Incomplete domestication of South American grain amaranth (*Amaranthis caudatus*) from its wild relatives

Markus G. Stetter, Thomas Müller and Karl J. Schmid
Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Fruwirthstr. 21, 70599 Stuttgart / Germany

Keywords

Amaranthus, genotyping-by-sequencing, genetic diversity, domestication, flow cytometry, orphan crop

Corresponding Author

Name: Karl Schmid, Prof. Dr.
Address: Institute of Plant Breeding, Seed Science and Population Genetics (350)
University of Hohenheim
Fruwirthstraße 21
D-70599 Stuttgart / Germany
Tel: +49 711 459 23487
Fax: +49 711 459 24458
E-Mail address: karl.schmid@uni-hohenheim.de

Running title: Incomplete domestication of *A. caudatus*
Abstract

Grain amaranth is a pseudo-cereal and an ancient crop of Central and South America. Of the three species of grain amaranth, *Amaranthus caudatus* is mainly grown in the Andean region. Several models of domestication were proposed including a domestication from the wild relatives *A. hybridus* or *A. quitensis*. To investigate the domestication history of *A. caudatus* and its relationship to the two wild relatives, we used genotyping-by-sequencing (GBS) to genotype 119 amaranth accessions from the Andean region. We determined the genome sizes of the three species and compared phenotypic variation in two domestication-related traits, seed size and seed color. We show that the population genetic analysis based on 9,485 SNPs revealed very little genetic differentiation between the two wild species, suggesting they are the same species, but a strong differentiation between wild and domesticated amaranths. *A. caudatus* has a higher genetic diversity than its wild relatives and about 10% of accessions showed a strong admixture between the wild and cultivated species suggesting recent gene flow. Genome sizes and seed sizes were not significantly different between wild and domesticated amaranths, although a genetically distinct cluster of Bolivian accessions had significantly larger seeds. Taken together our analysis suggests that grain amaranth is an incompletely domesticated species, either because it was not strongly selected or because high levels of gene flow from its sympatric wild relatives counteract the fixation of key domestication traits in the domesticated *A. caudatus*.
Incomplete Amaranth domestication

Introduction

The genus *Amaranthus* 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, 2000). The grain amaranths *Amaranthus caudatus*, *Amaranthus cruentus* and *Amaranthus hypochondriacus* originated from South and Central America. Amaranth is an ancient crop, 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 wild *Amaranthus* species *A. hybridus* and *A. quitensis*, which are considered as potential ancestors (Sauer, 1967). Of these, *A. quitensis* was tolerated or cultivated in Andean home gardens and used for coloring in historical times.

Past research on the domestication of major crop plants revealed that crops from different plant families have similar domestication syndromes that include larger seeds, loss of seed shattering, reduced branching, loss of seed dormancy and increased photoperiod insensitivity (Abbo et al., 2014; Hake & Ross-Ibarra, 2015). In addition to phenotypic changes, domestication strongly affected the structure of genetic diversity of domesticated plants and created a genetic signature of selection and drift because domestication is frequently associated with a strong genetic bottleneck (Doebley et al., 2006; Olsen & Wendel, 2013; Sang & Li, 2013; Nabholz et al., 2014). The history of amaranth domestication is still under discussion. Sauer (1967) proposed two scenarios based on the morphology and geographic distribution of the different species. The first model postulates three independent domestication events, in which *A. hypochondriacus* originated from *A. powellii*, *A. cruentus* from *A. hybridus*, and *A. caudatus* from *A. quitensis*. The second model proposes an initial domestication of *A. cruentus* from *A. hybridus* followed...
Incomplete Amaranth domestication

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. Another model based on SNP markers suggested that all three domesticated amaranths evolved from *Amaranthus hybridus*, but at multiple locations (Maughan et al. 2011). Most recently, Kietlinsky et al. (Kietlinski et al. 2014) proposed a single domestication *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. Taken together, the diversity of hypotheses indicates either a complex domestication history or insufficient data to strongly support a single model of domestication.

