Convergent seed color adaptation during repeated domestication of an ancient new world grain

Markus G Stetter[*,†,1], **Mireia Vidal-Villarejo**[†] **and Karl J Schmid**[†,1]
[*]Botanical Institute, University of Cologne, Cologne, Germany, [†]Dept. of Plant Breeding, Population Genetics and Seed Science, University of Hohenheim, Stuttgart, Germany

ABSTRACT Out of the almost 2,000 plants that have been selected as crops, only a few are fully domesticated, and many intermediates between wild plants and domesticates exist. Genetic constraints might be the reason why incompletely domesticated plants have few characteristic crop traits, and retained numerous wild plant features. Here, we investigate the incomplete domestication of an ancient grain from the Americas, amaranth. We sequenced 121 genomes of the crop and its wild ancestors to show that grain amaranth has been selected three times independently from a single wild ancestor, but has not been fully domesticated. While only few domestication traits have been adapted during amaranth domestication, the seed color converted from dark seeds to white seeds between the ancestor and the crops. To investigate the seed color adaptation in amaranth we produced a mapping population between wild and domesticated amaranth, a whole genome sequenced bulked segregant population and perform genome wide association mapping in our divers panel. All three methods agree on two quantitative trait loci controlling the trait. We identify a MYB-like transcription factor gene, a known regulator for seed color variation in other plant species, within one of the significant regions and shows that the trait was independently converted in Central and South America. We propose that a low effective population size at the time of domestication might have contributed to the lack of adaptation of complex domestication traits. Our results show how genetic constraints influenced domestication and might have set the fate of hundreds of crops.

KEYWORDS Domestication, Orphan crop, MYB transcription factor, amaranth, crop wild relatives

Significance

Early farmers cultivated hundreds of plant species and altered their morphological and physiological characteristics during domestication to be well suited for human consumption. Despite long cultivation histories, many crops remained of minor importance, because they were limited in their adaptation to agro-ecological systems. We investigate the domestication history of the ancient minor crop amaranth and show that it was independently domesticated three times from the same ancestor. In all three domesticates, white seed color, which is controlled by only two genes, was independently selected. In contrast, more complex domestication traits like seed size were not changed during domestication, and our analyses suggest that this may have resulted from a lack of functional genetic variation in the ancestor during domestication.

Introduction

Crop domestication was the foundation for modern human societies and has taken place almost simultaneously in different regions of the world. Around 2,000 crops have been cultivated and underwent artificial selection during adaptation to human-mediated environments (Harlan 1992). This intense selection changed characteristic plant traits that are favorable for human cultivation (Olsen and Wendel 2013). Major crops, such as maize, rice, and wheat, combine multiple domestication traits, which together are referred to as domestication syndrome (Hammer 1984), reflecting their adaptation to agro-ecological systems and human preferences (Olsen and Wendel 2013; Li *et al.* 2006; Doebley 2004; Peleg *et al.* 2011). The domestication syndrome includes production-related traits, like reduction of seed shattering, larger seed size, and the loss of seed dormancy, and traits that reflect human preference, including taste changes, reduced toxin content, and seed or fruit color variation (Hammer 1984; Lenser and Theißen 2013).

Crop domestication is a long and fluid process rather than an instantaneous event, and for this reason crops differ in their degree of domestication (Stetter *et al.* 2017a; Gaut *et al.* 2018). Although plants with different degrees of domestication have been cultivated by humans over similar time periods (MacNeish 1965), fully domesticated crops that combine

[1] Markus G Stetter: mgstetter@gmail.com
Karl Schmid: karl.schmid@uni-hohenheim.de

multiple domestication traits are a small minority compared to hundreds of cultivated crops with a weak domestication syndrome (Meyer *et al.* 2012).

A weak domestication syndrome in many minor crops cannot be explained by low selection pressure due to cultural preferences of humans alone, as minor crops were often of high importance for early cultures and have been cultivated for similar time scales as fully domesticated crops. Yet, the lack of adaptation to agro-ecological systems led to the decline of importance and production of many crops that were commonly grown in early human civilizations (Cubry *et al.* 2018; Meyer *et al.* 2012). Potential reasons for a lack of adaptation might be genetic constraints (Walsh and Blows 2009) or missing functional variation in the ancestor or during early domestication. Incompletely domesticated crops are therefore suitable models to investigate the role of genetic constraints and missing diversity, because the traits under selection are known and multiple incompletely domesticated species are available as replicates (Purugganan and Fuller 2009).

The pseudo-cereal grain amaranth is an example of a crop with a long cultivation history, but a weakly pronounced domestication syndrome. Traits, like reduced spines on inflorescences and increased seed number per inflorescence reflect domestication-related selection, but other typical domestication traits of grain crops, such as increased seed size and the loss of seed shattering are missing and therefore limit the yield potential of grain amaranth (Sauer 1967; Brenner *et al.* 2010; Stetter *et al.* 2017b). One conspicous domestication trait of cultivated amaranth is a change in seed color, because all wild amaranths have dark seeds, whereas most cultivated grain amaranth landraces have white seeds (Stetter *et al.* 2017b).

Three species of *Amaranthus* are cultivated as grain crops in South America (*Amaranthus caudatus* L.) and Central America (*A. hypochondriacus* L. and *A. cruentus* L.) (Brenner *et al.* 2010). They are closely related and were likely domesticated from the same wild ancestor, *A. hybridus L.*, but it remains unclear if three independent or a single domestications gave rise to the three grain amaranth species (Sauer 1967; Kietlinski *et al.* 2014; Stetter and Schmid 2017). The taxonomic differentiation of amaranth species is based on morphological traits, in particular on the apprearnace flower and inflorescence morphology (Sauer 1967). Archaeological findings show that amaranth seeds were collected already 8,000 years ago in South America (Arreguez *et al.* 2013) and 6,000 years ago in North America (MacNeish 1965). Historical accounts suggest that grain amaranths were of similar importance for the agriculture of pre-Colombian societies as maize (Sauer 1967), but amaranth cultivation rapidly declined after the Spanish conquest of the Americas. Nowadays, amaranth grains are gaining importance for human nutrition, due to their high concentration of essential amino acids which provide high nutritional value (Downton 1973; Rastogi and Shukla 2013) and low gluten content that make it a suitable cereal replacement for patients with celiac disease (Ballabio *et al.* 2011).

We investigate the domestication process of amaranth by combining whole genome re-sequencing of a diverse set of 121 genebank accessions of the three grain amaranth species and two wild relatives, with an F_2 mapping population from a cross of domesticated and wild individuals, and two bulk populations from advanced breeding material segregating for seed color. To characterize seed color as exemplary domestication trait, we combine genetic and evolutionary analyses to infer the role of selection and genetic constraints during domestication. Our results provide strong evidence that grain amaranth was independently domesticated three times and that white seed color was independently selected in each grain species. We identify a MYB-like transcription factor gene within a QTL region as potential regulator of seed coat color variation of grain amaranth. Based on the observation of low effective population size, we suggest that the domestication of grain amaranth was limited by the lack of functional diversity for complex domestication traits at the time of domestication.

Results and discussion

Resequencing and variant characterization

We resequenced a diverse set of 121 genebank accessions that represent the native distribution ranges of grain amaranth. The set included the three grain amaranths, caudatus (*A. caudatus* L., $n = 32$), cruentus (*A. cruentus* L. $n = 25$), and hypochondriacus (*A. hypochondriacus*, $n = 24$), the two wild relatives hybridus (*A. hybridus* L., $n = 19$) and quitensis (*A. quitensis* (Kunth), $n = 19$), and two individuals that were labeled as 'hybrids' or *A. powellii* L. in their passport data (Table S1). The average sequence coverage per accession was 9x and 94% of reads could be mapped to the hypochondriacus 2.1 reference genome (Lightfoot *et al.* 2017). After variant calling, we identified over 29 million biallelic variants with less than 30% missing data and from these we extracted a final data set of 17,157,381 million biallelic SNPs for further analysis.

Three independent domestication attempts from a single ancestor

In numerous crop species, domestication led to a reduction in genetic diversity due to population bottlenecks and strong directional selection (Hufford *et al.* 2012; Stein *et al.* 2018) Nucleotide diversity (π) in all three grain amaranths (caudatus: $\bar{\pi} = 1.50 \times 10^{-3}$; cruentus: $\bar{\pi} = 2.031 \times 10^{-3}$; hypochondriacus: $\bar{\pi} = 2.32 \times 10^{-3}$) was reduced by a factor of two to three relative to hybridus ($\bar{\pi} = 4.91 \times 10^{-3}$), although the wild relative quitensis ($\bar{\pi} = 1.57 \times 10^{-3}$) has similar values as the

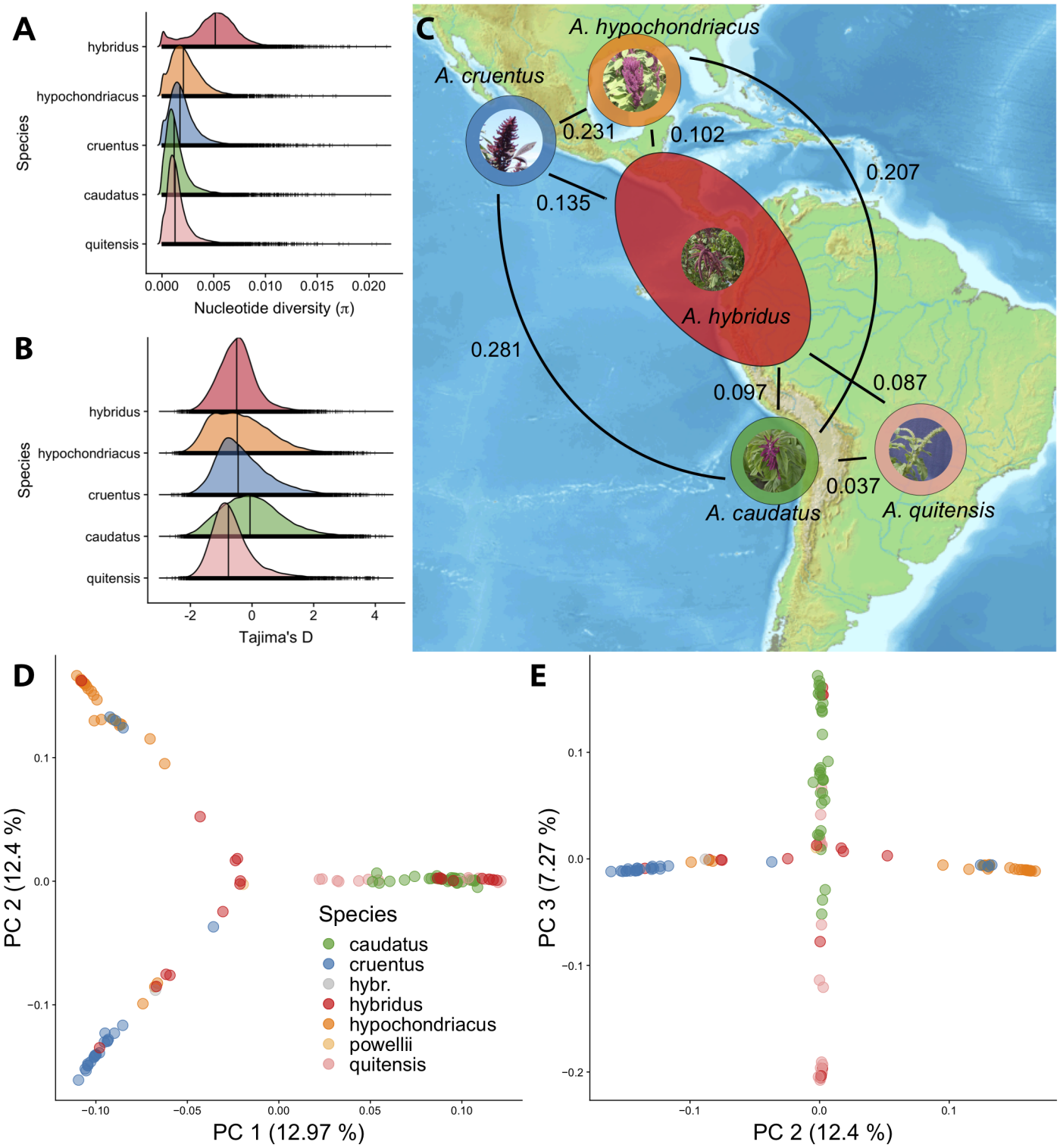

Figure 1 Distribution and diversity of amaranth. **A)** Nucleotide diversity (π) distribution of 10 kb windows, vertical line denotes mean. **B)** Distribution of Tajima's D in 10 kb windows, vertical line denotes mean. **C)** Circles show the approximate geographic distribution of each species (The distribution of hybridus overlaps with all three grain species). Values on the lines show genome wide pairwise F_{st} comparisons between species. **D)** PCA of 121 individuals that indicates the differentiation between Central and South American accessions on PC 1, and between the grain amaranths from Central America on PC 2. **E)** shows the differentiation between caudatus and quitensis on PC 3. Coloring follows the desciption in D. Crop species, caudatus, cruentus and hypochondriacus in green, blue and orange, and wild species, hybridus, quitensis in red and light red.

cultivated amaranths (Figure 1A). Using the fixation index, F_{st} as measure of genetic differentiation we tested different hypotheses of domestication. If all three grain amaranth species were domesticated in a single process, the level of genetic differentiation should be lower between the three domesticated than between the wild and domesticated amaranths. A single domestication is not supported by our data, because the three domesticated amaranths are more strongly differentiated from each other ($\bar{F}_{st} = 0.207 - 0.281$) than each domesticated amaranth from their wild ancestor, hybridus ($0.097 - 0.135$) (Figure 1C). This result is supported by a principal component analysis (PCA), which separates accessions into three groups. The first principal component (PC) separates South American (caudatus, quitensis, and hybridus from South America) from Central American accessions (cruentus, hypochondriacus, and hybridus from Central America). The second PC differentiates between the two Central American grain species, and the third PC differentiates between quitensis and the other South American species. The PCA analysis suggests three independent domestications, as each of the three grain amaranth species clusters with a subset of hybridus accessions that originate from the same geographic region as their genetically most similar cultivated amaranths (Figure 1 and S1). Consecutive domestications after initial domestication from hybridus would result in a closer relationship between two crop species than between each crop and hybridus. The F_{ST} value between the sympatric species caudatus and quitensis was only 0.037 (1C). Such a close genetic relationship was observed before and led to the hypothesis that quitensis is the direct ancestor of caudatus (Sauer 1967; Kietlinski *et al.* 2014) (Figure 1). F_{st} between hybridus and caudatus was higher than between caudatus and quitensis, but hybridus is geographically structured (Figure S3). South American hybridus is more closely related to caudatus and quitensis than caudatus to quitensis (Figure 1, S1 and S3). The previous analysis of a large sample of South American amaranths has shown the close relationship between hybridus and quitensis and suggested hybridus as ancestor of caudatus (Stetter *et al.* 2017b).

Model based inference of population structure suggested between three and seven clusters. It separated the three grain amaranths but clustered wild relatives with domesticated accessions from the same geographic regions (Figure S3 and S4). Taken together, the inferred population structure was highly consistent with a model of independent domestication of each of the three grain amaranths from geographically separated populations of hybridus.

