the phylogeny of the *Amaranthus* genus using genotyping by sequencing (GBS) of 94 genebank accessions representing 35 *Amaranthus* species and measured their genome sizes. SNPs were called by *de novo* and reference-based methods, for which we used the distant sugarbeet *Beta vulgaris* and the closely related *Amaranthus hypochondriacus* as references. SNP counts and proportions of missing data differed between methods, but the resulting phylogenetic trees were highly similar. A distance-based neighbor joing tree of individual accessions and a species tree calculated with the multispecies coalescent supported a previous taxonomic classification into three subgenera although the subgenus *A. Acnida* consists of two highly differentiated clades. The analysis of the Hybridus complex within the *A. Amaranthus* subgenus revealed insights on the history of cultivated grain amaranths. The complex includes the three cultivated grain amaranths and their wild relatives and was well separated from other species in the subgenus. Wild and cultivated amaranth accessions did not differentiate according to the species assignment but clustered by their geographic origin from South and Central America. Different geographically separated populations of *Amaranthus hybridus* appear to be the common ancestors of the three cultivated grain species and *A. quitensis* might be additionally be involved in the evolution of South American grain amaranth (*A. caudatus*). We also measured genome sizes of the species and observed little variation with the exception of two lineages that showed evidence for a recent polyploidization. With the exception of two lineages, genome sizes are quite similar and indicate that polyploidization did not play a major role in the history of the genus.

2. Introduction

The *Amaranthus* genus has a world-wide distribution and harbors between 50 and 70 species. The taxonomic differentiation of these species has proven difficult because only few traits are suitable for this purpose despite a high phenotypic diversity. In addition, there is a high level of phenotypic plasticity and a propensity to form interspecific hybrids and hybrid swarms (Brenner *et al*, 2013; Greizerstein & Poggio, 1994; Wassom & Tranel, 2005). Fertile hybrids can be obtained in crosses of distant species from different subgenera (Trucco *et al*, 2005). This disposition for natural hybridization further complicates the taxonomic differentiation of species.

Several species in the genus are of high economic importance and they include grain and vegetable crops as well as invasive weeds (Costea & DeMason, 2001; Sauer, 1967). The three species *A. cruentus*, *A. hypochondriacus* and *A. caudatus* were prehistorically cultivated in North, Central and South America for grain production. Together with their wild relatives *A. hybridus* and *A. quitensis* they form the Hybridus species complex and the latter two species have been suggested as ancestors of the three grain amaranth species, but the domestication history of amaranth is still under debate (Kietlinski *et al*, 2014; Sauer, 1967). *A. tricolor* is cultivated as leaf vegetable in Africa and Asia, in addition to *A. cruentus*, *A. dubius* and *A. hybridus*, which are also used as vegetable crops. Both seeds and leaves are high in micronutrients with a favorable amino acid composition (Rastogi & Shukla, 2013) and are therefore promoted as valuable crops for cultivation outside their native ranges. Appropriate cultivation conditions and protocols for efficient crosses allow to establish breeding programs to achieve this goal by breeding improved varieties of grain amaranths (Stetter *et al*, 2016). Weedy amaranths are the other group of economically and agronomically important species in the genus. The best known is Palmer amaranth (*A. palmeri*) because of its tolerance of the herbicide glyphosate. For example, yield losses in soybean fields due to Palmer amaranth infestation can range from 30 to 70 % (Bensch *et al*, 2003; Davis *et al*, 2015). Other weedy species of the genus include *A. tuberculatus*, *A. rudis* and *A. retroflexus*, which also lead to substantial yield losses in a diversity

of crops (Bensch *et al*, 2003; Steckel & Sprague, 2004). The genus harbors several species that were reported to be resistant against herbicides (www.weedscience.com) and are useful models for studying the evolution of herbicide resistance.

The taxonomy and phylogeny of the genus was investigated using phenotypic traits and genetic markers. The most recent taxonomic revision defined three subgenera that include *Amaranthus Albersia*, *Amaranthus Acnida* and *Amaranthus Amaranthus* (Costea & DeMason, 2001; Mosyakin & Robertson, 1996). Previous studies with different genetic marker systems could not identify a consistent phylogeny of the genus that corresponds with the taxonomic classification (Lanoue *et al*, 1996; Chan & Sun, 1997; Wassom & Tranel, 2005; Das, 2014). Due to the difficulty of differentiating *Amaranthus* species by phenotypic traits, a total number 70 named species may be an overestimate if different populations of the same or closely related subspecies as well as hybrids are classified as different species. Almost 40 species are currently stored in the US (USDA/ARS) and German (IPK Gatersleben) *ex situ* genebanks and are readily available for taxonomic and phylogenetic analyses. A phylogeny of these species based on genome-wide genetic markers has the potential to improve the taxonomic classification and evolution of the whole genus beyond the grain amaranths and their close relatives, which are currently the best studied species (Jimenez *et al*, 2013; Xu & Sun, 2001). The rapid development of sequencing technology allows to utilize genome-wide polymorphisms from different species for phylogenetic analysis. Reduced representation sequencing methods, such as genotyping by sequencing (GBS) can provide thousands of single nucleotide polymorphisms (SNPs) for genetic analysis (Elshire *et al*, 2011; Poland *et al*, 2012b) although for non- model species SNP detection can be challenging without a reference genome. In such species SNPs are identified by using the reference sequence of a different, but closely related species (Maughan *et al*, 2009a), or the *de novo* assembly of sequencing reads (Catchen *et al*, 2011, 2013). Despite these limitations, GBS and related RADseq approaches have been used for phylogenetic analyses of both closely and distantly related taxa (Ariani *et al*, 2016; Eaton & Ree, 2013; Harvey *et al*, 2016; Nicotra *et al*, 2016)

Several software tools were developed for phylogenetic analyses based

on biallelic markers. For example, SNAPP (SNP and AFLP Package for Phylogenetic analysis) infers species trees directly from biallelic markers by implementing a full multispecies coalescent model (Bryant *et al*, 2012). It integrates over all possible trees instead of sampling them explicitly, which results in a high statistical power, but is computationally expensive because it scales with the number of samples and markers (Paul *et al*, 2013).

The availability of a phylogenetic tree for a taxon allows to test hypotheses regarding phenotypic traits or other characters of interest. Species in the genus *Amaranthus* show variation in several traits such as reproductive system (monoecious vs. dioecious) and genome duplication. The latter process is commonly observed in plants and the genus *Amaranthus* is no exception because it is considered to be a paleoallotetraploid with a genome duplication between 18.4 and 34.0 Ma ago (Clouse *et al*, 2016). Haploid chromosome numbers reported for *Amaranthus* species are 16 and 17 (Greizerstein & Poggio, 1994, http://data.kew.org/cvalues), which indicates a cytological stability within the genus although there are several tetraploid species like *A. dubius* and *A. australis*, which likely have a different genome size or structure. Therefore, the variation of genome size within a genus is an interesting trait for analysis in the context of species formation and other phenotypic or ecological traits.

In this study we inferred the phylogeny of the genus *Amaranthus* using molecular markers and analyzed genome size variation to identify putative polyploidization events that may have played a role in speciation or influenced ecological traits. Of particular interest was the relationship of cultivated amaranths with their ancestors because the domestication history is not well understood. A genus-wide phylogeny may identify the ancestors of this ancient crop and allow to consider the evidence in the light of previous domestication models. Furthermore, knowledge of the evolutionary relationship between weedy amaranth species and their relatives allows to investigate the evolution of herbicide resistance. Previously, a diversity of molecular methods were used to infer a phylogeny of the *Amaranthus* genus that include seed proteins, RAPDs, AFLPs and SSRs (Chan & Sun, 1997; Khaing *et al*, 2013; Kietlinski *et al*, 2014). Most of these studies were applied to a subset of the species of the genus and gave inconsistent

results (reviewed by Trucco & Tranel, 2011). In this study, we inferred a molecular phylogeny using a significantly larger number of species than previous studies using thousands of genome-wide markers identified with GBS. To evaluate the robustness of the phylogenetic analysis we compared different SNP calling methods that rely on reference sequences of distant relatives or on a *de novo* assembly of sequenced regions.

3. Material and Methods

3.1. Plant Material

We obtained a total of 94 accessions representing 35 *Amaranthus* species from the USDA/ARS genebank and the German genebank at IPK Gatersleben (Table 2.1). Plants were grown under controlled conditions in standard gardening soil before leaves of young plantlets were collected for DNA and cell extraction. For genome size measurements all accessions were grown in two independent replicates.

Table 2.1.: List of samples included in this study

D	species	accession number	Genebank	Country
1	*A. acanthochiton*	PI 632238 *	USDA/ARS	USA
2	*A. acutilobus*	PI 633579	USDA/ARS	
3	*A. albus*	PI 608029	USDA/ARS	USA
4	*A. arenicola*	PI 667167	USDA/ARS	Mexico
5	*A. asplundii*	PI 604196 *	USDA/ARS	Ecuador
6	*A. blitoides*	PI 649301	USDA/ARS	USA
7	*A. blitum*	PI 490298	USDA/ARS	Kenya
8	*A. blitum*	PI 612860	USDA/ARS	USA
9	*A. californicus*	PI 595319	USDA/ARS	USA
15	*A. caudatus*	PI 511680 *	USDA/ARS	Argentina
26	*A. caudatus*	PI 642741	USDA/ARS	Bolivia
28	*A. caudatus*	PI 649230 †	USDA/ARS	Peru

ID	Species	Accession number	Genebank	Country
31	*A. caudatus*	PI 649235 †	USDA/ARS	Peru
34	*A. caudatus*	PI 511679 * †	USDA/ARS	Argentina
47	*A. caudatus*	PI 649217 †	USDA/ARS	Peru
50	*A. caudatus*	PI 511681 * †	USDA/ARS	Bolivia
51	*A. caudatus*	PI 649228 *	USDA/ARS	Peru
58	*A. caudatus*	PI 608019	USDA/ARS	Ecuador
64	*A. caudatus*	Ames 5302 †	USDA/ARS	Peru
66	*A. crassipes*	PI 649302	USDA/ARS	USA
67	*A. crispus*	PI 633582	USDA/ARS	
68	*A. cruentus*	PI 511714 *	USDA/ARS	Peru
76	*A. cruentus*	PI 667160	USDA/ARS	Guatemala
80	*A. cruentus*	PI 576481	USDA/ARS	Mexico
89	*A. cruentus*	PI 433228 * †	USDA/ARS	Guatemala
91	*A. cruentus*	PI 658728 †	USDA/ARS	Mexico
93	*A. cruentus*	PI 511876	USDA/ARS	Mexico
101	*A. cruentus*	PI 643037 †	USDA/ARS	Mexico
103	*A. deflexus*	PI 667169	USDA/ARS	Argentina
104	*A. dubius*	Ames 25792 *	USDA/ARS	Panama
105	*A. fimbriatus*	PI 605738	USDA/ARS	Mexico
106	*A. floridanus*	PI 553078	USDA/ARS	USA
107	*A. graecizans*	PI 173837	USDA/ARS	India
110	A. hybr.	PI 604571 †	USDA/ARS	Mexico
119	A. hybr.	PI 604564 †	USDA/ARS	Mexico
120	A. hybr.	PI 604566 †	USDA/ARS	Mexico
123	*A. hybridus*	Ames 5232 †	USDA/ARS	Peru
127	*A. hybridus*	PI 636180	USDA/ARS	Colombia
134	*A. hybridus*	PI 667156	USDA/ARS	Ecuador
137	*A. hybridus*	PI 604568 †	USDA/ARS	Mexico
138	*A. hybridus*	PI 604574	USDA/ARS	Mexico
140	*A. hybridus*	Ames 5335 *	USDA/ARS	Bolivia
141	*A. hypochondriacus*	PI 649587	USDA/ARS	Mexico
146	*A. hypochondriacus*	PI 633589	USDA/ARS	Mexico

ID	Species	Accession number	Genebank	Country
158	*A. hypochondriacus*	PI 604595 †	USDA/ARS	Mexico
160	*A. hypochondriacus*	PI 649529	USDA/ARS	Mexico
171	*A. hypochondriacus*	PI 652432	USDA/ARS	Brazil
175	*A. muricatus*	PI 633583	USDA/ARS	Spain
176	*A. palmeri*	PI 633593	USDA/ARS	Mexico
177	*A. polygonoides*	PI 658733	USDA/ARS	USA
178	*A. quitensis*	PI 511747	USDA/ARS	Ecuador
185	*A. quitensis*	PI 652426	USDA/ARS	Brazil
187	*A. quitensis*	PI 652428 †	USDA/ARS	Brazil
189	*A. quitensis*	PI 652422	USDA/ARS	Brazil
192	*A. quitensis*	PI 511736 * †	USDA/ARS	Bolivia
196	*A. quitensis*	Ames 5342	USDA/ARS	Peru
197	*A. retroflexus*	PI 603852	USDA/ARS	USA
198	*A. spinosus*	PI 500237	USDA/ARS	Zambia
199	*A. standleyanus*	PI 605739	USDA/ARS	Argentina
200	*A. tamaulipensis*	PI 642738	USDA/ARS	Cuba
201	*A. tricolor*	PI 603896	USDA/ARS	
202	*A. tuberculatus*	PI 604247	USDA/ARS	USA
203	*A. tuberculatus*	PI 603865	USDA/ARS	USA
204	*A. tuberculatus*	PI 603872	USDA/ARS	USA
206	*A. tuberculatus*	Ames 24593	USDA/ARS	USA
207	*A. viridis*	PI 654388	USDA/ARS	USA
208	*A. wrightii*	PI 632243	USDA/ARS	USA
209	*A. spinosus*	AMA 13	IPK	
210	*A. crispus*	AMA 14	IPK	
211	*A. graecizans*	AMA 24	IPK	
213	*A. lividus*	AMA 49	IPK	
216	*A. graecizans*	AMA 62	IPK	
217	*A. acutilobus*	AMA 63	IPK	
218	*A. albus*	AMA 65	IPK	Canada
219	*A. blitoides*	AMA 66	IPK	
221	*A. deflexus*	AMA 76	IPK	

ID	Species	Accession number	Genebank	Country
222	*A. viridis*	AMA 79	IPK	Peru
223	*A. dubius*	AMA 80	IPK	Rwanda
224	*A. lividus*	AMA 87	IPK	Rwanda
225	*A. powellii*	AMA 89	IPK	Rwanda
226	*A. retroflexus*	AMA 93	IPK	Mexico
227	*A. muricatus*	AMA 95	IPK	
228	*A. albus*	AMA 96	IPK	
229	*A. deflexus*	AMA 97	IPK	
233	*A. tricolor*	AMA 149	IPK	
235	A. hybr.	AMA 147 †	IPK	
238	*A. retroflexus*	AMA 105	IPK	China
240	*A. tricolor*	AMA 126	IPK	Cuba
242	*A. dubius*	AMA 140	IPK	Spain
243	*A. viridis*	AMA 175	IPK	
244	*A. powellii*	AMA 170	IPK	Germany
357	*A. tucsonensis*	PI 664490	IPK	USA
360	*A. australis*	PI 553076	IPK	USA
361	*A. australis*	PI 553077	IPK	USA

* Accessions not included in genome size measurements

† Accessions not included in SNAPP analysis

3.2. DNA Extraction and Sequencing

Genomic DNA was extracted with the Genomic Micro AX Blood Gravity kit (A&A Biotechnology, Poland) using CTAB extraction buffer for cell lysis (Saghai-Maroof *et al*, 1984). Double-digest genotyping by sequencing libraries (GBS) were constructed as described previously (Stetter *et al*, 2015). For each accession two samples with different barcodes were prepared to assure sufficient sequencing output per accession. Fragment sizes between 250 and 350 bp were selected with BluePippin (Sage Science, USA) and the resulting libraries were single-end sequenced to 100 bp on one lane of a Illumina HiSeq 2500 (Eurofins Genomics GmbH, Germany).

3.3. Data Preparation and Filtering

Raw sequence data were processed with a custom GBS analysis pipeline. First, reads were sorted into separate files according to their barcodes using Python scripts. Subsequently, read quality was assessed with fastQC (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`). Due to lower read quality towards the end of reads, they were trimmed to 90 bp. Low quality reads were excluded if they contained at least one N (undefined base) or if the quality score after trimming was below 20 in more than 10% of the bases. Replicated data per accession were combined and subsequently analyzed as one sample.

3.4. *De novo* and Reference-Based SNP Discovery

We used two different methods to call SNPs from the sequencing data, a *de novo* approach using Stacks 1.35 and an alignment to a reference genome. For the *de novo* approach we used the denovo_map.pl pipeline provided by Stacks to call SNPs directly from the processed data (Catchen *et al*, 2011, 2013). Highly repetitive GBS reads were removed in the ustacks program with option `-t`. Additionally, we analyzed data with two different minimum number of identical raw reads (`m` = 3 and `m` = 7) required to create a stack. These two settings resulted in different results in the SNP calling (Mastretta-Yanes *et al*, 2015) and we therefore used both settings for

comparison. Two mismatches were allowed between loci when processing a single individual, and four mismatches between loci when building the catalog, which is the set of non redundant loci based on all accessions and is used as reference for SNP calling. SNPs were called with the Stacks tool populations 1.35 with filtering for different levels of missing values.

In addition to the *de novo* approach we used the sugar beet (*Beta vulgaris*) RefBeet-1.2 (Dohm *et al*, 2014) and the *Amaranthus hypochondriacus* draft genome (Clouse *et al*, 2016) as reference genomes to align sequence reads with bwa mem (Li & Durbin, 2009). SNPs were called with samtools 1.2 (Li *et al*, 2009). The resulting SNPs were filtered for different levels of missing values at a locus with vcftools (Danecek *et al*, 2011).

Table 2.2.: Summary of four GBS datasets obtained by different SNP calling methods and parameters.

Name	Reference map	Tool	Mapped reads	SNPs	Missing (%)
refmap_hyp	Ahypochondriacus_1_0	BWA, Samtools	166,935,845 (74.8%)	2,978	5.2
refmap_beet	RefBeet-1_2	BWA, Samtools	57,766,877 (25.9%)	1,439	31.7
stacks_m3	*de novo* catalog	Stacks	223,104,991 (100.0%)	2,181	0.6
stacks_m7	*de novo* catalog	Stacks	223,104,991 (100.0%)	3,416	0.6

3.5. Phylogenetic Analysis Methods

We constructed a neighbor joining tree with 1000 bootstraps from the pairwise Euclidean distance between all 94 individuals based the four datasets using the R package ape (Paradis *et al*, 2004) and calculated an uncorrected neighbor joining network using the NeighborNet algorithm (Bryant & Moulton, 2004) with SplitsTree4 (Huson & Bryant, 2006).

We also used the multi-species coalescent implemented in SNAPP, which is part of the BEAST package, to infer species trees directly from unlinked biallelic markers (Bouckaert *et al*, 2014; Bryant *et al*, 2012). We reduced the number of individuals to a maximum of four per species be-

cause the SNAPP algorithm is computationally expensive. Additionally, we imputed the reference-map based datasets with beagle (Browning & Browning, 2016) before thinning all four datasets with vcftools (Danecek *et al*, 2011) to a distance of 100 bp which excludes multiple SNPs per GBS read. Since GBS loci are essentially randomly distributed throughout genome, we assumed that the assumption of unlinked biallelic markers was fulfilled after this filtering step. VCF files were converted to nexus format using a Python script and BEAST input files were created from these using BEAUti (Bouckaert *et al*, 2014). Mutation rates were calculated with BEAUti and default parameters were used for SNAPP. We conducted ten runs per dataset. Log files were analyzed with tracer 1.6 to examine convergence and converging log and tree files were combined using LogCombiner with 15% burn-in. The effective sample size (ESS) was adequate (> 200) for the important parameters but was lower for some θ values. We proceeded with the analysis as the low θ values should not influence the tree topology (Nicotra *et al*, 2016). TreeAnnotator was used to construct the 'Maximum clade credibility' tree and annotate it with posterior probabilities.

3.6. Genome Size Measurements and Phylogenetic Analysis

The genome sizes of 84 accessions representing 34 species were measured with flow cytometry and two independent replicates for each accession (Table 2.1). The tomato cultivar *Solanum lycopersicum* cv Stupicke was used as internal standard, due to its comparable genome size (DNA content = 1.96 pg; Dolezel *et al*, 1992). For the measurement, fresh leaves were cut up with a razor blade and cells were extracted with CyStain PI Absolute P (Partec, Muenster/Germany). Approximately 0.5 cm^2 of the sampled leaf was extracted together with a similar area of the tomato leaf in 0.5 ml of extraction buffer. The DNA content was determined with CyFlow Space (Partec, Muenster/Germany) flow cytometer and analyzed with FlowMax software (Partec, Muenster/Germany). For each sample, 10,000 particles

were measured. The DNA content was calculated as:

$$\text{DNA content 2C [pg]} = \text{genome size tomato} \times \frac{\text{fluorescence amaranth}}{\text{fluorescence tomato}} \tag{2.1}$$

and the genome size (in Mbp) was calculated as:

$$\text{genome size 1C [Mbp]} = (0.978 \times 10^3) \times \frac{\text{DNA content 2C [pg]}}{2} \tag{2.2}$$

The conversion from pg to bp was calculated with 1 pg DNA $= 0.978 \times 10^9$ bp (Dolezel *et al*, 2003). Means were calculated using R software (R Core Team, 2014) and an ANOVA was performed to infer differences in genome size for the species.

We combined the genomic data with the genome size measurements to study the genome size evolution. The 1 C genome sizes (Mbp) were mapped on the phylogeny using parsimony reconstruction in Mesquite 3.04 (http://mesquiteproject.org). In addition we used the fastAnc function from the phytools R package to conduct a Maximum Likelihood reconstruction of ancestral states (genome sizes) with default parameters (Revell, 2012). For this analysis we inferred the genome size of *A. acanthochiton* as the mean between its two closest related species (*A. blitum* and *A. lividus*) because fastAnc does not allow missing values. A Brownian motion model implemented in the fastBM function in phytools (Revell, 2012) was used to simulate the random evolution of genome size over the tree. 1000 simulations were run and by using the distribution of genome sizes for each branch in the phylogeny the 0.25% and 97.5% were used to conduct a two-tailed test whether observed genome sizes were significantly smaller or larger than simulated sizes.

3.7. Data availability

Sequence reads were submitted to the European Nucleic Archive (ENA) under accession number PRJEB18745. Analysis scripts, aggregated sequencing data and genome size raw data are available under Dryad (http://datadryad.org/) DOI:http://dx.doi.org/10.5061/dryad.1bv83.

4. Results

4.1. SNP Discovery

Until reference genomes for any species can be created on a routine basis, methods like genotyping by sequencing (GBS) are an efficient method to survey genome-wide diversity in non-model species. To compare the use of GBS with and without a reference sequence for phylogenetic reconstruction of the *Amaranthus* genus, we used different methods and reference sequences for SNP calling. The number of aligned reads differed strongly between the *Beta vulgaris* and *Amaranthus hypochondriacus* references. Only 25.9% of the reads aligned to sugar beet and 74.8% to *A. hypochondriacus* (Table 2.2), which resulted in different SNP numbers. We identified 23,128 SNPs with the sugar beet and 264,176 SNPs with the *A. hypochondriacus* reference genomes. GBS data have a high proportion of missing values and the number of SNPs retained depends on the allowed proportion of missing values per SNP (Figure 2.1). For example, if no missing values are allowed only one SNP remained with the sugar beet and 247 SNPs with the *A. hypochondriacus* reference.

The *de novo* assembly with Stacks allowed us to use all reads for SNP detection at the cost that resulting contigs are unsorted and without position information on a reference genome. The minimum number of identical raw reads required to create a stack influences the SNP detection (Mastretta-Yanes *et al*, 2015). With a minimum number of three reads ($m = 3$) we obtained 505,981, and with seven reads ($m = 7$) 371,690 SNPs. After filtering out loci with missing values, $m = 3$ retained 949 and $m = 7$ retained 1,605 SNPs. The total number of SNPs recovered was higher for $m = 3$, but the number of SNPs without missing values was higher for $m = 7$. The two parameter values ($m = 3$ and $m = 7$) resulted in the same number of SNPs if a proportion of 20 to 30 % missing values per site were allowed. With both parameter values the *de novo* approach resulted in more SNPs before filtering than the reference-based SNP datasets (Figure 2.1). We were able to retain a large number of SNPs if missing data in one individual per GBS locus were allowed, which corresponds to a cutoff of 2% missing values (Figure 2.1). For the phylo-

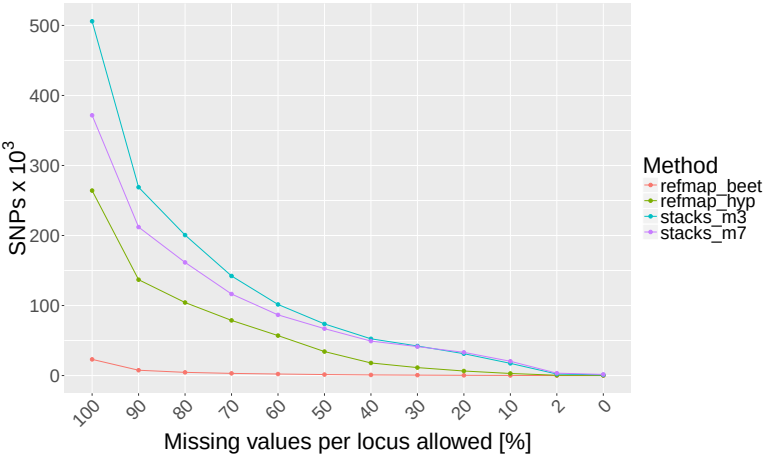

Figure 2.1.: Number of SNPs recovered at different levels of missing values allowed per locus. Data sets are labeled as follows: refmap_beet, reference mapping against sugar beet; refmap_hyp, reference mapping against *Amaranthus hypochondriacus*; stacks_m3, *de novo* assembly with Stacks using parameter value `m = 3` for minimal read coverage and stacks_m7, parameter value `m = 7`.

genetic analysis of the reference-based datasets we allowed 10% (sugar beet reference) and 50% missing values (*A. hypochondriacus* reference). The resulting total number of missing values ranged from 0.6% for the *de novo* to 31.7% for the dataset based on the sugarbeet reference (Table 2.2). For the consecutive analyses we used all four datasets but in the following we present only the results obtained with the SNP data from the mapping against the *A. hypochondriacus* reference and include the other results as supplementary information because the results from all four data sets are very similar.