Despite its long history of cultivation, 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, 2000). 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 has a different domestication syndrome or whether genetic constraints, a lack of genetic variation or (agri)cultural reasons led to a distinct domestication pattern compared to other crops. The phenotypic analysis of amaranth domestication is complicated by the taxonomic uncertainty of wild amaranth species. 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, 2000). 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 wild 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. 2012). Despite some biases associated with reduced representation sequencing, GBS and related methods are suitable and powerful approaches for studying inter-
Incomplete Amaranth domestication

specific 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 putative wild ancestors *A. quitensis* and *A. hybridus*, and compared patterns of genetic structure with two domestication-related phenotypic traits (seed color and hundred seed weight). We tested whether domestication led to a reduction of genetic diversity and larger seed size in domesticated amaranth, and clarified the taxonomic relationship and gene flow with the close relatives. Our results indicated that *A. caudatus* has a history of domestication that may be considered as incomplete and is consistent with models of multiple domestication.

Material and Methods

Plant material

A total of 119 South American amaranth accessions of three *Amaranthus* species were obtained from the USDA gene bank (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 (Figure S5). We selected the *A. caudatus* accessions based on the altitude of the collection site and focused on high-altitude populations (2,200 to 3,700 m). We further subdivided the species into populations according to their country of origin and included *A. caudatus* from Peru, Bolivia, *A. hybridus* from Peru, Bolivia, Ecuador, *A. quitensis* from Peru and Ecuador as well as hybrids from Peru and Bolivia. Accessions were planted in a field in Nürtingen (Germany), where a single young leaf of one representative plant per accession was sampled. From 12 accessions, three plants were sampled and sequenced individually for quality control.
Incomplete Amaranth domestication

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. (2012)]. 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 ddH₂O 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), T₄- DNA ligase (NEB), 4 µl ddH₂O and 20 µl of adapter mix containing 10 µl barcode adapter (0.3 ng/µl) and 10 µl common adapter (0.3 ng/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 5 µl 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 in 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 (25 pmol/µl of each primer) and 13 µl ddH₂O 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 30 µl 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).

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
Incomplete Amaranth domestication

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.

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 pipeline provided for 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 and the number of mismatches allowed between loci when processing a single individual was two. Four mismatches were allowed between loci when building the catalog. 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 further filtered with vcftools (Danecek et al. 2011), by allowing a maximum of 60% missing values per SNP position.

Inference of genetic diversity and population structure

Nucleotide diversity (\(\pi\)) weighted by coverage was calculated with a Python script implementing the formula of Begun et al. (2007) which corrects for different sampling depths of SNPs in sequencing data. The confidence interval of \(\pi\) was calculated by bootstrapping the calculation 10,000 times. Mean expected (\(H_{\text{exp}}\)) and observed (\(H_{\text{obs}}\)) heterozygosities based on SNPs were calculated with the R package adegenet 1.4-2 [Jombart & Ahmed, 2011]. The inbreeding coefficient (F) was calculated as:

\[
\frac{H_{\text{exp}} - H_{\text{obs}}}{H_{\text{exp}}}
\]
Incomplete Amaranth domestication

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 4 migration events, as preliminary runs indicates 4 migration events to be the highest number. The tree was bootstrapped 1,000 times.

Genome size

To compare genome sizes between *Amaranthus* species, we measured the genome size of 22 *A. caudatus*, 8 *A. hybridus* and 4 *A. quitensis* accessions. 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
Incomplete Amaranth domestication particles were measured each time. Two different plants were measured for each accession.

The DNA content was calculated as:

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

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

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

The conversion from pg to bp was calculated with 1pg DNA = 0.978 × 10⁹ bp \cite{Dolezel et al. 2003}. Means were calculated using R software \cite{Team} and an ANOVA was performed to infer differences in genome size for the species.

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 \cite[http://www.ars-grin.gov/cgi-bin/npgs/html/desclist.pl?159]{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 \cite[http://tarwi.lamolina.edu.pe/~fmendiburu/] {http://tarwi.lamolina.edu.pe/~fmendiburu/}

Results

SNP identification by GBS

To investigate genome-wide patterns of genetic diversity in cultivated amaranth and two putative ancestors, we genotyped a diverse panel of 119 amaranth accessions from the Andean region that we 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 \cite[Supporting information S2]{}. We tested for a lane effect of the Illumina flow cell, by sequencing...
Incomplete Amaranth domestication

12 individuals on each of the three lanes used to sequence all accessions. A subsequent analysis of variance (ANOVA) of the read number did not show a lane effect (Table S1). 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. 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 S1).