Common major domestication QTL in all three domesticated amaranths

We showed previously that white seed color is a domestication trait in South American caudatus (Stetter *et al.* 2017b) and investigated whether this is the case for the other two grain amaranth species by analyzing the seed color of all individuals in the diversity panel. All individuals from the wild species have dark seeds, whereas the majority (73 to 82 %) of all three grain amaranth accessions have white seeds (Figure S5). Since the population genetic analysis of our diversity panel suggests independent domestication of the three grain amaranths, we investigated if the seed color conversion was selected independently and identified the genetic basis of this trait.

The segregation ratio in traditional quantitative genetic analyses based on crosses between white and dark seeded amaranths suggested that seed color is controlled by two genetic loci and white seed color is recessive (Kulakow *et al.* 1985). To identify the genes that control seed color, we conducted a QTL analysis in a segregating F_2 population of a cross between a cultivated caudatus (white seeds) and wild quitensis (dark seeds), a bulked segregant analysis (BSA) of a population consisting of a diverse panel of cruentus and hypochondriacus individuals, and a genome wide association study (GWAS) with 116 accessions from the five species of the diversity panel. We identified the same two major QTL on chromosomes 3 and 9 in all three populations (Figures 2 and S6). The QTL on chromosome 3 is located in a relatively narrow interval (100 kb), whereas the QTL on chromosome 9 extends over a fairly large interval in the F_2 mapping and the bulk segregant analysis (Figure S6). Similar to the F_2 and BSA mapping results, SNP markers associated with seed color are distributed over a large region of ca. 600 kb on chromosome 9. Such a large interval is not expected in a GWAS on a genetically diverse panel of two wild and three domesticated amaranth species, even after correction for population structure. Nonetheless, the large linkage block observed in all three mapping populations indicates a suppression of recombination in this region that may be caused by a chromosomal inversion during domestication, leading to the joint inheritance of the large region (Griffiths *et al.* 2005).

We further tested whether white seed color originated once or multiple times during domestication by conducting a local PCA on the two QTL regions in the diverse GWAS panel (Figure 2A). A single evolutionary origin of the white seeds is expected to separate white and dark seeded accessions into two separate groups independent of the species. The analysis reveals three groups that correspond to the three grain amaranth species, in which the wild ancestors are nested in each group, similar to the genome-wide population structure, showing that the conversion from dark to white seeds occurred independently in each of the three grain amaranths (Figure 2A and S7).

MYB-like gene as candidate for convergent seed color conversion

To identify candidate genes controlling seed color, we investigated the annotation of genomic regions with significant SNPs in the GWAS. The QTL region on chromosome 3 harbors five genes and the large region on chromosome 9, 50 genes (Table S2). The latter region includes a MYB-like transcription factor gene. The MYB transcription factor family is

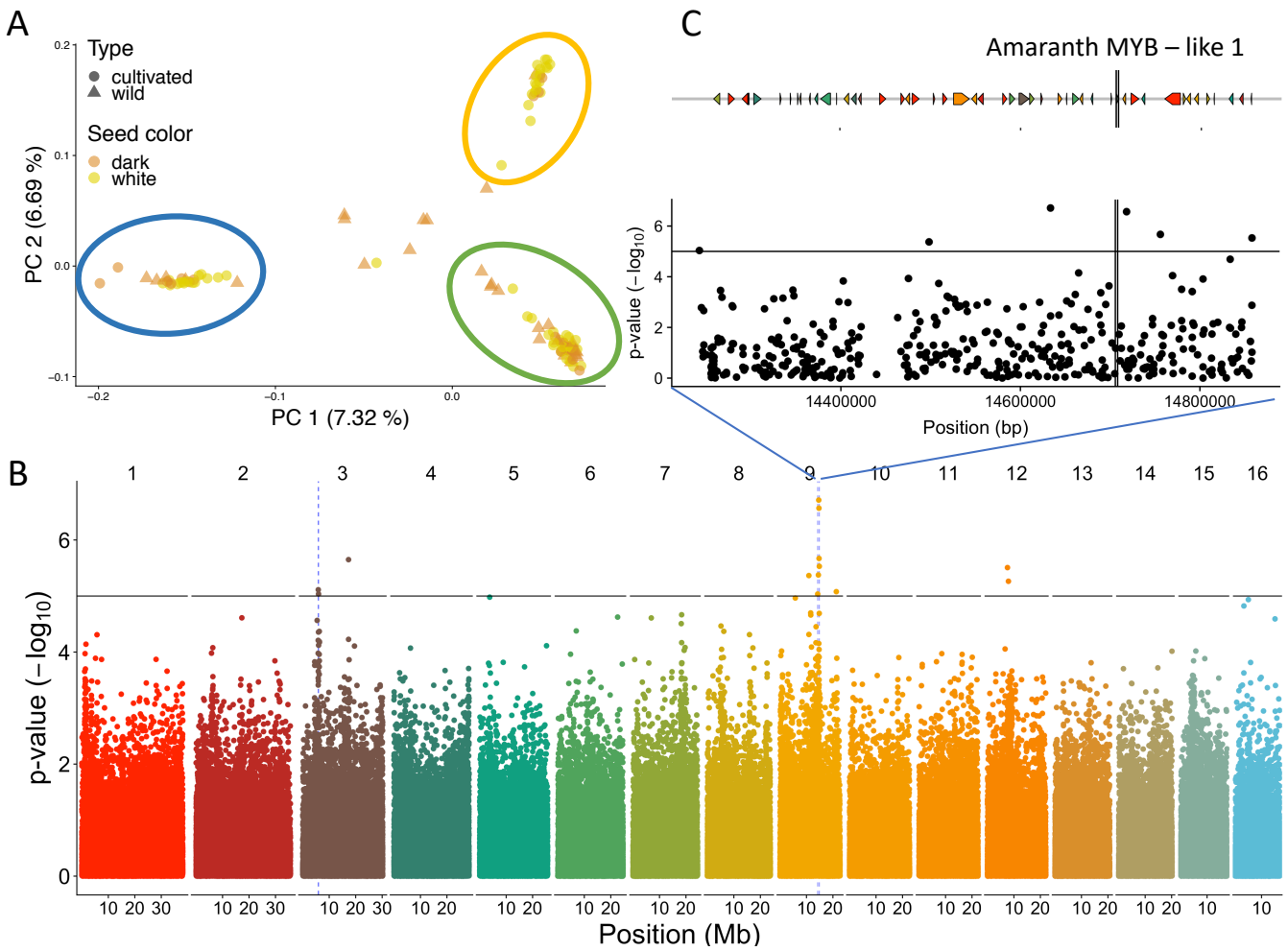

Figure 2 Genetic control of seed color variation **A)** Local PCA of QTL regions, using only SNPs within QTL regions for seed color as indicated by blue lines in B. Circles indicate the predominant crop species in the cluster, cruentus (blue), caudatus (green) and hypochondriacus (orange). **B)** Genome wide association mapping within 116 diverse amaranth accessions from the two wild amaranths and the three grain amaranth species. The horizontal line indicates a significance level of 10^{-5}. Dashed blue lines indicate QTL regions for seed color as identified in three independent populations. **C)** Zoom of QTL region on chromosome 9 and genes within this region. The candidate gene AmMYBL1 is marked by vertical lines.

widely known for its key regulatory function in color phenotypes across the plant kingdom (Gates *et al.* 2016) and their role in color adaptation during crop domestication has been shown for several grain crops (Paauw *et al.* 2019). The MYB gene on chromosome 9 is an ortholog of the Anthocyanin Regulatory C1 gene in maize (Cone *et al.* 1986), and the MYB2 gene in grape (Walker *et al.* 2007). The latter is responsible for red and green grape skin and the change from red to green grapes is caused by the insertion of a transposable element (TE) within the gene (Walker *et al.* 2007). In blood orange, an expression change of a MYB gene increases the anthocyanin content in the flesh and leads to the dark red flesh color (Butelli *et al.* 2012).

Despite the large number of genes in this region, the MYB transcription factor (Amaranth MYB-like 1, short: AmMYBL1) is a promising candidate gene for seed coat color. We annotated the unfiltered variants segregating at the gene and the 2kb upstream and downstream region to identify putative functional variants. Within this region we identify 9 missense variants, 6 splice site variants 1 stop gain variant and 2 frame shift variants (Figure S9). Further detailed molecular analyses are required to identify the causal variants determining seed color in each of the three grain amaranths. The sorted haplotype representation of the AmMYBL1 region shows that the three grain amaranths are differentiated in this region and that individuals with dark seeds have haplotypes more similar to the wild relatives than white-seeded grain amaranths (Figure S8A).

MYB transcription factors form a complex with two other transcription factors, bHLH and WD-repeat (Ramsay and Glover 2005), and we analyzed their presence in the QTL on chromosome 3 and 9, but we did not detect these based on similarity to ortholog sequences from other plants, which suggests that other genes in the QTL region on chromosome 3 influence seed color (Jun *et al.* 2015).

Genome wide signatures of selection

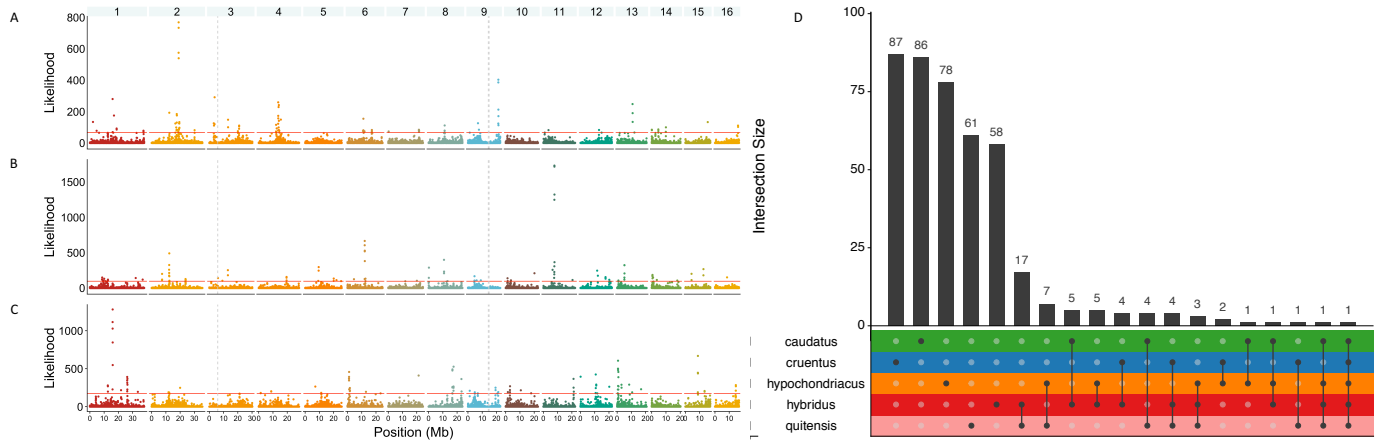

Figure 3 Selective sweeps during amaranth domestication. Selective sweeps along the genome in cultivated grain amaranths in **A)** caudatus (crop), **B)** cruentus (crop) and **C)** hypochondriacus (crop). The horizontal line shows the top 0.05 % likelihood scores within each species **D)** Overlap between sweep regions of the top 0.5% outlier regions in amaranth species (three crop species and two wild relatives). Lines between species show comparisons and bar height represents the number of overlaps.

Independent domestication and strong selection on domestication traits are expected to leave footprints of selection throughout the genome. In major crops, artificial selection during crop domestication led to a number of hard selective sweeps (Tian *et al.* 2009; Olsen *et al.* 2006). We used three different tests of selection in each of the five species to detect hards selective sweeps. To test whether selection affected the same or different genomic regions, we compared the overlap of selective sweep locations identified by SweeD (Pavlidis *et al.* 2013). The top 0.5 % outlier regions (17 of 99, $p \leq 7.073 \times 10^{-22}$) overlap in the two wild species quitensis and hybridus. The overlap of putative sweep regions was lower between wild and domesticated amaranths, but was also significant (4 of 99, $p \leq 0.002$). In contrast, outlier regions did not overlap between the three domesticated amaranths, and were unique to each species (Figure 3), which is consistent with the hypothesis of three independent domestications affecting different genes. We did not identify hard selective sweeps in the two genomic regions that harbor the seed color QTL in the three domesticated amaranths with SweeD (Figure S10) and RAiSD (Figure S12). To test whether soft selective sweeps my have occurred in the seed color QTL regions, we used the G2/G1 (Harris *et al.* 2018) test. This method identified a selective sweep in hypochondriacus within the QTL region on chromosome 9 (Figure S13). Out of the ten most significant outliers on the chromosome, two were located within the QTL region (Figure S13), and one of these had the highest G 123 value (0.4), suggesting a soft selective sweep. Comparing the ten most significant outliers from the other two sweep detection methods that only detect hard sweeps to those of G2/G1, which is able to detect both types of sweeps, provided further evidence for the rise of white seed color through a soft selective sweep (Figure S13). Such a result may be either explained by multiple new mutations in the causative region with a species during domestication, or selection from standing variation. Standing genetic variation for the recessive white seed color could persist at low frequency in the wild, despite potential selection against white seeds. The persistence of up to 27 % dark seeded accessions in grain amaranths further suggests the absence of a complete hard selective sweep during domestication.

Gene flow between wild and domesticated amaranths

Crops and their wild relatives can often intercross in the wild and exchange adaptive traits (Janzen *et al.* 2018). In a number of major crops such as maize (Hufford *et al.* 2013), barley (Poets *et al.* 2015), rice (Cubry *et al.* 2018; Meyer *et al.* 2016) and pearl millet (Burgarella *et al.* 2018) adaptive introgression from wild relatives has been observed. Adaptive introgression requires several generations of backcrossing to the crop in order to remove linkage drag (reviewed in Janzen *et al.* 2018).

