Isoform-resolved genome annotation enables mapping of tissue-specific betalain regulation and anthocyanin regulator co-option in amaranth

Tom S. Winkler[*], **Susanne K. Vollmer**[*†] **and Markus G Stetter**[*,1]
[*]Institute for Plant Sciences and Cluster of Excellence on Plant Sciences (CEPLAS), University of Cologne, Cologne, Germany, [†]Heinrich Heine University, Duesseldorf, Germany

Summary

• Betalains are coloring pigments produced only in some families of the order Caryophyllales, where they have functionally replaced anthocyanins as coloring pigments. While the betalain pathway itself is known, the tissue-specific regulation of the pathway remains mostly unclear.
• We first improve the high-quality *Amaranthus hypochondriacus* reference genome and create a highly complete genome annotation by combining *ab initio* prediction with long-read isoform sequencing data. We accurately annotate the betalain and anthocyanin pathway genes and their regulators in the genome of the betalain producing amaranth and further map the genetic basis of tissue color and tissue-specific regulation of the betalain pathway through a previously unknown MYB transcription factor, *AhMYB2*.
• Detailed haplotype analysis enabled us to identify two functional mutations that lead to a knock-out of the betalain pathway in natural accessions of amaranth. In addition, the down-regulation of *AhMYB2* in the flower leads to reduced expression of key betalain enzyme genes and loss of red flower color.
• Our improved amaranth reference genome represents the most complete genome of amaranth to date and is a valuable resource for betalain and amaranth research. The high similarity of the betalain flower regulator *AhMYB2* to anthocyanin regulators and a partially conserved interaction motif support the co-option of anthocyanin regulators for the betalain pathway as possible reason for the mutually exclusive nature of the two pigments.

KEYWORDS *Amaranthus hypochondriacus* (amaranth), Caryophyllales, flower color, betalains, flavonoids, MYB transcription factors

Introduction

Coloring pigments fulfill various functions in plants, including UV protection and pollinator attraction (Koes *et al.* 1994; Młodzińska *et al.* 2009). While anthocyanins are commonly observed as red pigments, the functionally similar but biochemically distinct betalains function as pigments in a small number of species within the Caryophyllales (Brockington *et al.* 2011; Jain and Gould 2015; Lopez-Nieves *et al.* 2018; Sunnadeniya *et al.* 2016). Like anthocyanins, betalains have been associated with photoprotection, abiotic stress resistance and antioxidant activity (Jain and Gould 2015). It is noteworthy that no plant species producing both anthocyanins and betalains has been identified, suggesting a replacement of widespread anthocyanins by betalains in multiple lineages of the Caryphyllales (Brockington *et al.* 2015; Timoneda *et al.* 2019). However, the transition and evolution of the mutually exclusive presence of these two pigment pathways remains poorly understood and requires further investigation.

Reference genome assemblies enable the identification of functional elements in genomes and allow to understand the evolution of these elements (Kersey 2019). In particular, high-quality gene annotations can give insights into gene families, biosynthetic pathways and provide an invaluable resource to study gene expression patterns and genotype-phenotype associations. The annotation and quantification of isoforms is challenging for computational annotation (Salzberg 2019). The inclusion of full-length transcript sequencing data improves accuracy and provides information about splicing-variation (Xia *et al.* 2019; Wang *et al.* 2016). Sequencing and assembling genomes are becoming common tasks and well-annotated reference genomes are available for most model plants. But these resources are still scarce for non-model species. As most betalain producing

[1] Dept. of Plant Sciences, University of Cologne, Cologne, Germany E-mail: m.stetter@uni-koeln.de

plants are non-model species, only few well-annotated reference genome sequences are available, hindering the study of the regulation of these pigments.

The anthocyanin biosynthesis pathway, as part of the flavonoid pathway, has been studied in detail and is well understood despite its high complexity and tissue-specific regulation (Deng and Lu 2017; Xu *et al.* 2015; Lai *et al.* 2013). Betalains and particularly their transcriptional regulation on the other hand, have received far less attention. Both tyrosine, as basis of the betalain pathway, and phenylalanine, as initial substrate of the general phenylpropanoid pathway, are derived from the same initial substrate in arogenate, but their downstream pathways are fundamentally different (Młodzińska *et al.* 2009; Lopez-Nieves *et al.* 2018). The biochemically rather simple betalain pathway initially requires the hydroxylation of L-tyrosine to L-3,4-dihydroxyphenylalanine (L-DOPA), and the conversion of L-DOPA to cyclo-DOPA. Both of these steps are carried out by cytochrome P450 enzymes encoded by CYP76AD genes (Hatlestad *et al.* 2012; Sunnadeniya *et al.* 2016). The common betalain chromophore betalamic acid is generated from L-DOPA through the activity of L-DOPA 4,5-dioxygenase (DODA; Christinet *et al.* 2004; Sheehan *et al.* 2020). While betalamic acid can spontaneously condense with amino acids to form yellow betaxanthins, glucosylation of intermediates is required for the production of red betacyanins (Gandía-Herrero and García-Carmona 2013; Timoneda *et al.* 2019). Glucosyl moieties can either be be added to cyclo-DOPA before condensation by cyclo-DOPA 5-O-glucosyltransferase (cDOPA5GT; Sasaki *et al.* 2004) or to Betanidin by 5- or 6-O-glucosyltransferases (Betanidin5GT and Betanidin6GT; Vogt *et al.* 1999; Vogt 2002). Despite the characterization of flavonoid and betalain pathway genes in different species, understanding of the conservation of both pathways and their regulation remains limited.

Although the main enzymes of the betalain pathway have been identified in recent years (Hatlestad *et al.* 2012; Christinet *et al.* 2004; Sasaki *et al.* 2004; Vogt *et al.* 1999; Vogt 2002), only few regulators of this pathway have been described (Hatlestad *et al.* 2015; Xie *et al.* 2023; Zhang *et al.* 2021; Zeng *et al.* 2023). For instance, MYB transcription factors are known to regulate the flavonoid and anthocyanin biosynthesis pathway and few have been identified to regulate betalain production (Deng and Lu 2017; Gonzalez *et al.* 2008; Lloyd *et al.* 2017). MYB transcription factors can be classified into subfamilies based on the number of adjacent repeats of the MYB DNA-binding domain (Jin and Martin 1999). Further classification of the R2R3 subfamily into conserved subgroups has been used to transfer experimentally validated regulatory functions to different species (Stracke *et al.* 2001; Yanhui *et al.* 2006; Stracke *et al.* 2014). Conserved subgroups participate in the regulation of the flavonoid biosynthesis in *Arabidopsis thaliana* (Stracke *et al.* 2007), grape (Kobayashi *et al.* 2002), strawberry (Schaart *et al.* 2013) and many other plant species (Naik *et al.* 2021; Lai *et al.* 2013; Brendolise *et al.* 2017). The MYB transcription factor that has been shown to regulate betalain biosynthesis in *Beta vulgaris* leaves was most similar to anthocyanin regulating transcription factors from other species (Hatlestad *et al.* 2015). This finding suggested the co-option of anthocyanin regulating transcription factors for the regulation of the chemically distinct but functionally similar betalain pathway (Hatlestad *et al.* 2015; Lloyd *et al.* 2017). However, the genetic basis and evolutionary history of this co-option remains unclear.

In this study, we generate a new reference genome for the pseudo-cereal crop amaranth, *A. hypochondriacus* L., which is part of the betalain producing Caryophyllales. We remove sequencing errors from the previous assembly (Lightfoot *et al.* 2017) and combine *ab initio* prediction and long-read isoform sequencing data to create a new, highly complete genome annotation for grain amaranth. We use this novel genome annotation to identify and annotate betalain and flavonoid color pathway genes in the genome of *A. hypochondriacus*. To identify putative regulators of these pigment pathways, we identify MYB transcription factors genome-wide and functionally assign them into conserved subgroups. Using our phylogenetic analysis, we identify multiple candidate MYBs for betalain pathway regulation in amaranth. We utilize different crosses to map genomic regions regulating amaranth tissue color and identify causal variants for color loss in a betalain pathway gene and the flower color regulator *AhMYB2* on color QTLs. *AhMYB2* has high similarity to anthocyanin regulators and regulates multiple enzymes of the betalain pathway. Understanding the tissue-specific regulation of betalain biosynthesis in amaranth can help to elucidate the evolution of this novel pigment pathway and the mutually exclusive nature of betalains and anthocyanins.

Materials and Methods

Improving the A. hypochondriacus reference genome

We first corrected indel errors in the *A. hypochondriacus* L. reference assembly version 2.1 (Lightfoot *et al.* 2017), using NextPolish (Hu *et al.* 2020). The reference genome included the assembly (version 2.0) and the annotation (version 2.1). We used the publicly available whole-genome paired-end sequencing reads from the second *A. hypochondriacus* genome assembly (Bioproject ID: PRJNA290113, SRA ID: SRR2106212) to polish the assembly (Lightfoot *et al.* 2017). To prepare the whole-genome sequencing (WGS) reads for usage with NextPolish, we removed all read pairs containing ambiguous bases using *bbduk.sh* from the BBTools suite v38.94 (https://jgi.doe.gov/data-and-tools/software-tools/bbtools/). To prepare the downloaded reference genome for genome polishing, we removed stretches of ambiguous bases, usually located at the beginning or end of scaffolds, using a custom script (https://github.com/T-Winkler/Ahyp_v2_2_publication.git). Ambiguous bases were merged back into the genome after genome polishing. We ran two cycles of polishing using the two NextPolish core modules, k-mer score chain and k-mer count (Hu *et al.* 2020). Before each polishing module, we mapped reads using bwa-mem2 v2.2.1 (Li 2013) and processed bam files using Samtools v1.13 (Li *et al.* 2009). After genome polishing, we masked the polished genome for repetitive elements in order to prepare the computational gene prediction, using Repeatmodeler v2.0.1 (Flynn *et al.* 2020) with the *-LTRStruct* option to create a database of repetitive elements from the polished assembly. We reclassified repetitive elements using RepeatClassifier v2.0.4 (Flynn *et al.* 2020) and used the repetitive element database to annotate repeats in the polished reference genome using Repeatmasker v4.1.1 (Chen 2004) with the *-e rmblast* option. We softmasked the polished genome using bedtools v2.29.2 (Quinlan and Hall 2010) *maskfasta* with the repetitive element annotation to generate the *A. hypochondriacus* genome assembly v2.2. To identify genomic regions showing mapping bias, we mapped the above described whole-genome sequencing reads to the polished reference genome using bwa-mem2 v2.2.1 (Li 2013) and calculated read depth using Samtools v1.13 (Li *et al.* 2009). We identified plastid contamination within the assembly by mapping the *B. vulgaris* mitochondrial genome (Kubo *et al.* 2000) and the *A. hypochondriacus* chloroplast genome (Chaney

et al. 2016) against the polished assembly using minimap2 v2.17 (Li 2018) with the *asm10* and *asm5* presets, respectively.

Plant material and genome annotation datasets

To guide computational prediction and augment the *ab initio* gene prediction, we used short- and long-read RNA sequencing data from a number of amaranth tissues. Both short- and long-read RNA sequencing data stems from the same accession (PI 558499) that was used to generate the *A. hypochondriacus* reference genome v2.1 (Lightfoot *et al.* 2017). The short-read RNA-seq data from eight different tissues/conditions from PI 558499 was obtained from the publicly available dataset (BioProject Accession PRJNA263128) generated by Clouse *et al.* (2016). The RNA-seq dataset consisted of 352,557,994 reads spanning a total of 31.7 Gb (average of 3.97 Gb per tissue type, 90 base pair (bp) paired-end HiSeq Illumina sequencing reads; Clouse *et al.* 2016).

For the generation of long-read RNA sequencing data, material was harvested from root, cotyledons, flower, pollen, and developing and mature seeds of *A. hypochondriacus* accession PI 558499. Root tissue and cotyledons of approximately 20 seedlings were collected 3 days after germination on wet filter paper. All other tissues were collected from one single plant, grown under controlled short-day conditions. Night and day temperatures were set to 19 °C and 24 °C, respectively. Plants were grown at 60 % air humidity under well-watered conditions. Leaf tissue was collected from a fully developed young leaf at the beginning of flowering. Flower tissue was collected from unopened flowers at the tip of the inflorescence. Pollen was collected by shaking the inflorescence inside a paper bag during full bloom. Pollen collection was repeated three times within one week. Developing seeds were extracted from the flower with forceps to avoid sampling flower tissue. Approximately 20 young seeds were collected at full size, but still soft. Mature seeds were fully ripe and stored before RNA extraction. All other samples were flash frozen in liquid nitrogen and stored at -80 °C until RNA extraction. We extracted RNA using the Invitrogen PureLink Plant RNA Reagent (Thermo Scientific™, Thermo Fisher Scientific Inc., Massachusetts, USA) following the manufacturers protocol.

PacBio isoform sequencing was conducted by the west german genome center (WGGC) using the extracted RNA samples from different tissues (Table S1). Circular consensus sequence (CCS) reads were processed using Isoseq3 v3.3.0 (https://github.com/PacificBiosciences/IsoSeq; Table S2). After primer removal and demultiplexing of the CCS reads, poly-A tails were trimmed and artificial concatemers were removed using *Isoseq3 refine* to create a set of full-length non-concatemer (FLNC) reads (requiring FL reads to include a poly-A tail >= 20 bp). We combined the FLNC reads from all sequenced tissues using Samtools v1.13 (Li *et al.* 2009) *merge*. We clustered full-length non-concatemer reads of the seven different tissues into a set of high quality isoforms with predicted accuracy $\geq$ 0.99 using *Isoseq3 cluster*.

We mapped the clustered high quality isoforms to the polished *A. hypochondriacus* reference genome using minimap2 v2.17 (Li 2018) with the *splice:hq* preset and the *–secondary=no* and *-uf* options. In order to remove redundant sequences, we collapse the mapped high quality isoforms into unique isoforms using the *collapse_isoforms_by_sam.py* script of cDNAcupcake v28.0.0 (https://github.com/Magdoll/cDNA_Cupcake). Only mapped high quality isoforms with a coverage of $\geq$ 0.95, identity $\geq$ 0.9, an identical intron chain and maximum 3′ and 5′ end differences of 1000 bp were collapsed into unique full-length transcripts. We obtained the number of clustered high quality isoforms supporting each unique full-length transcript and filtered possible 5′ degraded transcripts with the *cDNA cupcake* scripts *get_abundance_post_collapse.py* and *filter_away_subset.py*.

We used SQANTI3 v4.2 (Tardaguila *et al.* 2018) for quality control of the unique isoforms as well as potential sequencing error correction. SQANTI3 uses the genomic positions of the mapped unique full-length transcripts to correct possible sequencing errors using the reference genome sequence (Tardaguila *et al.* 2018). We predicted the coding potential of the corrected transcript sequences as well as putative open reading frames (ORFs) in the sequences using CPC2 v1.0.1 (Kang *et al.* 2017). Transcripts classified as protein-coding were required to be labeled "coding" with positive ORF integrity. To estimate the completeness of the generated transcripts, we used BUSCO (Simão *et al.* 2015) to compare the predicted protein sequences of the transcripts to the embryophyta orthoDBv10 dataset (*-l embryophyta_odb10* option), consisting of 1614 conserved single copy orthologs. The corrected transcript sequences were used to supplement the computational genome annotation. In order to assess the effect of genome polishing on the assembled long-read transcripts, we mapped the high quality isoforms to the unpolished reference genome (Lightfoot *et al.* 2017), processed the mapped reads as detailed above and evaluated the completeness of the transcript sets using BUSCO with the described settings. To compare the effect of reference genome based error correction on the transcripts, we predicted ORFs in the uncorrected transcript sequences that were mapped to the polished and unpolished reference genome using CPC2 and evaluated the completeness using BUSCO.