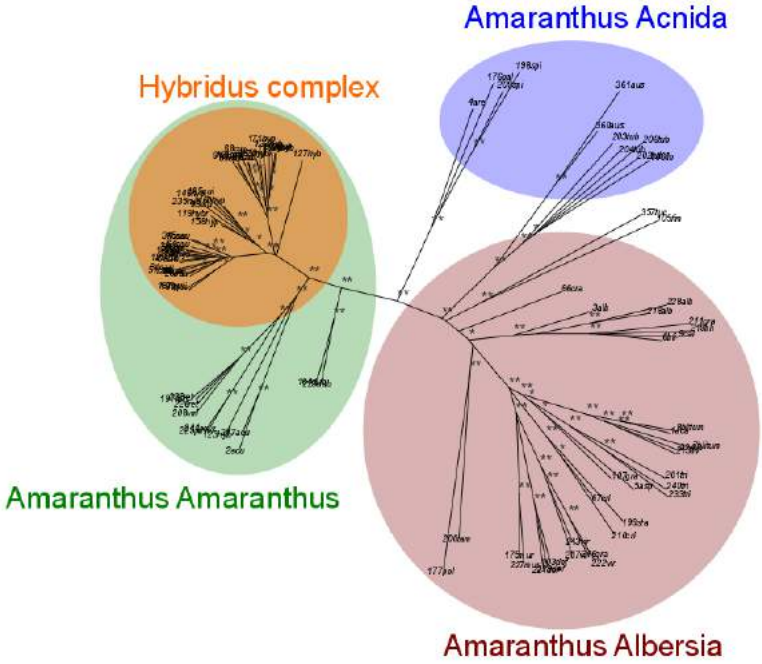

Figure 2.2.: Neighbor joining tree calculated from the Euclidean distances of 94 individuals representing 35 *Amaranthus* species. Single stars (*) indicate bootstrap values over 90% and double stars (**) indicate bootstrap values of 100%.

4.2. Phylogenetic Inference

Neighbor Joining Phylogeny

The neighbor joining phylogeny based on Euclidean distances of allelic states shows that most accessions cluster with other accessions from the same species (Figure 2.2). Within the Hybridus complex, however, there

is no strong separation of the species into different clusters. Based on the species names, four clades are expected, but only three are observed. The first consists of *A. caudatus*, *A. quitensis* and *A. hybridus* that all originated from South America. The second clade includes *A. cruentus*, *A. hypochondriacus*, *A. hybridus*, which originated from Mexico, one *A. quitensis* accessions from Brazil and two hybrid accessions likely formed from species of the Hybridus complex. The third clade is formed by *A. cruentus*, *A. hypochondriacus* and *A. hybridus*, as well as two hybrids, and one *A. dubius* individual (242_dub; Figure 2.3). The accessions in this clade originated from Mexico, with the exception of two accessions of *A. cruentus* from Guatemala and one from Peru, and one *A. hypochondriacus* accession from Brazil. The NeighborNet network confirms this pattern and in addition outlines the extent of conflicting phylogenetic signals among accessions that may reflect gene flow or hybridization (Figure 2.3). The three accessions of the leaf vegetable amaranth *A. tricolor* cluster closely and form a clade with other *Amaranthus* species (Figure 2.2).

Although the ability to resolve species level relationships seems to be limited with our data, the neighbor joining tree is consistent with the taxonomic classification into three subgenera that was previously defined using morphological traits (Figures 2.2 and S1). The phylogenies resulting from the four different SNP calling methods are highly similar and show that the tree topology of the genus is highly robust with respect to the SNP calling method (Figure S2).

Phylogeny Based on the Multispecies Coalescent

For inferring the phylogeny with the multispecies coalescent implemented in the SNAPP program we used a subset of individuals for two reasons. First, there were more individuals of the species from the Hybridus complex than of the other species which may bias the analysis, and second because the computation time scales exponentially with the number of individuals. Therefore we randomly sampled four individuals in those species with more than four genotyped accessions. The combined chain length without burn-in was 3,980,000 for the SNP data based on the *A. hypochondriacus* reference. The cloudogram derived from the SNAPP

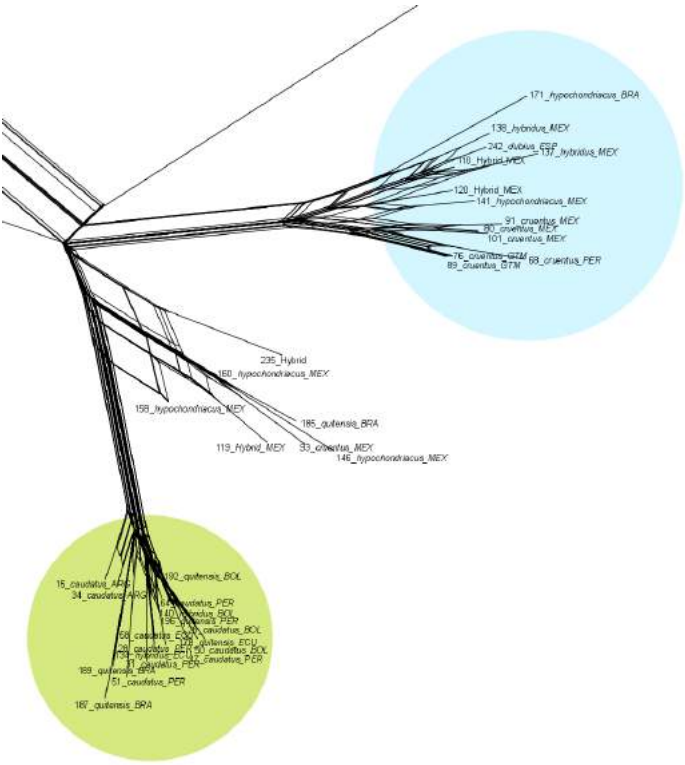

Figure 2.3.: Section of the NeighborNet network showing the Hybridus complex. The blue circle includes the Central American grain amaranths (*A. hypochondriacus*, *A. cruentus*) and the close wild relative *A. hybridus*. The green circle includes South American grain amaranth (*A. caudatus* and the potential ancestors (*A. hybridus* and *A. quitensis*). The country of origin according to the genebank passport information is shown at the end of the name of each accession. The whole network is shown in supplementary figure A.1.

analysis allows to identify the degree of uncertainty for several clades in the tree (Figure 2.4). For the group of species that include *A. tricolor* and *A. crispus* there was a high uncertainty between the species. Within the Hybridus complex the uncertainty was high among the cultivated *A. caudatus* and its putative wild ancestors *A. quitensis* and *A. hybridus*. In contrast, the split between these three South American species and the Central American species *A. cruentus* and *A. hypochondriacus* was strongly supported. Overall, the Hybridus complex is well separated from the other species (Figure 2.4 and 2.5).

4.3. Genome Size Evolution

The genome size measurements differed among the *Amaranthus* species although the range of variation was quite narrow (Table 3.4). Palmer amaranth has the smallest genome with a size of 421 Mbp, and *A. australis* the largest genome of 824 Mbp, which about twice the size of Palmer amaranth. Most species including the Hybridus complex had a genome size close to 500 Mbp (Table 3.4).

To test whether changes in genome sizes in the phylogeny reflect random evolution or non-neutral processes, we mapped the genome sizes to the phylogenetic tree obtained with SNAPP (Figures 2.5 and S3). There was a tendency for decreasing genome sizes within the *Amaranthus* subgenus, and a high variation of genome sizes within the *Acnida* subgenus because it included both the individuals with the smallest and largest genome sizes. Figure 2.5 further shows that *A. dubius* has a larger genome than the other species of the *Amaranthus Amaranthus* subgenus. Even though there were significant differences in genome size between species, the ancestral branches have wide confidence intervals and significantly differ in recent splits but not in early ones (Figures S4 and S5). The ancestral genome size was inferred by fastAnc as 569 Mbp, but with a large confidence interval of 416 Mbp to 722 Mbp that includes almost all empirical genome size measurements of the extant species. Using a Brownian motion model we tested whether genome sizes differed in individual branches of the phylogeny given the complete tree. Several branches in the tree differ from such a random process. The lineage leading to *A. tricolor* and *A. australis*

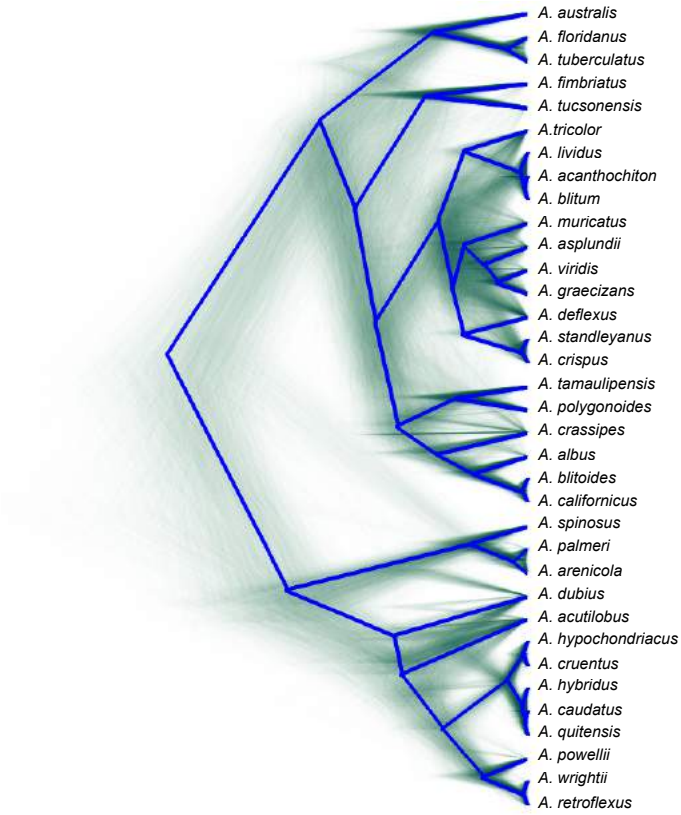

Figure 2.4.: Species tree of *Amaranthus* based on the multispecies coalescent calculated with SNAPP. The cloudogram (green lines) represents 3980 individual trees and the consensus tree is shown in blue color.

show significantly larger genome sizes that suggest that polyploidiziation likely influenced the genome sizes of these species. In contrast, the lin-

Table 2.3.: [
Genome sizes of *Amaranthus* species using flow cytometry.]Estimated genome size of *Amaranthus* species. n is the number of genotypes sampled per species.

species	n	Size (Mbp)	Standard Error	Lower CI	Upper CI
A. acutilobus	3	532.5	34.3	463.8	601.2
A. albus	3	530.3	33.4	463.2	597.3
A. arenicola	1	438.6	57.1	323.9	553.3
A. asplundii	1	535.0	57.1	420.2	649.7
A. australis	2	824.2	44.4	735.7	912.8
A. blitoides	3	521.9	33.4	454.8	588.9
A. blitum	2	748.8	40.6	667.2	830.4
A. californicus	1	547.9	57.1	433.2	662.6
A. caudatus	6	502.0	24.0	453.6	550.4
A. crassipes	1	512.5	62.4	388.1	637.0
A. crispus	2	576.0	40.6	494.4	657.6
A. cruentus	5	510.9	26.1	458.3	563.6
A. deflexus	3	640.2	33.4	573.1	707.2
A. dubius	2	711.9	40.6	630.3	793.5
A. fimbriatus	1	527.2	57.1	412.5	641.9
A. floridanus	1	658.2	57.1	543.5	772.9
A. graecizans	3	541.0	33.4	473.9	608.0
A. hybr.	3	508.0	33.4	440.9	575.0
A. hybridus	5	503.8	26.1	451.1	556.4
A. hybridus x *A. hypochondriacus*	1	523.8	57.1	409.1	638.5
A. hypochondriacus	5	506.4	26.1	453.7	559.0
A. lividus	2	685.8	40.6	604.2	767.4
A. muricatus	2	729.6	40.6	648.0	811.2
A. palmeri	1	421.8	57.1	307.1	536.5
A. polygonoides	1	512.3	57.1	397.6	627.0
A. powellii	2	512.3	40.6	430.7	593.9
A. quitensis	4	501.1	29.6	441.5	560.6
A. retroflexus	3	555.6	33.4	488.6	622.7
A. spinosus	2	471.6	40.6	390.0	553.2
A. standleyanus	1	502.9	57.1	388.2	617.6
A. tamaulipensis	1	524.9	57.1	410.2	639.6
A. tricolor	3	782.7	33.4	715.7	849.8
A. tuberculatus	4	675.6	27.0	621.4	729.8
A. tucsonensis	1	510.4	57.1	395.7	625.1
A. viridis	3	543.1	33.4	476.1	610.2
A. wrightii	1	534.3	57.1	419.6	649.0

eage leading to the weed *A. palmeri* has a significantly smaller genome size. The two clades of the *A. Acnida* subgenus consist of three species each. They are not only strongly separated according to the molecular phylogeny but also show different average genome sizes.

5. Discussion

5.1. Reference-Based Versus Reference-Free SNP Calling

Genotyping by Sequencing (GBS) identifies thousands of markers but usually requires a reference sequence for mapping sequence reads. *De novo* methods allow to call SNPs without a reference genome. We compared both approaches to determine their efficiency in SNP identification. With the distant sugar beet genome as a reference only 26% of the sequencing reads could be used for SNP calling because the sequence divergence between sugar beet and *Amaranthus* species is too high for an efficient mapping despite the high synteny between *Amaranthus* and sugar beet (Clouse *et al*, 2016). This resulted in a small number of SNPs available for phylogenetic analysis. In contrast, the *de novo* assembly used all data and the number of SNPs obtained was even larger than from the mapping against the *A. hypochondriacus* genome. The proportion of missing data was also highest with the evolutionary distant sugar beet reference genome. Comparisons of different values for the number of identical reads (`-m` parameter) in Stacks showed that a smaller number of identical reads produced more SNPs, but we obtained more SNPs without missing values when requiring a larger number of identical reads, in accordance to earlier studies (Mastretta-Yanes *et al*, 2015). A reference genome from the same or a closely related species combines the advantage of a larger SNP number with linkage information (Andrews *et al*, 2016). Since the level of evolutionary divergence within the genus is unknown and only one reference sequence from an amaranth species was available, we compared the different approaches. Taken together, a comparison of the four SNP calling approaches with different numbers of SNPs and different levels of

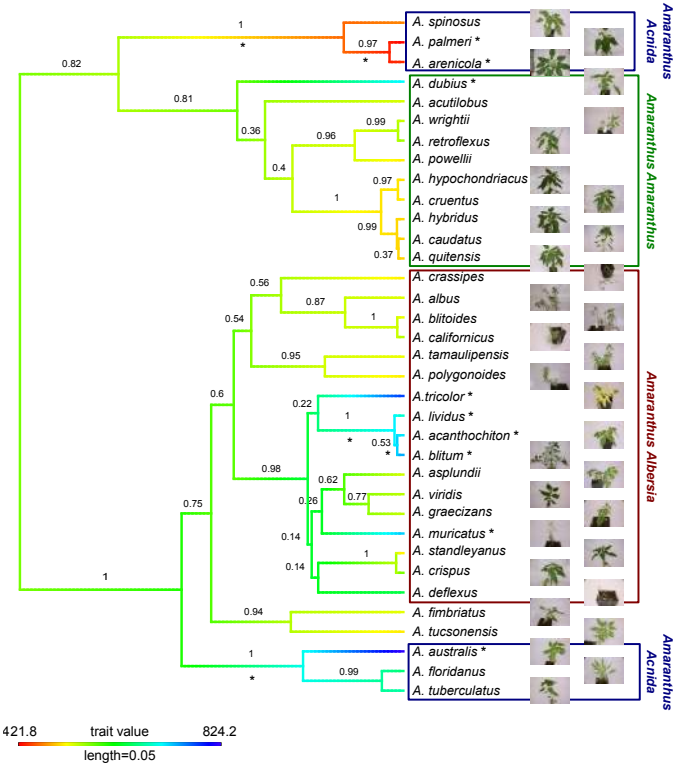

Figure 2.5.: Genome size evolution mapped onto consensus tree obtained with SNAPP. The branch labels show posterior probabilities of genome size estimates of interior nodes obtained with a Maximum Likelihood method implemented in the fastAnc function of the phytools R package. Branch colors show estimated genome sizes in Mbp. Stars (*) indicate deviation from random evolution of genome size at 95% confidence level based on a two-tailed test. Group labels annotate taxonomic subgenera.

missing data showed that the resulting neighbor joining tree of the genus was quite robust with respect to SNP calling parameters, because it did not differ strongly between datasets (Figure S1). A major disadvantage of the *de novo* approach is that information about physical map positions of SNPs is missing and it can not be tested whether SNPs are unlinked. To increase the chance that SNPs are unlinked, which is a requirement of the SNAPP algorithm, we used a double-digest protocol for GBS and filtered for one SNP per GBS locus, which should allow the reconstruction of the phylogeny using the multispecies coalescent method (Andrews *et al*, 2016; Bryant *et al*, 2012; DaCosta & Sorenson, 2016). Such an approach was shown to be suitable for the phylogentic reconstruction of Australian *Pelargonium* using RADseq data (Nicotra *et al*, 2016).

5.2. Phylogeny of the Whole *Amaranthus* Genus

The species-rich genus *Amaranthus* has been divided into the three subgenera, *Amaranthus*, *Acnida* and *Albersia*. Several studies investigated species relationships in the genus using molecular markers, but most included only few species and did not allow conclusions for the whole genus (Chan & Sun, 1997; Lanoue *et al*, 1996; Kietlinski *et al*, 2014; Xu & Sun, 2001). We included all species that are currently available as *ex situ* conserved germplasm and genotyped several accessions per species to evaluate their evolutionary relationship (Figure 2.2). As expected, most accessions from the same species clustered together, and the subdivision of the genus into three subgenera based on phenotypic traits is largely consistent with our molecular data, although we observed some notable exceptions which we discuss below.

The species tree obtained with SNAPP largely reflects the neighbor joining tree which is based on individual accessions, but the cloudogram of all sampled species trees indicates uncertainties in the positioning of species like *A. deflexus*, *A. tricolor* and *A. crispus* in the tree topology (Figure 2.4). In contrast, a clustering of the genus into four basal clades is strongly supported (Figures 2.4 and 2.5). We compared our phylogeny with the published taxonomy of the *Amaranthus* genus (Mosyakin & Robertson, 1996). The subgenera *Amaranthus Amaranthus* and *A. Albersia* show a

clear split at the root of the tree, but *A. Acnida* is split into two separate clades (Figure 2.5). The species of *A. Acnida* were categorized as dioecious and grouped based on this trait (Mosyakin & Robertson, 1996) although *A. palmeri* and *A. tuberculatus* were later described to be phylogenetically divergent (Wassom & Tranel, 2005). Another explanation for the observed split of *A. Acnida* species into two major groups may reflect the polyploid genomes of *A. tuberculatus*, *A. floridanus* and *A. australis* (see below). In our analysis, we treated all species as diploid and allowed only biallelic SNPs but polyploids may be characterized by high levels of heterozygosity (Figure S6) and harbor multiallelic SNPs, which are excluded from further analysis. Both factors may bias a phylogenetic inference. On the other hand, a high proportion of heterozygous loci would result in grouping the polyploid species in the same main branch as their ancestors or closest relatives. We conclude, however, that their grouping is correct because the posterior probabilities for the placement of these species in the phylogeny are very high.

5.3. Phylogenetic Analysis of the Hybridus Complex

The Hybridus complex contains the domesticated grain amaranths and putative ancestors such as *A. hybridus*. Previous studies suggested that the Hybridus complex comprises two clades (Adhikary & Pratt, 2015). We also identified the two clades, and in addition a third clade, which appears to be an intermediate of the other two other. It includes accessions from different species from Hybridus complex plus accessions that were labeled as 'hybrids' in the passport data and may have originated from interspecific hybridization. Interestingly, *A. hybridus* and *A. quitensis* accessions occur in all three clades (Figure 2.2), which may be explained by the geographic origin and geographic differentiation of these species. We previously suggested that *A. quitensis*, which is endemic to South America, and *A. hybridus* populations from the same region are a single species with a strong differentiation of geographically separated subpopulations within South America (Stetter *et al*, 2015). Since such a taxonomic grouping is

still under debate and *A. quitensis* might be a separate subspecies of *A. hybridus*, we treated them as separate species in the phylogenetic analysis as was done before (Coons, 1978, 1982; Kietlinski *et al*, 2014). A comparison of the position of individual *A. hybridus* and *A. quitensis* accessions in the neighbor joining tree with the species tree obtained with SNAPP showed that in the former, the two species are not strongly differentiated from each other (Figure 2.2) whereas they form independent lineages in the species tree, but are closely related and in a monophyletic group with the three grain amaranths (Figure 2.5). This may be explained by the fact that SNAPP uses pre-defined groups which forces the algorithm to separate the species and therefore does not allow to evaluate whether *A. quitensis* can be considered as a separate species or is a subspecies with a high level of admixture.

The taxonomic interpretation of species relationships in the Hybridus complex is further complicated by the geographic origin of the accessions used in this study and by the effects of domestication. Sauer (1967) suggested that both *A. hybridus* and *A. quitensis* may have been involved in the domestication of the grain amaranths. Our analysis is consistent with this notion because the three grain amaranths *A. caudatus*, *A. cruentus* and *A. hypochondriacus* and their wild relatives *A. hybridus* and *A. quitensis* are separated from the other amaranths. The species tree suggests that both wild species are more closely related to the South American *A. caudatus* than to the Central American *A. cruentus* and *A. hypochondriacus*, but the neighbor joining tree of individual accessions splits *A. hybridus* accessions by their geographic origin and clusters *A. hybridus* accessions collected in South America with the South American *A. caudatus* and *A. hybridus* accessions collected in Central America with *A. cruentus* and *A. hypochondriacus*, which also are native to Central America (Figure 2.3).

Most evidence published so far suggests that *A. hybridus* is the direct ancestor of all three cultivated grain amaranth species (Chan & Sun, 1997; Kietlinski *et al*, 2014; Park *et al*, 2014; Stetter *et al*, 2015). *A. quitensis* is closely related to *A. caudatus* (Park *et al*, 2014; Xu & Sun, 2001; Stetter *et al*, 2015) and a low support of the split between *A. caudatus* and *A. quitensis* (Figures 2.4 and 2.5) reflects gene flow (Stetter *et al*, 2015) or indicates that *A. quitensis* is an intermediate between the wild *A. hybridus*

and cultivated *A. caudatus* because it grows as weed in close proximity to grain amaranth fields and could have hybridized with *A. caudatus*. Another species for which a role in the domestication of grain amaranth was postulated is *A. powelli* (Sauer, 1967). In our analysis, as well as in a previous study, *A. powelli* is not closely related to the cultivated grain amaranths (Mallory *et al*, 2008) and therefore less likely a direct ancestor of *A. hypochondriacus* as proposed before (Park *et al*, 2014; Sauer, 1967; Xu & Sun, 2001).

Taken together, our analysis of the Hybridus complex is consistent with previous molecular phylogenies (Chan & Sun, 1997; Khaing *et al*, 2013) but we note that the GBS-based phylogenies show a weaker genetic differentation between the different species of the complex. In addition, both *A. caudatus* and *A. hypochondriacus* are more closely related to *A. hybridus* than to each other, which was observed before (Chan & Sun, 1997; Kietlinski *et al*, 2014). The *A. hybridus* accessions show a strong split along the North-South gradient (i.e., Central vs. South America), which supports the hypothesis that two different *A. hybridus* lineages were the ancestors of the three grain amaranths with a possible contribution of *A. quitensis* in the domestication of *A. caudatus* (Trucco & Tranel, 2011; Kietlinski *et al*, 2014; Adhikary & Pratt, 2015). Such a strong geographic pattern shows that in future studies requires a comprehensive geographic sampling to understand the evolutionary history of these species. Similar to the Hybridus complex other species of the genus should be sampled in greater detail to identify duplicated naming and issues with species delimitation. Species that are not yet included in germplasm collections should be collected and included in studies to determine the actual number of species in the genus.

5.4. Genome Size Evolution

The *Amaranthus* genes has undergone a whole genome duplication before speciation which was then followed by further duplication, chromosome loss and fusion events (Behera & Patnaik, 1982; Clouse *et al*, 2016). The mapping of genome size measurements onto the phylogeny revealed that the subgenus *Amaranthus* has a tendency towards smaller genomes,

whereas species in the *Albersia* clade show increased genome sizes (Figure 2.5). These patterns are not strong and uniform within groups, however, because *A. dubius* has a larger genome size than expected for the clade. It may result from a genome duplication and a subsequent speciation of *A. dubius*, which is tetraploid (Behera & Patnaik, 1982). The genome size of *A. dubius* is not exactly twice the size of closely related species and indicates a loss of DNA after duplication. A similar pattern was observed in the genus Chenopodium which also belongs to the *Amaranthaceae* (Kolano *et al*, 2016).