Inference of population structure

To infer the genetic relationship and population structure of A. caudatus and its putative ancestors, we used three different methods, ADMIXTURE, Discriminant Analysis of Principal Components (DAPC) and phylogenetic reconstruction with an uncorrected neighbor-joining network. The ADMIXTURE analysis with three predefined groups ($K = 3$), which corresponds to the number of Amaranthus species included in the study, did not cluster accessions by their species origin, but grouped the A. caudatus accessions into two distinct clusters and combined the two wild accessions A. hybridus and A. quitensis into a single cluster. This analysis indicates a clear separation between domesticated and the wild Amaranths, but the two wild amaranths appeared to be a single genetic group because with higher values of $K$ did not lead to subdivision of the two wild species into separate clusters that corresponds to the species assignment (Figure 1). Cross-validation showed that $K = 5$ was most consistent with the data (Supplementary Figure S2), which produced three different groups of A. caudatus accessions that included a few wild amaranth accessions, and two wild amaranth clusters that both consist of A. hybridus and A. quitensis accessions. The two wild amaranth clusters differ by the geographic origin because one cluster contains both A. hybridus and A. quitensis accessions from Peru and the other cluster from Ecuador. This indicates a strong geographic differentiation among wild ancestors.

The A. caudatus accessions clustered into three groups that also showed geographic differenti-
Incomplete Amaranth domestication

Figure 1: 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.

The first cluster consisted of individuals from Bolivia (Figures 2 and 1; 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 2 and 1; K = 5, yellow color), whereas the second cluster contains individuals distributed over a wide geographic range that extending from North to South Peru (K = 5, green color). Ten *A. caudatus* accessions from the Cuzco region clustered with the three accessions of wild amaranths from Peru (K = 5, blue color). These ten accessions showed admixture with the other cluster of wild amaranths and with a Peruvian cluster. 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 Peruvian wild amaranths (Figure 1). Taken together, the population structure inference with ADMIXTURE identified a clear separation between the wild and domesticated amaranth species and genetic differentiation among domesticated amaranths but also gene flow between populations.

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
Incomplete Amaranth domestication

Figure 2: 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 1). Species are indicated by shapes. \textit{A. caudatus} (□), \textit{A. hybridus} (△), \textit{A. quitensis} (▽) and hybrids between species (◦).

The principal component of the DAPC analysis which we used to cluster accessions based on their species explained 96% of the variation and separated the two wild species from the domest-
cated *A. caudatus* (Figure S3A). In a second DAPC analysis that was based on information on species and country of origin (Figure S3B) the first principal component explained 55% of the variation and separated most of the wild from the domesticated amaranths. The second principal component explained 35% of the variation and separated the Peruvian from the Bolivian *A. caudatus* accessions.

Figure 3: 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). It shows two distinct groups of mainly Peruvian *A. caudatus* accessions and a group of accessions with a wide geographic distribution (Figure 2 green color). The latter is more closely related to the Bolivian *A. caudatus* and the wild relatives. The strong network structure between these three groups suggests a high proportion of shared polymorphisms or a high level of recent gene flow. In contrast, the clade with *A. caudatus* accessions from Northern Peru are more separated from the other clades which indicates a larger evolutionary distance, less ongoing
gene flow with the wild ancestors or stronger selection (Figure 2; yellow color). They are split into two groups, of which the smaller includes only accessions with dark seeds. In a bifurcating phylogenetic tree, ten domesticated amaranth accessions clustered within the same clade as the wild species (Figure S4). The same clustering was also obtained with ADMIXTURE and $K = 7$ (Figure 1).

Table 1: Weir and Cockerham weighted $F_{ST}$ estimates between populations based on the taxonomic assignment of their passport data. The group of wild amaranth are *A. hybridus* and *A. quitensis* taken together.