Computational annotation of the polished A. hypochondriacus genome

We used BRAKER2 v2.1.6 (Brůna *et al.* 2021) to perform *ab initio* gene prediction using the softmasked *A. hypochondriacus* genome assembly v2.2. In a first run, we used as protein database 3,536,219 protein sequences from 117 Viridiplantae species from orthoDB v10 (Kriventseva *et al.* 2019), as well as the protein sequences of *Amaranthus cruentus* (Ma *et al.* 2021), as external evidence to guide the annotation. We used this preliminary run of BRAKER2 to guide mapping of the RNA-seq dataset described above to the reference genome v2.2 (Table S1). Quality control of downloaded fastq files was done using FastQC v0.11.9 (Andrews 2010). We indexed the polished reference genome using STAR v2.7.8a (Dobin and Gingeras 2015) with the following settings: *sjdbOverhang* 89, *genomeSAindexNbases* 13. We mapped the RNA-seq reads to the indexed genome using the STAR v2.7.8a (Dobin and Gingeras 2015) with default settings, using the protein-based BRAKER2 annotation as splice-junction database. To generate the final computational gene annotation, we ran BRAKER2 with reference genome v2.2, using as external evidence both the protein database as well as the previously mapped short-read RNA sequencing reads. After computational annotation of the polished genome, we extracted protein and coding sequence (CDS) fasta sequences from the BRAKER2 annotation using the AUGUSTUS script *getAnnoFastaFromJoingenes.py* (Stanke *et al.* 2006), filtering out all predicted genes with internal stop codons. We divided the predicted genes into three subcategories based on the amount of external support using the *selectSupportedSubsets.py* script of BRAKER2 (Brůna *et al.* 2021). The predicted genes were classified as "full support" if all introns

of multi-exons genes were supported by external evidence (protein or RNAseq alignments) or if both start and stop codon were supported by external evidence in the case of mono-exon genes. Genes were classified into the "any support" subset, if at least one intron or start or stop codon were supported by external evidence. Otherwise, predicted genes were classified into the "no support" subset. Since the predicted genes without external support contributed only marginally to the BUSCO score of the annotation and since it was unlikely for real genes in the genome to be covered neither by RNA sequencing reads nor by protein alignments, we excluded this subset from the annotation.

Combining full-length isoform and computational annotation

In order to combine the computational prediction of gene structures of BRAKER2 with sequenced full-length transcripts, we used TSEBRA v1.0.3 (Gabriel *et al.* 2021) to select gene structure annotations based on their external support. We used the supported dataset of predicted genes and the annotation file of external support hints from the BRAKER2 gene prediction together with the genomic positions of the coding sequence of the full-length transcripts, as predicted by CPC2, and the available long-read configuration file as input for TSEBRA. We used gffread v0.12.7 (Pertea and Pertea 2020) with the *-M* and *-K* options to discard redundant gene models, removing all genes which were fully contained in another gene model, to generate the genome annotation v2.2. We manually corrected the annotation of the *AmMYBl1* gene (AHp014591) based on alignment of previous complementary DNA (cDNA) Sanger sequencing results (generated from the dark seeded accession PI 604581) to the polished reference genome using minimap2 v2.17 (Li 2018). In order to estimate the completeness of the genome annotation v2.2, we used the annotated protein sequences as input for BUSCO v5.2.2 (Simão *et al.* 2015) with the Embryophyte reference set (*-l embryophyta_odb10* option). For each annotated isoform of the genome annotation v2.2, we assessed support by assembled Iso-seq transcripts using gffcompare v0.12.6 (Pertea and Pertea 2020). To compare the annotation completeness with the previously published genome annotations of *A. hypochondriacus* and *A. cruentus*, we submitted the annotated protein sequences from the two genome annotations to BUSCO v5.2.2 using the same parameters as detailed above. We visualized the genome annotation in a circos plot using the circlize R package (Gu *et al.* 2014), calculating gene and repetitive element density in 1 Mb windows along the genome.

Functional annotation

Functional annotation of annotated protein sequences was carried out using eggNOG-mapper v2.17 (online version; Cantalapiedra *et al.* 2021, *taxonomic scope*=Viridiplantae (33090), *E-value*=0.001, *score*=60, *percentage identity*=40, *query and subject coverage*=20, transferring all non-electronic gene onthology (GO) annotations) and Mercator4 v.5.0 (Schwacke *et al.* 2019). We additionally annotated sequences using Interproscan v5.56-89.0 (Jones *et al.* 2014), retrieving pathway and GO-term annotations. We identified betalain pathway genes using BLAST (Altschul *et al.* 1997) searches with previously described betalain pathway genes (Table S3). Searches were carried out against the annotated amaranth protein sequences using BLAST+ v2.9.0-2, requiring a high-scoring query coverage of 80% (*qcov_hsp_perc 80*) and using the best hit as candidate sequence. For phylogenetic analysis of the candidate sequences, we aligned the amaranth CYP76AD and DODA protein sequences with sequences of other species (Table S4 and S5) using Clustal Omega v1.2.4 (Sievers *et al.* 2011) with default settings and constructed neighbour joining phylogenetic trees using ClustalW v2.1 (Thompson *et al.* 1994) with 1000 bootstrap replicates. We used the annotated flavonoid pathway identification pipeline KIPEs v0.35 (Pucker *et al.* 2020) in order to identify candidate genes of the flavonoid pathway. Amaranth genes were defined as candidate genes if they included all of the conserved residues defined by KIPEs. If no candidate sequence with all conserved residues could be identified, instead the amaranth gene with the highest percentage similarity and less than three missing conserved residues was defined as candidate gene.

To identify potential MYB transcription factors, we searched the predicted protein sequences of the *A. hypochondriacus* genome annotation v2.2 for matches to the MYB DNA-binding domain HMM profile (Pfam ID: PF00249, downloaded 11.01.2022) using HMMer v3.3.2 (Eddy 2009). We filtered matching domains based on score (> 25), alignment length (> 20 amino acids) and accuracy (> 0.8) and divided genes with matching domains into the R2R3, 3R and 4R subfamilies based on the number of adjacent repeats of the MYB domain using a custom script. Domains were considered adjacent, if they were located within 15 amino acids of one another. To assign the identified MYB transcription factors to their established subgroups, we aligned amaranth MYB protein sequences with the MYB proteins of *A. thaliana* (Stracke *et al.* 2001) and *B. vulgaris* (Stracke *et al.* 2014) using Clustal Omega v1.2.4 (Sievers *et al.* 2011) with default settings. We constructed a neighbour joining phylogenetic tree with 1000 bootstrap replicates from the Clustal Omega multiple sequence alignment using ClustalW v2.1 (Thompson *et al.* 1994). MYB proteins were assigned to subgroups based on the subgroup assignment of the closest *A. thaliana* homolog, as reported by Dubos *et al.* (2010) and Li *et al.* (2016). For ambiguous placement in the phylogeny, additionally the subgroup assignment of the closest *B. vulgaris* homolog was considered (Stracke *et al.* 2014). For *A. thaliana* subgroup S8, we considered the classification also used by Tombuloglu (2020). We identified potential bHLH-interacting MYB transcription factors by searches in the protein sequence using the conserved sequence motif ([D/E]Lx_2[R/K]x_3Lx_6Lx_3R), described by Zimmermann *et al.* (2004). To investigate their sequence conservation, we aligned subgroup S6 anthocyanin regulators and BvMYB1-clade proteins using ClustalOmega with default settings. We quantified expression levels of amaranth betalain and flavonoid pathway genes and MYB transcription factors in different tissues using kallisto v0.48.0 (Bray *et al.* 2016) with sequence based bias correction and the previously described short-read datasets (Table S1).

Mapping betalain related quantitative trait loci

To map the genetic basis of tissue color differences in amaranth, we created two distinct mapping populations. For the color regulator bulk segregant analysis (BSA), we used an F_2 mapping population created from a cross of PI 686465, which represents an F_1 hybrid of PI 568125 and PI 568179 (Brenner 2019), with PI 538323. For the biosynthesis BSA, the F_2 mapping population was created from a cross between PI 686465 and PI 576485. Because PI 686465 and PI 538323 were not homozygous lines the mapping populations might carry multiple haplotypes for functional loci. Plants were grown well-watered at 60 % air humidity under controlled short-day conditions with night and day temperatures set to 19°*C* and 24°*C*, respectively. For the regulator BSA, plants were bulked (69 plants per bulk) based

on flower color, however all plants produced red colored cotyledons during the seedling stage. For the biosynthesis BSA, plants were bulked (80 plants per bulk) based on whole plant color and green plants never produced red coloring pigments. We extracted bulked DNA from the pools using the DNeasy Plant Pro Kit (Qiagen, Hilden, Germany) using the manufacturers protocol. Extracted DNA was sequenced by the Cologne Center for Genomics (CCG) on a Illumina TruSeq platform in 2 x 150 bp reads. Sequencing quality was assessed using fastQC (Andrews 2010). We mapped reads to the *A. hypochondriacus* reference genome v2.2 using bwa-mem2 v2.2.1 (Li 2013). We followed GATK v4.1.7 "best practices" (McKenna *et al.* 2010) to call and filtered SNPs from the mapped reads. Only biallelic SNPs from the 16 largest Scaffolds were included in the analysis and sites were filtered for a maximum of 5 % missing values using vcftools v0.1.16 (Danecek *et al.* 2011). We predicted the effect of called SNPs on protein coding genes using SnpEff (Cingolani *et al.* 2012). We carried out BSAs using the QTLseqR package (Mansfeld and Grumet 2018) with the following filtering settings: *refAlleleFreq* = 0.2, *minTotalDepth* = 50, *maxTotalDepth* = 100, *minSampleDepth* = 20 and *minGQ* = 99. We increased the strictness of the genome-wide false-discovery G′ threshold to 3.

To compare gene expression between the green and red bulks of both BSAs, we newly bulked individuals and extracted RNA from bulked flower tissue using the Invitrogen PureLink Plant RNA Reagent (Thermo Scientific™, Thermo Fisher Scientific Inc., Massachusetts, USA) following the manufacturers protocol from 29 plants of each bulk for the biosynthesis BSA and 24 plants from each bulk for the regulator BSA. Plants were grown under conditions as described above. Sampled flower tissue was immediately stored at $-80°C$ until RNA extraction. Sequencing libraries were created using the TruSeq stranded mRNA preparation kit and sequenced by the CCG on a Illumina TruSeq platform in stranded 2 x 100 bp reads. We removed adapter sequences using trimmomatic v0.39 (Bolger *et al.* 2014) *ILLUMINACLIP* and evaluated sequencing quality using fastQC (Andrews 2010). To quantify gene expression levels in green and red bulks, we used kallisto (Bray *et al.* 2016) with previously described settings. We defined a TPM threshold of 0.5 in at least one sample to define a gene as expressed.

To confirm the presence of the mutually exclusive non-synonymous and stop variants in the *AhCYP76AD2* gene of green plants from the biosynthesis BSA, we Sanger sequenced (LightRun, Eurofins Genomics) one individual with red leaves and four individuals with green leaves. We extracted genomic DNA from individuals as detailed above. PCR amplification was done using primers SV4 and SV5 (Table S6). Sequencing results were mapped to the *AhCYP76AD2* genomic sequence and visualized using CLC DNA Workbench 5.6.1 (QIAGEN).

Results

Improved A. hypochondriacus reference genome

A high-quality reference genome sequence is a prerequisite for numerous types of analyses. To improve the available *A. hypochondriacus* genome assembly and remove indel errors we polished the genome using 38.8 Gb of Illumina reads (Lightfoot *et al.* 2017). After filtering of reads with ambiguous basecalls, 384,876,330 reads with length of 100 bp remained. Genome polishing resulted in a length increase of 105,049 bp (from 403,889,442 bp to 403,994,491 bp) of the polished reference genome compared to the unpolished reference genome (Lightfoot *et al.* 2017). To identify repeats and prepare the polished reference genome for the computational prediction, we analyzed repetitive sequences using Repeatmodeler (Flynn *et al.* 2020) and Repeatmasker (Chen 2004). In total, 56.91 % (229,915,110 bp) of the assembly were classified as retroelements (20.57 %), DNA transposons (4.86 %), or simple repeats and low complexity regions (3.08 %; Table S7). These elements were distributed across the genome with increased density in regions with lower gene content (Figure 1). We softmasked the identified repetitive sequences for downstream analysis. We further identified and annotated regions with organellar sequence similarity. These regions could have resulted from erroneous assembly or could be the result of transfer into the genome (Wang *et al.* 2018). We annotated these regions as these might disturb a number of genomic analyses (Figure S1).

Assembly of full-length transcript sequencing data

In order to improve the annotation with full-length transcript sequencing evidence, we performed Pacbio isoform sequencing of seven different tissues of the amaranth reference accession (PI 558499, Table S1). A total of 5,306,288 CCS reads across the seven different tissues were generated, resulting in a total of 163,587 high-quality transcript sequences after filtering and clustering (Table S2). Of these high-quality transcripts, 94.23 % (154,949) could be uniquely mapped to the polished reference genome and were further collapsed into 53,638 unique isoforms. We further filtered possible 5′ degraded sequences and corrected sequences according to the polished reference genome, resulting in a set of 52,468 transcript sequences from 22,165 loci. Out of these, a total of 42,835 transcripts from 18,593 genes were predicted as coding with intact ORF. Correction of long-read transcript sequences using the polished reference genome resulted in a minor increase in reported BUSCO completeness from 90.7 % to 90.8 % (Table S8).

To evaluate the effect of genome polishing on long-read 'error correction', we mapped HQ long-read clustered isoforms from all tissues to the unpolished reference genome and collapsed the mapped isoforms using identical settings as above. Genome-based correction of long-read transcripts is expected to increase transcript accuracy by correcting possible sequencing errors in the long-reads with the reference genome sequence. A decreased accuracy of transcripts after genome-based correction instead indicates sequencing errors in the assembly not present in the transcripts. We 'corrected' the sequences of the collapsed isoforms using the unpolished reference genome and predict ORFs in both the uncorrected and corrected isoforms. For a total of 43,245 transcript sequences that were predicted to be coding before and after correction using the unpolished reference genome, 35.1 % (15,194) differed in predicted ORF length. While 12,289 transcripts had longer predicted ORFs without 'correction', 2,805 were predicted with longer ORF after 'correction' with the unpolished reference genome. ORFs of transcripts with differential ORF predictions before and after 'correction' differed by an average of 140 amino acids. SQANTI based correction of long-read transcripts using the unpolished reference genome strongly decreased the reported BUSCO score from 90.6 % to 74.6 % complete (Table S8). The large difference in BUSCO score and predicted ORF length between Iso-Seq data and unpolished genome indicate that the unpolished reference genome sequence included indels that disrupted the previous annotation. In contrast to that, reference genome based correction using our polished assembly resulted in an increase in reported BUSCO completeness from 90.7 % to 90.8 % (Table S8) with only 10.1 %

of transcripts differing in predicted ORF length after correction. These results demonstrate the successful removal of sequencing errors from the reference genome through genome polishing and the positive impact for gene annotation.