Our analysis of population structure did not indicate a strong isolation between wild relatives and domesticated amaranths, which may reflect extensive gene flow. To understand the extent and directions of gene flow, we generated

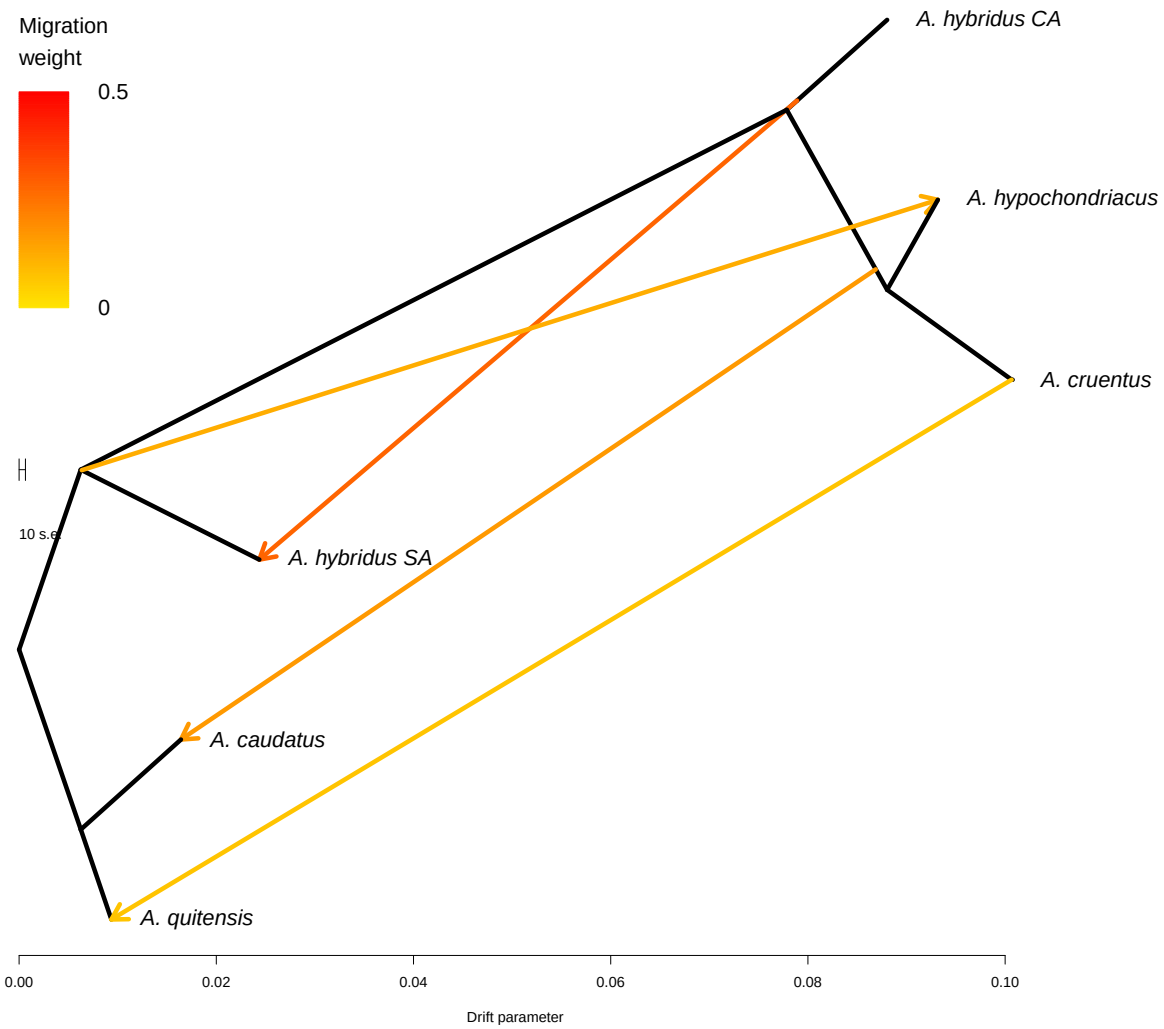

Figure 4 Allele frequency based species tree of wild (hybridus and quitensis) and domesticated amaranth (caudatus, cruentus and hypochondriacus). Arrows indicate gene flow between groups and their color indicates the intensity.

a population tree and estimated gene flow (Figure 4). The resulting unrooted tree is consistent with the population structure analysis, but also indicates a substantial level of gene flow between different groups. All three grain amaranths show traces of gene flow from or into wild relatives. The tree shows gene flow between Central and South American amaranths, which suggests a history of migration and exchange of genetic variation. The presence of gene flow between grain amaranths would have allowed the introgression of the white seed color variants on chromosome 3 and 9 into all three graim amaranths after an initial conversion. However, the local PCA and the haplotype representation do not group genetic variation in the two QTL regions or AmMYBl1 by seed color, but rather by their population (Figures S7 and S8). This result suggests that the seed color was converted independently in each grain amaranth and did not result from a single domestication of post-domestication introgression. This observation is in agreement with the genome wide signal of independent domestications. However, ongoing gene flow from wild into domesticated amaranths may have prevented the fixation of white seed color alleles in the grain amaranths, as we suggested previously for the South American caudatus (Stetter *et al.* 2017b).

Demographic history

The demographic history of crop species influences the extent and type of phenotypic variation available for selection during domestication. Polygenic domestication traits like grain size or seed shattering may require evolutionary changes in multiple loci (Huang *et al.* 2010; Xue *et al.* 2016) and more genetic variation to select the favorable combination for a

functional domestication trait. A potential reason for incomplete domestication in amaranth might be a lack of functional diversity for complex traits during domestication. We therefore reconstructed the demographic history and the split times for the different species.

The demographic analysis shows that all five species followed similar trajectories before the split, with minor differences in the timing of demographic changes (Figure S14). Hybridus has a larger effective population size (N_e) than the three domesticated species, which is consistent with the level of diversity of present populations (Figure 1A and B). Estimated split times of domesticated amaranths from their wild ancestor hybridus ranged from 50,000 generations for caudatus to 8,500 generations in hypochondriacus. In South America, the wild relative quitensis split more recently from hybridus (35,000 generations) than the domesticated caudatus. The population bottleneck in hybridus predated the split of the species and propagates into all species (Figure 5 and S14).

The similarity of population histories before domestication-related splits between species indicates a high confidence for the demographic model. The early date estimated as split time for caudatus might be confounded by introgression from Central American amaranth into caudatus (Figure 4) and a potentially inaccurate estimate of the mutation rate, which was taken from *A. thaliana*. While the mutation rate influences the absolute datings, it does not influence the relative timing between populations. The independent demographic analyses suggest that the bottleneck in amaranth predates its domestication, leading to reduced population size in the ancestor of the three crop species during the time of domestication (Figure 5).

A bottleneck may have caused a lack of functional diversity for complex domestication traits during the time of domestication. Polygenic domestication traits, such as seed shattering and seed size, require either the presence of standing variation at functional sites or a large number of new mutations during domestication. While new mutations can lead to rapid adaptation for traits with major regulator genes, such as seed color, which we show here to be controlled by two major QTL (Figure 2), they are less likely to drive polygenic adaptation with a large number of small effect alleles (Stetter *et al.* 2018). Hence, the small population size in the ancestor of all three grain amaranths that we observe here could potentially have prevented the adaptation of complex traits and might have hindered the complete domestication in amaranth.

Conclusion

In this study we revealed that amaranth was independently domesticated three times from a single ancestor (hybridus), twice in Central America and once in South America. Yet each of these attempts was incomplete, because key domestication traits like larger seeds are missing in all three grain amaranth species. We find phenotypic convergence for seed color conversion and identify the transcription factor MYBl1 as potential candidate gene for the seed coat color determination in amaranth. The seed color QTL regions did not show the signature of hard selective sweeps in the three domesticated amaranth species, similar to several domestication genes in other crops (Studer *et al.* 2011; Zhu *et al.* 2012). However, in hypochondriacus we identify signals of a soft selective sweep within the QTL region that contains AmMYBl1 (Figure S13). The white seed color trait in amaranth might have been selected from standing genetic variation in the wild ancestor, where the recessive white alleles are probably deleterious. In other crop ancestors dark seeds provide resistance to biotic stresses, but are selected against during domestication, because of human color preferences and pleiotropic effects with other domestication traits (Sweeney and McCouch 2007).

Hundreds of crops show a similar lack of complex domestication traits as amaranth. Our case study of an incomplete domestication indicates that the population size decline in the ancestor might have resulted in a lack of diversity to simultaneously select several quantitative domestication traits even under strong artificial selection by hummans. The demographic history has been reconstructed for a number of crops, which show bottlenecks during domestication, but have larger ancestral effective population size in the wild ancestors (Zhou *et al.* 2017; Beissinger *et al.* 2016; Wang *et al.* 2017; Cubry *et al.* 2018; Meyer *et al.* 2016). To further investigate this potential reason for incomplete domestication the demographic history of more incomplete domesticates should be studied and more potential reasons for a lack of adaptation, e.g., trait pleiotropy (Caspari 1952) and the balance between new beneficial mutations and genetic load (Morran *et al.* 2009) should be discovered. Understanding the domestication process, not only of major crops, but also the hundreds of locally grown crops is essential to adapt cropping systems and increase crop diversity for a sustainable future agriculture.

Acknowledgments

We thank Julie Jacquemin and the Ross-Ibarra lab for helpful discussions, Elisabeth Kokai-Kota for technical assistance in DNA sequencing and David Brenner, Fabian Freund, Jeffrey Ross-Ibarra, Katharina Böndel, Michelle Stitzer and Philipp Brand for comments on earlier versions of the manuscript. We thank David Brenner and the USDA amaranth germplasm collection for the permission to use amaranth images in figure 1. This work was supported by the F.W. Schnell Endowed Professorship of the Stifterverband to K.J.S and the Deutsche Forschungsgemeinschaft (DFG) grant STE 2654/1-1 to MGS.

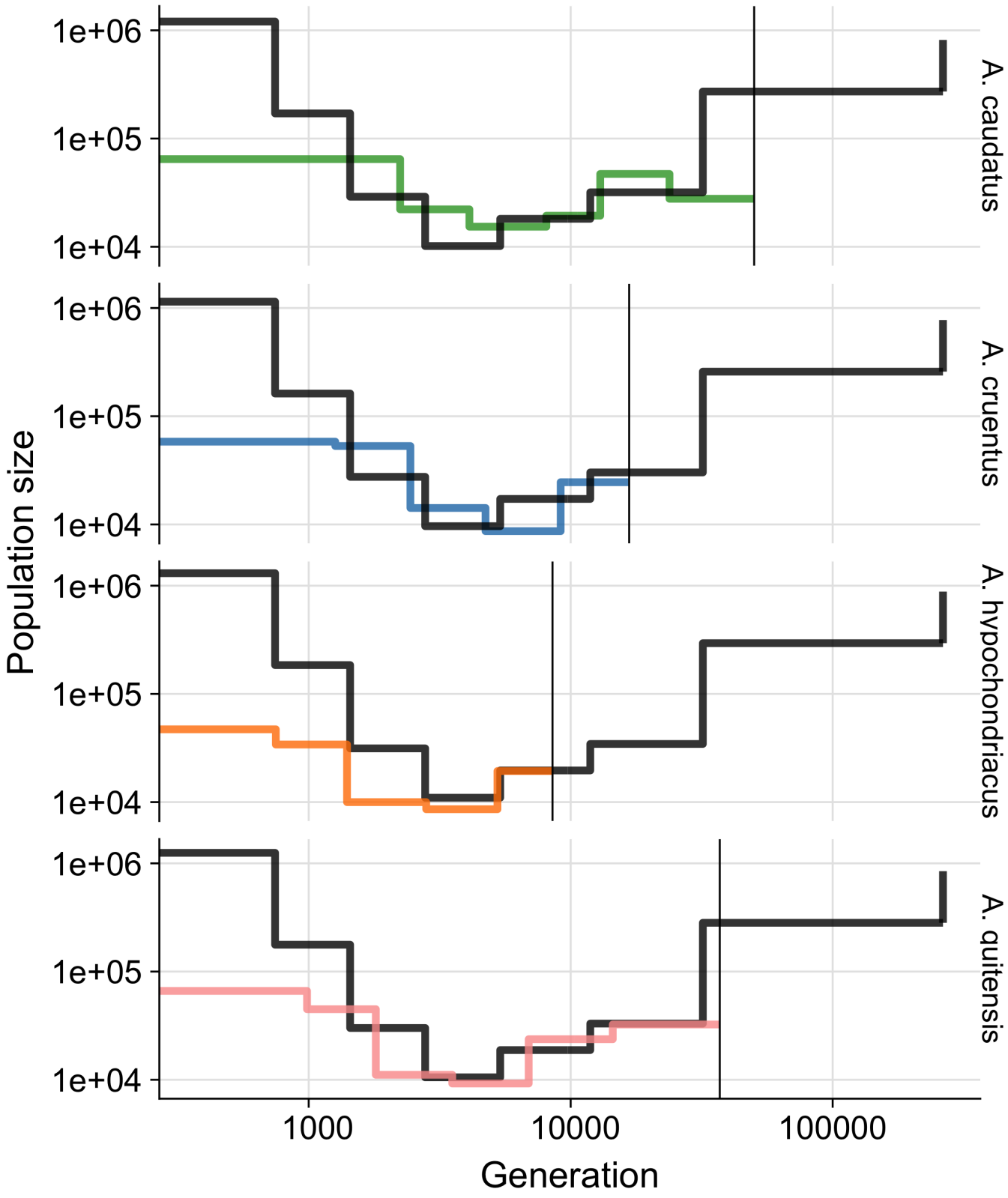

Figure 5 Demographic history of grain amaranth species. Black lines show population size estimates for hybridus. Vertical line shows the estimated split time between wild ancestor and domesticated amaranth in generations before present.

Author Contributions

M.G.S and K.J.S designed the experiments and supervised the research. M.G.S produced the mapping population, M.G.S conducted population genetic analyses, M.V. and M.G.S conducted bulk segregant and QTL mapping analysis. M.G.S wrote the draft of the manuscript. All authors edited, read and agreed to the final version of the manuscript.

Data Availability

Sequencing data is available from the European Nucleotide Archive (ENA) under the project numbers PRJEB30531 (whole genome diversity panel), PRJEB30532 (BSA data), and PRJEB30533 (F2 mapping population). Scripts used for the analysis are available through figshare (10.6084/m9.figshare.7600166)

Methods

Plant material and sequencing.

Geographically diverse amaranth accessions (Table S1) representing all three grain amaranth species, *A. caudatus* L., *A. cruentus* L. and *A. hypochondriacus* L., and three wild relatives, *A. hybridus* L., *A. quitensis* L. and *A. powellii* S. Watson, were obtained from USDA ARS genebank. DNA of young leafs from a single individuals per accession was extracted using DNeasy Plant Mini Kit (Qiagen, Netherlands) following manufacturers' instructions and sequencing libraries were prepared after quality control. Samples were sequences on an Illumina NextSeq with 2x150 bp with VIB Genomics Core Facility (Belgium) or an HiSeq4000 with 2x100 bp by Macrogen (Korea).

Data processing and filtering.

Sequencing reads were filtered and trimmed with `Trimmomatic v0.36` (Bolger *et al.* 2014) and read quality was assessed with `fastQC` (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). High quality reads were mapped to the *A. hypochondriacus* reference genome V2.1 (Lightfoot *et al.* 2017) using `BWA mem` (Li 2013). SNPs were called and filtered according to GATK "Best Practices" (McKenna *et al.* 2010). Additional filtering for a maximum of 30 % missing values at a site, and only SNPs from the 16 largest scaffolds of the reference genome (corresponding to the 16 chromosomes) was performed with `vcftools 0.1.14` (Danecek *et al.* 2011). For the GWAS analysis SNPs were filtered for a minor allele frequency >0.05.

Population genetic analysis.

We used filtered SNP data for all population genetic analyses. Population structure was inferred with `ADMIXTURE` (Alexander *et al.* 2009) with cross validation to determine cluster numbers most strongly supported by the data. Principle component analysis was performed after LD filtering in 10,000 bp windows with an LD threshold of 0.3 using SNPrelate (Zheng *et al.* 2012).

Nucleotide diversity (π) and Tajima's D (Tajima 1989) were calculated using `vcftools 0.1.14` (Danecek *et al.* 2011). F_{ST} values were calculated with SNPrelate (Zheng *et al.* 2012).

Mapping population and QTL mapping.