Chromosome numbers in the Hybridus complex species are variable. *A. cruentus* has 17, and the other species 16 chromosomes (Greizerstein & Poggio, 1994), although it does not seem to strongly influence genome sizes (Greizerstein & Poggio, 1994; Stetter *et al*, 2015, Table 3.4). For some species we observed a strong deviation in genome sizes from previously reported values. The genome sizes of *A. caudatus*, *A. cruentus* and *A. hypochondriacus* are within the previously reported range of 465 to 611 Mbp, but the genome sizes of *A. retroflexus*, *A. spinosus* and *A. tricolor* were about 200 Mb smaller than previous values. We also found that the five species of the Hybridus complex have similar genome sizes whereas previous measures from these species strongly differ from each other (Bennett & Smith, 1991; Bennett *et al*, 1998; Ohri *et al*, 1981, http://data.kew.org/cvalues). A strong variation in genome size was also observed in the dioecious *A. Acnida* subgenus. Previous molecular studies separated two members of this taxonomically defined subgenus *A. palmeri* and *A. tuberculatus* into different groups (Lanoue *et al*, 1996; Wassom & Tranel, 2005) and our phylogenetic analysis grouped the six species into two strongly separated clades of three species each, which differ by their average genome sizes. The genome size of *A. australis* is twice the size of *A. palmeri* and may result from a whole genome duplication (Mosyakin & Robertson, 1996). The closest relatives of *A. australis* are *A. florianus* and *A. tuberculatus*, which also have larger genome sizes than most species. This indicates that a polyploidization happened during the ancestral split of this group. In contrast, *A. palmeri* and its two closest relatives have the smallest genome sizes of the genus. The test for random evolution of genome size suggests that both clades deviate significantly from a model

of random evolution due to independ instances of genome duplication and sequence loss (Figure 2.5).

Genome size may correlate with ecological and life history characteristics (Oyama *et al*, 2008). For example, one could postulate that herbicide tolerant weedy amaranths have a smaller genome size because they are faster cycling than their non-resistant relatives. We found that the genome sizes of the weedy amaranths in the different subgenera are highly variable and there does not seem to be a strong relationship between resistance and genome size. For other traits like mating system the number of species in the genus is currently too limited to allow strong conclusions regarding the evolution of the genome sizes. In addition to polyploidization, genome size evolution is also driven by transposable element (TE) dynamics (Bennetzen & Wang, 2014). Since GBS data sample only a small part of the genome and only one draft genome is currently available from the genus, it is not possible to evaluate the role of TEs in genome size evolution of the genus with these data.

6. Conclusions

GBS is a suitable approach for the phylogenetic analysis of the *Amaranthus* genus and allows a high taxonomic resolution. The large number of SNPs obtained from the *de novo* assembly of GBS sequencing reads and the high congruence of phylogenetic trees based on reference-mapping and *de novo* assembly indicates that a reference genome is not required and allows to study the molecular phylogeny of distantly related and non-model species. The inferred phylogeny based on 35 species largely confirms the previous taxonomic classification into three subgenera but also identified highly differentiated groups within the tree taxonomically defined subgenera. In particular, the subgenus *A. Acnida* consists of two strongly different groups with very different genome sizes, which may warrant a taxonomic revision. The comparison of a coalescent species tree with a distance-based tree of multiple individual accessions from each species identified clades in which gene flow, hybridization or geographic differentiation influenced the genomic relationship of species. The species in

the Hybridus complex are closely related and were not separated along the species boundary, but are split into two main groups of accessions and species that reflect the geographically separated groups from South and Central America, respectively. The phylogeny of the genus further allowed to pinpoint the most likely ancestors and wild relatives of cultivated grain amaranths. In particular, *A. hybridus* appears to be the ancestor of all three crop amaranth species and the weed *A. quitensis* might be an intermediate between *A. hybridus* and *A. caudatus* or have contributed substantially to the domestication of *A. caudatus* by gene flow. The genome size measurements indicate that polyploidization events were rare in the genus. As in other plant taxa, further studies like the sequencing of the complete genomes of Amaranth species will be required to fully understand the relative importance of gene flow, hybridization and selection on the taxonomic relationships within the genus.

Acknowledgements

We thank the USDA and IPK genebanks for providing the seeds used in this study and to David Brenner, Amaranth Curator of the USDA genebank for advice and discussion. We are grateful to Elisabeth Kokai-Kota for technical assistance in the laboratory. This work was supported by the F. W. Schnell Endowed Professorship of the Stifterverband für Deutsche Wissenschaft to K. J. S. We also acknowledge support by the state of Baden-Württemberg through bwHPC.

3. Genomic and Phenotypic Evidence for an Incomplete Domestication of South American Grain Amaranth (*Amaranthus caudatus*)

Markus G. Stetter, Thomas Müller and Karl J. Schmid*
Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany

Published in
Molecular Ecology
doi: http://dx.doi.org/10.1111/mec.13974 Available as preprint on

BioRxiv
doi: http://dx.doi.org/10.1101/025866

* Corresponding author: karl.schmid@uni-hohenheim.de

1. Abstract

The domestication syndrom comprises phenotypic changes that differentiate crops from their wild ancestors. We compared the genomic variation and phenotypic differentiation of the two putative domestication traits seed size and seed color of the grain amaranth *Amaranthus caudatus*, which is an ancient crop of South America, and its two close wild relatives and putative ancestors *A. hybridus* or *A. quitensis*. Genotyping 119 accessions of the three species from the Andean region using genotyping-by-sequencing (GBS) resulted in 9,485 SNPs that revealed a strong genetic differentiation of cultivated *A. caudatus* from its two relatives. *A. quitensis* and *A. hybridus* accessions did not cluster by their species assigment but formed mixed groups according to their geographic origin in Ecuador and Peru, respectively. *A. caudatus* had a higher genetic diversity than its close relatives and shared a high proportion of polymorphisms with their wild relatives consistent with the absence of a strong bottleneck or a high level of recent gene flow. Genome sizes and seed sizes were not significantly different between *A. caudatus* and its relatives, although a genetically distinct group of *A. caudatus* from Bolivia had significantly larger seeds. We conclude that despite a long history of human cultivation and selection for white grain color, *A. caudatus* shows a weak genomic and phenotypic domestication syndrome and propose that it an incompletely domesticated crop species either because of weak selection or high levels of gene flow from its sympatric close undomesticated relatives that counteracted the fixation of key domestication traits.

2. Introduction

Research on the domestication of crop plants revealed that numerous traits can be affected by domestication, which results in so-called domestication syndromes. The type and extent of domestication syndromes depends on the life history and uses of crop plants (Meyer *et al*, 2012), although crops from distantly related plant families frequently show similar domestication phenotypes. For example, the 'classical' domestication syndrome, which includes larger seeds, loss of seed shattering, reduced branching, loss of seed dormancy and decreased photoperiod sensitivity, is observed in legumes and grasses (Abbo *et al*, 2014; Hake & Ross-Ibarra, 2015). Similar to phenotypic diversity, crops show variable genomic signatures of domestication because of differences in their biology and utilization by humans (Meyer *et al*, 2012). In particular, domestication affects the level and structure of genetic diversity in crops because selection and genetic drift contributed to strong genetic bottlenecks (Doebley *et al*, 2006; Olsen & Wendel, 2013; Sang & Li, 2013; Nabholz *et al*, 2014). The geographic expansion of domesticated crops provided the opportunity for gene flow with new crop wild relatives, which further contributed to genetic differentiation from wild ancestors. Such a diversity of phenotypic and genomic changes associated with domestication suggest that the classical model of a single domestication event in a short time span within a small geographic region may not apply to numerous crop plants like barley, apple and olive trees (Besnard & Rubio de Casas, 2016; Cornille *et al*, 2012; Poets *et al*, 2015). The motivation of the present study was to investigate both the phenotypic and genomic consequences of amaranth cultivation in the light of these concepts.

The genus *Amarantus* L. comprises between 50 and 75 species and is distributed worldwide (Sauer, 1967; Costea & DeMason, 2001). Four species are cultivated as grain amaranths or leaf vegetables (Sauer, 1967; Brenner *et al*, 2010). The grain amaranths *Amaranthus caudatus*, *Amaranthus cruentus* and *Amaranthus hypochondriacus* originated from South and Central America while *A. tricolor* is used as leafy vegetable in Africa. Amaranth is an ancient crop because archaeological evidence in Northern Argentina suggested that wild amaranth seeds were collected and used for

human consumption during the initial mid-Holocene (8,000 - 7,000 BP; Arreguez *et al*, 2013). In the Aztec empire, amaranth was a highly valued crop and tributes were collected from the farmers that were nearly as high as for maize (Sauer, 1967). Currently, amaranth is promoted as a healthy food because of its favorable composition of essential amino acids and high micronutrient content.

The three grain amaranth species differ in their geographical distribution. *A. cruentus* and *A. hypochondriacus* are most common in Central America, whereas *A. caudatus* is cultivated mainly in South America. In the Andean region, *A. caudatus* grows in close proximity to the two related (i.e., wild) species *A. hybridus* and *A. quitensis*, which are considered as potential ancestors (Sauer, 1967). *A. hybridus* has the widest distribution range from Central to South America while *A. quitensis* is restricted to the central part of South America. *A. quitensis* was tolerated and possibly cultivated in Andean home gardens and used for coloring in historical times.

The history of amaranth cultivation and extent of its domestication are still under discussion (Figure 3.1). Sauer (1967) proposed two domestication scenarios based on the morphology and geographic distribution of the different species. One scenario postulates three independent domestication events from three different wild ancestors, and another scenario proposes the domestication of *A. cruentus* from *A. hybridus* followed by a migration and intercrossing of *A. cruentus* with *A. powellii* in Central America and an intercrossing of *A. cruentus* with *A. quitensis* resulting in *A. caudatus* in South America. A third scenario was based on genetic markers and suggested that all three cultivated amaranths evolved from *Amaranthus hybridus*, but at multiple locations (Maughan *et al*, 2011). Most recently, Kietlinski *et al* (2014) proposed a single domestication of *A. hybridus* in the Andes or in Mesoamerica and a subsequent spatial separation of two lineages leading to *A. caudatus* and *A. hypochondriacus*, or two independent domestication events of *A. hypochondriacus* and *A. caudatus* from a single *A. hybridus* lineage in Central and South America (Figure 3.1C and D). A more recent analysis based on the phylogeny of the whole Amaranthus genus supports independent domestication of the South American *A. caudatus* and the two Central American grain amaranths from

two different, geographically separated lineages of *A. hybridus* as shown in Figure 1D (Stetter & Schmid, 2016).

Despite its long history of cultivation and the self-pollinating breeding system, the domestication syndrome of cultivated amaranth is remarkably indistinct because it still shows strong photoperiod sensitivity and has very small shattering seeds (Sauer, 1967; Brenner *et al*, 2010). Other crops like maize that were cultivated at a similar time period in the same region exhibit the classical domestication syndrome (Sang & Li, 2013; Lenser & Theißen, 2013). This raises the question whether amaranth is domesticated at all or has a different domestication syndrome, and if the latter is true whether genetic constraints, a lack of genetic variation or (agri-)cultural reasons determined its domestication syndrome. The phenotypic analysis of amaranth domestication is complicated by the taxonomic uncertainty of cultivated amaranth species and their close relatives. Although *A. quitensis* was suggested to be the ancestor of *A. caudatus*, the state of *A. quitensis* as a separate species is under debate. Sauer (1967) classified it as species, but later it was argued that it is the same species as *A. hybridus* (Coons, 1978; Brenner *et al*, 2010). However, until today *A. quitensis* is treated as separate species and since genetic evidence for the status of *A. quitensis* as a separate species is based on few studies with limited numbers of markers, this topic is still unresolved (Mallory *et al*, 2008; Kietlinski *et al*, 2014).

The rapid development of sequencing technologies facilitates the large-scale investigation of the genetic history of crops and their close relatives. Among available methods, reduced representation sequencing approaches such as genotyping-by-sequencing (GBS) allow a genome-wide and cost-efficient marker detection compared to whole genome sequencing (Elshire *et al*, 2011; Poland *et al*, 2012b). Despite some biases associated with reduced representation sequencing, GBS and related methods are suitable and powerful approaches for studying interspecific phylogenetic relationships (Cruaud *et al*, 2014) and intraspecific patterns of genetic variation in crop plants (Morris *et al*, 2013).

We used GBS and genome size measurements to characterize the genetic diversity and relationship of cultivated *A. caudatus* and its possible ancestors *A. quitensis* and *A. hybridus*, and compared patterns of genetic structure with two domestication-related phenotypic traits (seed color and

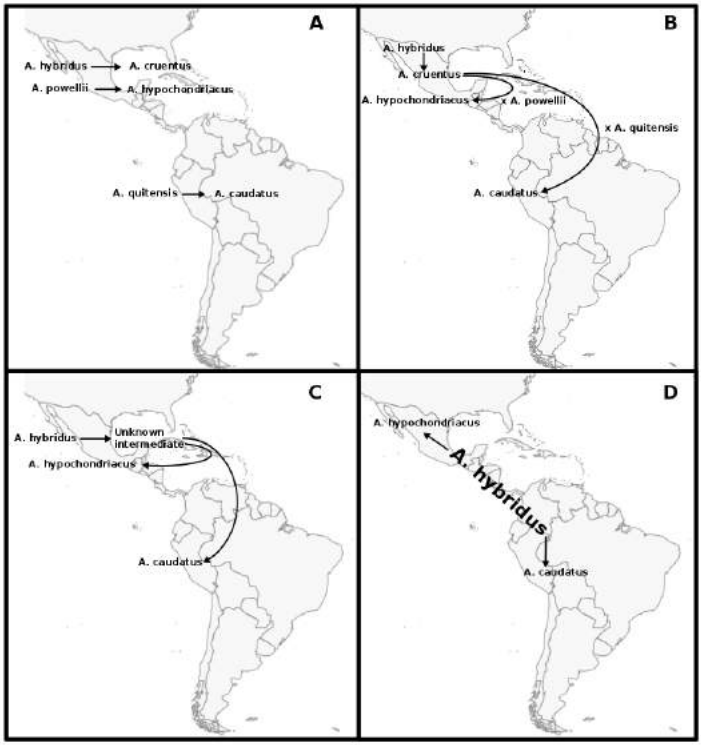

Figure 3.1.: Models of amaranth domestication. (A) Independent domestication of three grain amaranths from different close relatives (Sauer, 1967). (B) Initial domestication from *A. hybridus* and subsequent migration and hybridization with additional close relatives (Sauer, 1967). (C) Single domestication in the Andes or in Mesoamerica and subsequent spatial separation of two lineages leading to *A. caudatus* and *A. hypochondriacus* (Kietlinski *et al*, 2014). (D) Two domestication events from a single *A. hybridus* lineage spanning Central and South America (Kietlinski *et al*, 2014).

hundred seed weight). For this study, we focussed on the South American amaranth species, because *A. caudatus*, *A. quitensis* and South American accessions of *A. hybridus* form a clade that is strongly separated from the two Central American grain amaranths in a phylogenetic analysis of the whole genus (Stetter & Schmid, 2016). For this reason, we reasoned that the domestication of *A. caudatus*, which is native to South America, and its relationship to the sympatric relatives, *A. hybridus* and *A. quitensis* can be conducted independently of the Central American amaranth species. Our analysis includes a comparison of genetic diversity and seed-related traits like size and color between cultivated and wild amaranths and analyses the taxonomic relationship and gene flow among species. Our results indicate that *A. caudatus* has a history of domestication that may be considered as incomplete.

3. Material and Methods

3.1. Plant Material

A total of 119 amaranth accessions of three *Amaranthus* species originating from South America were obtained from the USDA genebank (`http://www.ars-grin.gov/npgs/searchgrin.html`). Of these accessions, 89 were classified as *A. caudatus*, 17 as *A. hybridus*, seven as *A. quitensis* and six as interspecific hybrids according to the passport information (Table S5). We selected *A. caudatus* accessions based on the altitude of the collection site and focused on high-altitude regions (2,200 to 3,700 m) where amaranth has been cultivated for thousands of years and survived until today since it fell into disuse after the Spanish conquest (Kauffman & Weber, 1990). Therefore, high-altitude accessions may represent a large proportion of the species-wide genetic diversity. The smaller sample sizes of *A. hybridus* and *A. quitensis* accessions reflect that fewer accessions of these species than of *A. caudatus* are available from the USDA genebank. However, the geographic origin of the two wild relatives covers the Andean highlands, which is the distribution range of *A. caudatus*, and we compared the population structure of the sample derived

from the genomic data with the passport information to test for consistency between the population structure and geographic origin. Accessions were planted in a field in Nürtingen (Germany), and one young leaf of one representative plant per accession was sampled to avoid the sampling of potential seed cross-contamination. We sampled and sequenced three plants each of 12 accessions independently for quality control.

3.2. Genome Size

To compare genome sizes among the three diploid *Amaranthus* species, we measured the genome size of 22 *A. caudatus*, 8 *A. hybridus* and 4 *A. quitensis* accessions. Genome size differences of individuals within species are expected to be low, and we therefore estimated species-specific genome sizes using 25% the total sample of *A. caudatus* and 50% of *A. hybridus* and *A. quitensis* accessions, respectively. Plants were grown for four weeks in the greenhouse before one young leaf was collected for cell extraction. A tomato cultivar (*Solanum lycopersicum* cv Stupicke) was used as internal standard because it has a comparable genome size that has been measured with high accuracy (DNA content = 1.96 pg; Dolezel *et al*, 1992). Fresh leaves were cut up with a razor blade and cells were extracted with CyStain PI Absolute P (Partec, Muenster/Germany). Approximately 0.5 cm^2 of the sample leaf was extracted together with similar area of tomato leaf in 0.5 ml of extraction buffer. The DNA content was determined with CyFlow Space (Partec, Muenster/Germany) flow cytometer and analyzed with FlowMax software (Partec, Muenster/Germany). For each sample, 10,000 particles were measured each time. Two different plants were measured for each accession. The DNA content was calculated as:

$$\text{DNA content 2C [pg]} = \text{genome size tomato} \times \frac{\text{fluorescence amaranth}}{\text{fluorescence tomato}}$$

and the genome size (in Mbp) was calculated as:

$$\text{genome size 1C [Mbp]} = (0.978 * 10^3) \times \frac{\text{DNA content 2C [pg]}}{2}$$

The conversion from pg to bp was calculated with 1pg DNA = 0.978 × 10^9 bp (Dolezel *et al*, 2003). Means were calculated using R software (R

Core Team, 2014) and an ANOVA was performed to infer differences in genome size for the species.

3.3. DNA Extraction and Library Preparation

Genomic DNA was extracted using a modified CTAB protocol (Saghai-Maroof *et al*, 1984). The DNA was dried and dissolved in 50-100 μl TE and diluted to 100 ng/μl for further usage. Two-enzyme GBS libraries were constructed with a modified protocol from the previously described two-enzyme GBS protocol (Poland *et al*, 2012b). DNA was digested with a mix of 2 μl DNA, 2 μl NEB Buffer 2 (NEB, Frankfurt/Germany), 1 μl ApeKI (4U/μl, NEB), 1 μl HindIII (20U/μl, NEB) and 14 μl ddH_2O for 2 hours at 37°C before incubating for 2 hours at 75°C. Adapters were ligated with 20 μl of digested DNA 5 μl ligase buffer (NEB), T_4- DNA ligase (NEB), 4 μl ddH_2O and 20 μl of adapter mix containing 10μl barcode adapter (0.3 ng/μl) and 10 μl common adapter (0.3ng/μl). Samples were incubated at 22°C for 60 minutes before deactivating ligase at 65°C for 30 minutes. Subsequently, samples were cooled down to 4°C. For each sequencing lane, 5μl of 48 samples with different barcodes were pooled after adapter ligation. Samples of the different species were randomized over the 3 pools and different barcode lengths. The 12 replicated samples were added to each pool. The pooled samples were purified with QIAquick PCR purification kit (Qiagen, Hilden/Germany) and eluted in 50 μl elution buffer before PCR amplification of the pools. The PCR was performed with 10 μl of pooled DNA, 25 μl 2x Taq Master Mix (NEB), 2 μl PCR primer mix (25pmol/μl of each primer) and 13 μl ddH_2O for 5 min at 72°C and 30 sec at 98°C before 18 cycles of 10 sec at 98°C, 30 sec at 65°C and 30 sec at 72°C after the 18 cycles 5 min of 72°C were applied and samples were cooled down to 4°C. Samples were purified again with QIAquick PCR purification kit (Qiagen) and eluted in 30μl elution buffer. Three lanes with 48 samples per lane were sequenced on an Illumina HighScan SQ with single end and 105 cycles on the same flow cell (see supporting data).

3.4. Data Preparation

Raw sequence data were filtered with the following steps. First, reads were divided into separate files according to the different barcodes using Python scripts. Read quality was assessed with fastQC (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`). Due to lower read quality towards the end of the reads, they were trimmed to 90 bp. Low quality reads were excluded if they contained at least one N (undefined base) or if the quality score after trimming was below 20 in more than 10% of the bases. Data from technical replicates were combined and individuals with less than 10,000 reads were excluded from further analysis (Table S5). The 12 replicated samples were used to detect a lane effect with an analysis of variance.

3.5. SNP Calling and Filtering

Since no high quality reference genome for *Amaranthus* sp. was available for read mapping, we used `Stacks 1.19`, for the *de novo* identification of SNPs in GBS data (Catchen *et al*, 2011, 2013). The SNP calling pipeline provided by Stacks `denovo_map.pl` was used to call SNPs from the processed data. Highly repetitive GBS reads were removed in the `ustacks` program with option `-t`. Additionally, the minimum number of identical raw reads required to create a stack was set to three (`m=3`) and the number of mismatches allowed between loci when processing a single individual was two (`M=2`). Four mismatches were allowed between loci when building the catalog (`n=4`). The catalog is a set of non redundant loci representing all loci in the accessions and used as reference for SNP calling. SNPs were called with the Stacks tool `populations 1.19` without filtering for missing data using option `-r 0`. One individual, PI 511754, was excluded from further analysis because it appeared to be misclassified. According to its passport information it belonged to *A. hybridus*, but with all clustering methods it was placed into a separate cluster consisting only of this individual, which suggested it belongs to a different species. Therefore, we repeated the SNP calling without this individual. The SNPs were filtered over the whole sample for missing data

with vcftools (Danecek *et al*, 2011) by allowing a maximum of 60% missing values per SNP position. Given the stringent filtering criteria for SNP calling and the restricted number of *A. quitensis* individuals, we did not filter SNPs by their minor allele frequency for further analysis.

3.6. Inference of Genetic Diversity and Population Structure

Nucleotide diversity (π) weighted by coverage was calculated with a Python script that implements the formula of Begun *et al* (2007) which corrects for different sampling depths of SNPs in sequencing data. The confidence interval of π was calculated by bootstrapping the calculation 10,000 times. To account for the difference in sampling between wild and cultivated amaranths, we sub-sampled *A. caudatus* 100 times with the the same number of individuals (23) as used for wild amaranth. The pairwise difference in π between *A. caudatus* and the close relatives was calculated for each site. Mean expected (H_{exp}) and observed (H_{obs}) heterozygosities based on SNPs were calculated with the R package `adegenet 1.4-2` (Jombart & Ahmed, 2011). The inbreeding coefficient (F) was calculated as:

$$\frac{H_{\text{exp}} - H_{\text{obs}}}{H_{\text{exp}}}$$

Weir and Cockerham weighted F_{ST} estimates were calculated with vcftools (Weir & Cockerham, 1984; Danecek *et al*, 2011). To infer the population structure, we used ADMIXTURE for a model-based clustering (Alexander *et al*, 2009) and conducted the analysis with different numbers of predefined populations ranging from $K = 1$ to $K = 9$ to find the value of K that was most consistent with the data using a cross-validation procedure described in the ADMIXTURE manual. To avoid convergence effects we ran ADMIXTURE 10 times with different random seeds for each value of K. As a multivariate clustering method, we applied discriminant analysis of principal components (DAPC) implemented in the R-package adegenet (Jombart *et al*, 2010; Jombart & Ahmed, 2011) and determined the number of principal components (PCs) used in DAPC with

the `optim.a.score` method. To investigate the phylogenetic relationship of the species, we calculated an uncorrected neighbor joining network using the algorithm NeighborNet (Bryant & Moulton, 2004) as implemented in the `SplitsTree4` program (Huson & Bryant, 2006). The Euclidean distance was calculated from the genetic data to construct a neighbor joining tree, which was bootstrapped 1,000 times with the `pegas` R-package (Paradis *et al*, 2004). The migration between genetic groups was modeled with TreeMix (Pickrell & Pritchard, 2012). For the TreeMix analysis we used the groups that were identified by ADMIXTURE (K = 5) without an outgroup, and allowed four migration events, as preliminary runs indicates four migration events to be the highest number. The tree was bootstrapped 1,000 times.

3.7. Seed Color and Hundred Seed Weight

For each accession we calculated the hundred seed weight (HSW) by weighting three samples of 200 seeds. Seed color was determined from digital images taken with a binocular (at 6.5x magnification) and by visual comparison to the GRIN descriptors for amaranth (`http://www.ars-grin.gov/cgi-bin/npgs/html/desclist.pl?159`). There were three colors present in the set of accessions, white, pink, which also indicates a white seed coat and dark brown. To infer how the species, assigned genetic groups or seed color influenced seed size, we conducted an ANOVA. Differences were tested with a LSD test implemented in the R package `agricolae` (`http://tarwi.lamolina.edu.pe/~fmendiburu/`).

4. Results

4.1. Genome Size Measurements

Although the genomic history of amaranth species still is largely unknown, genome sizes and chromosome numbers are highly variable within the genus *Amaranthus* (`http://data.kew.org/cvalues/`). We therefore tested whether a change in genome size by polyploidization or large-

scale insertions or deletions played a role in the speciation history of *A. caudatus* and the two relatives *A. quitensis* and *A. hybridus* by measuring the genome size of multiple individuals from all three species with flow cytometry. The mean genome size of *A. caudatus* was 501.93 Mbp, and the two relatives did not differ significantly from this value (Table 3.1) indicating that all measured individuals are diploid and that polyploidization did not play a role in the recent evolution of cultivated amaranth.

Table 3.1.: Genome size of representative group of individuals for each species. There are no significant differences between genome sizes ($p \leq 0.05$). The number of individuals per population is N and SD is the standard deviation for each parameter.