<table>
<thead>
<tr>
<th></th>
<th>$F_{ST}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. caudatus</em> x <em>A. hybridus</em></td>
<td>0.319</td>
</tr>
<tr>
<td><em>A. caudatus</em> x <em>A. quitensis</em></td>
<td>0.274</td>
</tr>
<tr>
<td><em>A. caudatus</em> x wild amaranth</td>
<td>0.322</td>
</tr>
<tr>
<td><em>A. hybridus</em> x <em>A. quitensis</em></td>
<td>0.041</td>
</tr>
<tr>
<td><em>A. caudatus</em> (PER) x <em>A. caudatus</em> (BOL)</td>
<td>0.132</td>
</tr>
</tbody>
</table>

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 the wild *A. hybridus* and *A. quitensis* species were 0.31 and 0.32, respectively (Table 1), 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 Peruvian (group 5) than to the Ecuadorian wild amaranths (group 4; Table S2). A tree constructed with TreeMix, which is based on allele frequencies within groups (Figure 4), suggests gene flow from the Peruvian *A. caudatus* (group 2) to Peruvian wild amaranth (group 5) and, with a lower confidence level, between wild amaranths from Ecuador (group 4) into Bolivian *A. caudatus* (group 1), as well as from Bolivian *A. caudatus* to Peruvian *A. caudatus*.
Incomplete Amaranth domestication (Group 2).

Figure 4: 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 wild amaranth form Ecuador and group 5 wild amaranth from Peru. 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.

**Analysis of genetic diversity**

We further investigated whether domestication reduced genetic diversity in *A. caudatus* compared to wild amaranths (Table 2). All measures of diversity were higher for the cultivated than the wild amaranths. For example, nucleotide diversity (\(\pi\)) was about two times higher in
Incomplete Amaranth domestication

*A. caudatus* than in the two wild species combined. The diversity values of the accessions classified as hybrids showed intermediate values between wild and domesticated populations supporting their hybrid nature. The inbreeding coefficient, $F$, was highest in the domesticated amaranth but did not differ from the wild amaranths combined. In contrast accessions classified as 'hybrids' and *A. quitensis* showed lower inbreeding coefficients. Within the groups of accessions defined by ADMIXTURE, genetic diversity differed substantially. The wild amaranths from Ecuador had the lowest ($\pi = 0.00031$) while the group from northern Peru showed the highest level of nucleotide diversity ($\pi = 0.00111$; Table S3). Figure 5 shows that even though the overall diversity of *A. caudatus* was higher a substantial proportion of sites were more diverse in the wild amaranths ($\pi_{caud} - \pi_{wild} < 0$; Figure 5).

Table 2: Genetic diversity parameters for the three putative *Amaranthus* species and the wild amaranth (*A. hybridus* and *A. quitensis*). $\pi$ is the nucleotide diversity over all sites, CI$_{\pi}$ is the 95% confidence interval of $\pi$, $H_{exp}$ the mean expected heterozygosity for the variant sites and SD$_{He}$ its standard deviation, $H_{obs}$ the mean observed heterozygosity and SD$_{Ho}$ its standard deviation. $F$ is the inbreeding coefficient and SD$_{F}$ its standard deviation.

<table>
<thead>
<tr>
<th>Population</th>
<th>$\pi$</th>
<th>CI$_{\pi}$</th>
<th>$H_{exp}$</th>
<th>SD$_{He}$</th>
<th>$H_{obs}$</th>
<th>SD$_{Ho}$</th>
<th>$F$</th>
<th>SD$_{F}$</th>
<th>$\theta_{w}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. caudatus</em></td>
<td>0.00117</td>
<td>$\pm$ 0.00002</td>
<td>0.175</td>
<td>0.167</td>
<td>0.049</td>
<td>0.140</td>
<td>0.688</td>
<td>0.462</td>
<td>0.00123</td>
</tr>
<tr>
<td><em>A. hybridus</em></td>
<td>0.00061</td>
<td>$\pm$ 0.00001</td>
<td>0.085</td>
<td>0.135</td>
<td>0.041</td>
<td>0.170</td>
<td>0.679</td>
<td>0.608</td>
<td>0.00073</td>
</tr>
<tr>
<td><em>A. quitensis</em></td>
<td>0.00059</td>
<td>$\pm$ 0.00001</td>
<td>0.076</td>
<td>0.169</td>
<td>0.040</td>
<td>0.170</td>
<td>0.451</td>
<td>0.763</td>
<td>0.00048</td>
</tr>
<tr>
<td>Wild amaranth</td>
<td>0.00062</td>
<td>$\pm$ 0.00002</td>
<td>0.090</td>
<td>0.140</td>
<td>0.041</td>
<td>0.166</td>
<td>0.681</td>
<td>0.591</td>
<td>0.00070</td>
</tr>
<tr>
<td>Hybrids</td>
<td>0.00091</td>
<td>$\pm$ 0.00001</td>
<td>0.112</td>
<td>0.179</td>
<td>0.060</td>
<td>0.173</td>
<td>0.436</td>
<td>0.645</td>
<td>0.00107</td>
</tr>
</tbody>
</table>
Incomplete Amaranth domestication