An F_2 mapping population was created from a single F_1 hybrid plant between 'Oscar Blanco' (PI 642741,*A. caudatus*, white seeds) and PI 490705 (*A. quitensis*, dark seeds). A total of 224 F_2 plants were genotyped using the original protocol for genotyping by sequencing (GBS) (Elshire *et al.* 2011) and their seed color was recorded at maturity after cultivation in a greenhouse. DNA of young leaves was extracted and GBS libraries were prepared as described in (Stetter *et al.* 2017b) and sequenced on a Illumina HiSeq4000 instrument with 2x100 bp by Macrogen. After read filtering, alignment the *A. hypochondriacus* reference genome V2.1 (Lightfoot *et al.* 2017) with `BWA mem` (Li 2013) and SNP calling with `samtools` (Li *et al.* 2009) and 206 samples were kept for further analysis. SNP data was filtered to allow a maximum of 70% missing values at a site and minor allele frequency $\leq$ 0.01. Linkage mapping was performed with `r/qtl` (Arends *et al.* 2010).

Bulk segregant analysis.

Leaf samples from a total of 200 plants with contrasting seeds colors (dark and white) were collected from the Hohenheim Amaranth Breeding Population in October 2016 in Stuttgart, Germany. This advanced population is an early breeding population (three mass selection cycles) consisting of segregating populations from spontaneous (interspecific) hybrids and genetic resources mainly consisting of *A. cruentus* and *A. hypochondriacus*. One hundred single plants for each seed color were collected from diverse parental backgrounds to form two bulks. DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, Netherlands) following manufacturers instructions and TruSeq DNA PCR-Free Library (Illumina, USA) sequencing libraries were prepared after quality control. Libraries were sequenced with 2x150 bp on a Illumina HiSeq X

instrument by Macrogen at 23 -28x coverage. After quality control with `FastQC`, reads were trimmed with `Trimmomatic v0.36` (Bolger *et al.* 2014) and mapped to the *A. hypochondriacus* reference genome V2.1 (Lightfoot *et al.* 2017) with `BWA mem` (Li 2013) and SNPs were called according to GATK "Best Practices" (McKenna *et al.* 2010). Bulk segregant analysis was performed with the G′ method (Magwene *et al.* 2011) implemented in QTLseqr R library Mansfeld and Grumet (2018) on a total of 2,656,505 sites after filtering for high quality SNPs.

Genome wide association mapping ad candidate gene identification

Seed colors were inferred visually from seed samples for the 121 diverse individuals and coded binary. We performed GWAS on 276,466 LD pruned SNPs that were used for the PCA for 116 unambiguous white and dark seed accessions (accessions with mixed seed colors were removed from the analysis) using the compressed mixed linear (Zhang *et al.* 2010) model implemented in GAPIT (Lipka *et al.* 2012). To identify independent or shared origin of the seed color QTLs we performed a PCA analysis on SNPs within the two QTL regions on chromosome 3 and 9 identified using mapping and GWAS. Gene annotations and sequences were extracted from the *A. hypochondriacus* reference genome V2.1 (Lightfoot *et al.* 2017). The impact of variants within AmMYBl1 region was annotated and summarized using snpEff (Cingolani *et al.* 2012). Haplotypes were plotted for the region around AmMYBl1 (chr 9: 1405000 - 14709000) and the entire QTL region on chromosome 3 (chr 3: 5881516 - 5966304) using haplostrips (Marnetto and Huerta-Sánchez 2017).

Selective sweep detection.

We studied the potential adaptation during amaranth domestication by identifying regions under selection using a composite likelihood ratio test (Nielsen *et al.* 2005) implemented in SweeD (Pavlidis *et al.* 2013) using 20 kb windows along the genome. We applied the test to each amaranth population and considered the top 5% of SNPs as outliers. To analyze convergence between the populations we compared the overlapping outliers between populations using upsetR (Lex *et al.* 2014) and tested if overlaps were significant through a 1,000,000 times repeated permutation unsing the sample.int function in R (R Core Team 2019). In addition, we analyzed each population using RAiSD (Alachiotis and Pavlidis 2018) with default parameters. To detect potential soft selective sweeps we calculated G12, G2/G1 and G123 statistics (Harris *et al.* 2018) which is able to detect hard and soft sweeps. We performed the analysis using public python scripts (https://github.com/ngarud/SelectionHapStats) and the three G statistics were calculated for each population and chromosome separately using a window size of 400 SNPs and a step size of 50 SNPs. Identification of G12 peaks was done using the default method implemented in H12peakFinder.py and selected afterwards the top 10 G12 peaks. We used as an indicator of a soft sweep a high value of both G2/G1 and G123 values on the top 10 G12 peaks and the comparison to the other two methods that detect hard selective sweeps.

Gene flow and reconstruction of demographic history.

We modeled the level of gene flow between amaranth populations with Treemix (Pickrell and Pritchard 2012) allowing for a maximum of four migration events. To infer the demographic history of domesticated and wild amaranth populations we ran `SMC++` on the folded site frequency spectrum with a maximum gap size of 1,000 bp for each population separately and in pairs of *A. hybridus* and the respective other species to infer split times. As there is currently no reliably mutation rate estimate available for the different *Amaranthus* species we used the estimate for *Arabidopsis thaliana* of 7×10^{-9} (Ossowski *et al.* 2010). It should be noted that an incorrectly estimated mutation rate may offset the timescale of the population size inference and the estimated split time.

References

Alachiotis, N. and P. Pavlidis, 2018 Raisd detects positive selection based on multiple signatures of a selective sweep and snp vectors. Communications biology **1**: 79.

Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Research .

Arends, D., P. Prins, R. C. Jansen, and K. W. Broman, 2010 R/qtl: high-throughput multiple QTL mapping. Bioinformatics **26**: 2990–2992.

Arreguez, G. A., J. G. Martínez, and G. Ponessa, 2013 *Amaranthus hybridus* L. ssp. *hybridus* in an archaeological site from the initial mid-Holocene in the Southern Argentinian Puna. Quaternary international **307**: 81–85.

Ballabio, C., F. Uberti, C. Di Lorenzo, A. Brandolini, E. Penas, *et al.*, 2011 Biochemical and immunochemical characterization of different varieties of amaranth (Amaranthus L. ssp.) as a safe ingredient for gluten-free products. Journal of Agricultural and Food Chemistry **59**: 12969–12974.

Beissinger, T. M., L. Wang, K. Crosby, A. Durvasula, M. B. Hufford, *et al.*, 2016 Recent demography drives changes in linked selection across the maize genome. Nature Plants **2**: 16084.

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114–2120.

Brenner, D., D. Baltensperger, P. Kulakow, J. Lehmann, R. Myers, *et al.*, 2010 Genetic resources and breeding of Amaranthus. Plant Breeding Reviews, Volume 19 pp. 227–285.

Burgarella, C., P. Cubry, N. A. Kane, R. K. Varshney, C. Mariac, *et al.*, 2018 A western Sahara centre of domestication inferred from pearl millet genomes. Nature Ecology & Evolution p. 1.

Butelli, E., C. Licciardello, Y. Zhang, J. Liu, S. Mackay, *et al.*, 2012 Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. The Plant Cell pp. tpc–111.

Caspari, E., 1952 Pleiotropic gene action. Evolution **6**: 1–18.

Cingolani, P., A. Platts, M. Coon, T. Nguyen, L. Wang, *et al.*, 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly **6**: 80–92.

Cone, K. C., F. A. Burr, and B. Burr, 1986 Molecular analysis of the maize anthocyanin regulatory locus C1. Proceedings of the National Academy of Sciences **83**: 9631–9635.

Cubry, P., C. Tranchant-Dubreuil, A.-C. Thuillet, C. Monat, M.-N. Ndjiondjop, *et al.*, 2018 The rise and fall of african rice cultivation revealed by analysis of 246 new genomes. Current Biology .

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, *et al.*, 2011 The variant call format and VCFtools. Bioinformatics **27**: 2156–2158.

Doebley, J., 2004 The genetics of maize evolution. Annu. Rev. Genet. **38**: 37–59.

Downton, W., 1973 *Amaranthus edulis*: a high lysine grain amaranth. World Crops **25**: 20.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, *et al.*, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS ONE **6**: e19379.

Gates, D. J., S. R. Strickler, L. A. Mueller, B. J. Olson, and S. D. Smith, 2016 Diversification of R2R3-MYB transcription factors in the tomato family Solanaceae. Journal of Molecular Evolution **83**: 26–37.

Gaut, B. S., D. K. Seymour, Q. Liu, and Y. Zhou, 2018 Demography and its effects on genomic variation in crop domestication. Nature plants p. 1.

Griffiths, A. J., S. R. Wessler, R. C. Lewontin, W. M. Gelbart, D. T. Suzuki, *et al.*, 2005 *An introduction to genetic analysis*. Macmillan.

Hammer, K., 1984 Das Domestikationssyndrom. Die Kulturpflanze **32**: 11–34.

Harlan, J. R., 1992 *Crops and Man*, volume 2. American Society of Agronomy.

Harris, A. M., N. R. Garud, and M. DeGiorgio, 2018 Detection and classification of hard and soft sweeps from unphased genotypes by multilocus genotype identity. Genetics **210**: 1429–1452.

Huang, X., T. Sang, Q. Zhao, Q. Feng, Y. Zhao, *et al.*, 2010 Genome-wide association studies of 14 agronomic traits in rice landraces. Nature Genetics **42**: 961.

Hufford, M. B., P. Lubinksy, T. Pyhäjärvi, M. T. Devengenzo, N. C. Ellstrand, *et al.*, 2013 The genomic signature of crop-wild introgression in maize. PLoS Genetics **9**: e1003477.

Hufford, M. B., X. Xu, J. Van Heerwaarden, T. Pyhäjärvi, J.-M. Chia, *et al.*, 2012 Comparative population genomics of maize domestication and improvement. Nature Genetics **44**: 808.

Janzen, G. M., L. Wang, and M. B. Hufford, 2018 The extent of adaptive wild introgression in crops. New Phytologist .

Jun, J. H., C. Liu, X. Xiao, and R. A. Dixon, 2015 The transcriptional repressor MYB2 regulates both spatial and temporal patterns of proanthocyandin and anthocyanin pigmentation in *Medicago truncatula*. The Plant Cell pp. tpc–15.

Kietlinski, K. D., F. Jimenez, E. N. Jellen, P. J. Maughan, S. M. Smith, *et al.*, 2014 Relationships between the weedy *Amaranthus hybridus* (Amaranthaceae) and the grain amaranths. Crop Science **54**: 220–228.

Kulakow, P., H. Hauptli, and S. Jain, 1985 Genetics of grain amaranths: I. Mendelian analysis of six color characteristics. Journal of Heredity **76**: 27–30.

Lenser, T. and G. Theißen, 2013 Molecular mechanisms involved in convergent crop domestication. Trends in Plant Science **18**: 704–714.

Lex, A., N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, 2014 UpSet: visualization of intersecting sets. IEEE transactions on visualization and computer graphics **20**: 1983–1992.

Li, C., A. Zhou, and T. Sang, 2006 Genetic analysis of rice domestication syndrome with the wild annual species, *Oryza nivara*. New Phytologist **170**: 185–194.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997 .

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics **25**: 2078–2079.

Lightfoot, D., D. E. Jarvis, T. Ramaraj, R. Lee, E. Jellen, *et al.*, 2017 Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution.

BMC Biology **15**: 74.
Lipka, A. E., F. Tian, Q. Wang, J. Peiffer, M. Li, *et al.*, 2012 GAPIT: genome association and prediction integrated tool. Bioinformatics **28**: 2397–2399.
MacNeish, R. S., 1965 The origins of American agriculture. Antiquity **39**: 87–94.
Magwene, P. M., J. H. Willis, and J. K. Kelly, 2011 The statistics of bulk segregant analysis using next generation sequencing. PLoS Computational Biology **7**: e1002255.
Mansfeld, B. N. and R. Grumet, 2018 QTLseqr: An R Package for Bulk Segregant Analysis with Next-Generation Sequencing. The Plant Genome **11**.
Marnetto, D. and E. Huerta-Sánchez, 2017 Haplostrips: revealing population structure through haplotype visualization. Methods in Ecology and Evolution **8**: 1389–1392.
McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research .
Meyer, R. S., J. Y. Choi, M. Sanches, A. Plessis, J. M. Flowers, *et al.*, 2016 Domestication history and geographical adaptation inferred from a SNP map of african rice. Nature genetics **48**: 1083.
Meyer, R. S., A. E. DuVal, and H. R. Jensen, 2012 Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. New Phytologist **196**: 29–48.
Morran, L. T., M. D. Parmenter, and P. C. Phillips, 2009 Mutation load and rapid adaptation favour outcrossing over self-fertilization. Nature **462**: 350.
Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Research **15**: 1566–1575.
Olsen, K., A. Caicedo, N. Polato, A. McClung, S. McCouch, *et al.*, 2006 Selection under domestication: evidence for a sweep in the rice *waxy* genomic region. Genetics .
Olsen, K. M. and J. F. Wendel, 2013 A bountiful harvest: genomic insights into crop domestication phenotypes. Annual Review of Plant Biology **64**: 47–70.
Ossowski, S., K. Schneeberger, J. I. Lucas-Lledó, N. Warthmann, R. M. Clark, *et al.*, 2010 The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science **327**: 92–94.
Paauw, M., R. Koes, and F. M. Quattrocchio, 2019 Alteration of flavonoid pigmentation patterns during domestication of food crops. Journal of Experimental Botany .
Pavlidis, P., D. Živković, A. Stamatakis, and N. Alachiotis, 2013 SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Molecular Biology and Evolution **30**: 2224–2234.
Peleg, Z., T. Fahima, A. B. Korol, S. Abbo, and Y. Saranga, 2011 Genetic analysis of wheat domestication and evolution under domestication. Journal of Experimental Botany **62**: 5051–5061.
Pickrell, J. K. and J. K. Pritchard, 2012 Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics **8**: e1002967.
Poets, A. M., Z. Fang, M. T. Clegg, and P. L. Morrell, 2015 Barley landraces are characterized by geographically heterogeneous genomic origins. Genome Biology **16**: 173.
Purugganan, M. D. and D. Q. Fuller, 2009 The nature of selection during plant domestication. Nature **457**: 843–848.
R Core Team, 2019 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Ramsay, N. A. and B. J. Glover, 2005 MYB–bHLH–WD40 protein complex and the evolution of cellular diversity. Trends in Plant Science **10**: 63–70.
Rastogi, A. and S. Shukla, 2013 Amaranth: a new millennium crop of nutraceutical values. Critical Reviews in Food Science and Nutrition **53**: 109–125.
Sauer, J. D., 1967 The grain amaranths and their relatives: a revised taxonomic and geographic survey. Annals of the Missouri Botanical Garden **54**: 103–137.
Stein, J. C., Y. Yu, D. Copetti, D. J. Zwickl, L. Zhang, *et al.*, 2018 Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. Nature Genetics **50**: 285.
Stetter, M. G., D. J. Gates, W. Mei, and J. Ross-Ibarra, 2017a How to make a domesticate. Current Biology **27**: R896–R900.
Stetter, M. G., T. Müller, and K. J. Schmid, 2017b Genomic and phenotypic evidence for an incomplete domestication of South American grain amaranth (*Amaranthus caudatus*). Molecular Ecology **26**: 871–886.
Stetter, M. G. and K. J. Schmid, 2017 Analysis of phylogenetic relationships and genome size evolution of the Amaranthus genus using GBS indicates the ancestors of an ancient crop. Molecular Phylogenetics and Evolution **109**: 80–92.
Stetter, M. G., K. Thornton, and J. Ross-Ibarra, 2018 Genetic architecture and selective sweeps after polygenic adaptation to distant trait optima. PLOS Genetics **14**: 1–24.
Studer, A., Q. Zhao, J. Ross-Ibarra, and J. Doebley, 2011 Identification of a functional transposon insertion in the maize domestication gene *tb1*. Nature Genetics **43**: 1160.
Sweeney, M. and S. McCouch, 2007 The complex history of the domestication of rice. Annals of Botany **100**: 951–957.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–595.