	N	DNA content (pg)	SD	genome size (Mbp)	SD
A. caudatus	22	1.026	0.026	501.93	12.74
A. hybridus	8	1.029	0.025	502.96	12.20
A. quitensis	4	1.021	0.016	499.07	7.91

4.2. SNP Identification by GBS

To investigate genome-wide patterns of genetic diversity in *A. caudatus* and its two closest relatives, we genotyped a diverse panel of 119 amaranth accessions from the three species that were initially collected in the Andean region and then obtained from the USDA genebank. The sequencing data generated with a two-enzyme GBS protocol consisted of 210 Mio. raw reads with an average of 1.5 Mio. reads per accession (Supporting information S2). We tested for a lane effect of the Illumina flow cell by sequencing the same 12 individuals on each of the three lanes used for sequencing of the whole sample. An ANOVA of the read number did not show a lane effect (Table B.1). Since a high-quality reference genome of an amaranth species was not available, we aligned reads *de novo* within the dataset to unique tags using Stacks (Catchen *et al*, 2011). The total

length of the aligned reads was 16.6 Mb, which corresponds to approximately 3.3 % of the *A. caudatus* genome. For SNP calling, reads of each individual were mapped to the aligned tags. SNPs were called with parameters described in Materials and Methods, which resulted in 63,956 SNPs and a mean read depth of 40.28 per site. Since GBS data are characterized by a high proportion of missing values, we removed SNPs with more than 60% of missing values. After this filtering step, we obtained 9,485 biallelic SNPs with an average of 35.3 % missing data for subsequent analyses (Figure B.1). The folded site frequency spectrum showed an expected distribution but *A. quitensis* had more sites with low frequency due to the restricted number of individuals (Figure B.2)

4.3. Inference of Population Structure

To infer the genetic relationship and population structure of the three amaranth species, we used three different methods that included ADMIXTURE, Discriminant Analysis of Principal Components (DAPC) and phylogenetic reconstruction with an uncorrected neighbor-joining network. The ADMIXTURE analysis with three predefined groups ($K = 3$) that corresponds to the number of *Amaranthus* species included in the study did not cluster accessions by their species, but combined the two relatives *A. hybridus* and *A. quitensis* into a single cluster and grouped the *A. caudatus* accessions into two distinct clusters. Higher values of K did not lead to subdivision of the two close relatives into separate groups that correspond to the species assignment (Figure 4.3), however, the they were split according to their geographic origin. Cross-validation showed that $K = 5$ was most consistent with the data (Figure B.3), which produced three different groups of *A. caudatus* accessions that included a few accessions from the close relatives, and two clusters that both consist of *A. hybridus* and *A. quitensis* accessions. These two clusters are not separated by the species assignment but by the geographic origin of accessions because the clusters consist of *A. hybridus* and *A. quitensis* accessions from Peru and Ecuador, respectively, which indicates a strong geographic differentiation among the close relatives.

The groups of *A. caudatus* accessions also showed a clear geographic

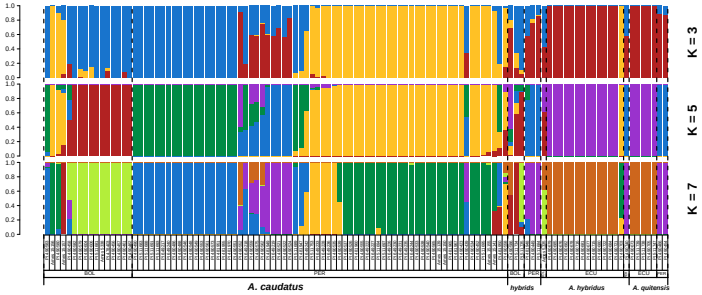

Figure 3.2.: Model based clustering analysis with different numbers of clusters (K=3, 5, 7) with ADMIXTURE. The clusters reflect the number of species in the study (K=3), the number of single populations (species per country of origin, K=7) and the optimal number as determined by cross validation (K=5). Individuals are sorted by species and country of origin (BOL=Bolivia, PER = Peru and ECU = Ecuador) as given by their passport data.

differentiation. The first cluster consisted of individuals from Bolivia (Figures 4.3 and 3.3; $K = 5$, red color). *A. caudatus* accessions from Peru were split into two clusters of which one predominantly represents a region from North Peru (Huari Province; Figures 4.3 and 3.3; $K = 5$, yellow color), whereas the second cluster contains individuals distributed over a wide geographic range that extends from North to South Peru ($K = 5$, green color). Ten *A. caudatus* accessions from the Cuzco region clustered with the three accessions of the close relatives from Peru ($K = 5$, blue color). These ten accessions showed admixture with the other cluster of *A. hybridus/A. quitensis* and with a Peruvian cluster of *A. caudatus*. Accessions that were labeled as 'hybrids' in their passport data, because they express a set of phenotypic traits of different species, clustered with different groups. 'Hybrids' from Bolivia were highly admixed, whereas 'hybrids' from Peru clustered with the close relatives from Peru (Figure 4.3). Taken together, the population structure inference with ADMIXTURE identified

a clear separation between the cultivated *A. caudatus* and its to close relatives, and a high level of genetic differentiation among cultivated amaranths with some evidence for gene flow between groups.

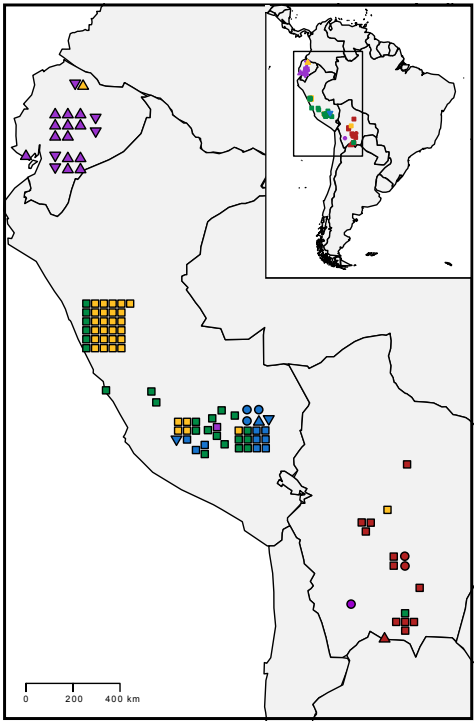

Figure 3.3.: Geographic distribution of accessions for which data was available from passport information. Locations are not exact geographic locations because location data was given as country province . Colors are given by ADMIXTURE with K=5 (Figure 4.3). Species are indicated by shapes. *A. caudatus* (□), *A. hybridus* (△), *A. quitensis* (▽) and hybrids between species (○)

The inference of population structure with a discriminant analysis of principal components (DAPC) and Neighbor-Joining network produced very similar results as ADMIXTURE. The first principal component of the DAPC analysis which we used to cluster accessions based on their species explained 96% of the variation and separated the cultivated *A. caudatus* from its two relatives (Figure B.4A). In a second DAPC analysis that was based on information on species and country of origin (Figure B.4B) the first principal component explained 55% of the variation and separated most cultivated amaranth accessions from the close relatives. The second principal component explained 35% of the variation and separated the Peruvian from the Bolivian *A. caudatus* accessions.

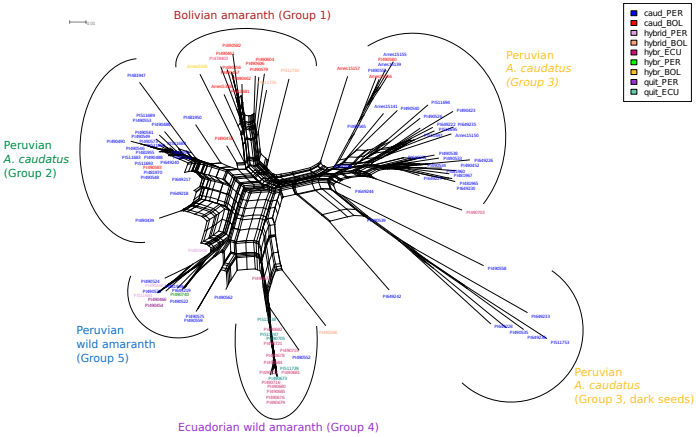

Figure 3.4.: Neighbor-joining network of 113 amaranth accessions from six potential populations. Different colors indicate the species and origin according to gene bank information. *A. caudatus* from Peru (blue) and from Bolivia (red), *A. hybridus* from Ecuador (magenta), from Peru (green) and Bolivia (yellow), *A. quitensis* from Ecuador (turquoise) and Peru (purple) and hybrids between species from Peru (salmon) and Bolivia (light orange). Arches show genetic clusters as inferred with ADMIXTURE ($K = 5$).

The phylogenetic network outlines the relationships between the different clusters (Figure 3.4). It shows two distinct groups of mainly Peruvian *A. caudatus* accessions and a group of accessions with a wide geographic distribution (Figure 3.3; green color). The latter is more closely related to the Bolivian *A. caudatus* and the close relatives. The strong network structure between these three groups suggests a high proportion of shared ancestral polymorphisms or a high level of recent gene flow. In contrast, *A. caudatus* accessions from Northern Peru are more strongly separated from the other groups (Figure 3.3; yellow color) and are split into two subgroups, of which the smaller one includes only accessions with dark seeds. In a bifurcating phylogenetic tree, ten cultivated amaranth accessions clustered within the same clade as the close relatives *A. quitensis* and *A. caudatus* (Figure B.5). The same clustering was also obtained with ADMIXTURE and $K = 7$ (Figure 4.3).

Table 3.2.: Weir and Cockerham weighted F_{ST} estimates between populations based on the taxonomic assignment of their passport data. The group of close relatives are *A. hybridus* and *A. quitensis* taken together.

	F_{ST}
A. caudatus x *A. hybridus*	0.319
A. caudatus x *A. quitensis*	0.274
A. caudatus x close relatives	0.322
A. hybridus x *A. quitensis*	0.041
A. caudatus (PER) x *A. caudatus* (BOL)	0.132

To quantify the level of genetic differentiation between the species and groups within *A. caudatus*, we estimated weighted F_{ST} values using the method of Weir and Cockerham (Weir & Cockerham, 1984). F_{ST} values between *A. caudatus* and *A. hybridus* and *A. quitensis* species were 0.31

and 0.32, respectively (Table 3.2), and 0.041 between *A. hybridus* and *A. quitensis* based on the taxonomic assignment. The latter reflects the high genetic similarity of the accessions from both species observed above. Within *A. caudatus* subpopulations, the F_{ST} between *A. caudatus* populations from Peru and Bolivia was 0.132, three times higher than between *A. hybridus* and *A. quitensis*. The above analyses suggested that some individuals may be misclassified in the passport information, and we therefore calculated F_{ST} values of population sets defined by ADMIXTURE. Although such F_{ST} values are upward biased, they allow to evaluate the relative level of differentiation between groups defined by their genotypes. The comparison of F_{ST} values showed that the three *A. caudatus* groups (groups 1-3) are less distant to the group of *A.quitensis/A.hybridus* accessions from Peru (group 5) than from Ecuador (group 4; Table B.2). A TreeMix analysis, which is based on allele frequencies within groups (Figure 3.5), suggests gene flow from the Peruvian *A. caudatus* (group 2) to *A. quitensis* and *A. hybridus* amaranths from Peru (group 5) and, with a lower confidence level, from *A. quitensis* and *A. hybridus* from Ecuador (group 4) into Bolivian *A. caudatus* (group 1), as well as from Bolivian *A. caudatus* to Peruvian *A. caudatus* (Group 2).

4.4. Analysis of Genetic Diversity

We further investigated whether domestication reduced genetic diversity in cultivated *A. caudatus* (Table 3.3). All measures of diversity were higher for *A. caudatus* than its relatives. For example, nucleotide diversity (π) was about two times higher in *A. caudatus* than in the two relatives combined. The diversity values of the accessions classified as 'hybrids' showed intermediate values between cultivated amaranth and its relatives supporting their hybrid nature. The inbreeding coefficient, F, was highest in the cultivated amaranth but did not differ from the two close relatives if they are combined.

In contrast, accessions classified as 'hybrids' and *A. quitensis* had lower inbreeding coefficients. Within the groups of accessions defined by ADMIXTURE, genetic diversity differed substantially. The close relatives from Ecuador had the lowest ($\pi = 0.00031$) while the group from north-

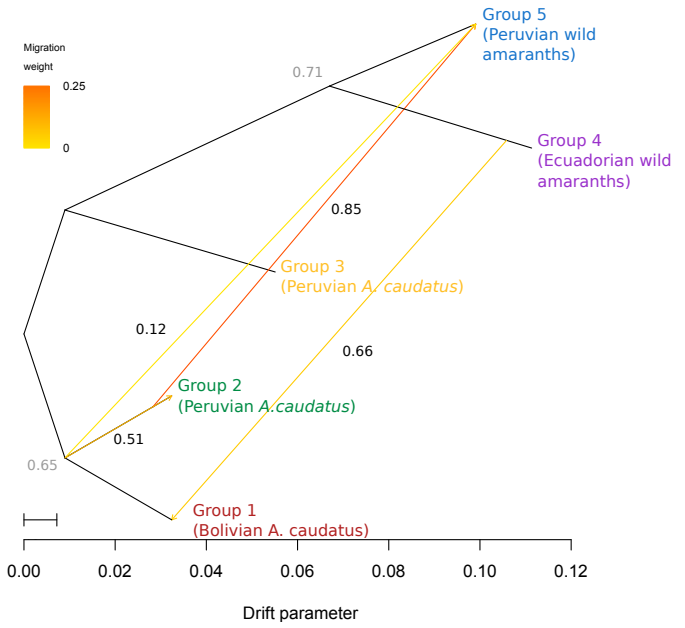

Figure 3.5.: Tree of five genetic clusters of South American amaranths inferred with TreeMix. The genetic clusters which were used to calculate the tree were inferred with ADMIXTURE. Groups 1 to 3 represent *A. caudatus* clusters from Peru and Bolivia, group 4 represents accessions of *A. quitensis* and *A. hybridus* from Ecuador and group 5 wild amaranth from Peru, respectively. The migration events are colored according to their weight. Numbers at branching points and on the migration arrow represent bootstrapping results based on 1,000 runs.

ern Peru showed the highest level of nucleotide diversity ($\pi = 0.00111$; Table B.3). Figure 3.6 shows that even though the overall diversity of *A. caudatus* was higher, a substantial proportion of sites were more diverse in the close relatives ($\pi_{caud} - \pi_{hyb/quit} < 0$; Figure 3.6).

Table 3.3.: Genetic diversity parameters for the cultivated *A. caudatus* and the close relatives *A. hybridus* and *A. quitensis*. π is the nucleotide diversity over all sites, CI_π is the 95% confidence interval of π, H_{exp} the mean expected heterozygosity for the variant sites and SD_{He} its standard deviation, H_{obs} the mean observed herterozygosity and SD_{Ho} its standard deviation. F is the inbreeding coeficient and SD_F its standard deviation.

Population	N	π	CI_π	H_{exp}	SD_{He}	H_{obs}	SD_{Ho}	F	SD_F	θ_w
A. caudatus	84	0.00117	± 0.00002	0.175	0.167	0.049	0.140	0.688	0.462	0.00123
A. hybridus	16	0.00061	± 0.00001	0.085	0.135	0.041	0.170	0.679	0.608	0.00073
A. quitensis	7	0.00059	± 0.00001	0.076	0.169	0.040	0.170	0.451	0.763	0.00048
Close relatives combined	23	0.00062	± 0.00002	0.090	0.140	0.041	0.166	0.681	0.591	0.00070
combined										
Hybrids	6	0.00091	± 0.00001	0.112	0.179	0.060	0.173	0.436	0.645	0.00107

4.5. Seed Color and Seed Size as Potential Domestication Traits

In grain crops, grain size and seed color are important traits for selection and likely played a central role in domestication of numerous plants (Abbo *et al*, 2014; Hake & Ross-Ibarra, 2015). To investigate whether these two traits are part of the domestication syndrome in grain amaranth, we compared the predominant seed color of the different groups of accessions and measured their seed size. The seeds could be classified into three colors, white, pink and brown. The white and pink types have both a white seed coat, but the latter has red cotyledons that are visible through the translucent seed coat. A substantial number of seed samples (19) from the genebank contained seeds of other color up to a proportion of 20%. One *A. caudatus* accession from Peru (PI 649244) consisted of 65% dark seeds and 35% white seeds in the sample. No accession from the two close relatives *A. hybridus* and *A. quitensis*, or from the hybrid accessions had white seeds, whereas the majority (74%) of *A. caudatus* accessions had white (70%) or pink (4%) seeds, and the remaining (26%) brown seeds (Figure 3.7 A).

We also compared the seed color of groups defined by ADMIXTURE

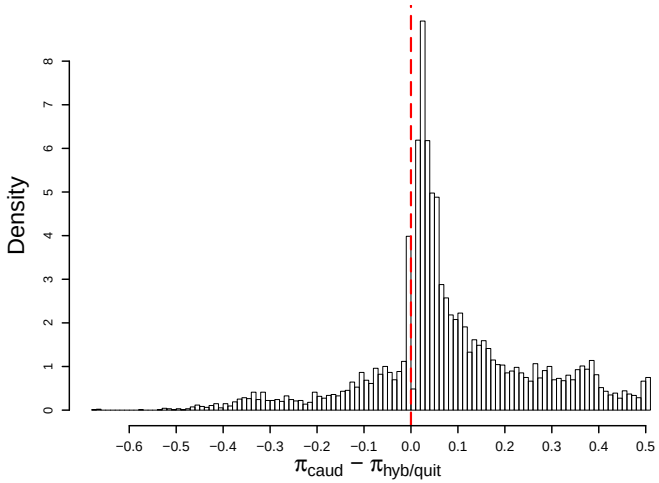

Figure 3.6.: Per site difference in nucleotide diversity (π) between cultivated amaranth (*A. caudatus*) and the close relatives *A. hybridus* and *A. quitensis*.

($K = 5$; Figure 4.3), which reflect their genetic relationship and may correct for mislabeling of accessions (Figure 3.7 B). No group had only white seeds, but clusters consisting mainly of *A. hybridus* and *A. quitensis* had no white seeds at all. In contrast to seed color, the hundred seed weight (HSW) of the different *Amaranthus* species did not significantly differ between cultivated *A. caudatus* and the two relatives. The mean HSW of *A. caudatus* was 0.056 g and slightly higher than the HWS of *A. hybridus* (0.051 g) and *A. quitensis* (0.050 g; Figure 3.7 C and Table B.4). Among the groups identified by ADMIXTURE ($K = 5$), one group showed a significantly higher HSW than the other groups, while the other four groups did not differ in their seed size. The group with the higher HSW consisted

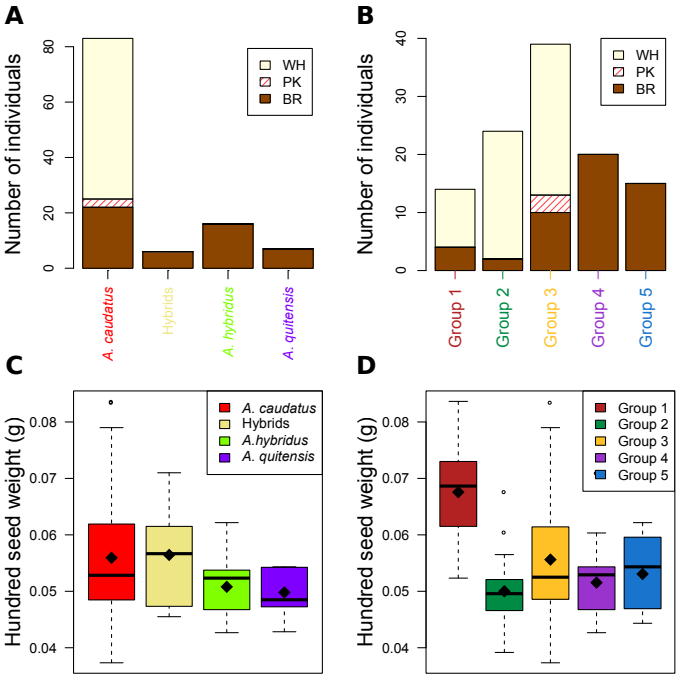

Figure 3.7.: Predominant seed color (**A,B**) and hundred seed weight (**C,D**) by *Amaranthus* species (**A,C**) and groups identified with ADMIXTURE for $K = 5$ (**B,D**) where group 1 (red) resembles *A. caudatus* from Bolivia, group 2 (green) and 3 (yellow) *A. caudatus* from Peru, group 4 (purple) represents the close relatives *A. quitensis* and *A. hybridus* from Ecuador and group 5 (blue) from Peru, respectively. Seed colors were white (WH), pink (PK) and dark brown (BR). While there were no significant differences in seed size between the species, group 1 had significantly higher hundred seed weight ($p \leq 0.05$) than the other groups.

mainly of Bolivian *A. caudatus* accessions and had a 21 % and 35 % larger HSW than the two groups consisting mainly of Peruvian *A. caudatus* accessions, respectively (Figure 3.7 D). An ANOVA also revealed that seed color has an effect on seed size because white seeds are larger than dark seeds (Table 3.4).

Table 3.4.: Analysis of variance for the hundred seed weight in dependence of the seed color and population as determined by ADMIXTURE

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Seed color	2	0.000657	0.0003285	4.657	0.0116 *
Group	4	0.003151	0.0007877	11.165	1.46e-07 ***
Seed color:Group	2	0.000042	0.0000209	0.297	0.7440
Residuals	103	0.007266	0.0000705		

5. Discussion

5.1. Genotyping-by-Sequencing of Amaranth Species

The genotyping of cultivated amaranth *A. caudatus* and two close relatives *A. quitensis* and *A. hybridus* revealed a strong genetic differentiation between both groups and a high level of genetic differentiation within cultivated *A. caudatus*. We based our sequence assembly and SNP calling on a *de novo* assembly of GBS data with Stacks because currently no high quality reference sequence of these species is available. Stacks allows SNP calling without a reference genome by constructing a reference catalog from the data and includes all reads in the analysis (Catchen *et al*, 2011). *De novo* assembled fragments without a reference are unsorted and can not be used to investigate genetic differentiation along along the genomic regions, but they are suitable for analysing genetic diversity and

population structure (Catchen *et al*, 2013). GBS produces a large number of SNPs (Poland *et al*, 2012b; Huang *et al*, 2014), albeit with a substantial proportion of missing values. Missing data lead to biased estimators of population parameters such as π and θ_w (Arnold *et al*, 2013) and need to be accounted for if different studies are compared. Additionally, variable error rates in different GBS data sets can inflate differentiation estimates (Mastretta-Yanes *et al*, 2015). The comparison of accessions and groups within a study is possible, however, if all individuals were treated with the same experimental protocol. We filtered out sites with high levels of missing values to obtain a robust dataset for subsequent population genomic analysis. The SNPs were called based on the total sample without accounting for the species which should not bias diversity estimates. Since a smaller sample from the close relatives may underestimate their diversity compared to cultivated *A. caudatus*, we compared diversity estimates by repeated random sampling of 23 out of 84 *A. caudatus* accessions and calculating π from the smaller sample. Diversity estimates of the smaller *A. caudatus* did not differ from the full sample and estimates were in all cases higher than in the close relatives (Figure B.6). We conclude that the different sample sizes of the two groups do not introduce a bias on diversity estimates.

5.2. Genetic Relationship of *A. quitensis* and *A. hybridus*

Coons (1978) suggested that *A. quitensis* is the same species as *A. hybridus*, but in the genebank passport data *A. quitensis* is still considered as a separate species. The taxonomic differentiation between the two species rests on two minor morphological trait, namely the shape of the tepals and the short utricles, which are very small and prone to misidentification (Sauer, 1967; Adhikary & Pratt, 2015). The high phenotypic similarity of *A. quitensis* and *A. hybridus* is supported by the GBS data because accessions from the two species are closely related. They are not separated by their species assigment but cluster into two groups that both consist of accessions from the two species and reflect their geographic origin from

Peru and Ecuador, respectively. The F_{ST} value between *A. quitensis* and *A. hybridus* was lower than between the two *A. caudatus* groups from Peru and Bolivia (Tables 3.2 and S2). The highly similar genome sizes of all three diploid species is consistent with genetic relationship inferred from the GBS data and indicates that large-scale genomic changes like polyploidization events did not occur in the recent history of these species. For comparison, other species in the genus *Amaranthus* have very different genome sizes due to variation in chromosome numbers and ploidy levels (Baohua & Xuejie, 2002; Rayburn *et al*, 2005).

In contrast to our analysis, Kietlinski *et al* (2014) found stronger evidence for a genetic differentiation between *A. hybridus* and *A. quitensis* based on the 11 SSR markers. However, their data also show that both species are distinct groups that do not cluster their species assignment but by geographic origin. These differences may result from the different marker types (SNPs vs. SSRs) and a different sample composition because our sample consists of accessions from the Andean region, whereas Kietlinski et al. included putative wild amaranth accessions with little geographic overlap between the two species. The groups of *A. hybridus* and *A. quitensis* accessions from Peru and Ecuador show a high level of differentiation ($F_{ST} = 0.579$; Table S2), which is similar to the differentiation between one of two Peruvian *A. caudatus* groups and the *A. hybridus/A.quitensis* accessions from Peru ($F_{ST} = 0.553$). Although the sample size of *A. quitensis* and *A. hybridus* is small, genetic differentiation between species should be stronger than between individuals within species in the ADMIXTURE and phylogenetic analyses. In summary, our analysis and the work of Kietlinski *et al* (2014) show that *A. quitensis* and *A. hybridus* do not have a simple genetic relationship that follows species assignment. The high level of intraspecific differentiation in both cultivated amaranth and their relatives is relevant for investigating domestication because the genetic distance between groups of cultivated amaranth is related to the geographic distance of the putative wild ancestors. Therefore, future studies of these two close relatives of the grain amaranths should include large number of accessions from the whole species range to model genetic differentiation within the two species as well as the relationship between species.