Figure 5: Per site difference in nucleotide diversity ($\pi$) between domesticated amaranth (A. caudatus) and wild amaranth (A. hybridus and A. quitensis).

Genome size in wild and cultivated amaranth

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/]. This raises the possibility that the domestication of A. caudatus was accompanied by polyploidization events as observed in other crops. We therefore tested whether a change in genome size played a role in the context of domestication 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 wild ancestors did not differ significantly from this value (Table 3) indicating that polyploidization did not play a role in the recent evolution of domesticated amaranth.
Incomplete Amaranth domestication

Table 3: Genome size of representative group of individuals for each species. There are no significant differences between genome sizes (p ≤ 0.05). The number of individuals per population is N and SD is the standard deviation for each parameter.

<table>
<thead>
<tr>
<th></th>
<th>DNA content (pg)</th>
<th>SD</th>
<th>genome size (Mbp)</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. caudatus</td>
<td>22</td>
<td>1.026</td>
<td>0.026</td>
<td>501.93</td>
</tr>
<tr>
<td>A. hybridus</td>
<td>8</td>
<td>1.029</td>
<td>0.025</td>
<td>502.96</td>
</tr>
<tr>
<td>A. quitensis</td>
<td>4</td>
<td>1.021</td>
<td>0.016</td>
<td>499.07</td>
</tr>
</tbody>
</table>

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 wild species or 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 6 A). We also compared the seed color of groups defined by ADMIXTURE (K = 5; Figure 1), which reflect genetic relationship and may correct for mislabeling of accessions (Figure 6 B). None of these groups had only white seeds, but clusters that mainly consist of accessions from the wild relatives 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 wild and cultivated amaranths. 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 5 C and Table S4). 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 mainly of Bolivian A. caudatus accessions and had a 21% and 35% larger HSW than the two groups consisting...
Incomplete Amaranth domestication

mainly of Peruvian *A. caudatus* accessions, respectively (Figure 6D). An ANOVA also revealed that seed color has an effect on seed size because white seeds are larger than dark seeds (Table 4).

![Figure 6](image)

**Figure 6:** Hundred seed weight (A,B) and predominant seed color (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 wild amaranth from Ecuador and group 5 (blue) wild amaranth from Peru. 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 ≤ 0.05) than the other groups.
Table 4: Analysis of variance for the hundred seed weight in dependence of the Seed color and Population as determined by ADMIXTURE

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seed color</td>
<td>2</td>
<td>0.000657</td>
<td>0.0003285</td>
<td>4.657</td>
<td>0.0116 *</td>
</tr>
<tr>
<td>Group</td>
<td>4</td>
<td>0.003151</td>
<td>0.0007877</td>
<td>11.165</td>
<td>1.46e-07 ***</td>
</tr>
<tr>
<td>Seed color:Group</td>
<td>2</td>
<td>0.000042</td>
<td>0.0000209</td>
<td>0.297</td>
<td>0.7440</td>
</tr>
<tr>
<td>Residuals</td>
<td>103</td>
<td>0.007266</td>
<td>0.0000705</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discussion