Annotation of the polished A. hypochondriacus reference genome increases completeness of annotated genes

In addition to the long-read RNA annotation we predicted protein-coding genes in the polished reference genome, using BRAKER2 (Brůna *et al.* 2021) with external evidence in the form of a protein database and short-read RNA sequencing reads from eight different tissues (Table S1). Computational gene prediction resulted in a total of 32,604 predicted gene isoforms, 67.5 % (22,012) with full support, 21.5 % (7,013) with partial support and 11 % (3,579) without external support. We excluded all predicted isoforms without support from RNA sequencing or protein alignments. The filtered computational annotation included 29,025 annotated isoforms from 25,405 genes with a BUSCO score of 97.9 % complete (Table S8).

We combined the BRAKER2 genes that were at least partially supported by extrinsic evidence with the predicted coding sequence of the full-length transcripts into the new genome annotation v2.2 using TSEBRA (Gabriel *et al.* 2021). Our final genome annotation has 23,817 annotated genes of which 99.12 % locate to the first 16 Scaffolds. We found the complete intron chain of 20,959 (74.66 %) annotated isoforms to be supported by full-length transcripts from long-read sequencing data, indicating high-quality of the annotation. As expected, regions with high repeat content show less annotated genes (Figure 1). While the number of predicted genes was comparable to the previously published annotation (Lightfoot *et al.* 2017) and lower than the number of predicted genes from *A. cruentus* (Ma *et al.* 2021), the mean CDS length and exon number was slightly increased (Table 1). While the previous *A. hypochondriacus* genome annotation included only a single primary isoform for each gene, our genome annotation v2.2 includes multiple annotated isoforms for 3,484 genes, resulting in a total of 28,074 annotated isoforms. In comparison to the genome annotation of *A. cruentus* (89.8 % complete) and the previously published genome annotation v2.1 (80.7 % complete), the genome annotation v2.2 shows a strong increase in completeness with a BUSCO score of 98.0 % complete (Table S8).

We carried out functional annotation of the new gene annotation using eggNOG-mapper (Cantalapiedra *et al.* 2021), Mercator (Schwacke *et al.* 2019), and Interproscan (Jones *et al.* 2014) and annotated 21,920 genes (92.04 %) with biological functions and metabolic pathways (Mercator) and features including PFAM domains (Interproscan) or gene ontologies (eggNOG-mapper and Interproscan) (Figure S2). We were able to assign genes into 5,460 of 5,783 (94.41 %) of available Mercator bins of conserved biological processes. In addition, 20,946 (95.56 %) of annotated genes included annotations from at least two sources (Figure S2). These annotations will allow the effective use of the *A. hypochondriacus* reference genome for quantitative genetic analysis including genome wide association studies and molecular work.

Color pathways in the A. hypochondraicus genome

While gene ontologies are useful for common pathways and functions, they do not include genes required for specialized metabolites that only exist in a small number of species, such as betalains. In order to identify candidate genes for the betalain pathway, we carried out BLAST searches using the sequences of previously characterized pathway genes of different species against our predicted *A. hypochondriacus* protein sequences. For the *CYP76AD* query genes, two main candidate genes were identified, one gene as best hit for *CYP76AD1* from *B. vulgaris* and *CYP76AD2* from *A. cruentus* and one gene as best hit for *CYP76AD5* and *CYP76AD6* from *B. vulgaris* (Table S3). For *DODA1*, two different genes were identified as candidates. The candidate gene more similar to the *B. vulgaris* gene was located next to the candidate gene for *CYP76AD1* and *CYP76AD2* on Scaffold 16 (Table S3). Since multiple betalain pathway genes underwent gene duplication and neofunctionalization (Brockington *et al.* 2015), we performed phylogenetic analysis of the multiple *CYP76AD* and *DODA* candidate genes with previously described betalain pathway genes (Table S4 and Table S5). We termed the identified *CYP76AD* genes *AhCYP76AD2* and *AhCYP76AD5* for clustering with the CYP76AD α-clade and β-clade, respectively (Figure S3). We investigated the identified sequences for residues described by Brockington *et al.* (2015) to be invariant in the CYP76AD α-clade (H111, T114, T131, L186, I207, N241, V275, numbering based on the *B. vulgaris* BvCYP76AD1 protein) and found all α-clade residues in AhCYP76AD2, but none in the β-clade protein AhCYP76AD5. Phylogenetic analysis of identified DODA candidate genes indicated that both belonged into the DODAα clade with potential DODA activity (Figure S4). According to their relationship to BvDODAα1, which is involved in betalain biosynthesis (Hatlestad *et al.* 2015), and BvDODAα2, which displayed only minor DODA activity *in planta* and *in vitro* (Bean *et al.* 2018), we termed the two DODA candidates AhDODAα1 and AhDODAα2. We found that AhDODAα1 included all functionally relevant residues described for BvDODAα1 by Bean *et al.* (2018), AhDODAα2 instead included the residues described for BvDODAα2 and therefore likely only exhibits marginal DODA activity. We further identified three different glucosyltransferases (*AhcDOPA5GT*, *AhBetanidin5GT* and *AhBetanidin6GT*) based on the published sequences from *A. hypochondriacus* and *Amaranthus tricolor* (Table S3). We found strong expression of *AhCYP76AD2* and *AhDODAα1* especially in the red flower of PI 558499, as well as in root and stem tissue (Figure S5). The expression patterns of glucosyltransferase genes differed between tissues. While *cDOPA5GT* showed constitutive expression in all assessed tissues, *AhBetanidin5GT* and *AhBetanidin6GT* were most strongly expressed in flower and leaf, respectively (Figure S5). We identified candidate genes for all main betalain pathway genes in the genome annotation and were able to annotate their genomic location.

The second major coloring pigment pathway we annotated was the flavonoid biosynthesis pathway that in other species includes the anthocyanin production. We identified flavonoid pathway genes using KIPEs (Pucker *et al.* 2020) and found copies of most flavonoid pathway genes in amaranth (Table S9). Few amaranth candidate genes were reported as best candidates for multiple flavonoid pathway genes. Since these ambiguous candidate genes could be assigned to specific flavonoid pathway genes with high confidence (displaying all conserved residues), no candidates for *CHI2*, *F3'5'H*, *FNS1* and *ANS* could be inferred in *A. hypochondriacus* (Table S9). Since the *ANR* candidate gene had low sequence similarity to known homologs and missed conserved residues, we defined the gene to missing in grain amaranth (Table S9). Classification into metabolic pathways using Mercator supported all genes identified by KIPEs, except for

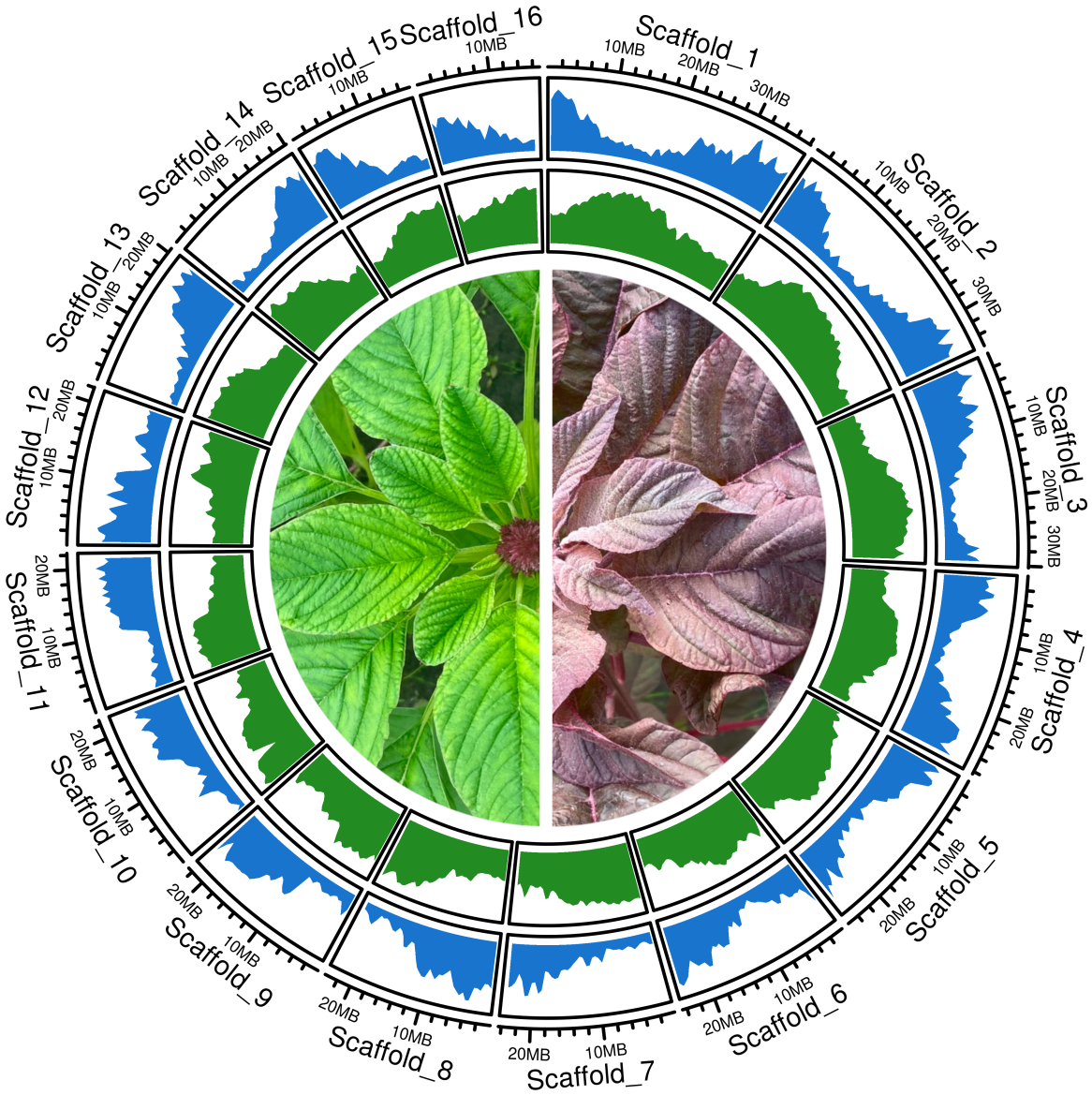

Figure 1 Circos plot of reference genome v2.2. The tracks denote gene density (blue) and repetitive element density (green) calculated in 1 Mb windows along the genome. We inserted sample pictures of green and red colored accessions into the centre.

Table 1 Comparison of genome annotation statistics between the *A. cruentus* genome (Ma *et al.* 2021), the published *A. hypochondriacus* genome v2.1 (Lightfoot *et al.* 2017) and the genome annotation derived from this study (v2.2). Coding sequence (CDS) and exon statistics for both previously published genome annotations were published in Ma *et al.* (2021). We calculated the BUSCO score using annotated protein sequences with identical parameters for all three genome annotations. Only the reported percentage of complete BUSCO genes (single copy and duplicated) is displayed.

	A. cruentus	*A. hypochondriacus* v2.1	*A. hypochondriacus* v2.2
Protein coding genes	25,477	23,883	23,817
Total isoforms	25,477	23,883	28,074
BUSCO score	89.8 %	80.7 %	98.0 %
Mean CDS length (bp)	1,141	1,066	1,258
Mean exons per gene	4.86	4.87	5.39
Mean exon length (bp)	235	219	233

additional *F3'H* candidates, and similarly inferred *F3'5'H*, *ANS* and *ANR* to be missing. We identified multiple gene copies displaying all conserved residues for the general phenylpropanoid pathway genes *PAL*, *4CL*, and *C4H* as well as for *F3'H*. We found the general phenylpropanoid pathway genes to be strongly expressed in most tissues (Figure S6). However the early flavonoid pathway gene *CHS* was mainly expressed in flower and water-stressed tissue (Figure S6), possibly limiting the the production of later flavonoid compounds in other tissues.

Location and conservation of the color regulating MYB transcription factor family

MYB transcription factors are involved in the regulation of key plant processes, including anthocyanin and betalain pigment biosynthesis (Dubos *et al.* 2010; Hatlestad *et al.* 2015; Lloyd *et al.* 2017; Liu *et al.* 2015). To identify potential regulators of the pigmentation pathways, we annotated MYB transcription factor genes in the genome of amaranth. In total, we identified 82 R2R3 MYB candidate genes (92 total isoforms), three 3R MYB candidate genes and one 4R MYB candidate gene in the genome of amaranth (2 isoforms; Table S10). The identified MYB transcription factors varied in length between 171 and 1017 amino acids (Table S10). Similar to *A. thaliana* and rice (Katiyar *et al.* 2012), most R2R3 MYB genes featured two (16 genes and isoforms) or three (60 genes and isoforms) exons (Table S10). For instance, AHp022773 (*AhMYB2*) had a conserved gene structure with three exons but two different isoforms (*AhMYB2.1* and *AhMYB2.2*) could lead to different proteins and add functional variation.

Out of the 86 identified MYB genes (R2R3, 3R and 4R), 67 could be assigned to established subgroups using a neighbour-joining phylogenetic tree (Figure 2). The identified 3R and 4R MYB genes clustered together with the 3R and 4R MYB genes of *A. thaliana* and *B. vulgaris*, respectively (Figure 2). Based on the subgroup classifications of the closest related *A. thaliana* and *B. vulgaris* MYB genes we inferred the subgroups for amaranth (Table S10). The conserved functions of subgroups across distant taxa enabled us to functionally assign MYBs to processes like stress responses, developmental processes or metabolite biosynthesis (Table S10). Only for subgroups S6 (anthocyanin biosynthesis), S12 (glucosinolate biosynthesis), S15 (epidermal cell development) and S19 (stamen development) no amaranth MYB genes were assigned (Figure 2). We classified four amaranth MYBs (AHp000789, AHp002277, AHp007908 and AHp009452) into subgroup S4, involved with transcriptional repression of the phenylpropanoid pathway (Wang *et al.* 2020). The amaranth MYBs AHp002686 and AHp017689 were classified into the early flavonoid pathway regulator subgroup S7. AHp018270 and the seed color regulation candidate AmMYBl1 (Stetter *et al.* 2020) were grouped into subgroup S5 involved with proanthocyanidin regulation. Even though no amaranth or *B. vulgaris* MYBs were classified into the anthocyanin biosynthesis subgroup S6, the betalain regulator BvMYB1 (Bv_jkkr; Hatlestad *et al.* 2015) was grouped together with three amaranth MYBs (AHp022773, AHp016530 and AHp016531) next to *A. thaliana* subgroup S6. The amaranth MYBs were assigned to the BvMYB1-clade, potentially associated with the regulation of betalain biosynthesis (Hatlestad *et al.* 2015). As candidate betalain pathway regulators, we named the transcription factors *AhMYB2* (AHp022773), *AhMYB3* (AHp016530) and *AhMYB4* (AHp016531). While we observed only low expression of *AhMYB3* and *AhMYB4* across all tissues, the two isoforms of *AhMYB2* were strongly expressed in flower tissue (Figure S7).