5.3. Diversity of South American Amaranth

In numerous crops, domestication was associated with a decrease in genome-wide levels of diversity due to bottleneck effects and strong artificial selection of domestication traits (Gepts, 2014). Under the assumption that the cultivated grain amaranth *A. caudatus* was domesticated, genetic diversity should be reduced compared to the two close relatives. In contrast, the overall genetic diversity in our sample of cultivated amaranths was higher than in the two close relatives. The distribution of diversity between the GBS fragments includes genomic regions with reduced diversity in *A. caudatus*, which may reflect selection in some genomic regions (Figure 3.6). Without a reference genome it is not possible to position reads on a map to identify genomic regions that harbor putative targets of selection based on an inference of the demographic history. Despite the indirect phenotypic evidence for selection, the higher genetic diversity of cultivated grain amaranth may result from a strong gene flow between cultivated amaranths and its relatives. Gene flow between different amaranth species has been observed before (Trucco *et al*, 2005) and is also consistent with the observation of six highly admixed accessions classified as 'hybrids' in the passport data and which appear to be interspecific hybrids (Figure 4.3 and Table 3.3). Gene flow between *A. caudatus* and the relatives *A. quitensis* and *A. hybridus* in different areas of the distribution range, not only from populations included in this study, could explain a higher genetic diversity in cultivated amaranth. This is also consistent with the strong network structure (Figure 3.4) and the TreeMix analysis (Figure 3.5). In summary, cultivated *A. caudatus* is unusual in its higher overall genetic diversity compared to populations of its putative wild ancestors originating from the same geographic region. The high genetic diversity of *A. caudatus* is in contrast to other domesticated plants and suggests that a domestication bottleneck in its cultivation history absent (i.e., no domestication), very weak or masked by recurrent gene flow. We consider these results to be robust because in comparison to previous work (Maughan *et al*, 2009b, 2011; Khaing *et al*, 2013; Jimenez *et al*, 2013; Kietlinski *et al*, 2014), our study includes a larger number of accessions and more genetic markers. This allowed us to assess the genetic diversity and population structure

of South American amaranth on a genome-wide basis, but we note that a more complete geographic sampling of all cultivated amaranths and their relatives is required for a complete understanding of amaranth history.

5.4. Evidence for a Weak Domestication Syndrome

Despite their long history of cultivation, diverse uses for food and feed and a high importance for the agriculture of ancient cultures (i.e., *A. hypochondriacus* during the Aztec period), grain amaranths do not display the classical domestication syndrome as strongly as other crops (Sauer, 1967). Cultivated amaranth shows morphological differentiation from putative wild ancestors like larger and more compact inflorescences (Sauer, 1967) and a level of genetic differentiation (Table 3.2) which is comparable to other domesticated crops and their wild relatives (Sunflower: F_{ST}=0.22 (Mandel *et al*, 2011); common bean: 0.1-0.4 (Papa *et al*, 2005), pigeonpea: 0.57-0.82 (Kassa *et al*, 2012)). However, the numerous amaranth flowers mature asynchronously and produce very small seeds that are shattered (Brenner *et al*, 2010). All putative wild amaranths have dark brown seeds, whereas the predominant seed color of cultivated grain amaranth is white, which suggests that selection for seed color played a role in the history of the latter. Dark-seeded accessions are present in all three groups of *A. caudatus* defined by the genotypic data, which indicates that white seed color is not a fixed trait. Seed sizes between cultivated amaranth and its relatives are not significantly different with the exception of white-seeded *A. caudatus* accessions from Bolivia (Figure 3.7), which have larger seeds. The larger seeds in this group and of white seeds in general (Table 3.4) indicates past selection for domestication-related traits, but only in specific geographic regions or in certain types of amaranth, and not in the whole cultivated crop species.

These findings suggest that some selection occurred in the history of amaranth cultivation that may reflect domestication. Possible explanations for the incomplete fixation of domestication traits in South American grain amaranth include weak selection, genetic constraints or ongoing gene flow. First, weak selection of putative domestication traits indicate that they were not essential for cultivation. Although white seeds are predominant

in cultivated amaranth and most likely a selected trait, other seed colors may have been preferred for different uses. Since amaranths are also an important leaf vegetable in Mexico, the grain use of *A. caudatus* may not have been a main target of selection during domestication, thereby allowing a diversity of seed traits due to weak or incomplete selection. Second, genetic constraints may limit phenotypic variation in domestication traits. In contrast to genes with strong pleiotropic effects or epistatic interactions, domestication genes that are part of simple molecular pathways, have minimal pleiotropic effects, and show standing functional genetic variation have a higher chance of fixation by selection (Doebley *et al*, 2006; Lenser & Theißen, 2013). Numerous genes with these characteristics were cloned and characterized in major crops like rice, barley and maize. They contribute to the distinct domestication syndrome such as a loss of seed shattering, larger seed size and compact plant architecture. The molecular genetics of amaranth domestication traits remains unknown, but the absence a strong domestication syndrome may reflect genetic constraints despite a long period of cultivation. A third explanation is ongoing gene flow between cultivated amaranth and its relatives that may prevent or delay the formation of a distinct domestication syndrome and contributes to the high genetic diversity (Table 3.3), similar seed size (Figure 3.7 C), and the presence of dark seeds (Figure 3.7) in cultivated amaranth. Both historical and ongoing gene flow are likely because amaranth has an outcrossing rate between 5% and 30% (Jain *et al*, 1982; Stetter *et al*, 2016). In South America, cultivated amaranth and its relatives are sympatric over wide areas and the latter were tolerated in the fields and home gardens with *A. caudatus* (Sauer, 1967), where they may have intercrossed. Gene flow between wild and domesticated plants has also been observed in maize and teosinte in the Mexican highlands, but did not have a major influence on the maize domestication syndrome (Hufford *et al*, 2013). Further support for ongoing gene flow in amaranth is given by the presence of hybrids and admixed accessions in our sample with evidence for genetic admixture and dark seeds that demonstrate the phenotypic effects of introgression. Since the dark seed color is dominant over white color (Kulakow *et al*, 1985) and *A. caudatus* is predominantly self-pollinating, dark seeds could have efficiently removed by selection despite gene flow. Additionally, amaranth

was grown in small plots in the Andean highlands, which favors fixation of traits (Kauffman & Weber, 1990). Thus, gene flow is a plausible explanation for the absence of a distinct domestication syndrome.

Although our sample does not include *A. hypochondriacus* or *A. cruentus* accessions, our data are consistent with the model by Kietlinski *et al* (2014) who proposed the domestication of *A. caudatus* and *A. hypochondriacus* from different *A. hybridus* lineages in Central and South America (Figure 3.1D). Gene flow between *A. caudatus* and its close relative *A. quitensis* in the Southern distribution range (Peru and Bolivia) may explain the higher genetic diversity of the latter despite a strong genetic differentiation.

6. Conclusions

The genotypic and phenotypic analysis of cultivated South American grain amaranth and its close relatives suggests that *A. caudatus* is an incompletely domesticated crop species. Key domestication traits such as the shape of inflorescences, seed shattering and seed size are rather similar between cultivated amaranths and their close relatives and there is strong evidence of ongoing gene flow between these species despite selection for domestication-related traits like white seeds. Grain amaranth is an ancient crop of the Americas but genomic and phenotypic signatures of domestication differ from other, highly domesticated crops that originated from single domestication events like maize (Hake & Ross-Ibarra, 2015). In contrast, the history of cultivated amaranth may include multiregional, multiple and incomplete domestication events with frequent and ongoing gene flow from sympatric relatives, which is more similar to the history of species like rice, apple or barley (Londo *et al*, 2006; Cornille *et al*, 2012; Poets *et al*, 2015). The classical model of a single domestication in a well-defined center of domestication may not sufficiently reflect the history of numerous ancient crops. Our study further highlights the importance of a comprehensive sampling to study the domestication of amaranth. The three cultivated amaranths and all close relatives should be included in further studies for a full understanding of amaranth domestication and its

broader implications for crop plant domestication.

Acknowledgments

We thank David Brenner (USDA-ARS) and Julie Jacquemin for discussions and Elisabeth Kokai-Kota for support with the GBS library preparation and sequencing. The work was funded by an endowment of the Stifterverband für die Deutsche Wissenschaft to K. J. S.

Data Accessibility

The original genomic data will be available on the European Nucleic Archive (ENA). Scripts and phenotypic raw data are available under Dryad (http://datadryad.org/).

Author Contributions

M.G.S. and K.J.S. designed research; M.G.S. and K.J.S. performed research; T.M. contributed analytic tools; M.G.S. analyzed data; and M.G.S. and K.J.S. wrote the paper.

Conflict of interest

The authors declare no conflict of interest.

4. Crossing Methods and Cultivation Conditions for Rapid Production of Segregating Populations in Three Grain Amaranth Species

Markus G. Stetter*, Leo Zeitler, Adrian Steinhaus, Karoline Kröner, Michelle Biljecki and Karl J. Schmid
Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany

Published in
Frontiers in Plant Science 2016,7:816
doi: http://dx.doi.org/10.3389/fpls.2016.00816

Available as preprint on
PeerJ preprints

doi: https://doi.org/10.7287/peerj.preprints.1951v1

* Corresponding author: markus.stetter@uni-hohenheim.de

1. Abstract

Grain amaranths (*Amaranthus* spp.) have been cultivated for thousands of years in Central and South America. Their grains are of high nutritional value, but the low yield needs to be increased by selection of superior genotypes from genetically diverse breeding populations. Amaranths are adapted to harsh conditions and can be cultivated on marginal lands although little is known about their physiology. The development of controlled growing conditions and efficient crossing methods is important for research on and improvement of this ancient crop. Grain amaranth was domesticated in the Americas and is highly self-fertilizing with a large inflorescence consisting of thousands of very small flowers. We evaluated three different crossing methods (open pollination, hot water emasculation and hand emasculation) for their efficiency in amaranth and validated them with genetic markers. We identified cultivation conditions that allow an easy control of flowering time by day length manipulation and achieved flowering times of four weeks and generation times of two months. All three different crossing methods successfully produced hybrid F_1 offspring, but with different success rates. Open pollination had the lowest (10%) and hand emasculation the highest success rate (74%). Hot water emasculation showed an intermediate success rate (26%) with a maximum of 94% success. It is simple to perform and suitable for a more large-scale production of hybrids. We further evaluated 11 single nucleotide polymorphism (SNP) markers and found that they were sufficient to validate all crosses of the genotypes used in this study for intra- and interspecific hybridisations. Despite its very small flowers, crosses in amaranth can be carried out efficiently and evaluated with inexpensive SNP markers. Suitable growth conditions strongly reduce the generation time and allow the control of plant height, flowering time and seed production. In combination, this enables the rapid production of segregating populations which makes amaranth an attractive model for basic plant research but also facilitates further the improvement of this ancient crop by plant breeding.

Keywords: Amaranth, hybridization, hot water emasculation, hand emasculation, genetic resources, marker assisted breeding

2. Introduction

Ancient crops from the Americas such as quinoa (*Chenopodium quinoa* willd.) or amaranth (*Amaranthus* spp. L.) are a valuable addition to the human diet because of their high nutritional value. These pseudocereals have a high protein content and are rich in lysine and other essential amino acids that are limited in other grains. (Vega-Gálvez *et al*, 2010; Rastogi & Shukla, 2013). In addition, these crops are well adapted to harsh environmental conditions and are therefore suitable for cultivation on marginal soils. Their yields are significantly lower than those of major crops due to a lack of plant breeding (Alemayehu *et al*, 2015), but the presence of a high genetic and phenotypic diversity in these species indicates an excellent potential for breeding and variety development (Brenner *et al*, 2010).

Grain amaranth originated from Central and South America, where it was of great importance in pre-columbian agriculture until its cultivation strongly declined after the Spanish conquest (Sauer, 1967; Brenner *et al*, 2010; Kauffman & Weber, 1990). Three species of *Amaranthus* are cultivated for grain production: *A. caudatus* L., *A. cruentus* L. and *A. hypochondriacus* L.. Amaranth expresses the C_4 carbon cycle, which is more common in grasses but rare in dicots. Despite a high genetic diversity (Stetter *et al*, 2015), breeding efforts in amaranth so far were limited to the selection of suitable genotypes from landraces. Amaranth is mainly self-pollinating and has numerous intricate flowers, which make crosses more difficult than in other crops. The ability to efficiently carry out crosses is an important requirement for plant research to understand genetic basis of relevant traits (Moose & Mumm, 2008; Olsen & Wendel, 2013). Crosses are equally important for plant breeding and are used to generate new genetic variation and to introgress exotic material into breeding populations.

In many crops, hybrid varieties are characterized by strongly increased yields (Duvick, 2001). The application of hybrid breeding is amaranth is also very promising, because a mid-parent heterosis of up to 88% has been reported (Lehmann *et al*, 1991). The ability to conduct crosses on a large scale with little effort is of central importance for the development and production of hybrid crop varieties. To use this potential in minor crops, an improvement of crossing methods is essential (Veerappan *et al*, 2014).

Several approaches for hybrid production are available, but for all methods the key step is to prevent self-fertilization by the male parent. This is either by using appropriate genetic self-incompatibility systems or by mechanical and chemical treatments that lead to male sterility. In several species, cytoplasmatic male sterility (CMS) systems prevent selfing of the female crossing partner (Laser & Lersten, 1972). To use CMS systems for breeding male sterile female parent and male parents with restorer genes are needed to allow seed production in the hybrid progeny. Additionally, a maintainer line is needed that allows multiplying the male sterile line without loosing the CMS. Male sterility has been reported in *A. hypochondriacus* but is not yet developed sufficiently to be used for breeding.(Peters & Jain, 1987). Mechanical emasculation methods are efficient if the male and female flower are well separated on the plant (e.g. as in maize) because then male flowers can be removed without interfering with the female inflorescence. In other crops like tomato and *Medicago*, anthers are removed before pollen shedding (Veerappan *et al*, 2014). Another physical method is the heat treatment of the flowers of the female parent to destroy the pollen, for example by a hot water treatment. Here, the temperature is crucial, as differences by few degrees can influence the efficiency of the emasculation (Otsuka *et al*, 2010; Mukasa *et al*, 2007; García-Yzaguirre & Carreres, 2008). Chemical gametocides are used in hermaphrodite crops for which no CMS systems are available or are too costly, for example in wheat (Dotlacil & Apltauerová, 1978). The grain amaranth species have male and female flowers on the same inflorescence where several female flowers are arranged circularly around a male flower (Figure 4.1). The flowers are less than 1 mm in diameter, which makes mechanical emasculation difficult. For this reason other emasculation methods such as a hot water treatment may be more efficient.

Frequently, crossing methods are not completely reliable and require the validation of progeny. Phenotypic traits with a dominant-recessive inheritance can be used to identify successful crosses. In amaranth, traits such as seed or leaf color differ between genotypes and are available for validation (Kulakow *et al*, 1985). For phenotypic traits to be useful, however, parents need to differ in at least one trait and the male parent needs to express the dominant allele. In contrast, molecular markers allow an ef-

ficient and early evaluation of crosses without restricting the combination of parents, and cost-efficient PCR-based marker systems are available for this purpose (Maughan *et al*, 2011).

Figure 4.1.: Flower morphology. Inflorescence of *A. caudatus* consisting of flower clusters in which a male flower in the center is surrounded by several female flowers.

For model plants it is important to take specific requirements of development into account. Amaranth shows a strong photoperiod sensitivity and starts to flower under short day conditions (Brenner *et al*, 2010). A single plant has the potential to produce several thousands of seeds and can therefore produce large populations. However, under field conditions amaranth plants are usually tall and require a significant amount of space for cultivation. If flowering time, plant size and seed production can be controlled in climate chambers and greenhouses, an efficient propagation of the plant may be possible.

In the work presented here we study the efficiency of three different crossing methods and describe environmental conditions in a controlled environment (growth chamber) to achieve efficient and rapid generation of progeny for genetic studies. We suggest a method for hybrid identification with cost efficient PCR- based markers. Subsequently, we apply our method to three species of amaranth to evaluate its potential for the wider application to species within the genus *Amaranthus*.

3. Methods

3.1. Plant Material and Growth Conditions

The amaranth accessions for testing the three crossing methods were selected to comprise accessions with green seedlings as female parent and accessions with red seedlings as male parent. Additionally, amaranth varieties were used to verify hybridization and the use of genetic markers (Table 4.2). Single seeds were planted in 7 x 7 cm pots in standard gardening soil. Plants were grown for 2 weeks under long day conditions (Table 4.1) before transferring them step-wise in weekly intervals to short day conditions (Table 4.1). This helped to synchronize flowering of different genotypes and spread workload for performing the crosses.

Table 4.1.: Growth conditions. Parameters for amaranth in growth chamber for long and short day conditions.

	day length	light intensity	temp day	temp night
Long day	16	150 mmol	35 °C	30 °C
Short day	8	150 mmol	30 °C	25 °C

3.2. Crossing Methods

We evaluated three different methods for crossing wild (*A. hybr.* and *A. hybridus*) and cultivated species of amaranth (Figure 4.2 and table 4.2). The first method was open pollination by fixing the flowers of the female and male parent to each other and protecting them with a pollen proof bag (Sealed Air, Germany) from cross pollination by other plants. The second method was a warm water treatment of the inflorescence during flower initiation of the first emerging flowers (García-Yzaguirre & Carreres, 2008). Female flowers were dipped into a water bath of 45 °C warm water for 10 min to emasculate the male flowers before proceeding as in the first

method. The water treatment was repeated after 7 days. The third method was hand emasculation. For this approach, female flowers that were already open and all male flowers were removed from the inflorescence. The tip of the inflorescence was also removed to prevent the emergence of new flowers. The emasculation was repeated after 7 days and any flowers that developed later were removed. For all three methods plants were shaken daily to increase pollen dispersal and to assure cross-fertilization.

Figure 4.2.: Crossing methods. Three crossing methods: (**A**) Hand emasculation by removal of male flowers from female plant. (**B**) Hot water emasculation by 10 min treatment with 45 °C water bath. (**C**) Fixing male and female flower to each other for better pollen transfer.

3.3. Success Evaluation and Statistical Analysis

Seeds of the female parent were harvested four weeks after crossing. For each cross 50 seeds were counted and planted in pots. Seedling color evaluation was performed two weeks after planting by counting green and red seedlings. The success rate was the ratio between red and green offspring. Data analysis to test the differences between methods and between crossing types was conducted with a Generalized Linear Model (GLM) with binomial variance and a logit link function that included the crossing method (α), the crossing type (β, Intra- and inter-specific) and the interaction as factors:

$$logit(\mu_{ij}) = log(\frac{\mu_{ij}}{1-\mu_{ij}}) = \eta_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}. \qquad (4.1)$$

The calculation was done with the R statistical package version 3.2.0 using the `stats` library.

3.4. DNA Extraction

For genotyping the DNA was extracted with EconoSpin® columns (Epoch Life Science Inc.) using 1% CTAB extraction buffer (Saghai-Maroof *et al*, 1984). Dry leaf samples were homogenized and incubated for 2 h at 50 °C in 400 μl 1% CTAB extraction buffer and 4 μl Proteinase K. After addition of 300 μl Ammonium acetate (7.5 M) and 300 μl Ethanol (96%), the samples were centrifuged for 1 min at full speed. Then 800 μl of the supernatant were transferred on a EconoSpin® column placed in the collection tube and centrifuged for 1 min. The flow through was discarded. The columns were washed twice with wash buffers from (Saghai-Maroof *et al*, 1984) before eluting DNA twice with 50 μl Tris-HCl (10 mM, pH 8).

Table 4.2.: Parental genotypes for 11 KASP marker assays. HEX and FAM are the fluorescence dyes associated with each allele. Markers were tested in 10 individuals of which 8 were used for crosses

ID	Name	species	AM17978	AM19584	AM19963	AM21336	AM21605	AM22029	AM22341	AM24451	AM24579	AM25548	AM26171
26	PI 642741	*A. caudatus*	FAM	HEX	HEX	HEX	HEX	HEX	FAM	FAM	-	HEX	HEX
34	PI 511679	*A. caudatus*	FAM	HEX	FAM	HEX	FAM	HEX	HEX	HEX	FAM	HEX	HEX
37	PI 649220	*A. caudatus*	FAM	HEX	FAM	HEX	FAM	HEX	FAM	FAM	FAM	HEX	HEX
117	PI 511684	*A. hybr.*	FAM	HEX	FAM	HEX	FAM	HEX	HEX	FAM	FAM	HEX	HEX
174	PI 649623	*A. hypochondriacus*	HEX	FAM	FAM	-	FAM	FAM	HEX	HEX	HEX	FAM	FAM
245	Baernkrafft	*A. cruentus*	FAM	FAM	FAM	FAM	FAM	HEX	HEX	HEX	HEX	FAM	HEX
246	C6	*A. cruentus*	FAM	HEX	FAM	FAM	FAM	HEX	HEX	HEX	HEX	FAM	HEX
247	Puerto Moutt	*A. cruentus*	FAM	FAM	FAM	FAM	FAM	HEX	HEX	HEX	HEX	FAM	HEX
248	Pastewny	*A. hybridus*	-	FAM	FAM	-	FAM	FAM	HEX	HEX	HEX	-	FAM
369	PI 511695	*A. caudatus*	FAM	HEX	HEX	HEX	HEX	HEX	FAM	FAM	FAM	HEX	HEX

3.5. Evaluation of Genetic Markers

Eleven KASP assays (LGC Berlin/Germany) were selected from Maughan *et al* (2011) to validate crosses. The assays were prepared with 5 μl DNA (10 ng/μl) and 5 μl genotyping mix and run on the LightCycler® 480 Instrument II (Roche Life Science) with standard settings as given by the KASP manual (LGC Berlin/Germany) and analyzed using the LightCycler® 480 Software. First, parental lines were evaluated to find polymorphic markers for each of the crosses. Later, these markers were used to validate the crosses. For a proof of concept we genotyped offspring that were evaluated before by their seedling color. Both offspring with green (selfed plants) and red (hybrids) were genotyped.

3.6. Additional Hybrid Production

The previously evaluated hand emasculation method was used to produce additional hybrids. Plants were grown as described above, but crossing partners were not restricted to different seedling colors. The success of the crosses was validated with SNP markers.

4. Results and Discussion

4.1. Cultivation and Life Cycle

In the field the generation time of the three grain amaranth species is approximately six months and leads to very tall plants with thousands of flowers. To reduce the generation time, plant height and number of flowers, we cultivated the plants under short day conditions (8 h) and high temperature (30 °C) which both induced early flowering four weeks after planting. Additionally, we controlled the initiation of flowering by transferring plants from long day (16 h, 35 °C) conditions to short day conditions. Under long day conditions the plants displayed strong vegetative growth and did not flower within ten weeks after planting, but started flowering approximately 14 days after a transfer to short day conditions. The step-wise transfer of plants from long to short-day conditions allows

the production of plants in different flowering stages, which greatly facilitates synchronous flowering for crosses between genotypes that differ in their flowering time. This treatment is further useful to produce male parents that are able to shed large amounts of pollen when females parents start flowering. As soon as four weeks after flowering, mature seeds could be harvested. By employing these treatments, very short generation times can be achieved that allow up to six generations per year, which is comparable to the model plant *Arabidopsis thaliana*. In addition, plant height and seed number per plant can be controlled by adjusting growth conditions. Long day conditions lead to more vegetative growth, later flowering and more seeds, while short day conditions result in small early flowers. This is useful for different applications, because crosses require only few flowers, whereas the resulting F_1 plants should produce larger amounts of seeds (e.g. for creating mapping populations).

4.2. Crossing Methods

We compared three crossing methods that included open pollination, hot water emasculation and hand emasculation. All three methods produced successful crosses, but the success rates and variances differed strongly between the methods (Table 4.3). The amount of seeds produced did not substantially differ between methods and the mother plants produced between 100 and 200 seeds. Open pollination between two plants under a single bag without emasculation of the female parent led to a mean success rate of 10% with a standard deviation (s.d.) of 0.05. The hot water treatment of the female parent led to a significantly increased success rate of 26% but with a very high deviation (s.d. = 0.35) and a minimal success rate of 0%. However, the maximal success after hot water treatment was 94%, which shows that the method has a high potential if the key conditions for a successful application can be identified. We sterilized flowers at 45 °C and an adaptation of temperature may contribute to a higher rate of success. In other species (e.g. Acacia, buckwheat and rice), different temperatures change the efficiency of emasculation (Otsuka *et al*, 2010; Mukasa *et al*, 2007; García-Yzaguirre & Carreres, 2008). A temperature of 45 °C for emasculation is rather high compared to other crops

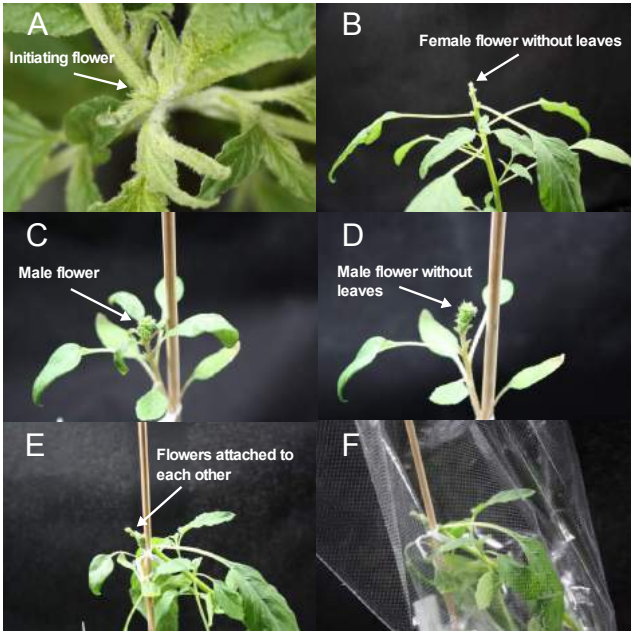

Figure 4.3.: Hand crossing procedure. (**A**) Flower initiation of female plant. (**B**) Female plant prepared for crossing. Leaves near the flower are removed. (**C**) Male crossing partner with first open male flowers. (**D**) Male plant prepared for crossing. Leaves near the flower are removed for improved pollen exchange. (**E**) Female and male crossing partners attached to each other. (**F**) Crossing partner are isolated with pollen proof bag to avoid contamination by foreign pollen.

(García-Yzaguirre & Carreres, 2008), but not too high because the amaranth plants still set seeds after this treatment and a further optimization may be achieved by varying the length of the heat treatment. Overall, hot water emasculation works with amaranth and, if it can be further improved, is suitable for application in the field to large numbers of plants.