Genotyping-by-sequencing of amaranth species

The genotyping of wild and cultivated amaranth accessions revealed a strong genetic differentiation between wild and cultivated amaranths and a high level of genetic differentiation within domesticated 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 an amaranth 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). Since de novo assembled fragments are not mapped to a reference, they are unsorted and do not allow to investigate differentiation along genomic regions but the data are suitable for the analysis of genetic diversity and population structure (Catchen et al., 2013). GBS produces a large number of SNPs (Poland et al., 2012; Huang et al., 2014), albeit with a substantial proportion of missing values. Missing data lead to biased estimators of population parameters such as $\pi$ and $\theta_w$ (Arnold et al., 2013) and need to be accounted for if different studies are compared. The comparison of accessions and groups within a study is possible, however, because 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. Compared to previous studies on amaranth genetic diversity (Maughan et al., 2009, 2011; Khaing et al., 2013; Jimenez et al., 2013; Kietlinski et al., 2014), our study combines a larger number of accessions and more genetic markers, which allowed us to assess the genetic diversity and population structure on a genome-wide basis.
A. quitensis and A. hybridus are not different species

The two wild relatives A. quitensis and A. hybridus do not appear to be separate species in our analyses but form two distinct subgroups of Peruvian and Ecuadorian wild amaranths that both consist of accessions from the two species. It was suggested before that A. quitensis is the same species as A. hybridus (Coons, 1978), but the passport information regarding the species of genebank accessions was not changed and A. quitensis is still considered as a separate species in these records. The taxonomic differentiation between the two species rests on a single morphological trait, namely the shape of the tepals, which are very small and prone to misidentification (Sauer, 1967). The high phenotypic similarity of A. quitensis and A. hybridus is supported by our analyses which showed that they are very closely related and mainly separated by their geographic origin, from Peru and Ecuador. The FST value between the two wild species was lower than between the two A. caudatus groups from Peru and Bolivia (Tables 1 and S2). A close relationship is also supported by the highly similar genome sizes of all three species, although the genus Amaranthus harbors species with very different genome sizes due to variation in chromosome numbers and ploidy levels (Baohua & Xuejie, 2002; Rayburn et al., 2005). In contrast to our results, a recent study found evidence for a genetic differentiation between A. hybridus and A. quitensis (Kietlinski et al., 2014). This discrepancy may result from the different composition of samples because our sample consists of accessions of both species from the Andean region whereas Kietlinski et al. (2014) used A. hybridus and A. quitensis accessions with little geographic overlap between the two species. Our FST values also indicate that Peruvian and Ecuadorian wild amaranths show a high level of differentiation (FST = 0.579; Table S2), which is similar to the differentiation between one of two Peruvian A. caudatus groups and the wild amaranths from Peru (FST = 0.553). In summary, under the assumption that the passport information of the wild amaranths is correct, we propose that A. quitensis and A. hybridus are a single species. The high level of intraspecific differentiation in both wild and cultivated amaranth is relevant for investigating domestication because the genetic distance between groups of cultivated amaranth is related to the geographic distance of the wild ancestors.
Incomplete Amaranth domestication

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]. In contrast, the overall genetic diversity in our sample of domesticated amaranths was higher than in the two wild 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 5). 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 domesticated grain amaranth may result from a strong gene flow between wild and domesticated amaranths. 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 1 and Table 2). Gene flow between *A. caudatus* and other *Amaranthus* species in different areas of the distribution range could explain a higher genetic diversity in the domesticated amaranth, which is also consistent with the strong network structure (Figure 3) and the TreeMix analysis (Figure 4). Taken together, cultivated *A. caudatus* is unusual in its higher overall genetic diversity compared to its putative wild ancestors, which is uncommon in domesticated crops.

Amaranth domestication syndrome

Despite its long history of cultivation, diverse uses for food and feed and its high importance during the Aztec period, grain amaranth does not display the classical domestication syndrome as strongly as other crops [Sauer 1967]. On one hand, domesticated amaranth shows morphological differentiation from wild amaranths like larger and more compact inflorescences [Sauer 1967] and a level of genetic differentiation (Table 1) which is comparable to the level of differentiation of other domesticated crops and their wild relatives (Sunflower: 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]). On the other hand, the individual flowers of a plant do not mature synchronously and produce...
Incomplete Amaranth domestication

very small seeds that are shattered (Brenner, 2000). In contrast to wild amaraths, which all have dark brown seeds, the predominant seed color of cultivated grain amaranth is white, which suggests that selection for seed color played a role in the history of A. caudatus. However, 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. Similarly, seed sizes between wild and domesticated amaraths are not significantly different with the exception of A. caudatus accessions with white seeds from Bolivia (Figure 6), which have larger seeds. The increased seed size in this group and in white seeds in general indicates past selection for domestication-related traits, but only in specific geographic regions or in certain types of amaranth, and not in the whole domesticated crop species.