Different flavonoid pathway regulating MYB proteins require the formation of the MBW protein complex with two additional transcription factors (bHLH and WDR) to carry out their function (Li 2014). The interaction between MYB proteins and the bHLH interaction partners for flavonoid pathway regulation is mediated by a conserved interaction motif in the R3 domain (Zimmermann *et al.* 2004; Millard *et al.* 2019). It has been suggest that, in contrast to anthocyanin regulators, the betalain regulating *BvMYB1* has lost this interaction motive and functions without known interaction partners (Hatlestad *et al.* 2015). We therefore evaluated the presence of the conserved bHLH-interaction motif ([D/E]Lx_2[R/K]x_3Lx_6Lx_3R) in all identified amaranth MYBs. Similar to inferences in other plant species (Zimmermann *et al.* 2004; Rajput *et al.* 2022), we detected the motif in the MYB domain of four amaranth MYBs, two of four genes of S4 (AHp002277 and Ahp009452), and both genes of S5 (*AmMYBl1* and AHp018270), indicating potential interaction with bHLH proteins and transcriptional regulation through the MBW complex. A slightly altered bHLH-interaction motif in which a conserved leucine residue was replaced by isoleucine ([D/E]Lx_2[R/K]x_3Lx_6Ix_3R) was found in two *A. hypochondriacus* MYBs, AHp009565 from the AtMYB5 clade and AhMYB2.1 from the BvMYB1 clade. The motif differentiates AhMYB2.1 from other BvMYB1-clade proteins, in which the first residues of the motif are not conserved.

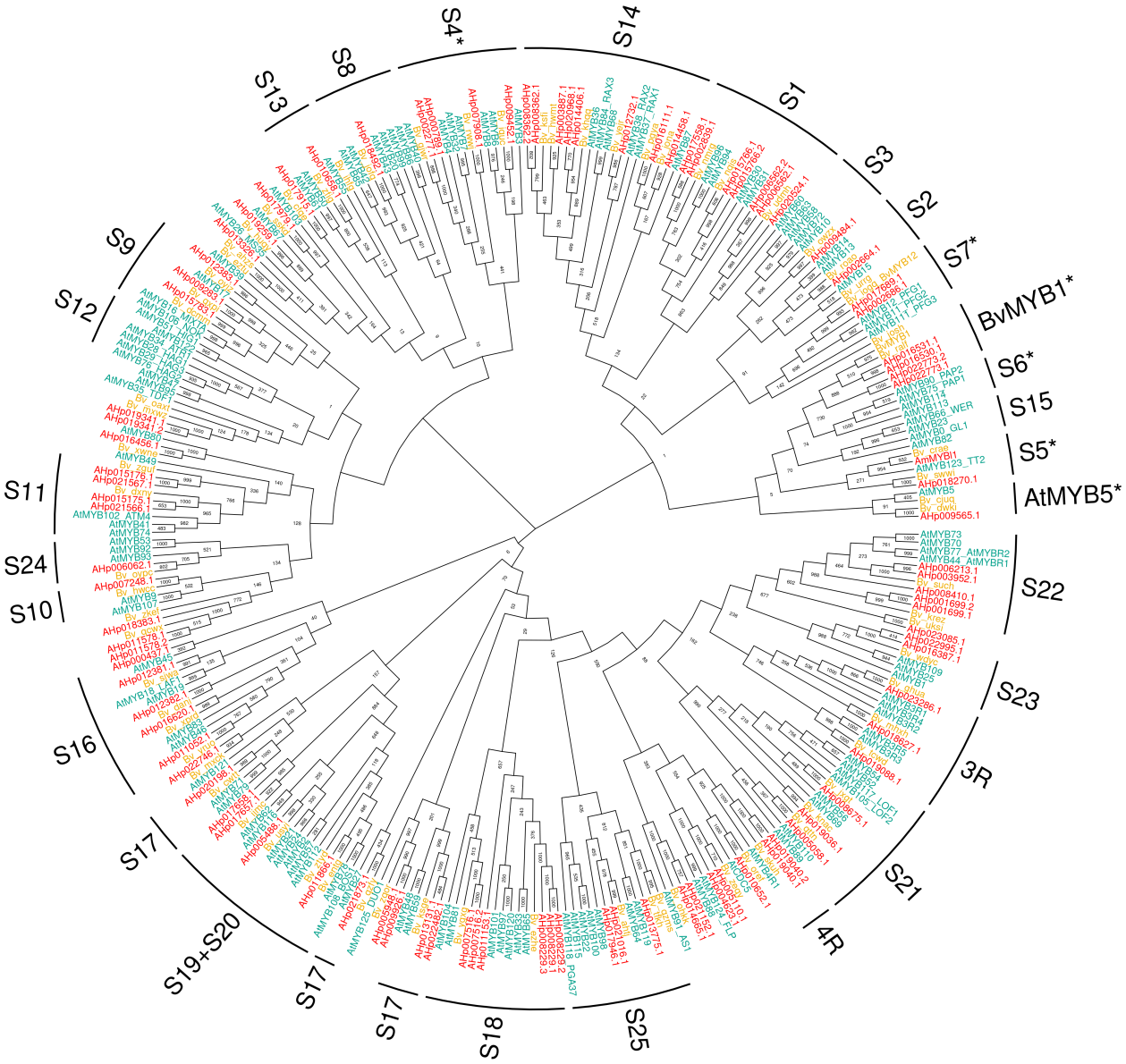

Figure 2 Neighbour-joining phylogenetic tree based on multiple sequence alignment of the R2R3, 3R and 4R MYB protein sequences of *A. hypochondriacus* (red), *A. thaliana* (green; Stracke *et al.* 2001) and *B. vulgaris* (yellow; Stracke *et al.* 2014). Bootstrap values from 1000 replicates were annotated in the phylogeny. Subgroup assignments were marked in the phylogeny, with subgroups including MYBs participating in flavonoid and betalain biosynthesis marked with asterisks.

Mapping pigment biosynthesis pathway genes

In nature pigmentation varies among individuals and even across plant organs. Identifying the evolutionary changes that impact those phenotypes is essential to understand the mutual exclusive nature of anthocyanins and betalains. To study the genetic basis of tissue color and its regulation in amaranth, we conducted two bulk segregant analyses (BSA) from different initial crosses. Plants of the biosynthesis BSA were either green (no red pigmentation in any visible tissue) or red (red in all visible organs). We pooled 80 green and red plants, respectively, to identify the mutations that likely led to a knock out of the whole pathway. A comparison of allele frequencies between bulks identified a large QTL on Scaffold 16, from 1.1 mb to 14.4 mb (from 1,080,628 bp to 14,358,251 bp; Figure 3 A and B). The two betalain pathway genes *AhCYP76AD2* and *AhDODAα1* were located on the QTL between the two peaks and represent likely candidate genes for loss of coloration (Figure 3 B). To investigate possible gene expression differences in betalain pathway genes, we quantified gene expression from bulked flower tissue for the green and red bulks of both BSAs using RNA sequencing. However, gene expression levels for *AhCYP76AD2* and *AhDODAα1* were similar between the red and green bulk of the biosynthesis BSA (Figure S8).

To identify functional variants which could lead to color loss, we used SnpEff (Cingolani *et al.* 2012) to annotate the effect of each variant which segregated in at least one of the two bulks. We searched betalain pathway genes for segregating 'high impact' variants and identified missense variants and a stop-gain variant in the *AhCYP76AD2* gene (Figure 3 C) that we previously annotated as the most likely functional homolog for this enzyme. We used the allele depth from whole-genome sequencing data from bulked tissue samples as index of the relative occurrence of reference and alternative allele in the red and green bulks. While most annotated variants occurred at similar frequencies in both bulks, two missense variants (F to I, T1258A and L to F, A1263T) and the stop gain variant at C1387T, all located on exon 2, differed strongly in relative allele depth between the two bulks (Figure 3 D and Figure S9). While C1387T was monomorph for the reference allele in the red bulk of the biosynthesis BSA (allele depth of 0/24), the alternative allele occurred in intermediate frequency in the green bulk (allele depth of 16/26). Similarly, the alternative alleles of T1258A and A1263T were found in intermediate frequency (relative allele depth 17/40 and 16/38, respectively) in the green bulk but were monomorph for the reference allele in the red bulk. The QTL included two peaks at 3.2 mb (3,205,818 bp) and 7.4 mb (7,425,588 bp) with a decreased G' value in between, which nonetheless did not go below the significance threshold (Figure 3 B). A potential explanation might be the presence of multiple causal variants in the bulks, given their multi-parent nature (see Methods).

Since the RNA-seq data had even higher depth in genes than the whole genome sequencing data, we investigated the presence and absence of the three variants in high coverage RNA sequencing data from the bulked floral tissue of additional red and green colored pools from the same cross. To resolve *AhCYP76AD2* haplotypes, we considered only read pairs which covered at least one of the non-synonymous variants and the stop variant C1387T. Inspection of read pairs originating from the green bulk revealed that either the alternative alleles of the non-synonymous variants (T1258A and A1263T) or the alternative allele of the stop variant C1387T were found in all read pairs from the green bulk (Figure 3 E). The presence of alternative alleles at T1258A and A1263T and at C1387T was almost mutually exclusive in green plants, with only two reads containing alternative alleles at C1387T and one of the missense variants. In contrast to this, less than half of all read pairs covering the *AhCYP76AD2* variants in the red bulk featured either alternative allele at T1258A, A1263T and C1387T (Figure S10), consistent with heterozygous allele frequencies at these variants. Sanger sequencing of four individuals from the green bulk and one individual from the red bulk confirmed the presence of mutually exclusive alleles in the sampled green bulk, with three individuals carrying the alternative alleles at T1258A and A1263T and one carrying the alternative allele at C1387T, while the red bulk individual featured only reference alleles (Figure S11). The presence of two mutually exclusive haplotypes can explain the decrease in G' value at the *AhCYP76AD2* locus in the biosynthesis BSA. Together, this suggests that we mapped two functional mutations that both disrupt the first enzyme in the betalain pathway, leading to loss of betalains in the whole plant.

Flower specific regulation of betalain enzymes by AhMYB2

We further intended to identify tissue-specific regulators of betalain pigmentation in amaranth. Therefore, we conducted a second BSA from a different initial cross and pooled 69 plants based on red or green flower color. All plants from both pools had red cotyledons and hence could produce betalains. In the regulator BSA, we identified a QTL for flower color at the beginning of Scaffold 16 up to 5.2 mb (1-5,244,109 bp; Figure 4 A and B). We found the newly annotated betalain pathway regulator candidate *AhMYB2* to be located at the center of the QTL of the regulator BSA on Scaffold 16 (Figure 4 B) While *AhCYP76AD2* and *AhDODAα1* were also covered by the QTL, we did not find large effect mutations differing between pools and both pools did synthesise betalains in cotyledones. We did not identify any "high impact" mutations in *AhMYB2*, but whole RNA sequencing of flower tissue from pools of diverging flower color individuals revealed a strongly reduced gene expression of both annotated *AhMYB2* isoforms in green flowers compared to red flowers. This indicates that the loss of *AhMYB2* expression is associated with the loss of flower color (Figure 4 C).

Comparing the expression of betalain pathway enzymes revealed reduced gene expression levels of *AhCYP76AD2* and *AhDODAα1* in green flowers (Figure 4 C and Figure S8). Other downstream genes of the pathway had no altered expression (Figure 4 C). Given that *AhMYB2* was most strongly expressed in flower tissue (Figure S7) and its expression is correlated with that of two betalain pathway genes, this transcription factor likely represents a flower specific regulator of the betalain pathway (Figure 4 C). We identified a slightly altered version of the bHLH-interaction motif found in S6 anthocyanin regulators in AhMYB2.1, in which a conserved leucine residue was replaced by isoleucine ([D/E]Lx_2[R/K]x_3Lx_6Ix_3R; Figure 4 D). Despite high similarity to anthocyanin regulators, the previously identified betalain MYB regulating tissue color in *B. vulgaris* leaves (BvMYB1) lacks the interaction motif and likely the ability to interact with known bHLH proteins (Hatlestad *et al.* 2015). The motif differentiates AhMYB2.1 from other BvMYB1-clade proteins, in which the first three motif residues are not fully conserved (Figure 4 D). The protein sequences of the two isoforms AhMYB2.1 and AhMYB2.2 differed only by four additional residues in AhMYB2.2, stemming from an earlier start of exon 3 (Figure 4 D). The additional residues in AhMYB2.2 disrupt the conserved R3 domain and the bHLH-interaction

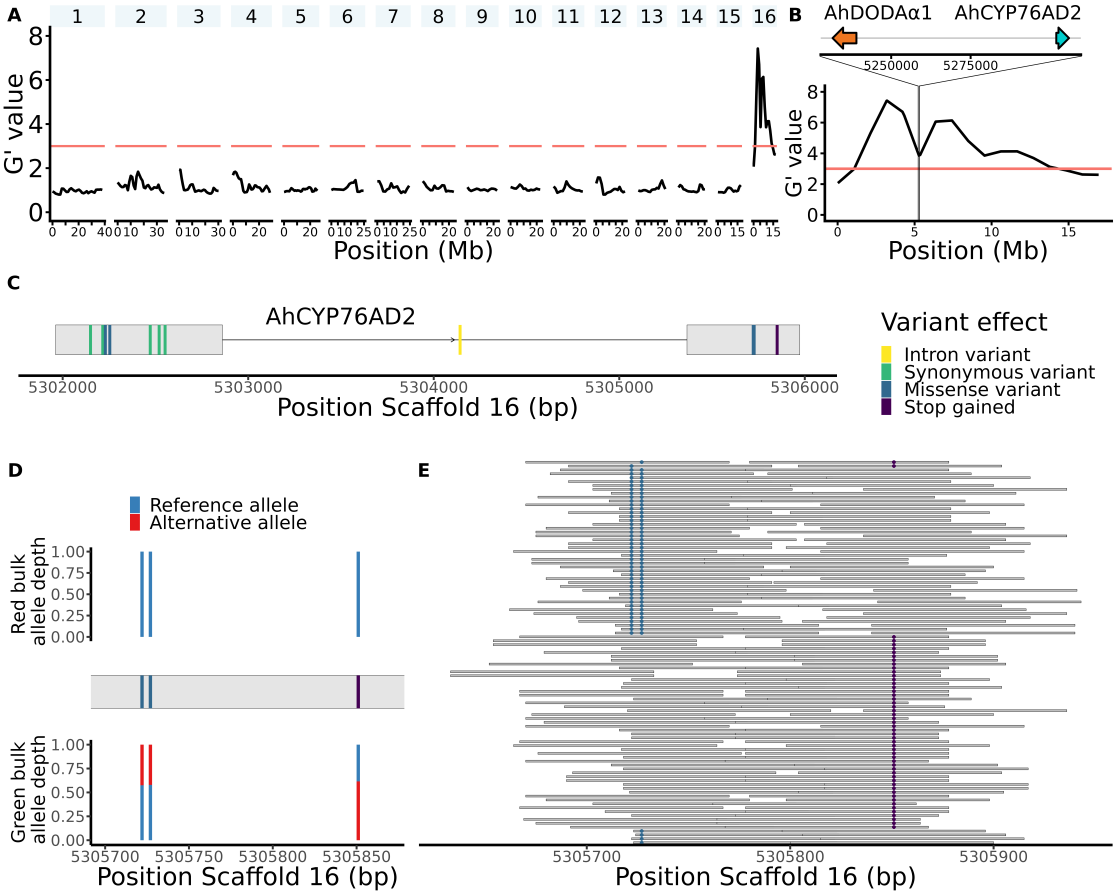

Figure 3 Biosynthesis BSA results genome-wide (A) and on Scaffold 16 (B) show QTL region on Scaffold 16. Marked by lines are the location of *AhDODAα1* (cyan) and *AhCYP76AD2* (orange) on the QTL. Variants segregating in whole-genome sequencing data from red and green bulk are displayed with assigned effect their position in the exon/intron structure of *AhCYP76AD2* (C). T1258A and A1263T (missense, dark blue) and C1387T (stop, purple) variants in the second exon of *AhCYP76AD2* and their relative allele depth are shown in (D). All RNA sequencing read pairs from green bulk flower tissue of the biosynthesis BSA covering both at least one missense and the stop variant, with marked occurrences of T1258A, A1263T and C1387T alternative alleles are displayed in (E).

motif, which might impair the function of this isoform. The altered interaction motif could impact MBW complex formation and presence of the motif in AhMYB2.1 could indicate a distinct mechanism of betalain regulation, requiring MBW complex formation similar to anthocyanin regulators.