The most elaborated and time consuming method we evaluated was hand emasculation (Figure 4.3). The mean success rate of 74% was the highest of the three methods and the deviation (s.d. = 0.29) was lower than of the heat treatment. The minimum success was comparable to free pollination, but the maximum success was up to 100%. Hand emasculation is difficult because amaranth has many small flowers and each male flower sheds enough pollen to pollinate a whole plant. Therefore it is critical to remove all male flowers from the female parent before flower dehiscence. The deviation can be decreased by keeping only few flower clusters per plant. We also tested whether intra- and inter-specific crosses are different in their efficiency, but there was no significant difference between intra- and inter-specific crosses (Table 4.3). This shows that inter-specific hybridization is possible, but as the two species are closely related this might not be the case for distant member of the *Amaranthus* genus.

Table 4.3.: Success rate of different crossing methods. Success rates and standard deviation (SD) for different crossing methods based on seedling color of 50 offspring per sample. The mean was calculated on basis of four to seven crosses (N) per method and crossing type. Intra-specific crosses were performed with *A. caudatus* (PI 511679 x PI 649220) and inter-specific crosses between *A. caudatus* (PI 511679) and *A. hybr.* (PI 511684). A Generalized Linear Model (GLM) with binomial variance and a logit link function were used to analyze differences between methods. Different letters show significant differences between methods. There was no significant difference between intra- and inter-specific crosses.

	Type	N	Mean (%)	SD (%)	Minimum (%)	Maximum (%)
Open pollination		7	10 ^{c}	5	4	18
	intra-specific	3	11	3	8	14
	inter-specific	4	10	6	4	18
Heat treatment		8	26 ^{b}	35	0	94
	intra-specific	4	26	45	0	94
	inter-specific	4	27	27	0	57
Hand emasculation		11	74 ^{a}	29	17	100
	intra-specific	4	80	20	50	94
	inter-specific	7	71	34	17	100

The comparison of the three methods shows that open pollination had low success rates, whereas heat treatment can be an effective and simple method for crosses if many seeds are required and simple morphological markers are available for the evaluation of offspring. Hand emasculation by well trained personnel shows the best performance. Since amaranth plants may produce thousands of seeds, a single successful cross can produce large F_2 populations, and the number of hand crosses needed can be kept low, which decreases the work load of the method substantially and makes it suitable for large projects.

4.3. Genetic Markers for Hybrid Identification

Since no crossing method provides a 100% success rates, unsuccessful crosses have to be excluded in early stages. Furthermore, crossing partners should not be limited by phenotypic differences in certain traits (e.g. different seedling color), but all possible combinations parents should be available. We therefore evaluated all accessions used in this study with 11 PCR-based SNP markers. The markers were the most polymorphic from a set of 411 KASP markers from Maughan *et al* (2011). Each marker was polymorphic between at least two lines and each cross segregated at least for one marker (Table 4.2). After evaluating the parental lines, we selected suitable markers to evaluate crosses.

First, we investigated progeny which had already been evaluated by their seedling color, because we expected green seedlings to be homozygous for the maternal allele since the green allele is recessive, and red seedlings to be heterozygous. For example, the application of marker AM22341 in a cross of PI511679 x PI649220 showed that green seedlings were homozygous for the allele of parent PI511679 and red seedlings were heterozygous for both parental alleles (Figure 4.4A). Frequently, the same marker can be used in several crosses, which allows the evaluation of more than one cross simultaneously (Figure 4.4B). This strongly decreases the work load and the cost of the evaluation. When working with homozygous parental lines a single maker is sufficient to validate successful crosses.

To test the effectiveness of the hand crossing method and the validation with genetic markers, we produced hybrids between amaranth genotypes

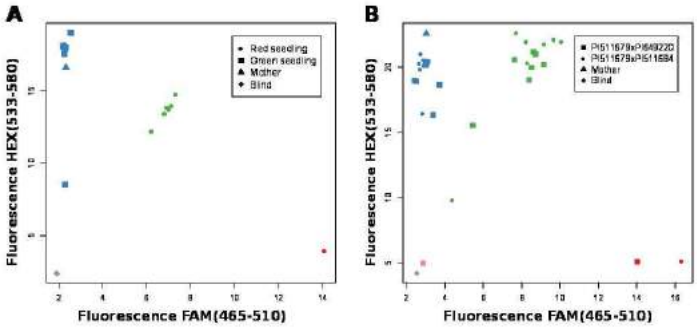

Figure 4.4.: SNP genotyping for known crosses. (**A**) Validation of PI511679xPI649220 with AM22341 and comparison with seedling color. (**B**) Validation of two crosses with AM24451.

from different species and validated them with the marker system. Although not all crosses produced hybrids, for most crosses the number of hybrids produced was high and less than 10 offspring had to be evaluated per cross (Table 4.4).

5. Conclusions

Ancient and underutilized crops greatly benefit from the ongoing revolution in genomics. However, to utilize this information for the improvement of minor crops, efficient crossing methods which are the basis of breeding programs need to be established. We developed crossing methods and genetic markers for hybrid identification in amaranth and showed that these can be used for crosses within and between species. We further showed that the life cycle and plant size of amaranth can be reduced substantially when light and temperature conditions are adapted. For genetic and physiological studies a short generation time is advantageous, which is a common characteristic of model organisms for basic research. Under the con-

Table 4.4.: Crosses of different amaranth varieties by hand emasculation and evaluation of success rates with SNP markers. All crosses are interspecific crosses between the three grain amaranths and/or their putative ancestors. The ID of mothers and fathers corresponds to Table 4.2.

	ID (Mother)	ID (Father)	Marker	Genotyped	Selfings	Successful crosses	Failed assay
1	34	245	AM19584	7	0	7	0
2	34	245	AM19584	6	2	2	2
3	34	245	AM19584	7	2	4	1
4	34	248	AM19584	7	2	4	1
5	34	248	AM19584	7	0	6	1
6	34	248	AM19584	7	0	4	3
7	245	26	AM19584	6	4	2	0
8	245	26	AM19584	8	2	3	3
9	247	248	AM22029	10	1	9	0
10	247	248	AM22029	9	9	0	0
11	248	245	AM22029	6	6	0	0
12	248	245	AM22029	8	5	0	3
13	248	245	AM22029	7	4	0	3

ditions described here, generation times as short as those of *A. thaliana* are possible (Meyerowitz & Pruitt, 1985). Additionally, the amount of seeds can be controlled, which allows the production of large offspring populations for genetic mapping. Furthermore, amaranth has a relatively small genome (500 Mbp) with a reference sequence, and a large number of genotyped genebank accessions are available (Clouse *et al*, 2016; Stetter *et al*, 2015). Taken together, these resources and the possibility of interspecific crosses make the grain amaranth species a very suitable model organism for studying fundamental processes such as adaptation, speciation, heterosis, C_4 photosynthetic metabolism, or domestication. The ability to conduct crosses from genetically diverse material facilitates the establishment of advanced breeding programs and the selection of improved genotypes using current breeding methods such as genomic selection will improve the value of this minor crop for agricultural production.

Conflict of Interest Statement

The authors declare that they have no competing interests.

Author Contributions

KJS and MGS designed the experiments. LZ, AS, KK and MB performed the crosses. MGS performed the genotyping and analyzed the data. KJS and MGS wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We would like to thank the Institute of Crop Science of the University of Hohenheim for providing the amaranth varieties (Baernkrafft, C6, Puerto Moutt and Pastewny) and the USDA-ARS for providing germplasm. This work was funded by an endowment of the Stifterverband der Deutschen Wissenschaft grant to KJS (Project H1705701502016296). MGS, LZ, AS, KK and MB were supported by the Humboldt Reloaded project at the University of Hohenheim, Germany (BMBF 01PL11003) .

5. General Discussion

This thesis extends the previous knowledge existing on amaranth and advances amaranth research into the genomic era. The techniques developed here provide new data and methods for future research and applied amaranth breeding. The results presented in this work also provide further insights into the domestication process of crops in general.

1. Amaranth as Model for Crop Domestication

Over the last 12,000 years, about 2,500 plants have been partially or fully domesticated. Domestication history differs between crops, but follows common patterns. The number of domestication events also differs between crops and ranges from one in maize to up to nine in olive (*Olea europaea ssp. europaea var. sativa*) (Matsuoka *et al*, 2002; Breton *et al*, 2009). Our studies on grain amaranth (*A. caudatus*, *A. cruentus* and *A. hypochondriacus*) suggest that there was more than one domestication attempt in amaranth. Early archaeological findings (Arreguez *et al*, 2013) and the high importance in early cultures (Brenner *et al*, 2010) suggest that amaranth fulfills the requirements for complete domestication. However, in contrast to the major crops, the domestication of amaranth appears to be incomplete, because the domestication syndrome is only weakly pronounced. While successful domestication processes have been studied more intensively, there is evidence that incomplete or failed domestication, as observed for grain amaranth, is a common phenomenon in crop and domestic animal evolution (Gaut *et al*, 2015). Examples of partially domesticated crops include carrot (Iorizzo *et al*, 2013) and apple (Cornille *et al*, 2012).

In animals some close relatives of our domestic animals have not been domesticated (e.g. zebra or antelope). Six main conditions have been described that have to be fulfilled for successful animal domestication. However, these conditions are mainly linked to character traits and therefore do not apply to plants (Diamond, 2002). In plants, weak selection, genetic constraints, or ongoing gene flow from wild relatives might be reasons for incomplete fixation of the domestication syndrome (Lenser & Theißen, 2013; Hufford *et al*, 2012). As discussed in chapter 3 there are diverse other possible reasons for such an incomplete domestication in amaranth. While the treemix analysis provides evidence for gene flow and the low phenotypic differentiation between wild and cultivated amaranth might be a sign for weak selection, our study did not analyze potential genetic constraints (chapter 3). Genetic constraints, such as a complex genetic architecture or little standing variation on domestication alleles, should be investigated to understand domestication in the three grain amaranth species. Several scenarios of domestication have been suggested for grain amaranth (Sauer, 1967; Kietlinski *et al*, 2014, Figure 3.1). The phylogeny of *Amaranthus* shows the clear relationship between the three grain amaranths and their two potential ancestors *A. hybridus* and *A. quitensis*.

The grain amaranth, *A. caudatus*, showed a close relationship to *A. hybridus*, but also to *A. quitensis*. Thus, *A. quitensis* might be directly or indirectly (through recurrent gene flow) involved in the domestication of *A. caudatus* as previously suggested by Sauer (1967). Although the previous hypothesis of a single domestication of *A. hypochondriacus* and *A. caudatus* from *A. hybridus* (Kietlinski *et al*, 2014) cannot be completely ruled out by our results, the phylogeny shows that *A. hypochondriacus* and *A. cruentus* are clearly separated from *A. caudatus*. Separate domestication events of South and Central American grain amaranth from two lineages of *A. hybridus* appear to fit better to our data and are supported by other studies (Kietlinski *et al*, 2014; Jimenez *et al*, 2013, chapter 3). Conclusive sampling from all three grain species and the two close wild relatives, and whole genome sequencing studies will be necessary to formally test the domestication models and identify the correct scenario of evolutionary history.

Amaranth is well-suited to serve as model to study incomplete domesti-

cation and potential experimental domestication, because it combines several features that are present in different crops. First, the three species of cultivated amaranth appear to have been domesticated independently in different regions of the Americas, although all three derive from the same wild species, *A. hybridus*. Secondly, there has been gene flow from *A. quitensis*, a wild species from South America, into *A. caudatus*. Gene flow with wild relatives outside the core area or a crop has been described as a common feature during the spread of crops after initial domestication (Gaut *et al*, 2015, chapter 3). A third advantage is that the genome of amaranth is relatively small (500 Mbp) and diploid, making it easy to study potential genetic constraints for domestication.

2. *Amaranthus ex-situ* Conservation

The conservation of crops and their wild relatives is important to provide material for future breeding programs aiming for the adaptation of crops to changing environments. To conserve the high genetic diversity of amaranth efficiently and improve the use of *ex-situ* collections by farmers and breeders, more accessions from different genebanks should be genotyped. This is of special importance for genebanks located in the centers of diversity, for which diversity of the material remains largely unknown, as information is scarce and access to the material is difficult. The genetic diversity is especially high there, because different landraces were cultivated in these regions over a long time (Brenner *et al*, 2010; Jimenez *et al*, 2013). The drop of genotyping prices with methods such as GBS, allows to genotype whole collections within genebanks (Song *et al*, 2015). There are over 6,300 *Amaranthus* accessions in the Genesys database that combines data from the largest genebanks worldwide (www.genesys-pgr.org). A genetic analysis of all these accessions would reveal duplicates and genetically closely related individuals, but also allow to categorize accessions into the correct species. Whole genome sequencing is still costly and not practical yet for thousands of accessions, but genotyping could be used to create core collections that represent the diversity of the genus or single species (Belaj *et al*, 2012). Such core collections can then be used

for in depth analysis (e.g. sequencing and phenotyping). Core collections of the grain amaranths would be useful for studies that aim to identify the genetic control of domestication and agronomic traits using genome wide association mapping.

Our crossing experiments showed the importance of isolating accessions during seed regeneration, because outcrossing rates are high and may disturb the purity of accessions (chapter 4). Even intercrossing between different *Amaranthus* species is possible and could lead to a loss of the original genetic material in collections. This should be taken into account when reproducing material in *ex-situ* collections.

3. Future of *Amaranthus* Research

The previous chapters of this work describe some of the diverse research questions that can be addressed within the genus *Amaranthus*. While in model organisms basic techniques have been developed over decades, few methods have been developed and tested in *Amaranthus*. Studying the relationship between *Amaranthus* species and the genetic diversity of South American grain amaranth, we applied modern sequencing methods and developed new genetic resources that will be useful for future evolutionary studies and amaranth breeding. Amaranth fulfills many requirements to be a suitable model organism for evolutionary biology and domestication, but some essential resources, presented in the following chapters, are still missing.

3.1. Reference Genome

Major crops, such as maize or rice, have almost complete reference sequences available (Bukowski *et al*, 2015; Kawahara *et al*, 2013). High quality reference genomes are essential for a correct mapping of sequencing reads, functional genomic studies and to study structural variation of genomes. Although a draft genome sequence is available for *A. hypochondriacus*, it is not yet a high quality reference genome. The sequence consists of over 3,000 scaffolds that have not yet been assembled into the

16 chromosomes of the species (Clouse *et al*, 2016). The dropping costs for whole genome sequencing and single molecule sequencing techniques provide the opportunity to supply a high quality reference genome for several *Amaranthus* species in a near future.

3.2. Transformation

Genetic transformation has become a key technique in molecular biology for functional genomics and crop breeding. The transfer of foreign DNA constructs allows the integration of novel traits, as well as the knock-out of genes. Both are of crucial importance to study the genetic control of phenotypes and to verify QTLs discovered with GWAS or by QTL mapping (Hansen & Wright, 1999). Since the development of genome editing (Ma *et al*, 2016), transformation has gained even further importance, because it is required to introduce the T-DNA construct carrying Cas9 and the single guide RNA (sgRNA) into the plant (see following section). In *Amaranthus* successful transformation has been reported using floral dip with *Agrobacterium*, but with a low success rate of maximal 1.8% (Munusamy *et al*, 2013). Nevertheless, this is a promising starting point for further improvement of transformation in amaranth. Applying transformation techniques to grain amaranth will allow more detailed analysis of domestication and production related genes.

Genome Editing

While random mutagenesis with chemical reagents has been used for decades to introduce new variation and study gene function in plants, a more recent method of directed mutagenesis, genome editing, has been applied to several organisms. The development of genome editing methods provided a major breakthrough in genome manipulation (Bibikova *et al*, 2002; Urnov *et al*, 2010). Different genome editing methods have been applied in plants. Zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) require complicated constructs and have therefore been mostly replaced by the simpler and most widely used CRISPR/Cas9 system (Ma *et al*, 2016). The CRISPR/Cas9 technology originates from a

bacterial defense system against viruses and plasmids. The system has been adapted to artificially induce a double strand break, by introducing Cas9 and the appropriate sgRNA into a cell (Doudna & Charpentier, 2014).

This method could be used to answer various questions in amaranth, for example the evolution of herbicide resistance, observed in several species of *Amaranthus*. Resistances to different modes of action of frequently used herbicides have been reported in *A. tuberculatus*. *A. tuberculatus* shows resistance to herbicides that inhibit photosystem II (PSII), ALS, PPO, and 5-enolypyruvyl-shikimate-3-phosphate synthase (EPSPS) (Patzoldt *et al*, 2005; Trucco & Tranel, 2011). Resistance to glyphosate (which inhibits EPSPS) has become widespread in *A. palmeri* in recent years and leads to substantial yield losses in field crops (Culpepper *et al*, 2006). Using the CRISPR/Cas9 system, modifications in herbicide resistance genes could help to understand the changes and processes underlying resistance evolution. Another interesting use of the method would be the introduction of traits of interest into cultivated grain amaranths. For example, herbicide resistance in grain amaranth would strongly reduce the crop management efforts and increase amaranth production. Genome editing could also have a key role in studying and potentially completing the domestication of grain amaranth species, once underlying genes for seed size, shattering and seed color are identified.

4. Future of Amaranth Breeding

4.1. Breeding Techniques

Previously, structure and resources for efficient amaranth breeding programs were lacking. The advance in genotyping techniques including the systems employed in this work, GBS and KASP, could serve in different stages of breeding programs (chapters 3 and 4). Genome wide genetic markers, as generated with GBS, can be used to define genetic pools from landraces and *ex situ* collections when setting up breeding programs (Boeven *et al*, 2016). An initial pool of individuals that serve as parents,

out of which lines can be selected, has to be established for traditional line breeding. Although this could be done randomly, initial genotyping is advantageous to maximize genetic diversity and to use restricted resources, such as labor and field space, efficiently.

The genus phylogeny presented in this work is the most complete species tree constructed for *Amaranthus* so far. It reveals the close relatives of crop amaranths and defines distant gene pools for breeding (chapter 2). While we were able to produce crosses between members of the Hybridus complex, and natural hybrids between distant wild species have been reported, gene pools have yet to be tested by intentional crosses between species. As defined by the phylogenetic relationship and the crossing patterns, the secondary gene pool of grain amaranth might be the Hybridus complex, while the tertiary gene pool would be the *A. Amaranthus* subgenus. Nevertheless, species from the other two subgenera might be potential crossing partners, too.

Hybrid breeding in plants depends on heterotic groups which are the parental gene pools that when crossed, lead to superior offspring. When defining heterotic groups for hybrid breeding, it is important that the two pools are genetically distant. Traditionally, this has been achieved by separately developing pools over decades, often unintentionally. In crops, where such a separation did not exist, genetic makers can be used for heterotic group definition to set up hybrid breeding programs (Boeven *et al*, 2016). In running breeding programs, it is important to maintain genetic diversity despite the strong selection to improve breeding pools. For this, genome wide molecular marker systems provide an efficient way to monitor genetic diversity. External genetic resources, including crop wild relatives, can be introgressed into breeding pools to increase and maintain genetic diversity. Crossing methods presented earlier in this work, in combination with the KASP marker system, can be used to produce hybrids, generate genetic variation and introgress favorable alleles from genetic resources into breeding pools (chapter 4).

The classical way to identify quantitative trait loci (QTLs) is to use a segregating population from two diverging parents. This method has been proven to be efficient for traits with a simple genetic structure. Major QTLs in combination with their genetically linked markers can be used in

breeding programs (Collard *et al*, 2005). Marker assisted selection (MAS) and marker assisted backcrossing can efficiently be used to breed for resistances and other monogenic traits, while reducing linkage drag (Collard *et al*, 2008, 2005). For QTL mapping, hand emasculation can be used to create the segregating population before genotyping. Multiplexing of GBS allows genotyping hundreds of individuals at low cost, to have high statistical power for QTL detection (Liu *et al*, 2014). The KASP system can be used to identify the desired allele in breeding material or to reduce linkage drag (Neelam *et al*, 2013). GBS reads are suitable to create specific primers for KASP assays, because the sequence up- and downstream the SNP is known (Boutet *et al*, 2016).

Until now amaranth cultivation and breeding mainly took place in less developed parts of the world, where the access to advanced resources is limited. The crossing methods described here do not require high technological inputs and are therefore well-suited to be used in the Andean highlands and rural areas of Central America. Low cost genetic marker systems like the PCR based KASP system could also be implemented in such areas to increase breeding progress and yields.

The controlled growth conditions for accelerated growth, we developed during this project, and short generation cycles could be of particular interest, when combined with the recent development of genomic prediction. This method allows to evaluate individuals on the basis of genetic information without having to test them in the field. Only the so called training set is phenotyped and genotyped in order to calibrate the genome wide model, while the rest of the population only has to be genotyped (Desta & Ortiz, 2014). In contrast to QTL studies, where single markers with high effects are identified, each marker contributes to a genomic-estimated breeding value in genomic prediction. Selection can be done on basis of these breeding values in an early growth stage after genotyping. While amaranth plants stay very small under the presented controlled conditions and do not express traits that are desired for production, there is enough material for DNA extraction and further reproduction (chapter 4). The generation time with our controlled growth conditions decreases to four to six weeks instead of six months. GBS has proven to provide many genome wide markers that can be used for genomic selection and the multiplexing

allows low cost genotyping for a high number of breeding lines (Poland *et al*, 2012a). Genomic prediction has not been experimented in amaranth so far, although we believe that the use of genomic prediction could largely accelerate the progress in amaranth breeding.

Figure 5.1.: Biomass amaranth breeding trial. Breeding lines surrounded by Bärnkrafft (only variety registered in Germany). Strongly increased height and biomass within three years of breeding.

4.2. Breeding Goals

Traditionally, amaranth has been used as vegetable and pseudo-cereal, but also, in a limited extend, for animal feed. The growing need for renewable energies leads to new production opportunities for amaranth. In Germany, biogas has a high importance within the energy mix of renewable energies. In contrast to wind and solar power, electricity production from biogas is

independent from direct weather conditions. In Germany biogas production from field crops is mainly based on maize, with 73 % of biomass crops (900,000 ha in 2015; FNR 2015, biogas.fnr.de). Amaranth would be a suitable addition to crop rotation systems for biomass production (Figure 5.1). For example, *A. caudatus* is photo period sensitive and grows very high under Central European conditions and as earlier described amaranth shows strong heterosis on biomass (Lehmann *et al*, 1991). Breeding efforts for higher dry matter content and biomass are needed to develop this special commercial use of amaranth.
Further characteristics of amaranth could be of commercial interest. The extraction of oil protein from the seeds could be future uses of amaranth, which is rich in lysine and high quality oil (Rastogi & Shukla, 2013). These components could be interesting for pharmaceutical products, food supplements and the cosmetic industry. For these uses further investigation of grain composition and processing technologies is needed. A dual use of grain and plant biomass would be especially desirable, as the high value grains could be used for protein isolation and oil production instead of biomass. The various methods presented in this work provide resources for breeding efforts aiming at these uses and could be implemented into breeding programs.

6. Concluding Remarks

In this work next generation sequencing was successfully employed to create the most comprehensive phylogeny of the *Amaranthus* genus so far. The phylogeny, including 35 species of the genus, and the species genome sizes will be an important resource for future studies of crop and weed amaranth species. The multispecies coalescent method allowed to narrow down the potential ancestors of grain amaranth and provides hypotheses of domestication scenarios that can be modeled and further investigated. Studying the genetic and phenotypic diversity of South American grain amaranth and its potential wild ancestors showed that amaranth domestication was incomplete. Low selective pressure on the grain amaranths or recurrent gene flow from *A. quitensis* might have hindered domestication, but provide high genetic diversity in South American grain amaranth and a genetic structure that can be used for breeding. We were further able to present different crossing methods to produce hybrids within and between amaranth species. For the evaluation of crosses we suggest a cost efficient single marker system that can be applied in future breeding programs. Overall we have genotyped more than 200 amaranth gene bank accessions on a genome wide basis using GBS. This is the largest genomic resource for amaranth research and conservation so far. Although this work presents major advances in amaranth research, further studies with larger sampling and whole genome data will be needed to reveal reasons for incomplete domestication. The genus *Amaranthus* is a suitable model for fundamental evolutionary questions in plants, for instance C_4 carbon fixation, heterosis or evolution of herbicide resistance. While these questions are far from being solved, future studies can benefit from the results and resources presented in this work.

7. Acknowledgments

First I would like to thank my supervisor Prof. Karl Schmid who allowed me to work on this interesting topic and develop own ideas. Karl gave me high scientific freedom and his guidance allowed me to work independently and acquire methodological and scientific knowledge. I would like to thank PD Dr. Andreas Börner and Prof. Dr. Uwe Ludewig for evaluating this work and participation in the final colloquium. I express my gratitude to Thomas Müller for his collaboration on the domestication paper and for inheriting some of his useful scripts and commands. My thanks go to Julie and Annette for reading and correcting this work. I thank all the Postdocs and PhD students of the 350b group for the help with methods, feedback in lab meetings and many discussions. Elisabeth Kokai-Kota and Caro Lichthardt I thank for their help in the lab. I acknowledge the help of Viola Abraham, Max Haupt, Bettina Scherer, Leo Zeitler and Franz Spanner for their help with greenhouse and field experiments and all the other members of the group for support in the last years. Bedanken möchte ich mich auch bei Herbet Stelz, Roger Lürig und allen Mitarbeiten der Versuchsstationen für die gute Zusammenarbeit. Auch bei Hannelore Brandt, Margit Lieb, Susanne Meyer, Hans-Peter Maurer und Willmar Leiser möchte ich mich herzlich für die Unterstützung und Materialien bedanken. The Humboldt reloaded project at the University of Hohenheim supported the work on amaranth crosses and the eight undergraduate students did a great job on making this project a success.

Special thanks to my office mates over the last years, Ivan, Dounia, Julie, Raul, Bianca and Mireia, for keeping the heating low, allowing me to open the window and sharing cookies. Patrick and Max, I would like to acknowledge for the supply of delicious coffee and interesting discussions on general plant breeding topics. I thank Fabian for sharing some of his enormous mathematical knowledge with me and for discussing current

society issues after work.