Tian, F., N. M. Stevens, and E. S. Buckler, 2009 Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. Proceedings of the National Academy of Sciences **106**: 9979–9986.

Walker, A. R., E. Lee, J. Bogs, D. A. McDavid, M. R. Thomas, *et al.*, 2007 White grapes arose through the mutation of two similar and adjacent regulatory genes. The Plant Journal **49**: 772–785.

Walsh, B. and M. W. Blows, 2009 Abundant genetic variation+ strong selection= multivariate genetic constraints: a geometric view of adaptation. Annual Review of Ecology, Evolution, and Systematics **40**: 41–59.

Wang, L., T. M. Beissinger, A. Lorant, C. Ross-Ibarra, J. Ross-Ibarra, *et al.*, 2017 The interplay of demography and selection during maize domestication and expansion. Genome Biology **18**: 215.

Xue, S., P. Bradbury, T. Casstevens, and J. B. Holland, 2016 Genetic architecture of domestication-related traits in maize. Genetics pp. genetics–116.

Zhang, Z., E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari, *et al.*, 2010 Mixed linear model approach adapted for genome-wide association studies. Nature Genetics **42**: 355–360.

Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie, *et al.*, 2012 A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics **28**: 3326–3328.

Zhou, Y., M. Massonnet, J. S. Sanjak, D. Cantu, and B. S. Gaut, 2017 Evolutionary genomics of grape (*Vitis vinifera* ssp. *vinifera*) domestication. Proceedings of the National Academy of Sciences p. 201709257.

Zhu, Y., N. C. Ellstrand, and B.-R. Lu, 2012 Sequence polymorphisms in wild, weedy, and cultivated rice suggest seed-shattering locus *sh4* played a minor role in Asian rice domestication. Ecology and Evolution **2**: 2106–2113.

Supplement

Supplementary Figures

Figure S1 PCA Interactive 3D representation of PCA showing PC1 - PC3.
Also available at https://www.cropevolution.org/files/PCA3D

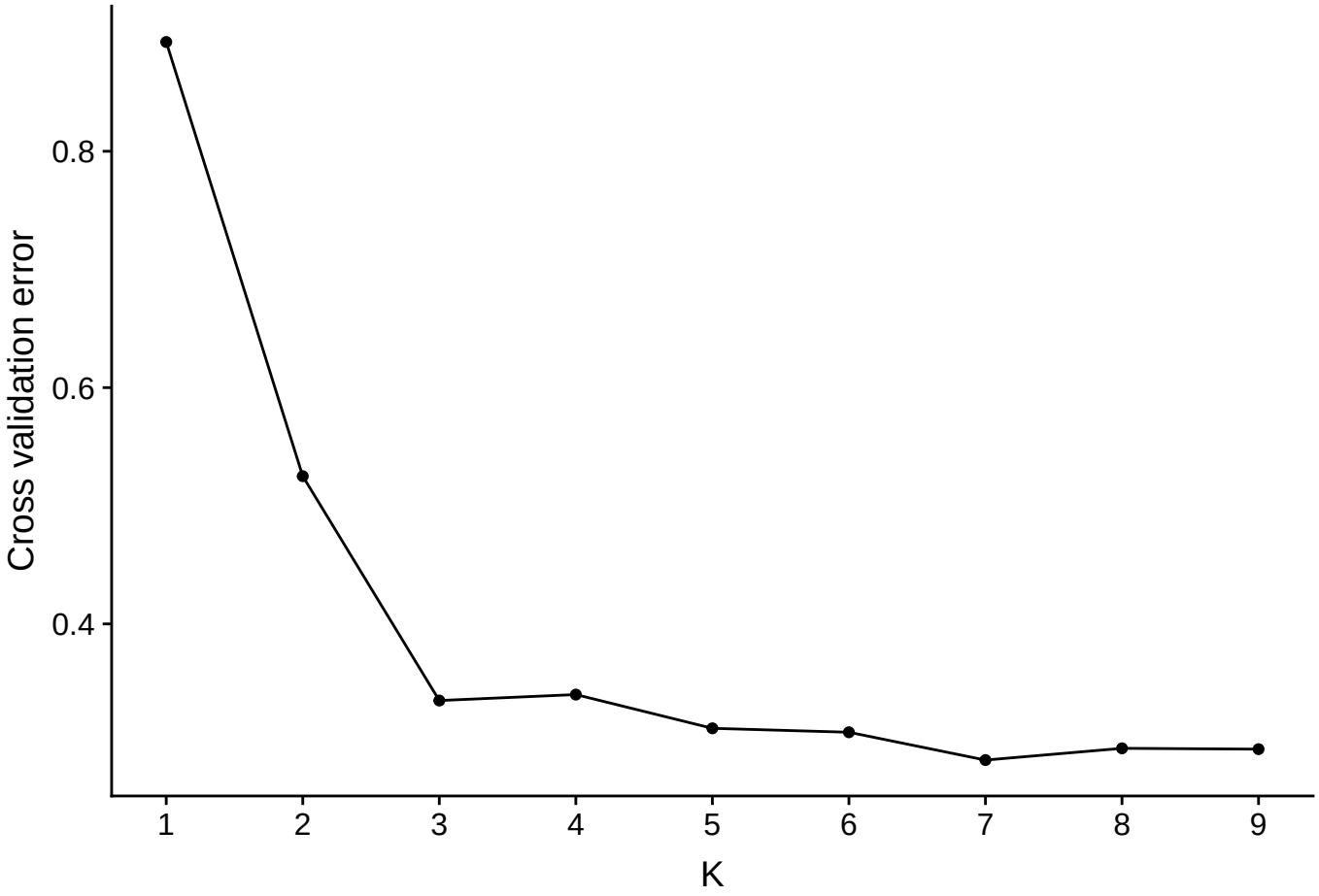

Figure S2 Cross validation of ADMIXTURE.

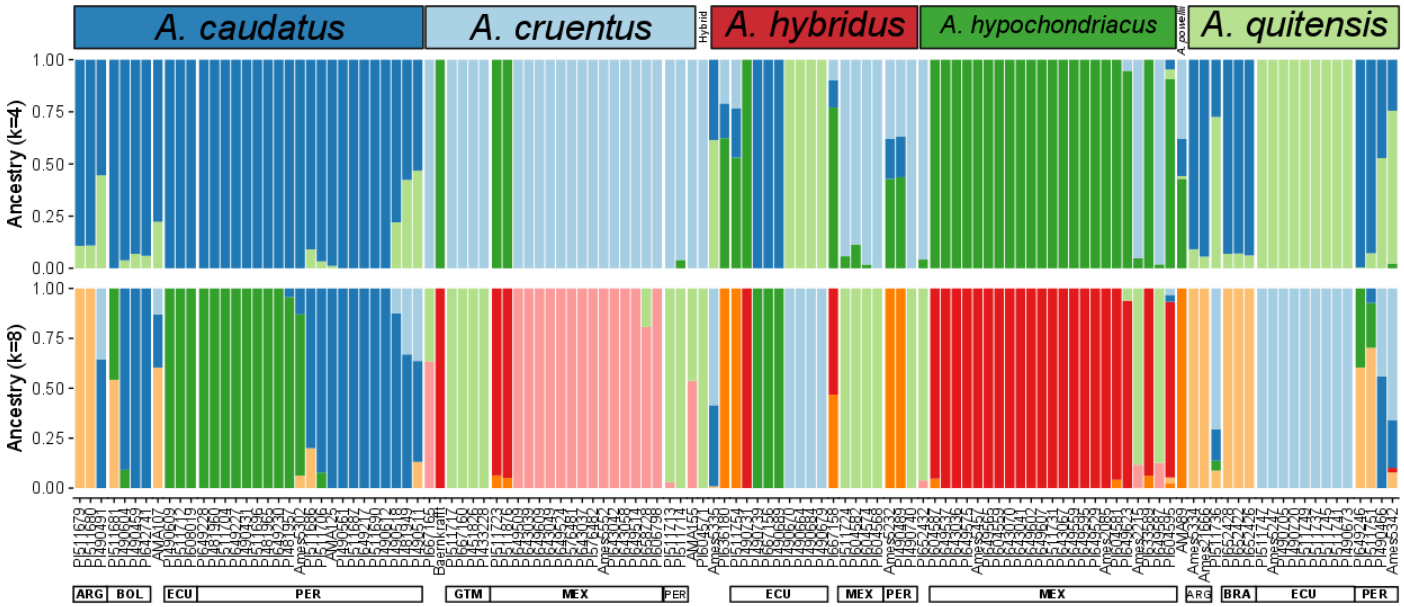

Figure S3 Individual based ancestry contribution calculated with ADMIXTURE for $K = 4$, for three potential crop populations and wild ancestors and $K = 8$ representing species by geographic origin.

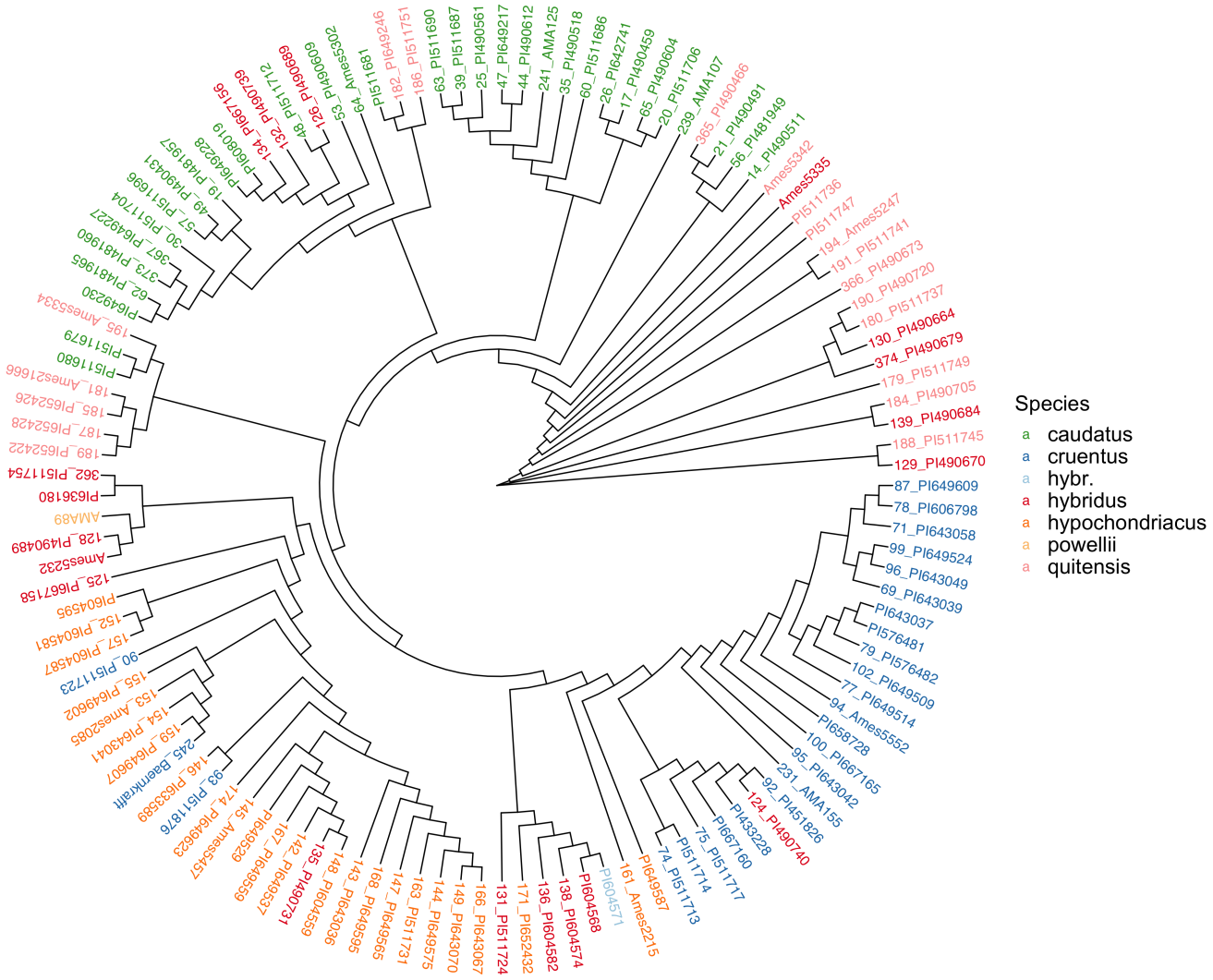

Figure S4 Neighbor joining tree from genomic SNPs of 121 diverse individuals. Colors represent the different species.

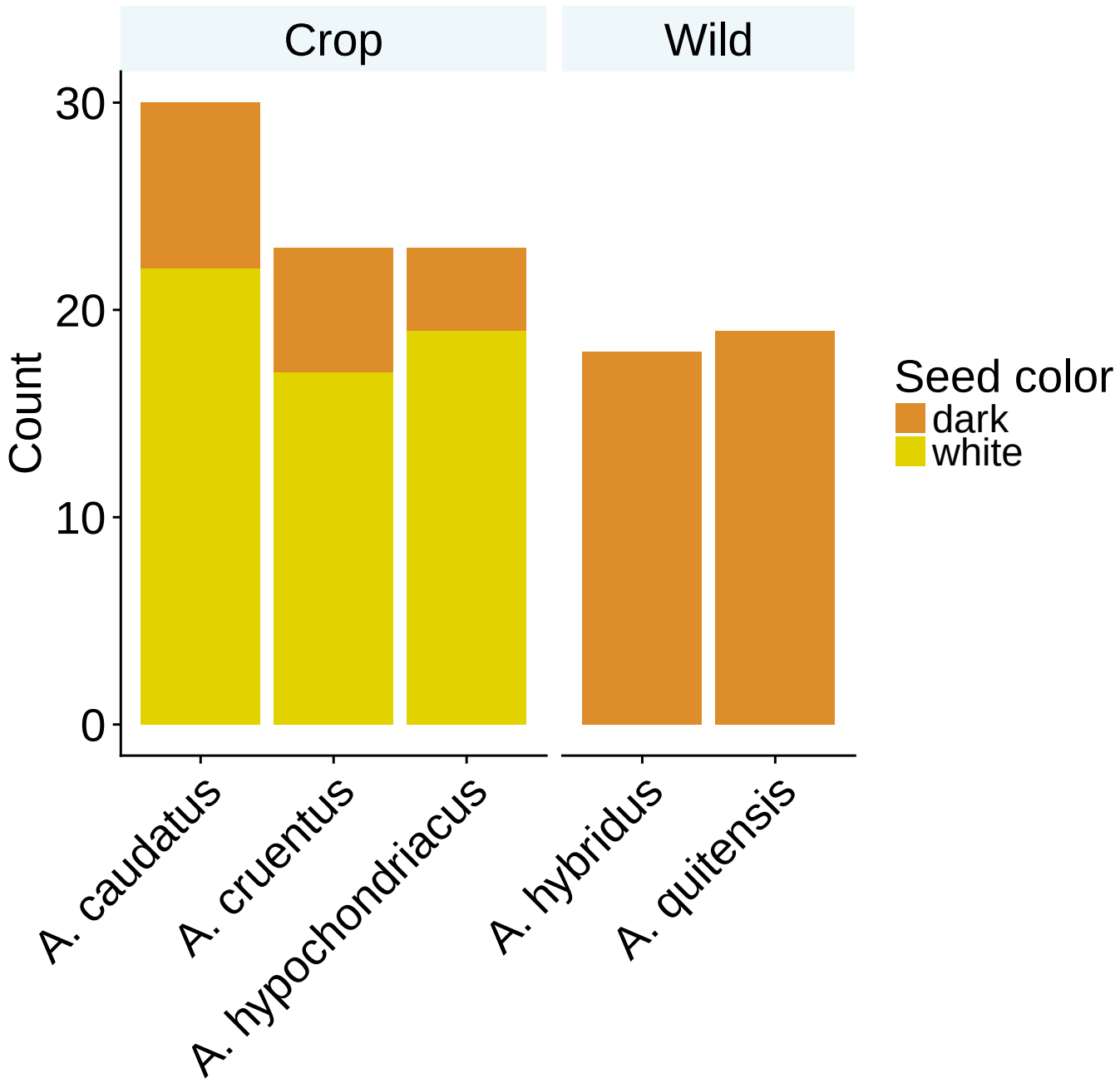

Figure S5 Seed color phenotypes of diversity panel by species.