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 may reflect that they were not essential for domestication. Although white seeds are predominant in cultivated amarath and unambiguously a domestication-related trait under selection, other seed colors may have been preferred for different uses with the consequence that genes for white seed color were not fixed. Second, domestication traits may experience genetic constraints that limit phenotypic variation. Genes controlling domestication traits that are part of simple molecular pathways, have minimal pleiotropic effects, and show standing functional genetic variation have a higher chance of fixation than genes with high pleiotropic or epistatic interactions (Doebley et al., 2006; Lenser & Theißen, 2013). Numerous genes with these characteristics were cloned and characterized in major crops like rice, barley and maize and shown to contribute to the distinct domestication syndrome such as a loss of seed shattering, larger seed size and compact plant architecture. Since the molecular genetic basis of domestication traits in amaranth is unknown, the lack of a strong domestication syndrome and a lack of fixation of putative domestication traits despite a long period of cultivation may result from genetic constraints which limited the origin and selection of domestication phenotypes. A third explanation is ongoing gene flow between wild and domesticated amaranth that may prevent or delay the formation of a distinct domestication syndrome and contributes to the high genetic diversity (Table 2), similar seed size (Figure 6C), and the presence of dark seeds (Figure 6) in cultivated amaranth. Both historical and ongoing
Conclusions

The genotypic and phenotypic analysis of wild and domesticated South American grain amaranths suggest 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 wild and cultivated amaranths and there is strong evidence of ongoing gene flow from wild to domesticated populations. Gene flow are likely because amaranth has an outcrossing rate between 5% and 30%. Amaranth domestication and gene flow are likely because amaranth has an outcrossing rate between 5% and 30%.

In South America, wild and domesticated amaranths are sympatric over wide areas and the wild A. hybridus and A. quitensis 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 that demonstrate the phenotypic effects of introgression. Since the dark seed color is dominant over white color (Kulakow et al., 1985), dark seeds could have efficiently removed by selection despite gene flow. Therefore, gene flow is not the only explanation for the lack of a distinct domestication syndrome.

Our data are consistent with the model by Kietlinski et al. (2014) who proposed a single domestication of A. caudatus and A. hypochondriacus in Central America followed by migration to South America of A. caudatus. The second model of Kietlinski et al. of two independent domestication events from a single A. hybridus lineage that ranged from Central America to the Andes is supported under the assumption that domestication occurred in South Peru because of the strong genetic differentiation between Ecuadorian and Peruvian wild amaranths in the same distribution range (Peru and Bolivia). The Peruvian group of wild amaranths inferred with ADMIXTURE comprises A. quitensis and A. hybridus, but also A. caudatus accessions, the latter may represent accessions from the center of domestication.

Further support for ongoing gene flow in amaranth is given by the presence of hybrids and admixed accessions in our sample that demonstrate the phenotypic effects of introgression. Since the dark seed color is dominant over white color (Kulakow et al., 1985), dark seeds could have efficiently removed by selection despite gene flow. Therefore, gene flow is not the only explanation for the lack of a distinct domestication syndrome.
Incomplete Amaranth domestication

its wild ancestor despite selection for domestication traits like white seeds. Although grain amaranth is an ancient crop of the Americas, 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 wild 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. All three domesticated amaranths, A. caudatus, A. cruentus and A. hypochondriacus, as well as all wild relatives throughout the whole distribution range should be included in further studies to fully understand and model the domestication history of Central and South American amaranth.

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.
Incomplete Amaranth domestication

References


Incomplete Amaranth domestication


Incomplete Amaranth domestication


Incomplete Amaranth domestication


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.


**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.