Bei meinen Freunden, Leidensgenossen und ehemaligen Mitbewohnern möchte ich mich für die super Zeit in Hohenheim bedanken. Mein Dank gilt auch Meike und Paul, meinen Lieblingshipster, die mein Draht zum coolen Leben in der City waren und mir etwas Ablenkung vom Doktorandenleben geboten haben. Ganze besonders danke ich Meike für die vielen Kaffees in der Cafete und die Spaziergänge im Park. I would also like to thank my good friends Steve, Fernando and Quiz for their support.

Ich möchte mich bei meiner ganzen Familie bedanken, die mir immer mit Rat und Tat zur Seite standen und von denen mir jede/r auf ihre/seine Weise ein Vorbild ist. Bleibt wie ihr seid. Ganz besonderer Dank gilt natürlich meinen Eltern, die mich immer unterstützt haben und mir Vorbilder in Sachen Ehrgeiz, Fleiß und Durchhaltevermögen und vielem mehr sind. Most of all, I would like to thank Julie for supporting me during the process of this work with great food, an open ear and good advice.

Bibliography

Abbo S, van Oss RP, Gopher A, Saranga Y, Ofner I, Peleg Z (2014) Plant domestication versus crop evolution: a conceptual framework for cereals and grain legumes. *Trends Plant Sci.*, **19**, 351 – 360.

Adhikary D, Pratt DB (2015) Morphologic and taxonomic analysis of the weedy and cultivated *Amaranthus hybridus* species complex. *Syst. Bot.*, **40**, 604–610.

Alemayehu FR, Bendevis MA, Jacobsen SE (2015) The potential for utilizing the seed crop amaranth (*Amaranthus* spp.) in East Africa as an alternative crop to support food security and climate change mitigation. *J Agron Crop Sci*, **201**, 321–329.

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, **19**, 1655–64.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.*, **17**, 81–92.

Ariani A, Berny Mier y Teran JC, Gepts P (2016) Genome-wide identification of SNPs and copy number variation in common bean (*Phaseolus vulgaris* L.) using genotyping-by-sequencing (GBS). *Mol. Breed.*, **36**, 87.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.*, **22**, 3179–90.

Arreguez GA, Martínez JG, Ponessa G (2013) *Amaranthus hybridus* L. ssp. *hybridus* in an archaeological site from the initial mid-Holocene in the Southern Argentinian Puna. *Quat. Int.*, **307**, 81–85.

Baohua S, Xuejie Z (2002) Chromosome numbers of 14 species in *Amaranthus* from China. *Acta Phytotaxonomica Sinica*, **40**, 428–432.

Begun DJ, Holloway AK, Stevens K, *et al* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.*, **5**, e310.

Behera B, Patnaik SN (1982) Genome analysis of *Amaranthus dubius* mart. ex thell. through the study of *Amaranthus spinosus* x *A. dubius* hybrids. *Cytologia*, **47**, 379–389.

Beissinger TM, Wang L, Crosby K, Durvasula A, Hufford MB, Ross-Ibarra J (2016) Recent demography drives changes in linked selection across the maize genome. *Nat. Plants*, **2**, 16084.

Belaj A, Dominguez-García MdC, Atienza SG, *et al* (2012) Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. *Tree Genet. Genomes*, **8**, 365–378.

Bennett MD, Leitch IJ, Hanson L (1998) DNA amounts in two samples of angiosperm weeds. *Ann. Bot.*, **82**, 121–134.

Bennett MD, Smith JB (1991) Nuclear DNA amounts in angiosperms. *Phil. Trans. Roy. Soc. B*, **334**, 309–345.

Bennetzen JL, Wang H (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Ann. Rev. Plant Biol.*, **65**, 505–530.

Bensch CN, Horak MJ, Peterson D (2003) Interference of redroot pigweed (*Amaranthus retroflexus*), Palmer amaranth (*A. palmeri*), and common waterhemp (*A. rudis*) in soybean. *Weed Sci.*, **51**, 37–43.

Besnard G, Rubio de Casas R (2016) Single vs multiple independent olive domestications: the jury is (still) out. *New Phytol.*, **209**, 466–470. 2014-18870.

Bibikova M, Golic M, Golic KG, Carroll D (2002) Targeted chromosomal cleavage and mutagenesis in drosophila using zinc-finger nucleases. *Genetics*, **161**, 1169–1175.

Boeven PHG, Longin CFH, Würschum T (2016) A unified framework for hybrid breeding and the establishment of heterotic groups in wheat. *Theor. Appl. Genet.*, **129**, 1–15.

Bouckaert R, Heled J, Kühnert D, *et al* (2014) BEAST 2: A software platform for bayesian evolutionary analysis. *PLOS Comp. Biol.*, **10**, 1–6.

Boutet G, Alves Carvalho S, Falque M, *et al* (2016) SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RIL population. *BMC Genomics*, **17**, 121.

Brenner DM, Baltensperger DD, Kulakow PA, *et al* (2010) Genetic resources and breeding of *Amaranthus*. In *Plant Breeding Reviews*, pp. 227–285. John Wiley & Sons, Inc.

Brenner DM, Johnson WG, Sprague CL, Tranel PJ, Young BG (2013) Crop–weed hybrids are more frequent for the grain amaranth 'Plainsman' than for 'D136-1'. *Gen. Res. Crop Evol.*, **60**, 2201–2205.

Breton C, Terral JF, Pinatel C, Médail F, Bonhomme F, Bervillé A (2009) The origins of the domestication of the olive tree. *Comptes Rendus - Biologies*, **332**, 1059–1064.

Browning BL, Browning SR (2016) Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.*, **98**, 116–126.

Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A (2012) Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.*, **29**, 1917–1932.

Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.*, **21**, 255–65.

Bukowski R, Guo X, Lu Y, *et al* (2015) Construction of the third generation *Zea mays* haplotype map. *bioRxiv*, p. 026963.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Mol. Ecol.*, **22**, 3124–40.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping Loci de novo from short-read sequences. *G3*, **1**, 171–82.

Chan KF, Sun M (1997) Genetic diversity and relationships detected by isozyme and RAPD analysis of crop and wild species of *Amaranthus*. *Theor. Appl. Genet.*, **95**, 865–873.

Clouse JW, Adhikary D, Page JT, *et al* (2016) The amaranth genome: Genome, transcriptome, and physical map assembly. *Plant Genome*, **9**.

Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK (2005) An introduction to markers, quantitative trait loci (qtl) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, **142**, 169–196.

Collard BCY, Mackill DJ, B PTRS (2008) Marker-assisted selection : an approach for precision plant breeding in the twenty-first century. *Phil. Trans. R. Soc. B*, **363**, 557–572.

Coons M (1982) Relationships of *Amaranthus caudatus*. *Econ. Bot.*, **36**, 129–146.

Coons MP (1978) The status of *Amaranthus hybridus* L. in South America. *Cienc. Nat.*, **18**.

Cornille A, Gladieux P, Smulders MJM, *et al* (2012) New insight into the history of domesticated apple: Secondary contribution of the european wild apple to the genome of cultivated varieties. *PLoS Genet*, **8**, 1–13.

Costea M, DeMason D (2001) Stem morphology and anatomy in *Amaranthus* L. (*Amaranthaceae*), taxonomic significance. *J. Torrey Bot. Soc.*, **128**, 254–281.

Cruaud A, Gautier M, Galan M, *et al* (2014) Empirical assessment of RAD sequencing for interspecific phylogeny. *Mol. Biol. Evol.*, **31**, 1272–4.

Culpepper aS, Grey TL, Vencill WK, *et al* (2006) Glyphosate-resistant palmer amaranth (*Amaranthus palmeri*) confirmed in georgia. *Weed Sci.*, **54**, 620–626.

DaCosta JM, Sorenson MD (2016) ddRAD-seq phylogenetics based on nucleotide, indel, and presence–absence polymorphisms: Analyses of two avian genera with contrasting histories. *Mol. Phyl. Evol.*, **94**, 122–135.

Danecek P, Auton A, Abecasis G, *et al* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–8.

Das S (2014) Domestication, phylogeny and taxonomic delimitation in underutilized grain *Amaranthus* (Amaranthaceae) - a status review. *Feddes Rep.*, pp. 1–10.

Davis AS, Schutte BJ, Hager AG, Young BG (2015) Palmer amaranth (*Amaranthus palmeri*) damage niche in Illinois soybean is seed limited. *Weed Sci.*, **63**, 658–668.

Desta ZA, Ortiz R (2014) Genomic selection: Genome-wide prediction in plant improvement. *Trends Plant Sci.*, **19**, 592–601.

Diamond J (2002) Evolution, consequences and future of plant and animal domestication. *Nature*, **418**, 700–707.

Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. *Cell*, **127**, 1309–21.

Dohm JC, Minoche AE, Holtgräwe D, *et al* (2014) The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature*, **505**, 546–9.

Dolezel J, Bartos J, Voglmayr H, Greilhuber J (2003) Nuclear DNA content and genome size of trout and human. *Cytometry A*, **51**, 127–8; author reply 129.

Dolezel J, Sgorbati S, Lucretti S (1992) Comparison of three DNA fliioroclirome for flow cytometric estimatioe of nuclear DNA conteet in plants. *Physiol. Plant.*, **85**, 625–631.

Dotlacil L, Apltauerová M (1978) Pollen sterility induced by ethrel and its utilization in hybridization of wheat. *Euphytica*, **27**, 353–360.

Doudna JA, Charpentier E (2014) The new frontier of genome engineering with CRISPR-Cas9. *Science*, **346**.

Duvick DN (2001) Biotechnology in the 1930s: the development of hybrid maize. *Nat. Rev. Genet.*, **2**, 69–74.

Eaton DAR, Ree RH (2013) Inferring phylogeny and introgression using RADseq data: An example from flowering plants (*Pedicularis*: *Orobanchaceae*). *Syst. Biol.*, **62**, 689–706.

Elshire RJ, Glaubitz JC, Sun Q, *et al* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.

Fu YB, Peterson GW, Dong Y (2016) Increasing genome sampling and improving snp genotyping for genotyping-by-sequencing with new combinations of restriction enzymes. *G3*, **6**, 845–856.

García-Yzaguirre A, Carreres R (2008) Efficiency of different hybridization methods in single crosses of rice for pure line breeding. *Span. J. Agric. Res.*, **6**, 395–400.

Gaut BS, Díez CM, Morrell PL (2015) Genomics and the contrasting dynamics of annual and perennial domestication. *Trends Genet.*, **31**, 709 – 719.

Gepts P (2014) The contribution of genetic and genomic approaches to plant domestication studies. *Curr. Opin. Plant Biol.*, **18**, 51–9.

Greizerstein EJ, Poggio L (1994) Karyological studies in grain amaranths. *Cytologia*, **59**, 25–30.

Hake S, Ross-Ibarra J (2015) Genetic, evolutionary and plant breeding insights from the domestication of maize. *eLife*, **4**, 1–8.

Hansen G, Wright MS (1999) Recent advances in the transformation of plants. *Trends Plant Sci.*, **4**, 226–231.

Harvey MG, Smith BT, Glenn TC, Faircloth BC, Brumfield RT (2016) Sequence capture versus restriction site associated dna sequencing for shallow systematics. *Syst. Biol.*, **65**, 910–924.

Huang YF, Poland JA, Wight CP, Jackson EW, Tinker NA (2014) Using genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat. *PLoS One*, **9**, e102448.

Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J (2013) The genomic signature of crop-wild introgression in maize. *PLos Genet.*, **9**, e1003477.

Hufford MB, Xu X, van Heerwaarden J, *et al* (2012) Comparative population genomics of maize domestication and improvement. *Nat. Genet.*, **44**, 808–11.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, **23**, 254–67.

Iorizzo M, Senalik DA, Ellison SL, *et al* (2013) Genetic structure and domestication of carrot (*Daucus carota subsp. sativus*) (*Apiaceae*). *Am. J. Bot.*, **100**, 930–938.

Jain S, Hauptil H, Vaidya K (1982) Outcrossing rate in grain amaranths. *J. Hered.*, **73**, 71–72.

Jimenez FR, Maughan PJ, Alvarez A, *et al* (2013) Assessment of genetic diversity in Peruvian amaranth (*Amaranthus caudatus* and *A. hybridus*) germplasm using single nucleotide polymorphism markers. *Crop Sci.*, **53**, 532.

Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, **27**, 3070–1.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.*, **11**, 94.

Kassa MT, Penmetsa RV, Carrasquilla-Garcia N, *et al* (2012) Genetic patterns of domestication in pigeonpea (*Cajanus cajan* L. millsp.) and wild cajanus relatives. *PLoS One*, **7**.

Kauffman CS, Weber LE (1990) Grain Amaranth. In *Advances in new crops* (edited by J Janick, JE Simon), pp. 127–139. Timber Press, Portland.

Kawahara Y, de la Bastide M, Hamilton JP, *et al* (2013) Improvement of the *Oryza sativa* nipponbare reference genome using next generation sequence and optical map data. *Rice*, **6**, 1–10.

Khaing AA, Moe KT, Chung JW, Baek HJ, Park YJ (2013) Genetic diversity and population structure of the selected core set in *Amaranthus* using SSR markers. *Plant Breed.*, **132**, 165–173.

Kietlinski KD, Jimenez F, Jellen EN, Maughan PJ, Smith SM, Pratt DB (2014) Relationships between the weedy *Amaranthus hybridus* (*Amaranthaceae*) and the grain amaranths. *Crop Sci.*, **54**, 220.

Kolano B, McCann J, Orzechowska M, Siwinska D, Temsch E, Weiss-Schneeweiss H (2016) Molecular and cytogenetic evidence for an allotetraploid origin of *Chenopodium quinoa* and *C. berlandieri* (*Amaranthaceae*). *Mol. Phyl. Evol.*, **100**, 109–123.

Kulakow P, Hauptli H, Jain S (1985) Genetics of grain amaranths I. Mendelian analysis of six color characteristics. *J. Hered.*, **76**, 27–30.

Lanoue KZ, Wolf PG, Browning S, Hood EE (1996) Phylogenetic analysis of restriction-site variation in wild and cultivated *Amaranthus* species (*Amaranthaceae*). *Theor. Appl. Genet.*, **93**, 722–32.

Laser K, Lersten N (1972) Anatomy and cytology of microsporogenesis in cytoplasmic male sterile angiosperms. *Bot. Rev.*, **38**, 425–454.

Lehmann J, Clark R, Frey K (1991) Biomass heterosis and combining ability in interspecific and intraspecific matings of grain amaranths. *Crop Sci.*, **31**, 1111–1116.

Lenser T, Theißen G (2013) Molecular mechanisms involved in convergent crop domestication. *Trends Plant Sci.*, **18**, 704 – 714.

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

Li H, Handsaker B, Wysoker A, *et al* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lin T, Zhu G, Zhang J, *et al* (2014) Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.*, **46**, 1220—-1226.

Liu H, Bayer M, Druka A, *et al* (2014) An evaluation of genotyping by sequencing (GBS) to map the breviaristatum-e (ari-e) locus in cultivated barley. *BMC Genomics*, **15**, 104.

Londo JP, Chiang YC, Hung KH, Chiang TY, Schaal BA (2006) Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 9578–9583.

Lu F, Lipka AE, Glaubitz J, *et al* (2013) Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. *PLoS Genet*, **9**, 1–14.

Ma X, Zhu Q, Chen Y, Liu YG, Y-g L (2016) CRISPR/Cas9 platforms for genome editing in plants: Developments and applications. *Mol. Plant*, **9**, 961–974.

Mallory MA, Hall RV, McNabb AR, Pratt DB, Jellen EN, Maughan PJ (2008) Development and characterization of microsatellite markers for the grain amaranths. *Crop Sci.*, **48**, 1098.

Mandel JR, Dechaine JM, Marek LF, Burke JM (2011) Genetic diversity and population structure in cultivated sunflower and a comparison to its wild progenitor, *Helianthus annuus* L. *Theor. Appl. Genet.*, **123**, 693–704.

Mastretta-Yanes a, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC (2015) Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Mol. Ecol. Res.*, **15**, 28–41.

Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez G J, Buckler E, Doebley J (2002) A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 6080–6084.

Maughan P, Smith S, Fairbanks D, Jellen E (2011) Development, characterization, and linkage mapping of single nucleotide polymorphisms in the grain amaranths (*Amaranthus sp.*). *Plant Genome*, **4**, 92.

Maughan PJ, Yourstone SM, Jellen EN, Udall JA (2009a) Snp discovery via genomic reduction, barcoding, and 454-pyrosequencing in amaranth. *The Plant Genome*, **2**, 260–270.

Maughan PJ, Yourstone SM, Jellen EN, Udall Ja (2009b) SNP discovery via genomic reduction, barcoding, and 454-pyrosequencing in amaranth. *Plant Genome*, **2**, 260.

Meyer RS, DuVal AE, Jensen HR (2012) Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *New Phytol.*, **196**, 29–48.

Meyer RS, Purugganan MD (2013) Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.*, **14**, 840–52.

Meyerowitz EM, Pruitt RE (1985) *Arabidopsis thaliana* and Plant Molecular Genetics. *Science (New York, N.Y.)*, **229**, 1214–1218.

Moose SP, Mumm RH (2008) Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiol.*, **147**, 969–977.

Morris GP, Ramu P, Deshpande SP, *et al* (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 453–458.

Mosyakin SL, Robertson KR (1996) New infrageneric taxa and combinations in *Amaranthus* (*Amaranthaceae*). *Ann. Bot. Fenn.*, **33**, 275–281.

Mukasa Y, Suzuki T, Honda Y (2007) Emasculation of Tartary buckwheat (*Fagopyrum tataricum* Gaertn.) using hot water. *Euphytica*, **156**, 319–326.

Munusamy U, Nor S, Abdullah A, Aziz MA, Khazaai H (2013) Female reproductive system of *Amaranthus* as the target for agrobacterium -mediated transformation. *Adv. Biosci. Biotechnol.*, **4**, 188–192.

Nabholz B, Sarah G, Sabot F, *et al* (2014) Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (*Oryza glaberrima*). *Mol. Ecol.*, **23**, 2210–27.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Mol. Ecol.*, **22**, 2841–2847.

Neelam K, Brown-Guedira G, Huang L (2013) Development and validation of a breeder-friendly KASPar marker for wheat leaf rust resistance locus lr21. *Mol. Breed.*, **31**, 233–237.

Nicotra AB, Chong C, Bragg JG, *et al* (2016) Population and phylogenomic decomposition via genotyping-by-sequencing in Australian *Pelargonium*. *Mol. Ecol.*, **25**, 2000–2014.

Ohri D, Nazeer MA, M P (1981) Cytophotometric estimation of nuclear DNA in some ornamentals. *Nucleus*, **24**, 39–42.

Olsen KM, Wendel JF (2013) A bountiful harvest: genomic insights into crop domestication phenotypes. *Annu. Rev. Plant Biol.*, **64**, 47–70.

Otsuka J, Yamaguchi S, Chigira O, Kato K (2010) Application of hot water emasculation to *Acacia auriculiformis* for controlled pollination. *Journal of Forest Research*, **15**, 210–216.

Oyama RK, Clauss MJ, Formanová N, *et al* (2008) The shrunken genome of *Arabidopsis thaliana*. *Plant Sys. Evol.*, **273**, 257–271.

Papa R, Acosta J, Delgado-Salinas a, Gepts P (2005) A genome-wide analysis of differentiation between wild and domesticated *Phaseolus vulgaris* from Mesoamerica. *Theor. Appl. Genet.*, **111**, 1147–1158.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Park YJ, Nishikawa T, Matsushima K, Minami M, Tomooka N, Nemoto K (2014) Molecular characterization and genetic diversity of the starch branching enzyme (SBE) gene from *Amaranthus*: the evolutionary origin of grain amaranths. *Mol. Breed.*, **34**, 1975–1985.

Patzoldt WL, Tranel PJ, Hager AG (2005) A waterhemp (*Amaranthus tuberculatus*) biotype with multiple resistance across three herbicide sites of action. *Weed Sci.*, **53**, 30–36.

Paul S, Fe S, Paul S (2013) Phylogenetic signal variation in the genomes of *Medicago* (*Fabaceae*). *Syst. Biol.*, **62**, 424–438.

Peters I, Jain S (1987) Genetics of grain amaranths III. Gene-cytoplasmic male sterility. *J. Hered.*, **78**, 251–256.

Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLos Genet.*, **8**.

Poets AM, Fang Z, Clegg MT, Morrell PL (2015) Barley landraces are characterized by geographically heterogeneous genomic origins. *Genome Biol.*, **16**, 173.

Poland J, Endelman J, Dawson J, *et al* (2012a) Genomic selection in wheat breeding using genotyping-by-sequencing. *Genome Biol.*, **5**, 103–113.

Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012b) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One*, **7**, e32253.

R Core Team (2014) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rastogi A, Shukla S (2013) Amaranth: a new millennium crop of nutraceutical values. *Crit. Rev. Food Sci. Nutr.*, **53**, 109–25.

Rayburn AL, McCloskey R, Tatum TC, Bollero GA, Jeschke MR, Tranel PJ (2005) Genome size analysis of weedy *Amaranthus* species. *Crop Sci.*, **45**, 2557–2562.

Revell LJ (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Meth. Ecol. Evol.*, **3**, 217–223.

Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci. USA*, **81**, 8014–8018.

Sang T, Li J (2013) Molecular genetic basis of the domestication syndrome in cereals. In *Cereal Genomics II* (edited by PK Gupta, RK Varshney), pp. 319–340. Springer Netherlands, Dordrecht.

Sauer J (1967) The grain amaranths and their relatives: a revised taxonomic and geographic survey. *Ann. Missouri Bot. Gard.*, **54**, 103–137.

Song Q, Hyten DL, Jia G, *et al* (2015) Fingerprinting soybean germplasm and its utility in genomic research. *G3*, **5**, 1999–2006.

Steckel LE, Sprague CL (2004) Common waterhemp (*Amaranthus rudis*) interference in corn. *Weed Sci.*, **52**, 359–364.

Stetter MG, Müller T, Schmid K (2015) Incomplete domestication of South American grain amaranth (*Amaranthus caudatus*) from its wild relatives. *bioRxiv*, p. doi:10.1101/025866.

Stetter MG, Schmid KJ (2016) Phylogenetic relationships and genome size evolution within the genus *Amaranthus* indicate the ancestors of an ancient crop. *bioRxiv*.

Stetter MG, Zeitler L, Steinhaus A, Kroener K, Biljecki M, Schmid KJ (2016) Crossing methods and cultivation conditions for rapid production of segregating populations in three grain amaranth species. *Front. Plant. Sci.*, **7**, 816.

Trucco F, Jeschke MR, Rayburn AL, Tranel PJ (2005) *Amaranthus hybridus* can be pollinated frequently by *A. tuberculatus* under field conditions. *Heredity*, **94**, 64–70.

Trucco F, Tranel PJ (2011) *Wild crop relatives: Genomic and breeding resources*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD (2010) Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.*, **11**, 636–646.

Veerappan V, Kadel K, Alexis N, *et al* (2014) Keel petal incision: a simple and efficient method for genetic crossing in *Medicago truncatula*. *Plant Methods*, **10**, 11.

Vega-Gálvez A, Miranda M, Vergara J, Uribe E, Puente L, Martínez EA (2010) Nutrition facts and functional potential of quinoa (*Chenopodium quinoa* willd.), an ancient Andean grain: a review. *J. Sci. Food Agric.*, **90**, 2541–2547.

Wassom JJ, Tranel PJ (2005) Amplified fragment length polymorphism-based genetic relationships among weedy *Amaranthus* species. *J. Hered.*, **96**, 410–416.

Weir B, Cockerham C (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Xu F, Sun M (2001) Comparative analysis of phylogenetic relationships of grain amaranths and their wild relatives (*Amaranthus*; *Amaranthaceae*) using internal transcribed spacer, amplified fragment length polymorphism, and double-primer fluorescent intersimple sequence. *Mol. Phyl. Evol.*, **21**, 372–387.

A. Analysis of Phylogenetic Relationships and Genome Size Evolution of the *Amaranthus* Genus Using GBS Indicates the Ancestors of an Ancient Crop - Supplementary Information

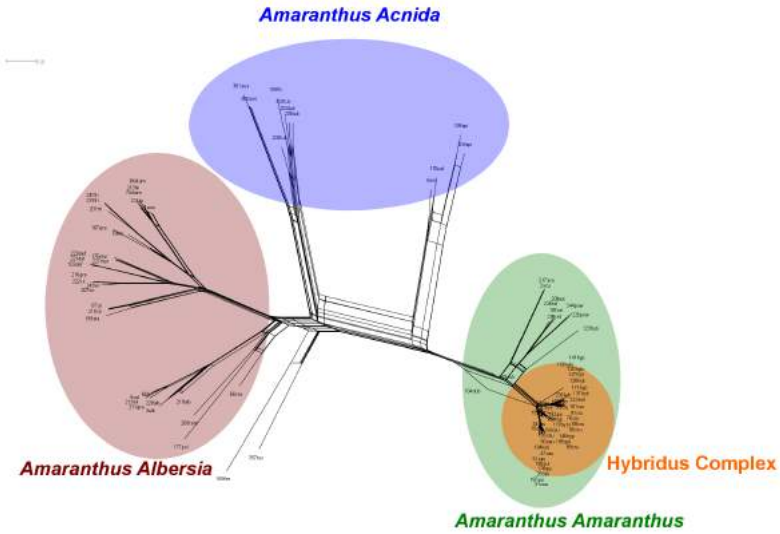

Figure A.1.: NeighborNet network representing 35 species of the genus *Amaranthus* and their subgenus membership.