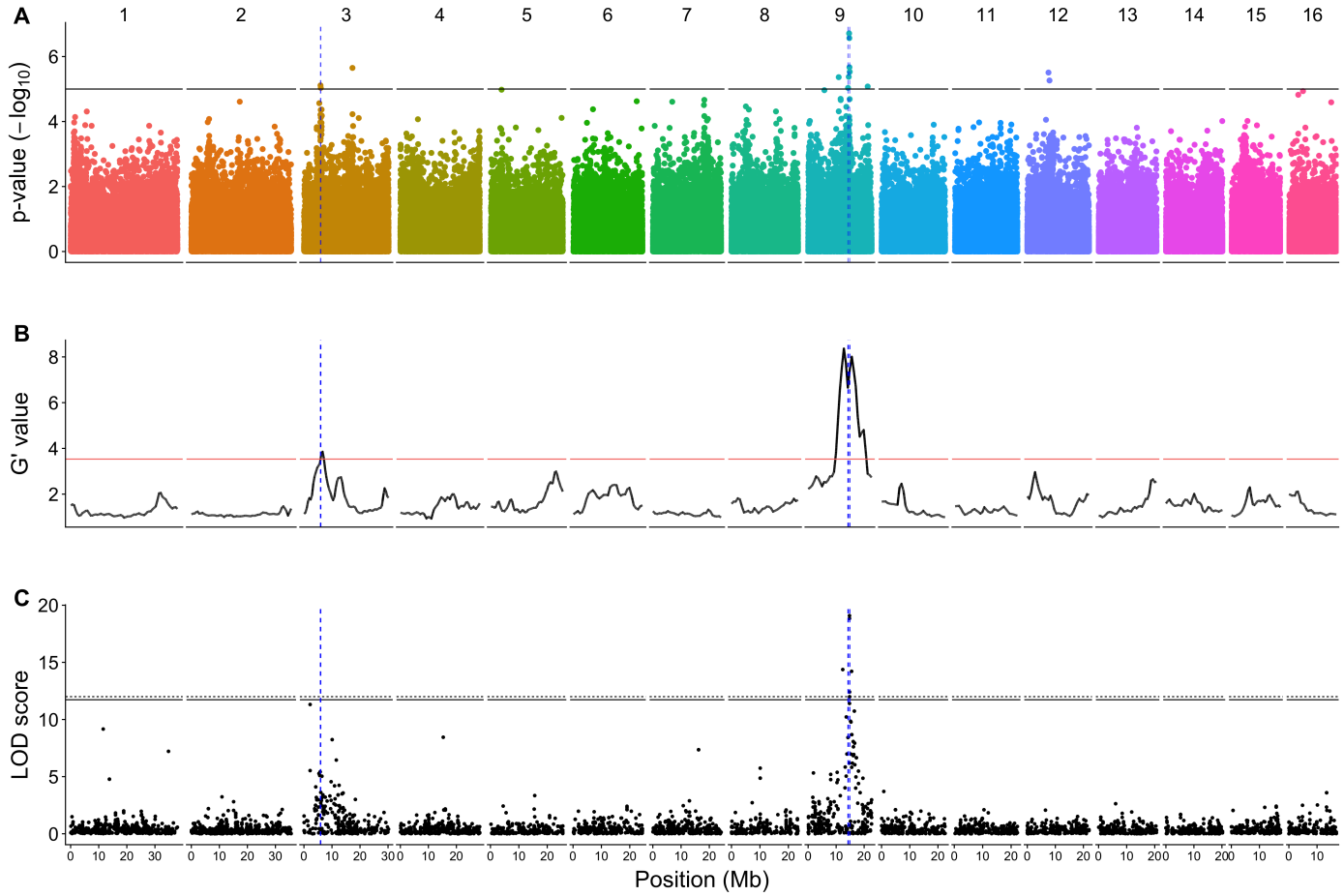

Figure S6 Comparison of QTL regions across mapping populations. A) Genome wide association mapping in 116 diverse germplasm accessions, including crop and wild species B) Results of a bulked segregant analysis for seed color in 100 bulked samples of dark seeded individuals and 100 samples with white seeds from breeding population consisting of hypochondriacus and cruentus. C) QTL mapping results for seed color of F_2 mapping population between caudatus (white seeds) and quitensis (dark seeds).

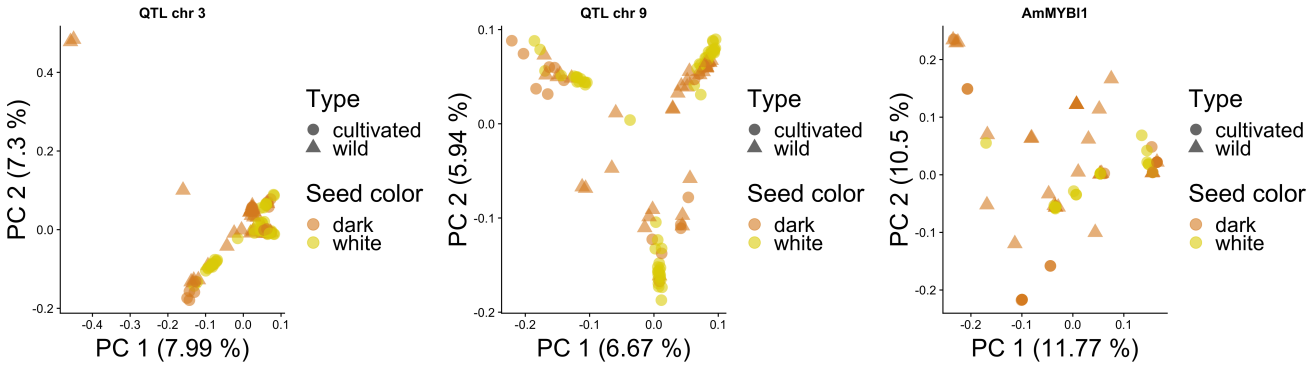

Figure S7 Local PCA for each QTL region. A) QTL on chromosome 3 using 127SNPs after LD pruning, B) QTL on chromosome 9 using 933 SNPs after LD pruning in the region and C) AmMYBl1 region with 2 kb upsream and downstream

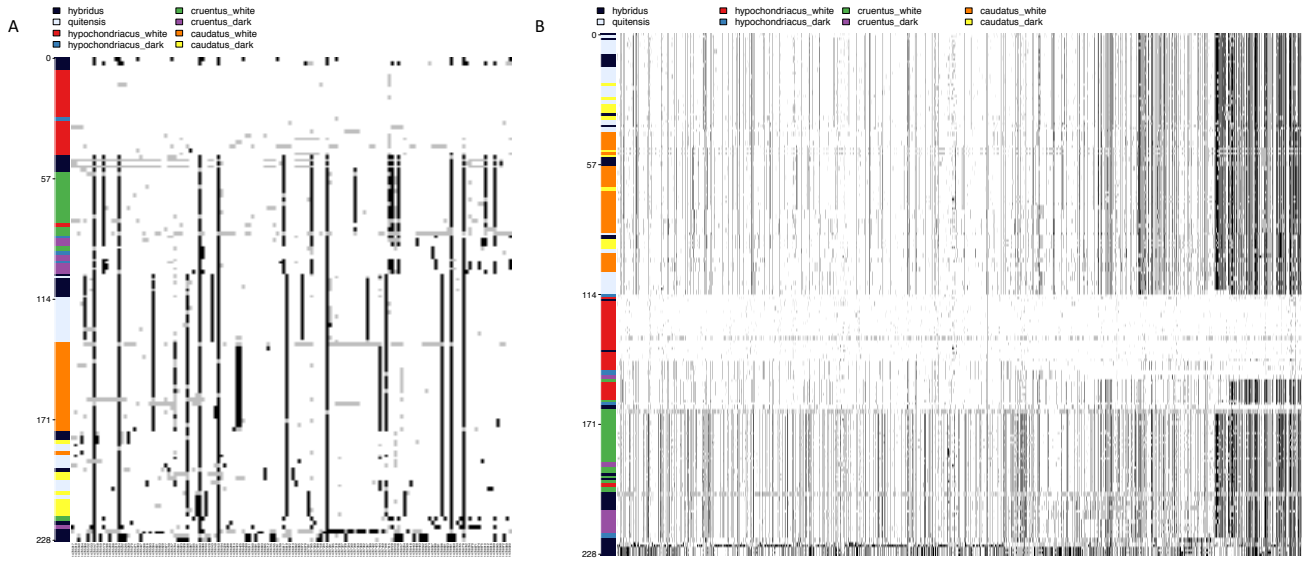

Figure S8 Haplotype representation of SNPs in A) AmMYBl1 region with 2 kb upsream and downstream of the gene and B) QTL region on chromosome 3. White bars represent the reference allele, black bars the alternative allele and grey bars represent missing SNP values

Number variants by type

Type	Total
SNP	375
MNP	0
INS	53
DEL	54
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	482

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	2	0.175%
LOW	8	0.698%
MODERATE	9	0.785%
MODIFIER	1,127	98.342%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	9	69.231%
SILENT	4	30.769%

Missense / Silent ratio: 2.25

Number of effects by type and region

Type

Type (alphabetical order)	Count	Percent
downstream_gene_variant	551	47.788%
frameshift_variant	2	0.173%
intergenic_region	322	27.927%
intron_variant	145	12.576%
missense_variant	9	0.781%
splice_region_variant	6	0.52%
stop_gained	1	0.087%
synonymous_variant	4	0.347%
upstream_gene_variant	113	9.801%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	551	48.08%
EXON	14	1.222%
INTERGENIC	322	28.098%
INTRON	141	12.304%
SPLICE_SITE_REGION	5	0.436%
UPSTREAM	113	9.86%

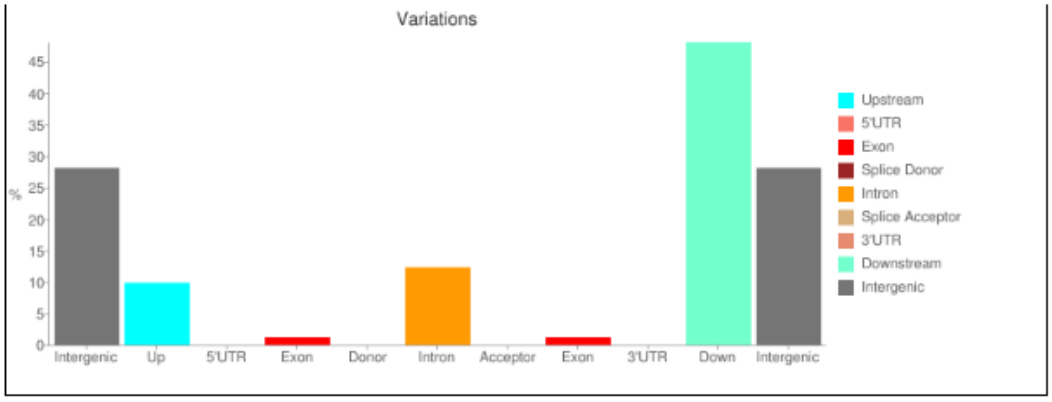

Figure S9 Summary of variant impacts in AmMYBl1 including 2 kb up- and downstream of the gene.