Discussion

High quality reference genome annotation for functional genomics

We improved the available reference genome of *A. hypochondriacus* by genome polishing and subsequent genome annotation. While the *A. hypochondriacus* assembly was highly contiguous (Lightfoot *et al.* 2017) with 16 chromosome-scale Scaffolds, even single nucleotide insertions or deletions in coding sequences could have a large impact on gene prediction through frame shifts. The increased completeness of our genome annotation v2.2 compared to the previous annotation likely stems from the removal of a large number of small insertions and deletions in the previous assembly, which impaired protein prediction. Higher similarity of long-read transcript sequences to the polished reference genome and higher overall annotation completeness after genome polishing demonstrates the successful removal of small errors in the assembly (Table S8). By using the polished genome for genome annotation and adding additional full-length transcript sequencing data, we were able to produce the most complete genome annotation of amaranth to date (Clouse *et al.* 2016; Lightfoot *et al.* 2017; Ma *et al.* 2021; Wang *et al.* 2023) and included the annotation of multiple isoforms for 3,484 genes (Table 1).

Alternative splicing is involved in regulatory processes and greatly expands on the coding capacity of eukaryotes through the production of alternative isoforms (Wang *et al.* 2016; Kornblihtt *et al.* 2013). Our full-length transcript sequencing data allowed for accurate annotation of different splice variants in grain amaranth. These alternative splice variants will enable future functional studies on isoform-specific gene expression, as shown for the newly annotated *AhMYB2* gene for which we found one stronger expressed, and likely functional splice variant (Figure 4 D). This splicing variation potentially allows for additional functions or differential regulation of the gene in other organs (Huang *et al.* 2022; Kornblihtt *et al.* 2013). While we were able to annotate multiple isoforms for over 3 thousand genes, additional isoform sequencing in further tissues and accessions will be needed to represent the complete splicing variation in the species.

The reported completeness of the new genome annotation represents the most accurate genome annotation for amaranth to date (Clouse *et al.* 2016; Lightfoot *et al.* 2017; Wang *et al.* 2023; Ma *et al.* 2021) and is comparable in accuracy to well annotated model plant genomes (Dohm *et al.* 2014). It represents a valuable resource that will allow more accurate studies of regulatory processes and support functional genetics and future breeding efforts in amaranth.

Conservation of the flavonoid and betalain pigment pathways in amaranth, but loss of anthocyanin pathway genes

While betalains and anthocyanins appear to be mutually exclusive, the flavonoid pathway upstream of anthocyanin production is reported to be intact in betalain producing species and loss or downregulation of only specific anthocyanin genes appears to lead to the loss of anthocyanins in the Caryophyllales (Sakuta *et al.* 2021; Pucker *et al.* 2022). To study the conservation of the flavonoid pathway in amaranth, we identified and annotated genes of the general phenylpropanoid and flavonoid pathway using our improved genome annotation. We found no close homolog of *FNS1*, corresponding to previous reports of *FNS1* homologs being frequently absent outside of Apiaceae species (Pucker *et al.* 2020; Gebhardt *et al.* 2007). The amaranth genome did not contain a *F3'5'H* homolog, which is absent in various plant species (Table S9; Wessinger and Rausher 2012). Still, we found the flavonoid pathway to be generally intact in amaranth with loss of anthocyanidin synthase (*ANS*) and key residues in anthocyanidin reductase (*ANR*) supporting the complete loss of anthocyanin production in amaranth.

While the genes of the betalain pathway have been identified in different amaranth species in previous studies (Chang *et al.* 2021; Zheng *et al.* 2016; Casique-Arroyo *et al.* 2014), no comprehensive annotation of the pathway in *A. hypochondriacus* was available that would enable candidate gene association in mapping studies (Table S3). The characterization of functional betalain genes was complicated by multiple ancient gene duplications in the Caryophyllales, which gave rise to specific CYP76AD and DODA lineages involved in betalain production (Brockington *et al.* 2015; Sheehan *et al.* 2020). Phylogenetic analysis showed that *AhCYP76AD2* and *AhDODAα1* belong to their respective α-clades and are likely functioning in betalain biosynthesis (Figure S3 and S4; Sunnadeniya *et al.* 2016; Brockington *et al.* 2015). In contrast to *AhCYP76AD2*, *AhCYP76AD5* from the CYP76AD β-clade is likely only capable to catalyze the conversion of L-DOPA to cyclo-DOPA (Brockington *et al.* 2015). Similar to previous reports (Ma *et al.* 2021; Lightfoot *et al.* 2017), we found *AhCYP76AD2* and *AhDODAα1* to be co-localized on a conserved betalain locus on Scaffold 16 (Figure 3 B; Lightfoot *et al.* 2017; Ma *et al.* 2021). The biosynthesis BSA identified this locus to be associated with red plant pigmentation (Figure 3), further supporting the involvement of the locus in betalain biosynthesis. We identified mutually exclusive alleles in the *AhCYP76AD2* gene, associated with complete loss of red tissue color (Figure 3). Studies on loss of anthocyanin meditated flower color show that, while functional (loss-of-function in pathway genes) and regulatory mutations (leading to e.g. gene downregulation in specific tissues) occur at comparable frequencies, regulatory mutations leading to color loss appear to be preferentially fixed in natural populations, likely because of pleiotropic effects of pathway genes in vegetative tissue (Wessinger and Rausher 2012; Gates *et al.* 2018). In our population, betalain pathway genes showed even higher expression in the green bulk compared to the red bulk (Figure S8). Since betalain genes were still expressed in green tissue, causal mutations for the loss of tissue color likely led to a knock-out of *AhCYP76AD2* (Figure 3). While the knock-out allele at C1387T shows high evidence to lead to a loss of *AhCYP76AD2* function and the loss of betalain biosynthesis, the missense alleles at T1258A and A1263T might lead to the loss of function by changed protein folding or through the replacement of functionally important residues (Figure 3 C and D). The presence of independent functional mutations and the complete lack of red color might indicate dispensability of the betalain pathway, at least in individual populations. Studies on the fitness effect of betalain loss and the prevalence of knock-out mutations and their possible pleiotropic effects on plant fitness in natural amaranth populations are currently lacking.

Through mapping betalain and anthocyanin genes, we were able to find conserved patterns in both pathways, but the degradation of key anthocyanin enzymes. However, the loss of an-

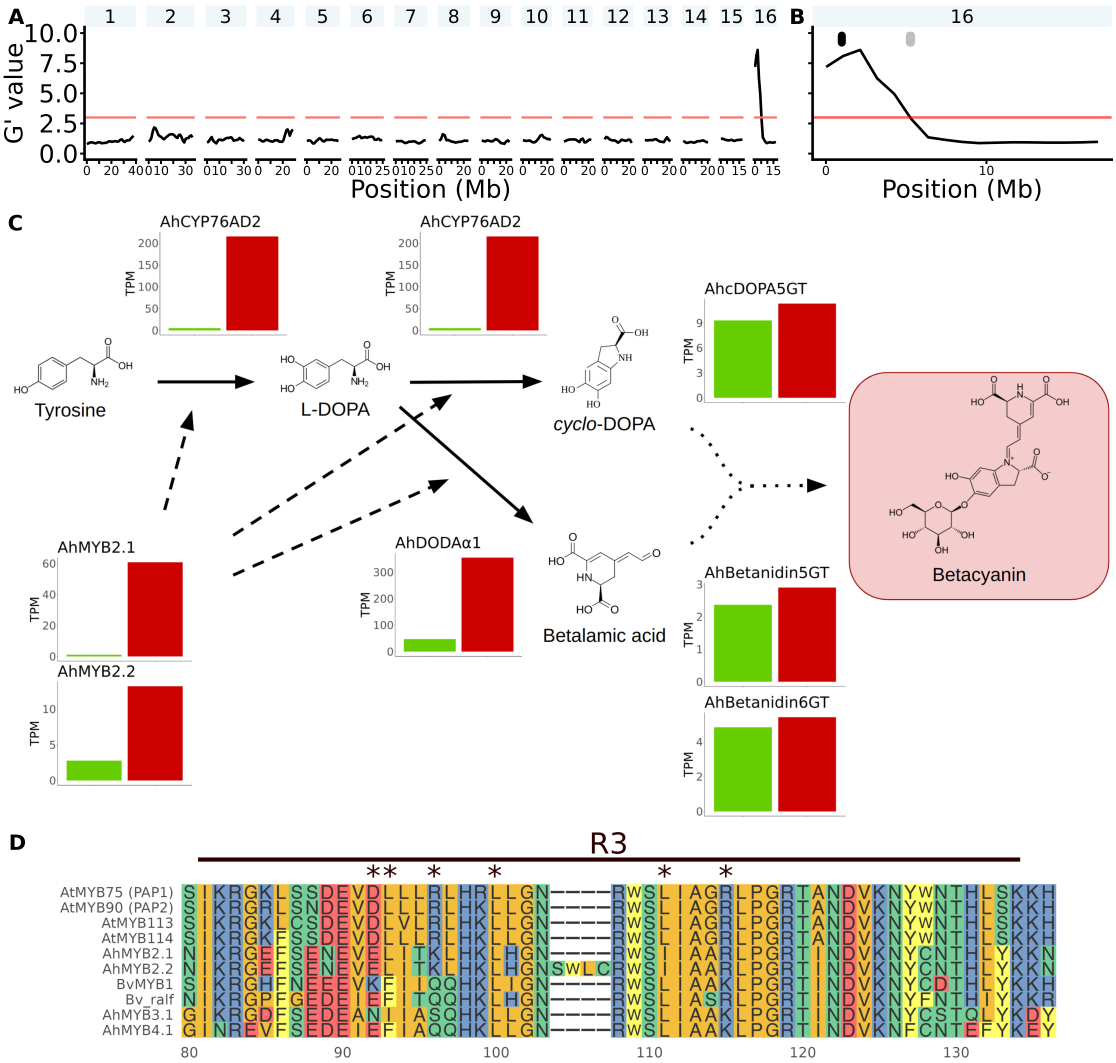

Figure 4 Regulator BSA results genome-wide (A) and for Scaffold 16 only (B) show the identified QTL region (1-5,244,109 bp) on Scaffold 16. The QTL region includes the betalain pathway genes *AhCYP76AD2* and *AhDODAα1* (grey) as well as the regulator *AhMYB2* (black) (B). A proposed model of regulation of early betalain pathway gene expression by *AhMYB2* in a simplified pathway model alongside relative gene expression [TPM] in flower tissue for betalain pathway genes in green and red flower bulk is shown in (C). Multiple enzymatic reactions and paths leading to red betacyanin production were simplified with pointed arrows. The potential betalain pathway genes *AhMYB3*, *AhMYB4* and *AhCYP76AD5* were not expressed. (D) Alignment of the R3 domain of *A. thaliana* anthocyanin regulators with BvMYB1-clade proteins. Residues of the bHLH interaction motif ([D/E]Lx$_2$[R/K]x$_3$Lx$_6$Lx$_3$R) are marked with asterisks. Residue 111 of the alignment shows the imperfectly conserved interaction motif residue in AhMYB2.1.

thocyanin genes does not have to be causal for mutual exclusiveness, since e.g., previous downregulation or replacement of anthocyanin genes would remove selective constraints to maintain the genes. Since other flavonoids can still be synthesised in betalain producing species (Iwashina 2015), the emergence of betalain biosynthesis alone is likely not sufficient to explain total replacement of anthocyanins and the observed mutual exclusiveness of the pigments across the large phylogenetic range of the Caryophyllales. To understand the subsequent order of the rise of mutual exclusiveness between betalains and anthocyanins, it might important to understand the regulation of both pathways in the same study system.

AhMYB2 is a novel regulator of the betalain pathway

Despite the characterisation of the betalain biosynthesis pathway in recent years (Hatlestad *et al.* 2012; Vogt *et al.* 1999; Christinet *et al.* 2004) and the identification of few regulators (Xie *et al.* 2023; Zhang *et al.* 2021; Zeng *et al.* 2023; Hatlestad *et al.* 2015), knowledge about tissue-specific regulation of the pathway is limited. We identified the novel amaranth flower color regulator *AhMYB2* in our regulator mapping population (Figure 4). Strongly reduced gene expression of *AhMYB2* and the pathway genes *AhCYP76AD2* and *AhDODAα1* in green flowered plants of the regulator BSA (Figure 4 C) leads us to propose *AhMYB2* as transcriptional regulator of the early betalain pathway. This is supported by the assignment of AhMYB2 to the BvMYB1-clade from phylogenetic analysis (Figure 2) and high expression only in the red flower tissue of PI 558499 (Figure S7). While we annotated two different *AhMYB2* isoforms and both showed reduced expression in green flower in the regulator BSA, the additional residues in the conserved R3 domain of AhMYB2.2 could disrupt the function of the gene and would leave AhMYB2.1 as only functional isoform (Figure 4 D). Flavonoid pathway regulating MYB transcription factors commonly control tissue-specific pigment accumulation and frequently determine specificity of the MBW complex (Rajput *et al.* 2022; Lin-Wang *et al.* 2010; Xu *et al.* 2015). While *BvMYB1* was shown to regulate the betalain pathway in *B. vulgaris* leaves, *AhMYB2* appears to regulate betalain biosynthesis in flower, instead. Similar to the tissue-specific regulation by flavonoid pathway MYBs, other BvMYB1-clade proteins might represent regulators in other tissues. While we did not observe high expression for *AhMYB3* and *AhMYB4* in any sampled tissue, the genes might be active in different conditions or genotypes and should be targets of characterization in future studies (Figure S7). The identification of *AhMYB2* flower-specific color regulator improves the understanding of the tissue-specific regulation of the early betalain pigment pathway.

The flavonoid pathway in plants is regulated by conserved subgroups of MYB transcription factors (Gonzalez *et al.* 2008; Lloyd *et al.* 2017; Liu *et al.* 2015). We found the conservation of phenylpropanoid pathway (S4), early flavonoid pathway (S7) and proanthocyanidin (S5 and AtMYB5-like) MYB regulators in amaranth, in accordance with the conservation of the flavonoid enzymes (Figure 2, Table S10). Phylogenetic assignment of the flower color regulator AhMYB2 and other amaranth MYBs into the BvMYB1-clade (Hatlestad *et al.* 2015; Stracke *et al.* 2014) next to S6 anthocyanin regulators, however, indicates a function of these MYBs in betalain pathway regulation and further supports the proposed neofunctionalization of former anthocyanin regulators for betalain pathway regulation (Figure 2 and 4, Table S10; Hatlestad *et al.* 2015; Timoneda *et al.* 2019). While MBW complex interaction is conserved in S6 anthocyanin regulators across species (Zimmermann *et al.* 2004; Xu *et al.* 2015), the ability to interact with bHLH proteins in the MBW complex might have been lost in the betalain regulator BvMYB1 (Hatlestad *et al.* 2015). In contrast to other *A. hypochondriacus* and *B. vulgaris* BvMYB1-clade MYBs (Hatlestad *et al.* 2015), AhMYB2.1 includes an altered version of the conserved bHLH-interaction motif in the R3 domain (Figure 4 D), which it shares with the amaranth MYB AHp009565. The homolog of AHp009565 in *A. thaliana*, AtMYB5, contains an canonical interaction motif (Zimmermann *et al.* 2004), suggesting that the altered bHLH-interaction motif might still enable the amaranth MYBs to interact with either known or novel bHLH proteins. Through possible MBW complex formation, AhMYB2 might regulate the betalain pathway in a novel mechanisms more similar to known anthocyanin regulators.