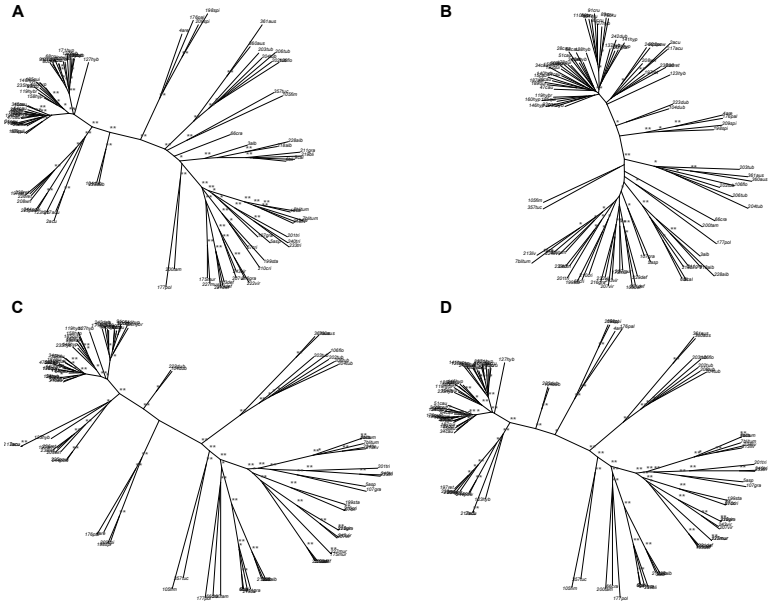

Figure A.2.: Neighbor joining trees from euclidean distance of different datasets. A) refmap_hyp, B) refmap_beet, C) stacks_m3 and D) stacks_m7

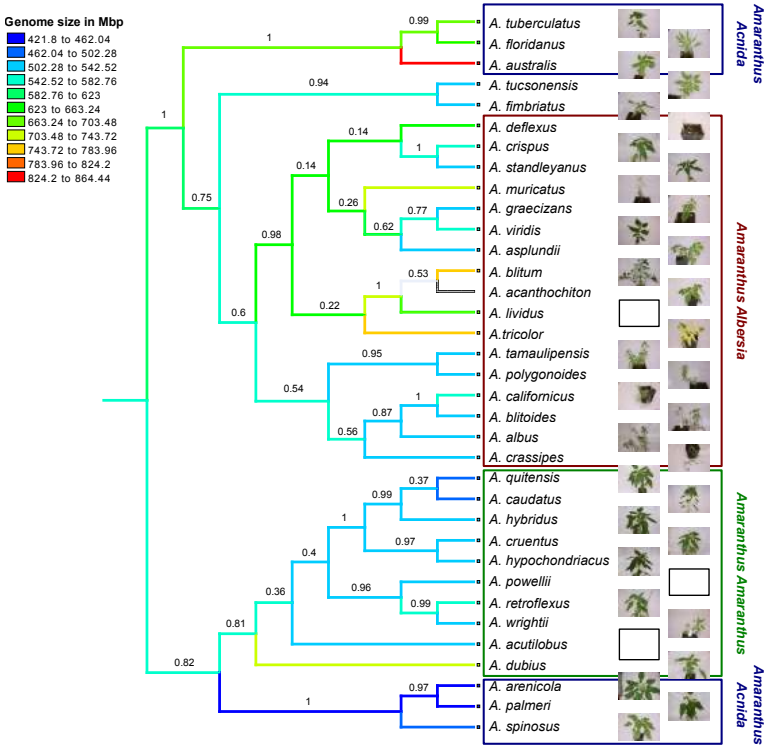

Figure A.3.: Genome size evolution on consensus tree generated with SNAPP. Branch labels show posterior probabilities. Branch colors show genome size evolution in Mbp as constructed with parsimony method. Group labels annotate taxonomic subgenera.

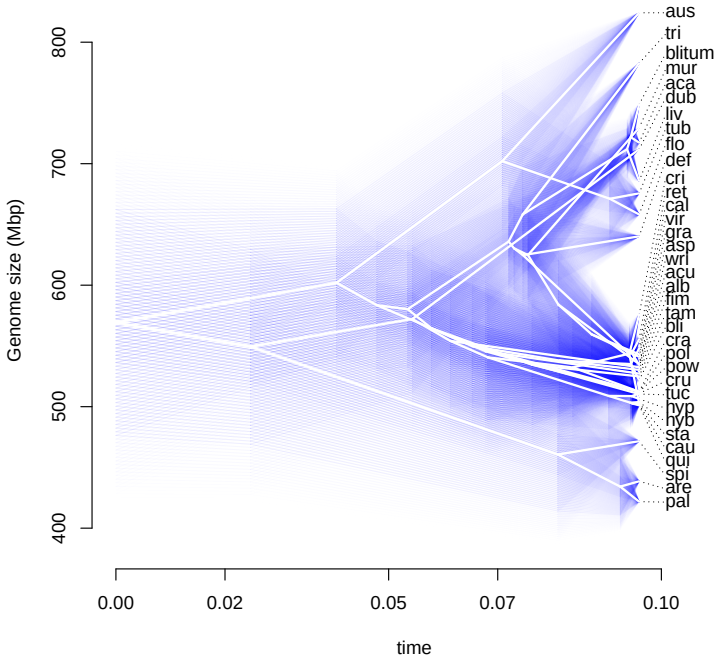

Figure A.4.: Genome sizes with 95% confidence intervals

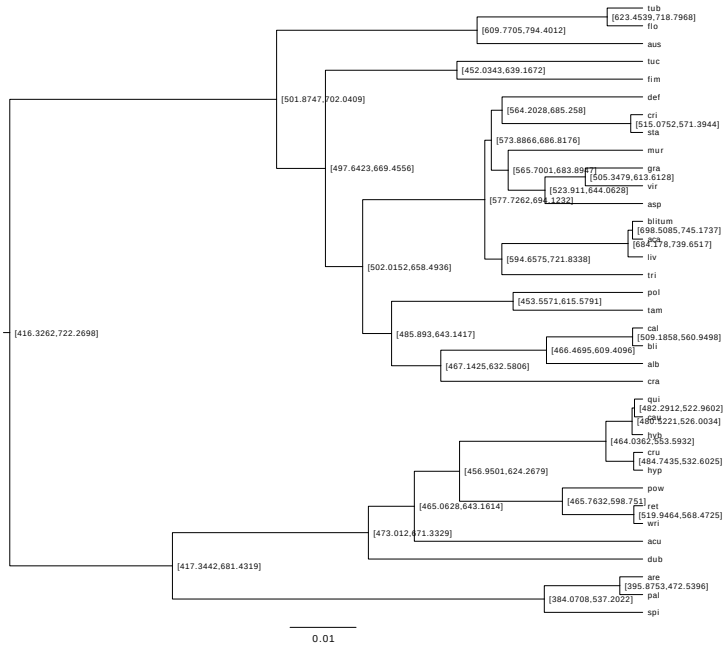

Figure A.5.: Confidence intervals of ancestral genome sizes mapped on consensus tree

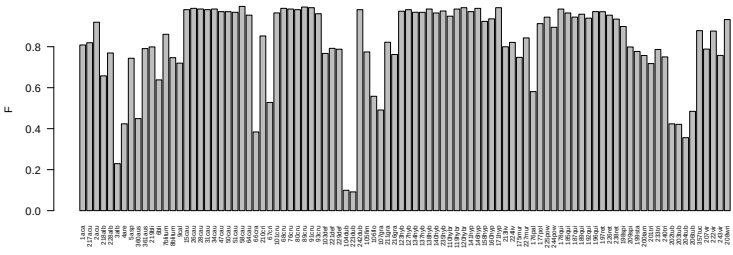

Figure A.6.: Inbreeding coefficient by individual

B. Genomic and Phenotypic Evidence for an Incomplete Domestication of South American Grain Amaranth (*Amaranthus caudatus*) - Supplementary Information

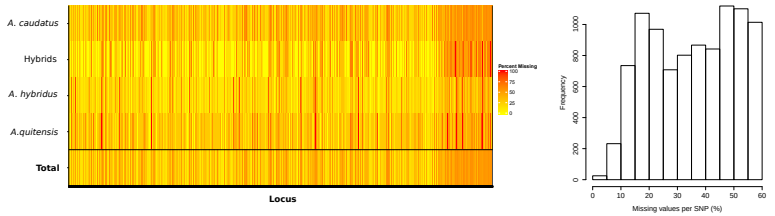

Figure B.1.: Percentage of missing values in the data set separated by population (**A**) and total distribution of missing values in the dataset (**B**).

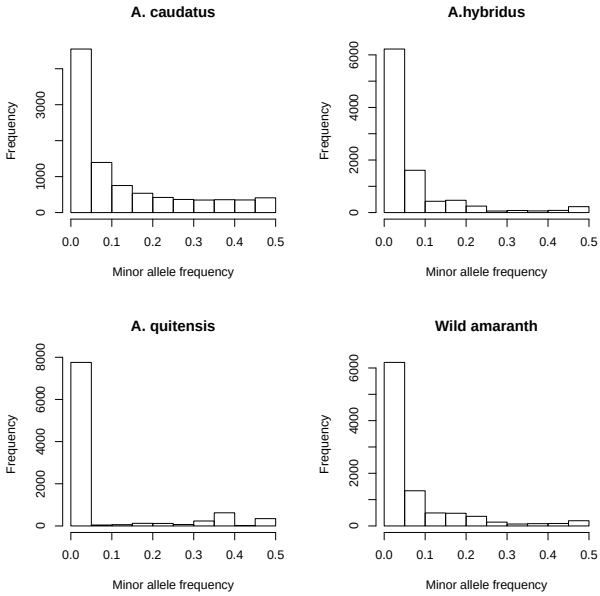

Figure B.2.: Minor allele frequency distribution for the different amaranth species and the close relatives combined

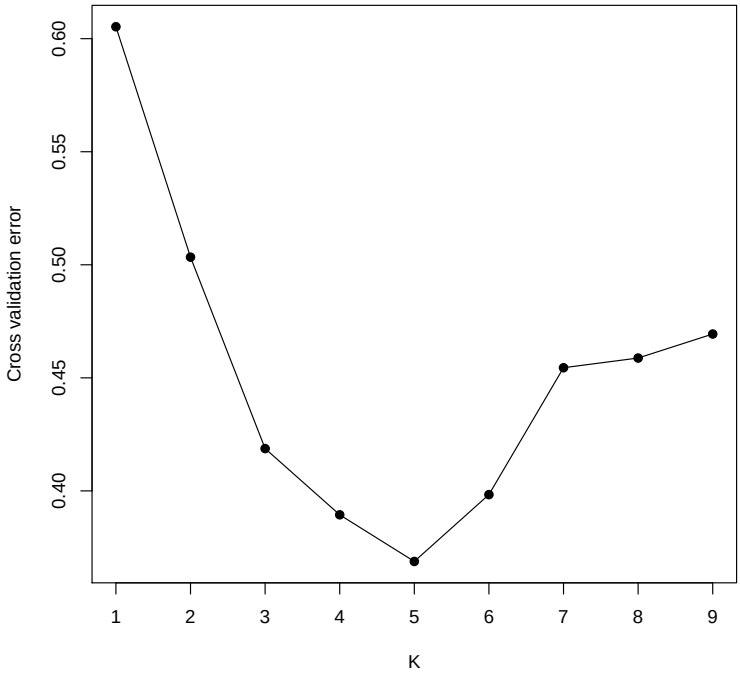

Figure B.3.: Cross validation error of ADMIXTURE with the hypothetical number of groups between one and nine.

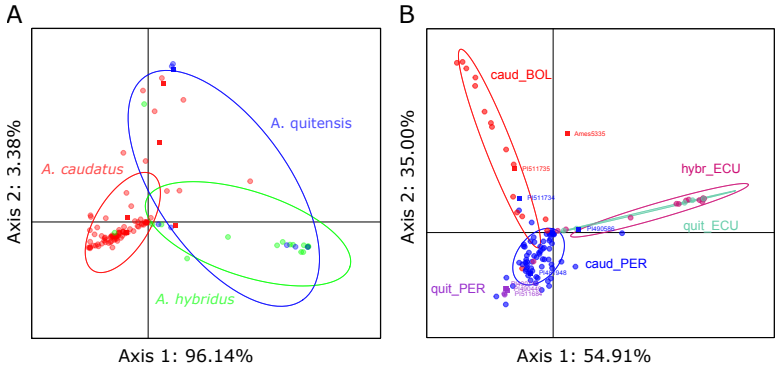

Figure B.4.: Population structure of amaranth populations determined by discriminant analysis of principal components (DAPC) for (**A**) three populations representing the three species *Amaranthus caudatus*, *A. hybridus* and *A. quitensis* and (**B**) for five populations representing the populations of each species in the three countries. *A. caudatus* from Peru (caud_PER), Bolivia (caud_BOL, *A. hybridus* from Ecuador (hybr_ECU) and *A. quitensis* from Peru (quit_PER) and Ecuador (quit_ECU). Squares represent individuals that were not of these populations and were assigned by DAPC.

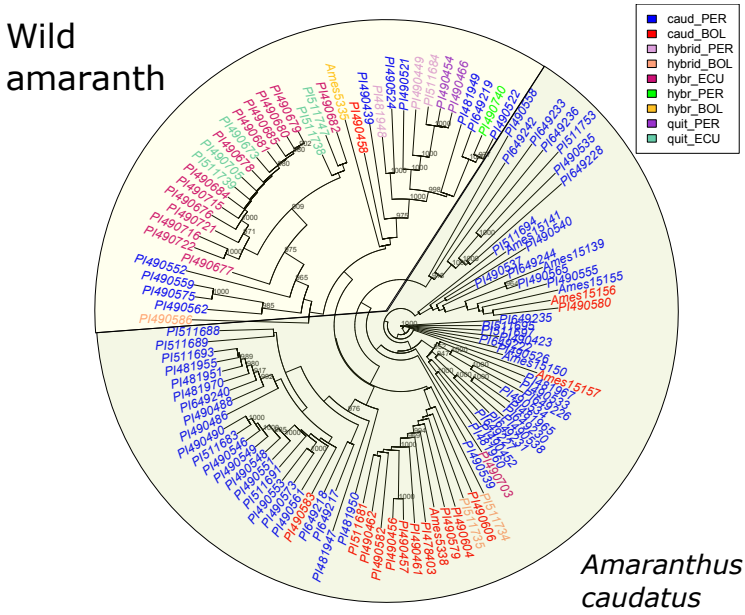

Figure B.5.: Neighbor-joining unrooted phylogenetic tree of 113 amaranth accessions from different potential populations defined by their species and country of origin according to the passport data. Bootstrap values higher than 900 for 1000 bootstraps displayed.

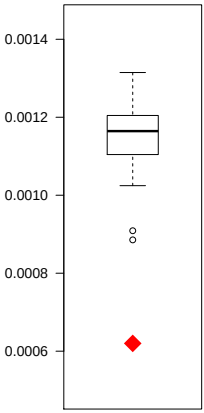

Figure B.6.: Nucleotide diversity of 100 random sub-samples of *A. caudatus* with the same number of individuals as used for combined close relatives *A. hybridus* and *A. quitensis* (23). Red symbol indicates the nucleotide diversity of wild amaranth. The mean diversity was significantly different at 0.05 level ($p = 2.2 * 10^{-16}$)

Table B.1.: Analysis of variance table of the read number depending on the seqeuencing lane. There was no significant relation between the read number and the lane on which a sample was sequenced.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Lane	2	0.31	0.155	0.035	0.966
Residuals	33	147.29	4.463		

Table B.2.: Weir and Cockerham weighted F_{ST} estimates between populations identified by ADMIXTURE for K=5. Group 1 (red) resembles *A. caudatus* from Bolivia, group 2 (green) and 3 (yellow) *A. caudatus* from Peru, group 4 (purple) represents wild amaranth form Ecuador and group 5 (blue) wild amaranth from Peru.

	Group2	Group3	Group4	Group5
Group1	0.464	0.350	0.656	0.476
Group2		0.464	0.762	0.553
Group3			0.549	0.433
Group4				0.579

Table B.3.: Nucleotide diversity of populations inferred by ADMIXTURE with K=5 with 95% confidence interval (CI_π). Group 1 (red) resembles *A. caudatus* from Bolivia, group 2 (green) and 3 (yellow) *A. caudatus* from Peru, group 4 (purple) represents wild amaranth form Ecuador and group 5 (blue) wild amaranth from Peru.

	N	π	CI_π
Group 1	15	0.00067	$\pm$ 0.00001
Group 2	24	0.00040	$\pm$ 0.00001
Group 3	39	0.00111	$\pm$ 0.00002
Group 4	20	0.00031	$\pm$ 0.00001
Group 5	15	0.00058	$\pm$ 0.00001

Table B.4.: Analysis of variance for the hundred seed weight in dependence of the species. There was no significant influence of the species on the hundred seed weight.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species	3 0.000553	0.0001844	1.886	0.136	
Residuals	108	0.010562	0.0000978		

Table B.5.: List of Accessions Used in this Study

No	Accession ID	Species	Origin
1	Ames 15156	*Amaranthus caudatus*	Bolivia
2	Ames 15157	*Amaranthus caudatus*	Bolivia
3	Ames 5338	*Amaranthus caudatus*	Bolivia
4	PI 478403	*Amaranthus caudatus*	Bolivia
5	PI 490456	*Amaranthus caudatus*	Bolivia
6	PI 490457	*Amaranthus caudatus*	Bolivia
7	PI 490458	*Amaranthus caudatus*	Bolivia
8	PI 490461	*Amaranthus caudatus*	Bolivia
9	PI 490462	*Amaranthus caudatus*	Bolivia
10	PI 490579	*Amaranthus caudatus*	Bolivia
11	PI 490580	*Amaranthus caudatus*	Bolivia
12	PI 490582	*Amaranthus caudatus*	Bolivia
13	PI 490583	*Amaranthus caudatus*	Bolivia
14	PI 490606	*Amaranthus caudatus*	Bolivia
15	PI 511681	*Amaranthus caudatus*	Bolivia
16	PI 490604	*Amaranthus caudatus*	Bolivia
17	PI 481947	*Amaranthus caudatus*	Peru
18	PI 481949	*Amaranthus caudatus*	Peru
19	PI 481950	*Amaranthus caudatus*	Peru
20	PI 481951	*Amaranthus caudatus*	Peru
21	PI 481955	*Amaranthus caudatus*	Peru
22	PI 481959 †	*Amaranthus caudatus*	Peru
23	PI 481960 *	*Amaranthus caudatus*	Peru
24	PI 481965	*Amaranthus caudatus*	Peru
25	PI 481967	*Amaranthus caudatus*	Peru
26	PI 481970	*Amaranthus caudatus*	Peru
27	PI 490423	*Amaranthus caudatus*	Peru
28	PI 490439	*Amaranthus caudatus*	Peru
29	PI 490452	*Amaranthus caudatus*	Peru
30	PI 490486	*Amaranthus caudatus*	Peru
31	PI 490488	*Amaranthus caudatus*	Peru
32	PI 490490 *	*Amaranthus caudatus*	Peru
33	PI 490518 †	*Amaranthus caudatus*	Peru
34	PI 490519 †	*Amaranthus caudatus*	Peru
35	PI 490521	*Amaranthus caudatus*	Peru
36	PI 490522	*Amaranthus caudatus*	Peru
37	PI 490524	*Amaranthus caudatus*	Peru
38	PI 490526	*Amaranthus caudatus*	Peru
39	PI 490533	*Amaranthus caudatus*	Peru

No	Accession ID	Species	Origin
40	PI 490534 *	*Amaranthus caudatus*	Peru
41	PI 490535	*Amaranthus caudatus*	Peru
42	PI 490537	*Amaranthus caudatus*	Peru
43	PI 490538	*Amaranthus caudatus*	Peru
44	PI 490539	*Amaranthus caudatus*	Peru
45	PI 490540	*Amaranthus caudatus*	Peru
46	PI 490546	*Amaranthus caudatus*	Peru
47	PI 490547 †	*Amaranthus caudatus*	Peru
48	PI 490548	*Amaranthus caudatus*	Peru
49	PI 490549	*Amaranthus caudatus*	Peru
50	PI 490551	*Amaranthus caudatus*	Peru
51	PI 490552	*Amaranthus caudatus*	Peru
52	PI 490553	*Amaranthus caudatus*	Peru
53	PI 490555 *	*Amaranthus caudatus*	Peru
54	PI 490558	*Amaranthus caudatus*	Peru
55	PI 490559	*Amaranthus caudatus*	Peru
56	PI 490561	*Amaranthus caudatus*	Peru
57	PI 490562	*Amaranthus caudatus*	Peru
58	PI 490565	*Amaranthus caudatus*	Peru
59	PI 490569 †	*Amaranthus caudatus*	Peru
60	PI 490573	*Amaranthus caudatus*	Peru
61	PI 490575	*Amaranthus caudatus*	Peru
62	PI 511683	*Amaranthus caudatus*	Peru
63	PI 511688	*Amaranthus caudatus*	Peru
64	PI 511689	*Amaranthus caudatus*	Peru
65	PI 511691	*Amaranthus caudatus*	Peru
66	PI 511693	*Amaranthus caudatus*	Peru
67	PI 511694	*Amaranthus caudatus*	Peru
68	PI 511695 *	*Amaranthus caudatus*	Peru
69	PI 511697	*Amaranthus caudatus*	Peru
70	PI 511753	*Amaranthus caudatus*	Peru
71	Ames 15150	*Amaranthus caudatus*	Peru
72	PI 649222 *	*Amaranthus caudatus*	Peru
73	PI 649226	*Amaranthus caudatus*	Peru
74	PI 649227 *	*Amaranthus caudatus*	Peru
75	PI 649228	*Amaranthus caudatus*	Peru
76	PI 649230	*Amaranthus caudatus*	Peru
77	PI 649231	*Amaranthus caudatus*	Peru
78	PI 649233 *	*Amaranthus caudatus*	Peru
79	PI 649235	*Amaranthus caudatus*	Peru
80	PI 649236	*Amaranthus caudatus*	Peru
81	PI 649240	*Amaranthus caudatus*	Peru

No	Accession ID	Species	Origin
82	PI 649242	*Amaranthus caudatus*	Peru
83	PI 649244	*Amaranthus caudatus*	Peru
84	Ames 15139	*Amaranthus caudatus*	Peru
85	Ames 15141	*Amaranthus caudatus*	Peru
86	Ames 15155	*Amaranthus caudatus*	Peru
87	PI 649217	*Amaranthus caudatus*	Peru
88	PI 649218 *	*Amaranthus caudatus*	Peru
89	PI 649219	*Amaranthus caudatus*	Peru
90	PI 490586	Hybrid	Bolivia
91	PI 511735	Hybrid	Bolivia
92	PI 511734	Hybrid	Bolivia
109	PI 481948 *	Hybrid	Peru
110	PI 490449	Hybrid	Peru
111	PI 511684	Hybrid	Peru
93	Ames 5335	*Amaranthus hybridus*	Bolivia
94	PI 490676	*Amaranthus hybridus*	Ecuador
95	PI 490677	*Amaranthus hybridus*	Ecuador
96	PI 490678	*Amaranthus hybridus*	Ecuador
97	PI 490679	*Amaranthus hybridus*	Ecuador
98	PI 490680	*Amaranthus hybridus*	Ecuador
99	PI 490681	*Amaranthus hybridus*	Ecuador
100	PI 490682	*Amaranthus hybridus*	Ecuador
101	PI 490684	*Amaranthus hybridus*	Ecuador
102	PI 490685	*Amaranthus hybridus*	Ecuador
103	PI 490703	*Amaranthus hybridus*	Ecuador
104	PI 490715	*Amaranthus hybridus*	Ecuador
105	PI 490716	*Amaranthus hybridus*	Ecuador
106	PI 490721	*Amaranthus hybridus*	Ecuador
107	PI 490722	*Amaranthus hybridus*	Ecuador
108	PI 511754 ‡	*Amaranthus hybridus*	Ecuador
112	PI 490740	*Amaranthus hybridus*	Peru
113	PI 490673 *	*Amaranthus quitensis*	Ecuador
114	PI 490705	*Amaranthus quitensis*	Ecuador
115	PI 511738	*Amaranthus quitensis*	Ecuador
116	PI 511739	*Amaranthus quitensis*	Ecuador
117	PI 511747	*Amaranthus quitensis*	Ecuador
118	PI 490466 *	*Amaranthus quitensis*	Peru
119	PI 490454	*Amaranthus quitensis*	Peru

* Individuals used for technical replications.
† Individuals removed due to low read number (less than 10000 reads)
‡ Excluded from analysis because it was a very strong outlier

C. Curriculum vitae

Markus G. Stetter

Born: 13.06.1988 in Ehingen (Donau)/ Germany

Education

- Research associate, University of Hohenheim, Institute of Plant Breeding, Seed Science and Population Genetics 2016 - present
- PhD student, University of Hohenheim, Institute of Plant Breeding, Seed Science and Population Genetics 2013 - 2016
- M.Sc. Crop Science (Major: Plant Breeding), University of Hohenheim 2013
- B.Sc. Agricultural Science, University of Hohenheim 2011

Publications

- **Stetter MG** and Schmid K. Analysis of phylogenetic relationships and genome size evolution of the *Amaranthus* genus using GBS indicates the ancestors of an ancient crop (2017) *Molecular Phylogenetics and Evolution*bioRxiv
- **Stetter MG** Benz M and Ludewig U. Increased root hair density by loss of WRKY6 in *Arabidopsis thaliana* (2017) *PeerJ* PeerJ Preprints 4:e2267v1
- **Stetter MG**, Mueller T and Schmid K. Genomic and phenotypic evidence for an incomplete domestication of South American grain amaranth (*Amaranthus caudatus*) (2016) *Molecular Ecology* bioRxiv
- **Stetter MG**, Zeitler L, Steinhaus A, Kroener K, Biljecki M and Schmid KJ. Crossing methods and cultivation conditions for rapid production of segregating populations in three grain amaranth species (2016) *Front. Plant Sci.* 7:816
- Vatter T, Neuhaeuuser B, **Stetter MG** and Ludewig U. Regulation of length and density of Arabidopsis root hairs by ammonium and nitrate (2015) *Journal of Plant Research* 128,5:839-848
- **Stetter MG**, Schmid K and Ludewig U. Uncovering genes and ploidy Involved in the high diversity in root hair density, length and response to local scarce phosphate in *Arabidopsis thaliana* (2015) *PloS one* 10, 3:e0120604