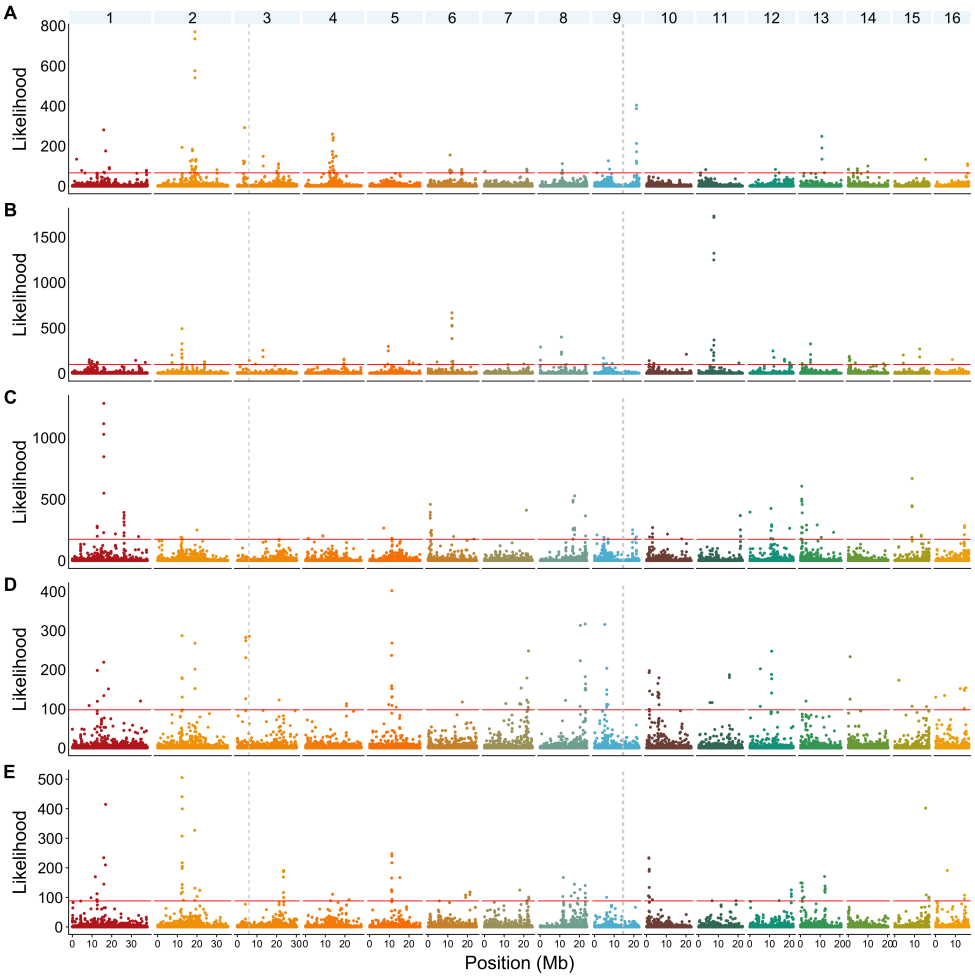

Figure S10 Hard selective sweeps identified with SweeD in cultivated and wild amaranth populations along the genome. A) caudatus, B) cruentus, C) hypochondriacus, D) hybridus, E) quitensis. Horizontal line shows top 0.5% of cutoff for outliers

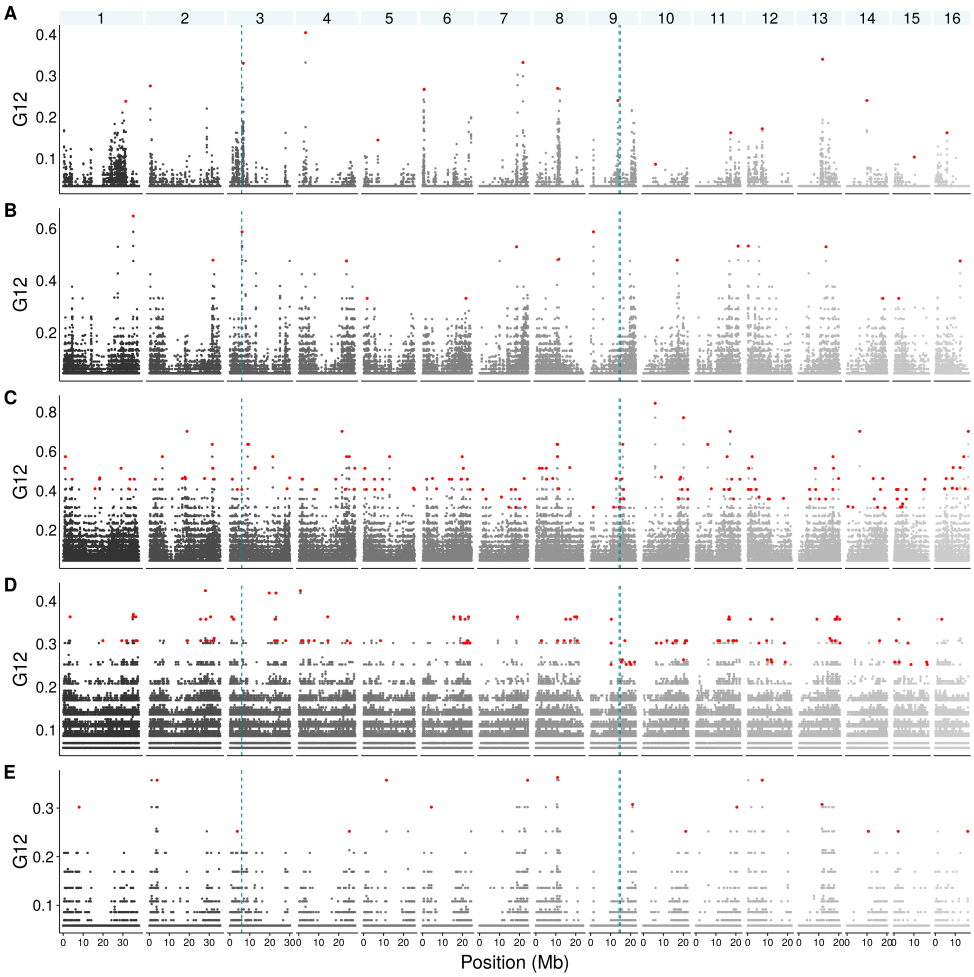

Figure S11 Hard and soft selective sweeps as identified with the G2/G1 method in cultivated and wild amaranth populations along the genome. A) caudatus, B) cruentus, C) hypochondriacus, D) hybridus, E) quitensis. In red, highlighted the top 10 G12 peaks in each chromosome.

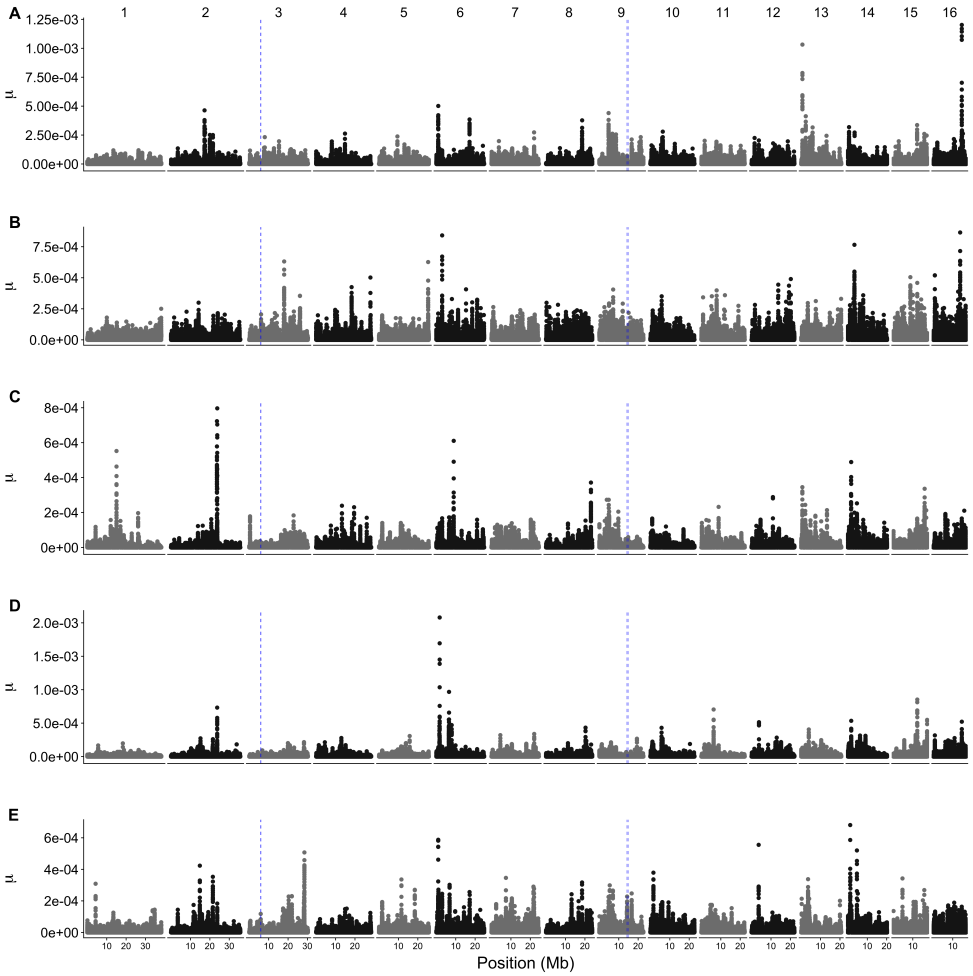

Figure S12 Hard selective sweeps as identified with RAiSD method in cultivated and wild amaranth populations along the genome. A) caudatus, B) cruentus, C) hypochondriacus, D) hybridus, E) quitensis. Horizontal line shows top 0.5% of cutoff for outliers

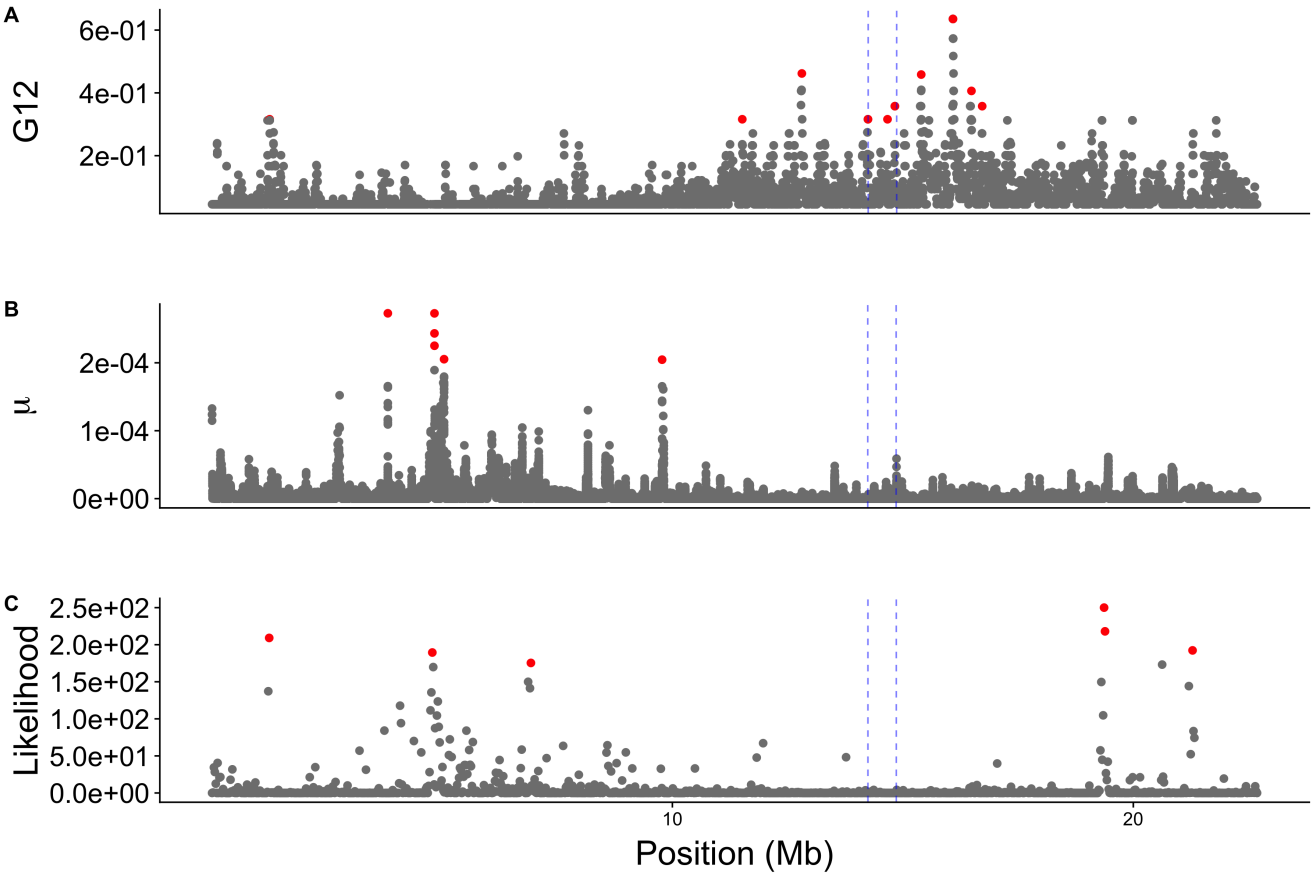

Figure S13 Top 10 selective sweep hits in hypochondriacus on Chromosome 9 using three selection tests, A) G2/G1, B) RAiSD and C) SweeD. G2/G1 identifies hard and soft selective sweeps while the other two methods only identify hard sweep signals.

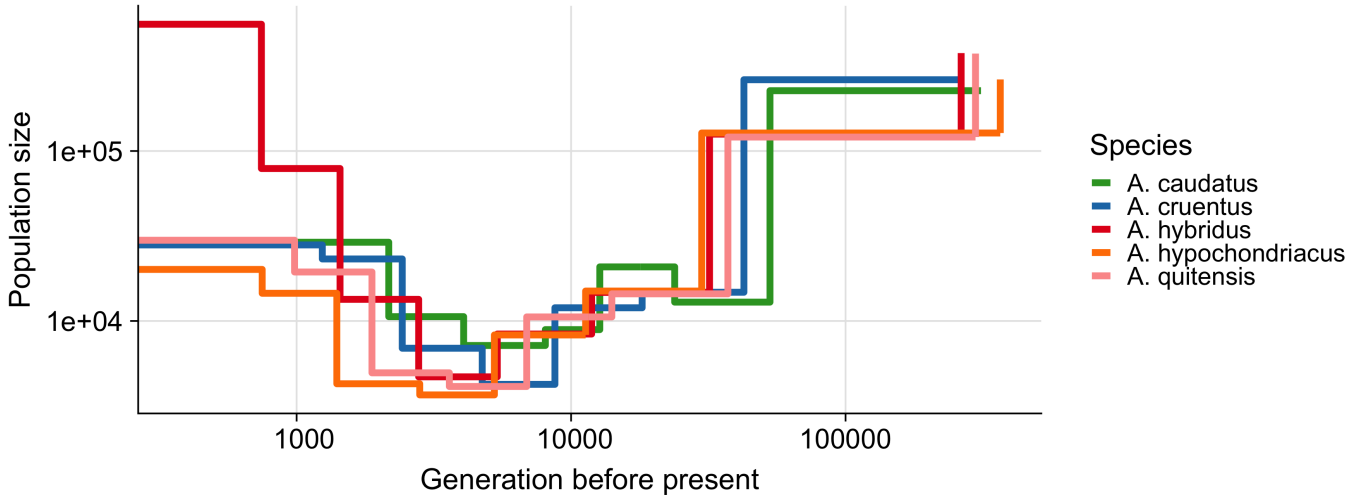

Figure S14 Demographic history for individual amaranth populations.

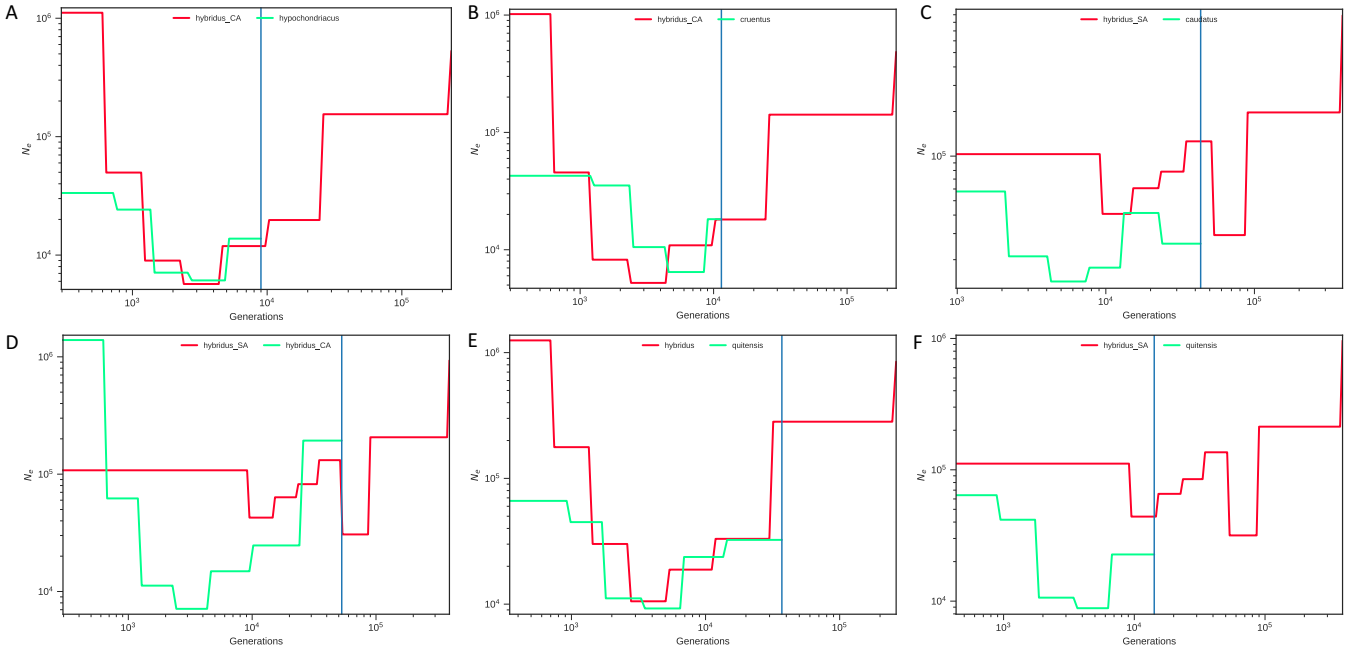

Figure S15 Split times of crops species from sepatate hybridus (SA= South America; CA = Central America) populations.