The rise of mutually exclusive pigment pathways

Within the Caryophyllales, betalain producing clades are intermixed with anthocyanin producing lineages, leading recent studies to propose multiple evolutionary origins of betalains (Sheehan *et al.* 2020; Timoneda *et al.* 2019). Despite this polyphyletic origin of betalain production, the biosynthesis of betalains and anthocyanins appears to be mutual exclusive in all plants (Clement and Mabry 1996). A deregulated form of the arogenate dehydrogenase gene has been suggested to have enabled the shift from anthocyanins to betalains, by increasing carbon flux of the common substrate arogenate towards the betalain precursor tyrosine and thereby decreasing the availability of phenylalanine for flavonoid (including anthocyanin) production (Lopez-Nieves *et al.* 2018; Timoneda *et al.* 2019). Yet, changes in substrate availability cannot sufficiently explain the mutual exclusiveness of the two pigments, since many other flavonoids are still readily synthesised in betalain producing species (Iwashina 2015). Both, expression differences in anthocyanin biosynthesis genes and loss of anthocyanin transport and flavonoid pathway genes contribute to prevent anthocyanin production in betalain species (Shimada *et al.* 2007; Pucker *et al.* 2022). Such regulatory or functional changes of anthocyanin genes can explain mechanisms how production of both betalains and anthocyanins is never observed, but not the evolutionary reason for alternation. Step-wise models of the shift from anthocyanin to betalain biosynthesis commonly include intermediate phases with both anthocyanin and betalain production with only a later loss of the functionally equivalent anthocyanins (Timoneda *et al.* 2019). A potential key to explain the mutual exclusiveness might be to focus instead on the common components utilized by both pathways, which includes the MYB regulators (Hatlestad *et al.* 2015). The co-option of MBW complex interacting anthocyanin MYB transcription into betalain regulation followed by loss of key residues for bHLH-interaction could have relieved anthocyanin genes of their key transcriptional regulators (Timoneda *et al.* 2019; Hatlestad *et al.* 2015). In this context, the altered bHLH-interaction motif identified in AhMYB2 could represent an interesting intermediate step in the proposed regulator shift. Further characterization of *AhMYB2* and other betalain regulators might help to understand the co-option of former anthocyanin regulators into the regulation of the novel betalain pigment pathway, the molecular prerequisites for this change and the mutual exclusiveness of the two pathways.

Acknowledgments

We thank David M. Brenner from Iowa State University and a Curator in the US National Plant Germplasm System for providing F_2 mapping populations for pigmentation gene identification. We thank Roswitha Lentz for help with DNA & RNA extraction and Kerstin Becker and Karl Koehrer from the Genomics & Transcriptomics Labor (BMFZ) at HHU for help with the PacBio sequencing. Analysis was conducted on the Regional Computing Center of the University of Cologne (RRZK) High Performance Computing (HPC) system CHEOPS (Funding number: INST 216/512/1FUGG). We acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC-2048/1 – Project ID 390686111. TW was supported by the University of Cologne Forum for Plant Biology and Politics of Nutrition (BiPoN).

Data availability

All sequencing data generated in this study are deposited under BioProject ID PRJEB65083. Code used for the study is available at https://github.com/T-Winkler/Ahyp_v2_2_publication.git. The amaranth genome and annotation v2.2 can be accessed through CoGe under Genome id66602 https://genomevolution.org/coge/GenomeInfo.pl?gid=66602. All files, including functional annotations, can be accessed at https://doi.org/10.6084/m9.figshare.23946765.

Competing interests

None declared.

Author contributions

TSW and MGS planned and designed the study. TSW and MGS performed experiments and analysed the data. SKV amplified and sequenced *AhCYP76AD2* haplotypes. TSW prepared the figures. TSW and MGS wrote the manuscript. All authors read and approved the manuscript.

ORCID

Tom S. Winkler: https://orcid.org/0009-0008-2483-8488
Susanne K. Vollmer: https://orcid.org/0009-0005-8643-1470
Markus G Stetter: https://orcid.org/0000-0001-7136-0589

References

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25**: 3389–3402.

Andrews, S., 2010 FastQC: A quality control tool for high throughput sequence data.

Bean, A., R. Sunnadeniya, N. Akhavan, A. Campbell, M. Brown, *et al.*, 2018 Gain-of-function mutations in beet *DODA2* identify key residues for betalain pigment evolution. New Phytologist **219**: 287–296.

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

Bray, N. L., H. Pimentel, P. Melsted, and L. Pachter, 2016 Near-optimal probabilistic RNA-seq quantification. Nature biotechnology **34**: 525–527.

Brendolise, C., R. V. Espley, K. Lin-Wang, W. Laing, Y. Peng, *et al.*, 2017 Multiple Copies of a Simple MYB-binding Site confers Trans-regulation by Specific Flavonoid-Related R2R3 MYBs in Diverse Species. Frontiers in plant science **8**: 1864.

Brenner, D. M., 2019 Registration of DB 199313, Cytoplasmic Male Sterile Grain Amaranth Genetic Stock. Journal of Plant Registrations **13**: 251–253.

Brockington, S. F., R. H. Walker, B. J. Glover, P. S. Soltis, and D. E. Soltis, 2011 Complex pigment evolution in the Caryophyllales. New Phytologist **190**: 854–864.

Brockington, S. F., Y. Yang, F. Gandia-Herrero, S. Covshoff, J. M. Hibberd, *et al.*, 2015 Lineage-specific gene radiations underlie the evolution of novel betalain pigmentation in Caryophyllales. New Phytologist **207**: 1170–1180.

Brůna, T., K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky, 2021 BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics and Bioinformatics **3**: 108.

Cantalapiedra, C. P., A. Hernández-Plaza, I. Letunic, P. Bork, and J. Huerta-Cepas, 2021 eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution **38**: 5825–5829.

Casique-Arroyo, G., N. Martinez-Gallardo, L. Gonzalez de la Vara, and J. P. Delano-Frier, 2014 Betacyanin Biosynthetic Genes and Enzymes Are Differentially Induced by (a)biotic stress in *Amaranthus hypochondriacus*. PLoS One **9**: e99012.

Chaney, L., R. Mangelson, T. Ramaraj, E. N. Jellen, and P. J. Maughan, 2016 The complete chloroplast genome sequences for four *Amaranthus* species (Amaranthaceae). Applications in Plant Sciences **4**: 1600063.

Chang, Y.-C., Y.-C. Chiu, N.-W. Tsao, Y.-L. Chou, C.-M. Tan, *et al.*, 2021 Elucidation of the core betalain biosynthesis pathway in *Amaranthus tricolor*. Scientific Reports **11**: 1–13.

Chen, N., 2004 Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics **5**: 4.10.1–4.10.14.

Christinet, L., F. X. Burdet, M. Zaiko, U. Hinz, and J.-P. Zrÿd, 2004 Characterization and Functional Identification of a Novel Plant 4, 5-Extradiol Dioxygenase Involved in Betalain Pigment Biosynthesis in *Portulaca grandiflora*. Plant Physiology **134**: 265–274.

Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen, *et al.*, 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. fly **6**: 80–92.

Clement, J. and T. Mabry, 1996 Pigment Evolution in the Caryophyllales: a Systematic Overview. Botanica Acta **109**: 360–367.

Clouse, J., D. Adhikary, J. Page, T. Ramaraj, M. Deyholos, *et al.*, 2016 The Amaranth Genome: Genome, Transcriptome, and Physical Map Assembly. The Plant Genome **9**.

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

Deng, Y. and S. Lu, 2017 Biosynthesis and Regulation of Phenylpropanoids in Plants. Critical Reviews in Plant Sciences **36**: 257–290.

Dobin, A. and T. R. Gingeras, 2015 Mapping RNA-seq reads with STAR. Current Protocols in Bioinformatics **51**: 11–14.

Dohm, J. C., A. E. Minoche, D. Holtgräwe, S. Capella-Gutiérrez, F. Zakrzewski, *et al.*, 2014 The genome of the recently do-

mesticated crop plant sugar beet (*Beta vulgaris*). Nature **505**: 546–549.

Dubos, C., R. Stracke, E. Grotewold, B. Weisshaar, C. Martin, *et al.*, 2010 MYB transcription factors in *Arabidopsis*. Trends in Plant Science **15**: 573–581.

Eddy, S. R., 2009 A new generation of homology search tools based on probabilistic inference. In *Genome Informatics 2009: Genome Informatics Series Vol. 23*, pp. 205–211, World Scientific.

Flynn, J. M., R. Hubley, C. Goubert, J. Rosen, A. G. Clark, *et al.*, 2020 RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences **117**: 9451–9457.

Gabriel, L., K. J. Hoff, T. Brůna, M. Borodovsky, and M. Stanke, 2021 TSEBRA: Transcript Selector for BRAKER. BMC bioinformatics **22**: 1–12.

Gandía-Herrero, F. and F. García-Carmona, 2013 Biosynthesis of betalains: yellow and violet plant pigments. Trends in plant science **18**: 334–343.

Gates, D. J., B. J. Olson, T. E. Clemente, and S. D. Smith, 2018 A novel R3 MYB transcriptional repressor associated with the loss of floral pigmentation in *Iochroma*. New Phytologist **217**: 1346–1356.

Gebhardt, Y. H., S. Witte, H. Steuber, U. Matern, and S. Martens, 2007 Evolution of Flavone Synthase I from Parsley Flavanone 3 β-Hydroxylase by Site-Directed Mutagenesis. Plant Physiology **144**: 1442–1454.

Gonzalez, A., M. Zhao, J. M. Leavitt, and A. M. Lloyd, 2008 Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in *Arabidopsis* seedlings. The Plant Journal **53**: 814–827.

Gu, Z., L. Gu, R. Eils, M. Schlesner, and B. Brors, 2014 *circlize* implements and enhances circular visualization in R. Bioinformatics **30**: 2811–2812.

Hatlestad, G. J., N. A. Akhavan, R. M. Sunnadeniya, L. Elam, S. Cargile, *et al.*, 2015 The beet Y locus encodes an anthocyanin MYB-like protein that activates the betalain red pigment pathway. Nature Genetics **47**: 92–96.

Hatlestad, G. J., R. M. Sunnadeniya, N. A. Akhavan, A. Gonzalez, I. L. Goldman, *et al.*, 2012 The beet R locus encodes a new cytochrome P450 required for red betalain production. Nature genetics **44**: 816–820.

Hu, J., J. Fan, Z. Sun, and S. Liu, 2020 NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics **36**: 2253–2255.

Huang, C.-K., W.-D. Lin, and S.-H. Wu, 2022 An improved repertoire of splicing variants and their potential roles in Arabidopsis photomorphogenic development. Genome Biology **23**: 1–28.

Iwashina, T., 2015 Flavonoid Properties in Plant Families Synthesizing Betalain Pigments. Natural Product Communications **10**: 1934578X1501000675.

Jain, G. and K. S. Gould, 2015 Are betalain pigments the functional homologues of anthocyanins in plants? Environmental and Experimental Botany **119**: 48–53.

Jin, H. and C. Martin, 1999 Multifunctionality and diversity within the plant *MYB*-gene family. Plant molecular biology **41**: 577–585.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, *et al.*, 2014 InterProScan 5: genome-scale protein function classification. Bioinformatics **30**: 1236–1240.

Kang, Y.-J., D.-C. Yang, L. Kong, M. Hou, Y.-Q. Meng, *et al.*, 2017 CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Research **45**: W12–W16.

Katiyar, A., S. Smita, S. K. Lenka, R. Rajwanshi, V. Chinnusamy, *et al.*, 2012 Genome-wide classification and expression analysis of *MYB* transcription factor families in rice and *Arabidopsis*. BMC genomics **13**: 1–19.

Kersey, P. J., 2019 Plant genome sequences: past, present, future. Current opinion in plant biology **48**: 1–8.

Kobayashi, S., M. Ishimaru, K. Hiraoka, and C. Honda, 2002 *Myb*-related genes of the Kyoho grape (*Vitis labruscana*) regulate anthocyanin biosynthesis. Planta **215**: 924–933.

Koes, R. E., F. Quattrocchio, and J. N. Mol, 1994 The flavonoid biosynthetic pathway in plants: Function and evolution. BioEssays **16**: 123–132.

Kornblihtt, A. R., I. E. Schor, M. Alló, G. Dujardin, E. Petrillo, *et al.*, 2013 Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nature reviews Molecular cell biology **14**: 153–165.

Kriventseva, E. V., D. Kuznetsov, F. Tegenfeldt, M. Manni, R. Dias, *et al.*, 2019 OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Research **47**: D807–D811.

Kubo, T., S. Nishizawa, A. Sugawara, N. Itchoda, A. Estiati, *et al.*, 2000 The complete nucleotide sequence of the mitochondrial genome of sugar beet (*Beta vulgaris L.*) reveals a novel gene for $tRNA^{Cys}$ (GCA). Nucleic acids research **28**: 2571–2576.

Lai, Y., H. Li, and M. Yamagishi, 2013 A review of target gene specificity of flavonoid R2R3-MYB transcription factors and a discussion of factors contributing to the target gene selectivity. Frontiers in biology **8**: 577–598.

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

Li, H., 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**: 3094–3100.

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

Li, S., 2014 Transcriptional control of flavonoid biosynthesis: Fine-tuning of the MYB-bHLH-WD40 (MBW) complex. Plant Signaling & Behavior **9**: e27522.

Li, X., C. Xue, J. Li, X. Qiao, L. Li, *et al.*, 2016 Genome-wide Identification, Evolution and Functional Divergence of MYB Transcription Factors in Chinese White Pear (*Pyrus bretschneideri*). Plant and Cell Physiology **57**: 824–847.

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**: 1–15.

Lin-Wang, K., K. Bolitho, K. Grafton, A. Kortstee, S. Karunairetnam, *et al.*, 2010 An R2R3 MYB transcription factor associated with regulation of the anthocyanin biosynthetic pathway in Rosaceae. BMC plant biology **10**: 1–17.

Liu, J., A. Osbourn, and P. Ma, 2015 MYB Transcription Factors as Regulators of Phenylpropanoid Metabolism in Plants. Molecular Plant **8**: 689–708.

Lloyd, A., A. Brockman, L. Aguirre, A. Campbell, A. Bean, *et al.*, 2017 Advances in the MYB–bHLH–WD repeat (MBW) Pigment Regulatory Model: Addition of a WRKY Factor and Co-option of an Anthocyanin MYB for Betalain Regulation. Plant and Cell Physiology **58**: 1431–1441.

Lopez-Nieves, S., Y. Yang, A. Timoneda, M. Wang, T. Feng, *et al.*, 2018 Relaxation of tyrosine pathway regulation underlies the evolution of betalain pigmentation in Caryophyllales. New Phytologist **217**: 896–908.

Ma, X., F. E. Vaistij, Y. Li, W. S. Jansen van Rensburg, S. Harvey, *et al.*, 2021 A chromosome-level *Amaranthus cruentus* genome assembly highlights gene family evolution and biosynthetic gene clusters that may underpin the nutritional value of this traditional crop. The Plant Journal **107**: 613–628.

Mansfeld, B. N. and R. Grumet, 2018 QTLseqr: An R Package for Bulk Segregant Analysis with Next-Generation Sequencing. The plant genome **11**: 180006.

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 **20**: 1297–1303.

Millard, P. S., K. Weber, B. B. Kragelund, and M. Burow, 2019 Specificity of MYB interactions relies on motifs in ordered and disordered contexts. Nucleic Acids Research **47**: 9592–9608.

Młodzińska, E. *et al.*, 2009 Survey of Plant Pigments: Molecular and Environmental Determinants of Plant Colors. Acta Biologica Cracovienca Series Botanica **51**: 7–16.

Naik, J., R. Rajput, B. Pucker, R. Stracke, and A. Pandey, 2021 The R2R3-MYB transcription factor MtMYB134 orchestrates flavonol biosynthesis in *Medicago truncatula*. Plant Molecular Biology **106**: 157–172.

Pertea, G. and M. Pertea, 2020 GFF utilities: GffRead and GffCompare. F1000Research **9**.

Pucker, B., F. Reiher, and H. M. Schilbert, 2020 Automatic Identification of Players in the Flavonoid Biosynthesis with Application on the Biomedicinal Plant *Croton tiglium*. Plants **9**: 1103.

Pucker, B., N. Walker-Hale, W. C. Yim, J. C. Cushman, A. Crumm, *et al.*, 2022 Evolutionary blocks to anthocyanin accumulation and the loss of an anthocyanin carrier protein in betalain-pigmented Caryophyllales. bioRxiv pp. 2022–10.

Quinlan, A. R. and I. M. Hall, 2010 Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**: 841–842.

Rajput, R., S. Tyagi, J. Naik, B. Pucker, R. Stracke, *et al.*, 2022 The R2R3-MYB gene family in *Cicer arietinum*: genome-wide identification and expression analysis leads to functional characterization of proanthocyanidin biosynthesis regulators in the seed coat. Planta **256**: 67.

Sakuta, M., A. Tanaka, K. Iwase, M. Miyasaka, S. Ichiki, *et al.*, 2021 Anthocyanin synthesis potential in betalain-producing Caryophyllales plants. Journal of plant research **134**: 1335–1349.

Salzberg, S. L., 2019 Next-generation genome annotation: we still struggle to get it right. Genome biology **20**: 1–3.

Sasaki, N., T. Adachi, T. Koda, and Y. Ozeki, 2004 Detection of UDP–glucose: *cyclo*-DOPA 5-O-glucosyltransferase activity in four o'clocks (*Mirabilis jalapa L.*). FEBS letters **568**: 159–162.

Schaart, J. G., C. Dubos, I. Romero De La Fuente, A. M. van Houwelingen, R. C. de Vos, *et al.*, 2013 Identification and characterization of MYB-bHLH-WD40 regulatory complexes controlling proanthocyanidin biosynthesis in strawberry (Fragaria×ananassa) fruits. New phytologist **197**: 454–467.

Schwacke, R., G. Y. Ponce-Soto, K. Krause, A. M. Bolger, B. Arsova, *et al.*, 2019 Mapman4: a Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. Molecular plant **12**: 879–892.

Sheehan, H., T. Feng, N. Walker-Hale, S. Lopez-Nieves, B. Pucker, *et al.*, 2020 Evolution of L-DOPA 4, 5-dioxygenase activity allows for recurrent specialisation to betalain pigmentation in Caryophyllales. New Phytologist **227**: 914–929.

Shimada, S., H. Otsuki, and M. Sakuta, 2007 Transcriptional control of anthocyanin biosynthetic genes in the Caryophyllales. Journal of Experimental Botany **58**: 957–967.

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, *et al.*, 2011 Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology **7**: 539.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**: 3210–3212.

Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack, *et al.*, 2006 AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Research **34**: W435–W439.

Stetter, M. G., M. Vidal-Villarejo, and K. J. Schmid, 2020 Parallel Seed Color Adaptation during Multiple Domestication Attempts of an Ancient New World Grain. Molecular Biology and Evolution **37**: 1407–1419.

Stracke, R., D. Holtgräwe, J. Schneider, B. Pucker, T. Rosleff Sörensen, *et al.*, 2014 Genome-wide identification and characterisation of R2R3-MYB genes in sugar beet (*Beta vulgaris*). BMC Plant Biology **14**: 249.

Stracke, R., H. Ishihara, G. Huep, A. Barsch, F. Mehrtens, *et al.*, 2007 Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the *Arabidopsis thaliana* seedling. The Plant Journal **50**: 660–677.

Stracke, R., M. Werber, and B. Weisshaar, 2001 The R2R3-MYB gene family in *Arabidopsis thaliana*. Current Opinion in Plant Biology **4**: 447–456.

Sunnadeniya, R., A. Bean, M. Brown, N. Akhavan, G. Hatlestad, *et al.*, 2016 Tyrosine Hydroxylation in Betalain Pigment Biosynthesis is Performed by Cytochrome P450 Enzymes in Beets (*Beta vulgaris*). PloS one **11**: e0149417.

Tardaguila, M., L. de la Fuente, C. Marti, C. Pereira, F. J. Pardo-Palacios, *et al.*, 2018 SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Research **28**: 396–411.

Thompson, J. D., D. G. Higgins, and T. J. Gibson, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research **22**: 4673–4680.

Timoneda, A., T. Feng, H. Sheehan, N. Walker-Hale, B. Pucker, *et al.*, 2019 The evolution of betalain biosynthesis in Caryophyllales. New Phytologist **224**: 71–85.

Tombuloglu, H., 2020 Genome-wide identification and expression analysis of R2R3, 3R-and 4R-MYB transcription factors during lignin biosynthesis in flax (*Linum usitatissimum*). Genomics **112**: 782–795.

Vogt, T., 2002 Substrate specificity and sequence analysis define a polyphyletic origin of betanidin 5-and 6-*O*-glucosyltransferase from *Dorotheanthus bellidiformis*. Planta **214**: 492–495.

Vogt, T., R. Grimm, and D. Strack, 1999 Cloning and expression of a cDNA encoding betanidin 5-O-glucosyltransferase, a betanidin-and flavonoid-specific enzyme with high homology

to inducible glucosyltransferases from the Solanaceae. The Plant Journal **19**: 509–519.

Wang, B., E. Tseng, M. Regulski, T. A. Clark, T. Hon, *et al.*, 2016 Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nature communications **7**: 11708.

Wang, H., D. Xu, S. Wang, A. Wang, L. Lei, *et al.*, 2023 Chromosome-scale *Amaranthus tricolor* genome provides insights into the evolution of the genus *Amaranthus* and the mechanism of betalain biosynthesis. DNA Research **30**: dsac050.

Wang, W., M. Schalamun, A. Morales-Suarez, D. Kainer, B. Schwessinger, *et al.*, 2018 Assembly of chloroplast genomes with long-and short-read data: a comparison of approaches using *Eucalyptus pauciflora* as a test case. BMC genomics **19**: 1–15.

Wang, X.-C., J. Wu, M.-L. Guan, C.-H. Zhao, P. Geng, *et al.*, 2020 *Arabidopsis* MYB4 plays dual roles in flavonoid biosynthesis. The Plant Journal **101**: 637–652.

Wessinger, C. A. and M. D. Rausher, 2012 Lessons from flower colour evolution on targets of selection. Journal of Experimental Botany **63**: 5741–5749.

Xia, E., F. Li, W. Tong, H. Yang, S. Wang, *et al.*, 2019 The tea plant reference genome and improved gene annotation using long-read and paired-end sequencing data. Scientific data **6**: 122.

Xie, F., C. Chen, J. Chen, J. Chen, Q. Hua, *et al.*, 2023 Betalain biosynthesis in red pulp pitaya is regulated via HuMYB132: a R-R type MYB transcription factor. BMC Plant Biology **23**: 1–14.

Xu, W., C. Dubos, and L. Lepiniec, 2015 Transcriptional control of flavonoid biosynthesis by MYB-bHLH-WDR complexes. Trends in Plant Science **20**: 176–185.

Yanhui, C., Y. Xiaoyuan, H. Kun, L. Meihua, L. Jigang, *et al.*, 2006 The MYB transcription factor superfamily of *Arabidopsis*: expression analysis and phylogenetic comparison with the rice MYB family. Plant molecular biology **60**: 107–124.

Zeng, J., J. Chen, K. Shah, F. Xie, C. Chen, *et al.*, 2023 Identification of HuSPL family and key role of HuSPL12 in regulation of betalain biosynthesis in pitaya. Physiologia Plantarum p. e13923.

Zhang, L., C. Chen, F. Xie, Q. Hua, Z. Zhang, *et al.*, 2021 A novel WRKY transcription factor HmoWRKY40 associated with betalain biosynthesis in pitaya (*Hylocereus monacanthus*) through regulating *HmoCYP76AD1*. International Journal of Molecular Sciences **22**: 2171.

Zheng, X., S. Liu, C. Cheng, R. Guo, Y. Chen, *et al.*, 2016 Cloning and expression analysis of betalain biosynthesis genes in *Amaranthus tricolor*. Biotechnology Letters **38**: 723–729.

Zimmermann, I. M., M. A. Heim, B. Weisshaar, and J. F. Uhrig, 2004 Comprehensive identification of *Arabidopsis thaliana* MYB transcription factors interacting with R/B-like BHLH proteins. The Plant Journal **40**: 22–34.

Supplement

Table S1 Summary of tissue of origin, read count, and source of long-read and short-read RNA sequencing data used to generate genome annotation v2.2. For long-read data, the number of reads refers to CCS reads. For short-read data, the number of reads refers to read pairs.

Tissue	Sequencing type	Reads	Uniquely mapped (%)	Source
Pollen	long-read	637,760	-	this study
Root	long-read	752,123	-	this study
Root	short-read	22,656,771	93.13	Clouse *et al.* (2016)
Cotyledones	long-read	928,752	-	this study
Cotyledones	short-read	22,291,775	89.50	Clouse *et al.* (2016)
Flower	long-read	691,694	-	this study
Flower	short-read	21,501,431	94.93	Clouse *et al.* (2016)
Leaf	long-read	478,352	-	this study
Leaf	short-read	22,045,070	94.28	Clouse *et al.* (2016)
Developing seed	long-read	371,693	-	this study
Developing seed	short-read	21,768,342	96.31	Clouse *et al.* (2016)
Mature seed	long-read	1,445,914	-	this study
Mature seed	short-read	22,242,512	92.17	Clouse *et al.* (2016)
Stem	short-read	22,162,620	95.61	Clouse *et al.* (2016)
Water-stressed tissues	short-read	21,610,476	95.37	Clouse *et al.* (2016)

Table S2 Number of PacBio Iso-Seq CCS and FLNC reads for all sequenced tissues. We combined the FLNC reads from all tissues before clustering FLNC reads into a set of HQ isoforms from all tissues.

	CCS reads	FLNC reads	HQ clustered
Root	752,123	734,212	
Cotyledone	928,752	895,195	
Flower	691,694	676,684	
Leaf	478,352	467,382	
Pollen	637,760	624,434	
Developing seed	371,693	363,574	
Mature seed	1,445,914	1,404,972	
Combined	5,306,288	5,166,453	163,587

Table S3 Summary of blast results for betalain pathway genes. GenBank or Uniprot entries for query genes are noted alongside their original publication. For each query, the best hit in the annotated *A. hypochondriacus* protein sequences and the percentage identity as reported by blast is displayed.

Pathway gene	Query GenBank	Query Uniprot	Reference	Query species	Best hit	Percentage identity
CYP76AD1	-	C76AD_BETVU	Hatlestad *et al.* (2012)	*B. vulgaris*	AHp023148	84.14
CYP76AD2	HQ656025.1	-	Hatlestad *et al.* (2012)	*A. cruentus*	AHp023148	99.80
CYP76AD5	-	A0A0U1YLX5_BETVU	Sunnadeniya *et al.* (2016)	*B. vulgaris*	AHp000674	83.43
CYP76AD6	-	A0A0U1YJA0_BETVU	Sunnadeniya *et al.* (2016)	*B. vulgaris*	AHp000674	82.16
DODA1	HQ889614.1	-	Casique-Arroyo *et al.* (2014)	*A. hypochondriacus*	AHp010386	100.00
DODA1	-	DOD1U_BETVU	Hatlestad *et al.* (2012)	*B. vulgaris*	AHp023147	78.10
cDOPA5GT	KJ136018.1	-	Casique-Arroyo *et al.* (2014).	*A. hypochondriacus*	AHp007219	100.00
Betanidin5GT	KJ136017.1	-	Casique-Arroyo *et al.* (2014)	*A. hypochondriacus*	AHp001663	99.79
Betanidin6GT	KP174812.1	-	Zheng *et al.* (2016)	*A. tricolor*	AHp005940	92.96

Table S4 Sequences used for phylogenetic analysis of amaranth CYP76AD genes. Species, gene name and GenBank accession of input genes for phylogenetic analysis.

Species	Gene name	GenBank accession
Arabidopsis thaliana	AtCYP76C2	NP_182081
Arabidopsis thaliana	AtCYP94B3	NP_190421
Basella alba	BaCYP76AD13	AJD87469
Bougainvillea peruviana	BpCYP76AD1-1	BCD59210
Bougainvillea peruviana	BpCYP76AD1-2	BCD59211
Bougainvillea peruviana	BpCYP76AD1-3	BCD59212
Bougainvillea peruviana	BpCYP76AD2	BCD59213
Beta vulgaris	BvCYP76AD1	AET43290
Beta vulgaris	BvCYP76AD5	AJD87473
Beta vulgaris	BvCYP76AD6	AJD87474
Chenopodium quinoa	CqCYP76AD1	XP_021769302
Mirabilis jalapa	MjCYP76AD3	AET43292
Mirabilis jalapa	MjCYP76AD7	AJD87463
Mirabilis jalapa	MjCYP76AD15	AJD87471

Table S5 Sequences used for phylogenetic analysis of amaranth DODA genes. Species, gene name and GenBank accession of input genes for phylogenetic analysis.

Species	Gene name	GenBank accession
Beta vulgaris	BvDODAα1	MN153194
Beta vulgaris	BvDODAα2	MN153195
Beta vulgaris	BvDODAβ	MN153198
Chenopodium quinoa	CqDODA1α1	XP_021769303
Mirabilis jalapa	MjDOD	AB435372.1
Mesembryanthemum crystallinum	McDODAα 1	MN136225
Mesembryanthemum crystallinum	McDODAβ	MN136227

Table S6 All primers used for sequencing in the study.

Name	Function	Sequence
SV4	*AhCYP76AD2* genotyping	gcctaaagatgcacaaatactcg
SV5	*AhCYP76AD2* genotyping	TCAATGcctaagggatggaatg

Table S7 Repetitive element content of the polished *A. hypochondriacus* genome assembly v2.2.