Supplementary Tables

Table S1 Accessions sequenced in this study and seed color denominated from the same seed bag as sequenced individuals. Accessions denoted as 'not clear' had seeds of both seed colors in the seed lot distributed by the USDA germplasm collection and were removed from the GWAS analysis

No	Accession number	Species	Origcty	Seed color
1	PI 490491	caudatus	ARG	dark
2	PI 511712	caudatus	ECU	dark
3	PI 608019	caudatus	ECU	dark
4	PI 649228	caudatus	PER	dark
5	PI 490511	caudatus	PER	dark
6	PI 490518	caudatus	PER	dark
7	PI 481949	caudatus	PER	dark
8	PI 511686	caudatus	PER	dark
9	PI 649217	caudatus	PER	not clear
10	PI 511690	caudatus	PER	not clear
11	PI 511680	caudatus	ARG	white
12	PI 511679	caudatus	ARG	white
13	PI 490459	caudatus	BOL	white
14	PI 511681	caudatus	BOL	white
15	PI 490604	caudatus	BOL	white
16	PI 642741	caudatus	BOL	white
17	AMA 107	caudatus	CHN	white
18	PI 490609	caudatus	ECU	white
19	PI 649230	caudatus	PER	white
20	PI 511704	caudatus	PER	white
21	PI 649227	caudatus	PER	white
22	AMA 125	caudatus	PER	white
23	PI 481957	caudatus	PER	white
24	PI 511706	caudatus	PER	white
25	PI 490561	caudatus	PER	white
26	PI 511687	caudatus	PER	white
27	PI 490612	caudatus	PER	white
28	PI 490431	caudatus	PER	white
29	PI 511696	caudatus	PER	white
30	PI 481965	caudatus	PER	white
31	Ames 5302	caudatus	PER	white
32	PI 481960	caudatus	PER	white
33	PI 511717	cruentus	GTM	dark
34	PI 667160	cruentus	GTM	dark
35	PI 658728	cruentus	MEX	dark
36	PI 511714	cruentus	PER	dark
37	PI 511713	cruentus	PER	dark
38	AMA 155	cruentus	USA	dark
39	PI 667165	cruentus	BRA	not clear
40	PI 643049	cruentus	MEX	not clear
41	Baernkrafft	cruentus	DEU	white
42	PI 433228	cruentus	GTM	white
43	PI 451826	cruentus	GTM	white
44	PI 643039	cruentus	MEX	white
45	PI 643058	cruentus	MEX	white
46	PI 649514	cruentus	MEX	white
47	PI 606798	cruentus	MEX	white
48	PI 576482	cruentus	MEX	white
49	PI 576481	cruentus	MEX	white
50	PI 649609	cruentus	MEX	white
51	PI 511723	cruentus	MEX	white
52	PI 511876	cruentus	MEX	white
53	Ames 5552	cruentus	MEX	white
54	PI 643042	cruentus	MEX	white
55	PI 649524	cruentus	MEX	white
56	PI 643037	cruentus	MEX	white
57	PI 649509	cruentus	MEX	white
58	PI 604571	hybr.	MEX	dark
59	PI 636180	hybridus	COL	dark
60	PI 490684	hybridus	ECU	dark
61	PI 511754	hybridus	ECU	dark
62	PI 490679	hybridus	ECU	dark
63	PI 490689	hybridus	ECU	dark
64	PI 490670	hybridus	ECU	dark
65	PI 490664	hybridus	ECU	dark
66	PI 490739	hybridus	ECU	dark
67	PI 667156	hybridus	ECU	dark
68	PI 490731	hybridus	ECU	dark
69	PI 667158	hybridus	GTM	dark
70	PI 604582	hybridus	MEX	dark
71	PI 604568	hybridus	MEX	dark
72	PI 604574	hybridus	MEX	dark
73	PI 511724	hybridus	MEX	dark
74	Ames 5232	hybridus	PER	dark
75	PI 490740	hybridus	PER	dark
76	PI 490489	hybridus	PER	dark
77	Ames 5335	hybridus	BOL	not clear
78	PI 652432	hypochondriacus	BRA	dark
79	PI 604581	hypochondriacus	MEX	dark
80	PI 649602	hypochondriacus	MEX	dark
81	PI 604587	hypochondriacus	MEX	dark
82	Ames 2085	hypochondriacus	MEX	not clear
83	PI 649607	hypochondriacus	MEX	white
84	PI 649587	hypochondriacus	MEX	white
85	PI 649537	hypochondriacus	MEX	white
86	PI 643036	hypochondriacus	MEX	white
87	PI 649575	hypochondriacus	MEX	white
88	Ames 5457	hypochondriacus	MEX	white
89	PI 633589	hypochondriacus	MEX	white
90	PI 649565	hypochondriacus	MEX	white
91	PI 604559	hypochondriacus	MEX	white
92	PI 643070	hypochondriacus	MEX	white
93	PI 643041	hypochondriacus	MEX	white
94	PI 604595	hypochondriacus	MEX	white
95	PI 649529	hypochondriacus	MEX	white
96	Ames 2215	hypochondriacus	MEX	white
97	PI 511731	hypochondriacus	MEX	white
98	PI 643067	hypochondriacus	MEX	white
99	PI 649559	hypochondriacus	MEX	white
100	PI 649595	hypochondriacus	MEX	white
101	PI 649623	hypochondriacus	MEX	white
102	AMA 89	powellii	RWA	dark
103	Ames 21666	quitensis	ARG	dark
104	Ames 5334	quitensis	ARG	dark
105	PI 511736	quitensis	BOL	dark
106	PI 652426	quitensis	BRA	dark
107	PI 652428	quitensis	BRA	dark
108	PI 652422	quitensis	BRA	dark
109	PI 511747	quitensis	ECU	dark
110	PI 511749	quitensis	ECU	dark
111	PI 511737	quitensis	ECU	dark
112	PI 490705	quitensis	ECU	dark
113	PI 511745	quitensis	ECU	dark
114	PI 490720	quitensis	ECU	dark
115	PI 511741	quitensis	ECU	dark
116	Ames 5247	quitensis	ECU	dark
117	PI 490673	quitensis	ECU	dark
118	PI 649246	quitensis	PER	dark
119	PI 511751	quitensis	PER	dark
120	Ames 5342	quitensis	PER	dark
121	PI 490466	quitensis	PER	dark

Table S2 Candidate genes for seed color in grain amaranth

Chromosome	Start	End	Name	Strand	Description
3	5881516	5896384	AH005257	+	Similar to CCT1: T-complex protein 1 subunit alpha (Arabidopsis thaliana)
3	5897584	5905365	AH005258	+	Similar to MLO13: MLO-like protein 13 (Arabidopsis thaliana)
3	5906387	5930826	AH005259	-	Similar to NUP160: Nuclear pore complex protein NUP160 (Arabidopsis thaliana)
3	5933604	5946653	AH005260	-	Similar to CMO: Choline monooxygenase, chloroplastic (Amaranthus tricolor)
3	5959962	5966304	AH005261	-	Similar to Ank3: Ankyrin-3 (Mus musculus)
9	14260388	14267067	AH014529	-	Similar to NYC1: Probable chlorophyll(ide) b reductase NYC1, chloroplastic (Oryza sativa subsp. japonica)
9	14276572	14283316	AH014530	+	Protein of unknown function
9	14291900	14298118	AH014531	-	Protein of unknown function
9	14298147	14298790	AH014532	-	Protein of unknown function
9	14298908	14299399	AH014533	-	Protein of unknown function
9	14304669	14312794	AH014534	+	Similar to CSLG2: Cellulose synthase-like protein G2 (Arabidopsis thaliana)
9	14333735	14334892	AH014535	+	Similar to rmnd5a: Protein RMD5 homolog A (Xenopus laevis)
9	14345025	14345715	AH014536	+	Similar to OFP8: Transcription repressor OFP8 (Arabidopsis thaliana)
9	14353195	14353923	AH014537	+	Protein of unknown function
9	14355443	14356829	AH014538	-	Similar to TRX2: Thioredoxin H2 (Arabidopsis thaliana)
9	14366001	14367256	AH014539	-	Similar to TRX2: Thioredoxin H2 (Arabidopsis thaliana)
9	14372147	14375412	AH014540	-	Similar to HSFC1: Heat stress transcription factor C-1 (Arabidopsis thaliana)
9	14378777	14389557	AH014541	-	Similar to KIN13B: Kinesin-like protein KIN-13B (Arabidopsis thaliana)
9	14396256	14396438	AH014542	+	Protein of unknown function
9	14404615	14409777	AH014543	-	Similar to Os01g0583100: Probable protein phosphatase 2C 6 (Oryza sativa subsp. japonica)
9	14412914	14418029	AH014544	-	Similar to CFIS1: Pre-mRNA cleavage factor Im 25 kDa subunit 1 (Arabidopsis thaliana)
9	14420597	14423542	AH014545	-	Similar to At4g06634: Uncharacterized zinc finger protein At4g06634 (Arabidopsis thaliana)
9	14443999	14450734	AH014546	+	Protein of unknown function
9	14467556	14471512	AH014547	+	Protein of unknown function
9	14472750	14477103	AH014548	-	Similar to UBA1C: UBP1-associated proteins 1C (Arabidopsis thaliana)
9	14480031	14487685	AH014549	+	Protein of unknown function
9	14503360	14505015	AH014550	+	Protein of unknown function
9	14513815	14518528	AH014551	+	Protein of unknown function
9	14525887	14543181	AH014552	+	Similar to XI-K: Myosin-17 (Arabidopsis thaliana)
9	14545885	14551450	AH014553	-	Similar to SPBC776.05: Uncharacterized membrane protein C776.05 (Schizosaccharomyces pombe (strain 972 / ATCC 24843))
9	14552847	14558739	AH014554	-	Similar to AL5: PHD finger protein ALFIN-LIKE 5 (Arabidopsis thaliana)
9	14580201	14584739	AH014555	+	Protein of unknown function
9	14587900	14593548	AH014556	+	Similar to NFYA6: Nuclear transcription factor Y subunit A-6 (Arabidopsis thaliana)
9	14598265	14609302	AH014557	+	Similar to BRXL4: Protein Brevis radix-like 4 (Arabidopsis thaliana)
9	14610048	14614953	AH014558	+	Similar to NLP3: Omega-amidase, chloroplastic (Arabidopsis thaliana)
9	14622402	14624168	AH014559	+	Similar to At1g54200: Protein BIG GRAIN 1-like B (Arabidopsis thaliana)
9	14641501	14645682	AH014560	+	Similar to VPS2.1: Vacuolar protein sorting-associated protein 2 homolog 1 (Arabidopsis thaliana)
9	14650590	14651111	AH014561	-	Similar to ATL53: Putative RING-H2 finger protein ATL53 (Arabidopsis thaliana)
9	14657567	14664508	AH014562	+	Similar to MSL10: Mechanosensitive ion channel protein 10 (Arabidopsis thaliana)
9	14665596	14668727	AH014563	-	Similar to RPA1A: Replication protein A 70 kDa DNA-binding subunit A (Arabidopsis thaliana)
9	14677913	14679912	AH014564	-	Similar to EFL4: Protein ELF4-LIKE 4 (Arabidopsis thaliana)
9	14700127	14700888	AH014565	+	Similar to UP3: Stress-response A/B barrel domain-containing protein UP3 (Arabidopsis thaliana)
9	14705869	14708414	AH014566	+	Similar to C1: Anthocyanin regulatory C1 protein (Zea mays)
9	14713047	14716244	AH014567	-	Similar to RPS3A: 40S ribosomal protein S3-1 (Arabidopsis thaliana)
9	14722269	14730995	AH014568	+	Protein of unknown function
9	14734261	14737531	AH014569	-	Similar to GRF5: Growth-regulating factor 5 (Arabidopsis thaliana)
9	14759660	14776842	AH014570	-	Protein of unknown function
9	14780031	14782777	AH014571	+	Similar to OPR3: 12-oxophytodienoate reductase 3 (Solanum lycopersicum)
9	14783927	14788106	AH014572	-	Similar to REM16: B3 domain-containing protein REM16 (Arabidopsis thaliana)
9	14792503	14796787	AH014573	-	Similar to REM10: B3 domain-containing protein REM10 (Arabidopsis thaliana)
9	14805984	14809221	AH014574	+	Similar to OPR3: 12-oxophytodienoate reductase 3 (Solanum lycopersicum)
9	14814160	14814522	AH014575	-	Similar to NFYB6: Nuclear transcription factor Y subunit B-6 (Arabidopsis thaliana)
9	14830784	14835140	AH014576	-	Similar to GALS1: Galactan beta-1,4-galactosyltransferase GALS1 (Arabidopsis thaliana)
9	14842490	14846361	AH014577	-	Similar to AMC5: Metacaspase-5 (Arabidopsis thaliana)
9	14855046	14855666	AH014578	-	Similar to ABP19A: Auxin-binding protein ABP19a (Prunus persica)

Supplementary Notes

Species delimitation

For six accessions, the species assignment in the genebank passport data did not match the results of the genetic analysis (Table S3) and may either represent misclassified accessions or mixup of seeds during seed handling and regeneration. These should be reconsidered for morphological classification. A possible misclassification is of particular interest interesting for the only amaranth variety ever registered in Germany, which considered to be *A. cruentus*, but according to the genetic data appears to be *A. hypochondriacus*.

Table S3 Accessions in which the species assignment in the genebank passport data is not supported by the genetic analysis.

Individual	Genebank annotation	Genomic group
PI604587	*A. hypochondriacus*	*A. cruentus*
Ames2215	*A. hypochondriacus*	*A. cruentus*
PI604595	*A. hypochondriacus*	*A. cruentus*
Baernkrafft (German variety)	*A. cruentus*	*A. hypochondriacus*
PI511723	*A. cruentus*	*A. hypochondriacus*
PI511876	*A. cruentus*	*A. hypochondriacus*