Classification	Genome covered [bp]	Genome covered [%]
DNA/CMC-EnSpm	4,098,198	1.01
DNA/MULE-MuDR	5,723,062	1.42
DNA/PIF-Harbinger	1,168,228	0.29
DNA/TcMar-Stowaway	1,239,206	0.31
DNA/Zisupton	800,423	0.2
DNA/hAT-Ac	2,943,534	0.73
DNA/hAT-Tag1	254,129	0.06
DNA/hAT-Tip100	1,342,775	0.33
RC/Helitron	2,071,001	0.51
LINE/CRE	1,014,983	0.25
LINE/L1	8,821,523	2.18
LINE/Penelope	21,568	0.01
LINE/R1	18,480	0
LINE/RTE	504,828	0.12
LINE/RTE-BovB	2,580,488	0.64
LTR	163,433	0.04
LTR/Copia	39,486,653	9.77
LTR/Gypsy	19,135,979	4.74
LTR/Pao	50,089	0.01
LTR/Unknown	10,497,276	2.6
SINE/5S	76,086	0.02
SINE/U	39,697	0.01
SINE/tRNA	697,466	0.17
SINE/tRNA-CR1	38,625	0.01
rRNA	21,280	0.01
tRNA	30,256	0.01
Low complexity	937,694	0.23
Simple repeat	11,495,793	2.85
Unknown	114,642,357	28.38
Total sum	229,915,110	56.91

Table S8 BUSCO score summary of different data sets against the Embryophyta reference set. Iso-Seq FLNC reads were mapped against the unpolished or polished reference genome and corrected using the unpolished or polished reference genome sequence. For Iso-Seq transcript sets, BUSCO was run in protein mode using predicted ORFs. For genome annotations, BUSCO was run in protein mode on predicted protein sequences.

Dataset	Species	Data source	Complete	Complete (single copy)	Complete (duplicated)	Fragmented	Missing
Genome annotation v2.1	*A. hypochondriacus*	Lightfoot *et al.* (2017)	80.7	78.5	2.2	8.8	10.5
Genome annotation	*A. cruentus*	Ma *et al.* (2021)	89.8	87.2	2.6	2.5	7.7
Iso-Seq transcripts (uncorrected)	*A. hypochondriacus*	this study	90.7	58.9	31.8	2.0	7.3
Iso-Seq transcripts (corrected)	*A. hypochondriacus*	this study	90.8	57.3	33.5	1.4	7.8
Iso-Seq transcripts (unpolished genome, uncorrected)	*A. hypochondriacus*	this study	90.6	58.7	31.9	2.0	7.4
Iso-Seq transcripts (unpolished genome, corrected)	*A. hypochondriacus*	this study	74.6	46.7	27.9	11.0	14.4
Braker2 annotation	*A. hypochondriacus*	this study	97.9	80.9	17.0	0.8	1.3
Genome annotation v2.2	*A. hypochondriacus*	this study	98.0	83.4	14.6	0.7	1.3

Table S9 Amaranth flavonoid pathway candidate genes identified using KIPEs (Pucker *et al.* 2020). Displayed are candidate gene names from genome annotation v2.2 alongside the percentage similarity, the percentage of conserved residues and the number of missing conserved residues as reported by KIPEs. Missing pathway genes (defined by having no unique candidate genes or three or more missing conserved residues) were marked with asterisks and defined as unidentified in *A. hypochondriacus*.

Pathway gene	Candidate gene	Percentage similarity	Percentage conserved residues	Missing conserved residues
PAL	AHp012752	79.49	100.0	0
PAL	AHp021980	79.40	100.0	0
C4H	AHp013217	81.06	100.0	0
C4H	AHp022384	80.66	100.0	0
C4H	AHp022382	72.71	100.0	0
4CL	AHp014409	64.73	100.0	0
4CL	AHp020962	65.36	100.0	0
CHS	AHp004305	79.78	94.29	2
CHI1	AHp009962	59.97	100.0	0
CHI2*	no candidate	-	-	-
F3′H	AHp017497	68.15	100.0	0
F3′H	AHp022122	36.88	100.0	0
F3′H	AHp003152	38.80	100.0	0
F3′H	AHp022120	37.24	100.0	0
F3′H	AHp022123	25.78	100.0	0
F3′5′H*	no candidate	-	-	-
FNS1*	no candidate	-	-	-
DFR	AHp009303	61.99	96.43	1
F3H	AHp006454	79.68	100.0	0
FLS	AHp008991	62.03	100.0	0
LAR	AHp017409	57.70	100.0	0
ANS*	no candidate	-	-	-
ANR*	AHp001635	37.14	86.96	3

Table S10 MYB transcription factors identified in *A. hypochondriacus*. Genes were assigned into subgroups based on their phylogenetic placement in comparison to *A. thaliana* and *B. vulgaris* MYBs. Subgroup functions were assigned according to conserved subgroup functions (see Materials and Methods).

Gene name	Gene id	Transcript id	Scaffold	Start	End	Strand	Subfamily	Protein length	Exon number	Subgroup	Functional assignment
-	AHp000437	AHp000437.1	Scaffold_1	4347523	4351157	+	R2R3	270	3	S16	Photomorphogenesis
-	AHp000789	AHp000789.1	Scaffold_1	10090637	10092364	+	R2R3	255	2	S4	Repressors of phenylpropanoids, sinapate, lignin
-	AHp001699	AHp001699.1	Scaffold_1	29115865	29116587	+	R2R3	241	1	S22	Defense, stress response
-	AHp001699	AHp001699.2	Scaffold_1	29115865	29116656	+	R2R3	239	2	S22	Defense, stress response
-	AHp002277	AHp002277.1	Scaffold_1	35637528	35639576	-	R2R3	272	2	S4	Repressors of phenylpropanoids, sinapate, lignin
-	AHp002664	AHp002664.1	Scaffold_2	1693532	1697261	+	R2R3	274	3	S2	Abiotic stress response, flower morphogenesis
-	AHp002686	AHp002686.1	Scaffold_2	1885905	1892082	-	R2R3	317	3	S7	Flavonol, flavonoid pathway
-	AHp002839	AHp002839.1	Scaffold_2	3368079	3370731	+	R2R3	322	3	S1	Defense, stress response
-	AHp003887	AHp003887.1	Scaffold_2	26854106	26855787	+	R2R3	309	3	S14	Axillary meristem, root growth
-	AHp003952	AHp003952.1	Scaffold_2	27861374	27862381	-	R2R3	336	1	S22	Defense, stress response
-	AHp004623	AHp004623.1	Scaffold_2	34693851	34698394	-	R2R3	454	12	-	-
-	AHp005058	AHp005058.1	Scaffold_3	3408139	3411828	+	R2R3	359	4	S21	Cell wall, lignin, axillary meristem
-	AHp005110	AHp005110.1	Scaffold_3	3994528	4000152	+	R2R3	454	12	-	-
-	AHp005488	AHp005488.1	Scaffold_3	8254983	8258030	+	R2R3	269	3	S20	Defense, stress response
-	AHp005948	AHp005948.1	Scaffold_3	14823654	14827853	-	R2R3	307	3	-	-
-	AHp006062	AHp006062.1	Scaffold_3	18197344	18199249	-	R2R3	171	4	S24	-
-	AHp006213	AHp006213.1	Scaffold_3	24208195	24209124	+	R2R3	310	1	S22	Defense, stress response
-	AHp006562	AHp006562.1	Scaffold_4	605153	607206	+	R2R3	316	3	S1	Defense, stress response
-	AHp006562	AHp006562.2	Scaffold_4	605153	608233	+	R2R3	178	4	S1	Defense, stress response
-	AHp007248	AHp007248.1	Scaffold_4	11044442	11046030	+	R2R3	359	3	S10	-
-	AHp007516	AHp007516.1	Scaffold_4	18919997	18922556	+	R2R3	540	3	S18	Anther devlopment, stress response
-	AHp007516	AHp007516.2	Scaffold_4	18919997	18922556	+	R2R3	535	3	S18	Anther devlopment, stress response
-	AHp007908	AHp007908.1	Scaffold_4	24322377	24324732	-	R2R3	296	3	S4	Repressors of phenylpropanoids, sinapate, lignin
-	AHp008229	AHp008229.1	Scaffold_4	27535327	27540827	-	R2R3	353	3	S18	Anther devlopment, stress response
-	AHp008229	AHp008229.2	Scaffold_4	27535327	27540827	-	R2R3	375	4	S18	Anther devlopment, stress response
-	AHp008229	AHp008229.3	Scaffold_4	27536451	27540827	-	R2R3	310	2	S18	Anther devlopment, stress response
-	AHp008362	AHp008362.1	Scaffold_5	579584	581724	+	R2R3	223	4	S14	Axillary meristem, root growth
-	AHp008362	AHp008362.2	Scaffold_5	579584	581724	+	R2R3	345	3	S14	Axillary meristem, root growth
-	AHp008410	AHp008410.1	Scaffold_5	1052626	1053435	-	R2R3	270	1	S22	Defense, stress response
-	AHp008675	AHp008675.1	Scaffold_5	4050096	4052532	+	R2R3	348	2	S21	Cell wall, lignin, axillary meristem
-	AHp009283	AHp009283.1	Scaffold_5	19207042	19209867	+	R2R3	301	3	S9	Trichome branching, petal morphogenesis
-	AHp009452	AHp009452.1	Scaffold_5	21916376	21917921	+	R2R3	315	3	S4	Repressors of phenylpropanoids, sinapate, lignin
-	AHp009484	AHp009484.1	Scaffold_5	22269940	22274870	+	R2R3	305	3	S3	Phenylpropanoid pathway, lignin
-	AHp009565	AHp009565.1	Scaffold_5	23545578	23547934	+	R2R3	348	2	AtMYB5-like	Development, proanthocyanidin biosynthesis
-	AHp009926	AHp009926.1	Scaffold_6	2944929	2948000	+	R2R3	330	3	-	-
-	AHp010652	AHp010652.1	Scaffold_6	17732890	17739448	-	R2R3	974	4	-	-
-	AHp010658	AHp010658.1	Scaffold_6	17860532	17862437	+	R2R3	420	3	S13	Mucilage, lignin, stomatal closure
-	AHp011052	AHp011052.1	Scaffold_6	22330077	22334786	-	R2R3	411	2	S16	Photomorphogenesis
-	AHp011153	AHp011153.1	Scaffold_6	23290312	23291464	-	R2R3	300	2	S18	Anther devlopment, stress response
-	AHp011578	AHp011578.1	Scaffold_7	10152532	10155066	-	R2R3	334	3	-	-
-	AHp011578	AHp011578.2	Scaffold_7	10153374	10155066	-	R2R3	334	2	-	-
-	AHp011866	AHp011866.1	Scaffold_7	16613289	16616126	+	R2R3	314	3	S20	Defense, stress response
-	AHp012381	AHp012381.1	Scaffold_7	23842541	23847434	+	R2R3	254	3	S16	Photomorphogenesis
-	AHp012382	AHp012382.1	Scaffold_7	23851791	23855945	-	R2R3	289	3	S16	Photomorphogenesis
-	AHp012383	AHp012383.1	Scaffold_7	23881163	23885856	+	R2R3	250	3	S9	Trichome branching, petal morphogenesis
-	AHp012732	AHp012732.1	Scaffold_8	3120126	3122087	+	R2R3	375	3	S14	Axillary meristem, root growth
-	AHp013131	AHp013131.1	Scaffold_8	6814523	6815510	+	R2R3	255	3	S17	-
-	AHp013326	AHp013326.1	Scaffold_8	9237650	9242324	-	R2R3	403	3	-	-
-	AHp013775	AHp013775.1	Scaffold_8	16914029	16918169	-	R2R3	411	3	S25	Embryogenesis, seed maturation
-	AHp014406	AHp014406.1	Scaffold_9	12055158	12058160	+	R2R3	350	3	S14	Axillary meristem, root growth
-	AHp014458	AHp014458.1	Scaffold_9	12904501	12907181	-	R2R3	242	3	S14	Axillary meristem, root growth
AmMYBl1	AHp014591	AHp014591.1	Scaffold_9	14711535	14714080	+	R2R3	283	4	S5	Proanthocyanidin biosynthesis
-	AHp014665	AHp014665.1	Scaffold_9	15476010	15477107	-	R2R3	366	1	-	-
-	AHp015175	AHp015175.1	Scaffold_9	20978046	20979995	-	R2R3	392	3	S11	Defense, stress response
-	AHp015176	AHp015176.1	Scaffold_9	20993200	20994999	-	R2R3	359	3	S11	Defense, stress response
-	AHp015766	AHp015766.1	Scaffold_10	12455733	12458734	+	R2R3	351	3	S1	Defense, stress response
-	AHp015766	AHp015766.2	Scaffold_10	12455733	12458734	+	R2R3	341	3	S1	Defense, stress response
-	AHp015783	AHp015783.1	Scaffold_10	12825187	12827217	-	R2R3	452	3	S9	Trichome branching, petal morphogenesis
-	AHp016111	AHp016111.1	Scaffold_10	16835722	16838292	-	R2R3	325	3	S14	Axillary meristem, root growth
-	AHp016387	AHp016387.1	Scaffold_10	20282177	20286204	+	R2R3	406	2	S23	-
-	AHp016456	AHp016456.1	Scaffold_10	21051032	21053433	+	R2R3	318	4	-	-
AhMYB3	AHp016530	AHp016530.1	Scaffold_10	21762621	21764924	-	R2R3	216	3	BvMYB1-like	Betalain biosynthesis
AhMYB4	AHp016531	AHp016531.1	Scaffold_10	21780565	21783050	-	R2R3	228	3	BvMYB1-like	Betalain biosynthesis
-	AHp016620	AHp016620.1	Scaffold_10	22596330	22599342	-	R2R3	341	2	S16	Photomorphogenesis
-	AHp017558	AHp017558.1	Scaffold_11	17970779	17972903	+	R2R3	293	3	S1	Defense, stress response
-	AHp017657	AHp017657.1	Scaffold_11	18887019	18889370	-	R2R3	289	3	S20	Defense, stress response
-	AHp017658	AHp017658.1	Scaffold_11	18908562	18910154	-	R2R3	248	3	S20	Defense, stress response
-	AHp017689	AHp017689.1	Scaffold_11	19220858	19226127	-	R2R3	341	3	S7	Flavonol, flavonoid pathway
-	AHp017915	AHp017915.1	Scaffold_11	21654178	21657013	-	R2R3	362	3	-	-
-	AHp017946	AHp017946.1	Scaffold_11	22019498	22021190	-	R2R3	294	3	S25	Embryogenesis, seed maturation

Gene name	Gene id	Transcript id	Scaffold	Start	End	Strand	Subfamily	Protein length	Exon number	Subgroup	Functional assignment
-	AHp017979	AHp017979.1	Scaffold_12	139570	144579	-	R2R3	358	3	-	-
-	AHp018270	AHp018270.1	Scaffold_12	3175247	3177848	-	R2R3	188	3	S5	Proanthocyanidin biosynthesis
-	AHp018383	AHp018383.1	Scaffold_12	4698125	4699671	-	R2R3	346	2	-	-
-	AHp018492	AHp018492.1	Scaffold_12	5906028	5907870	-	R2R3	306	2	S8	Lignin
-	AHp018627	AHp018627.1	Scaffold_12	7669283	7672257	+	3R	513	7	3R	-
-	AHp019036	AHp019036.1	Scaffold_12	15482171	15486341	-	R2R3	356	3	S21	Cell wall, lignin, axillary meristem
-	AHp019040	AHp019040.1	Scaffold_12	15614348	15629179	+	4R	900	13	4R	-
-	AHp019040	AHp019040.2	Scaffold_12	15614348	15629179	+	4R	902	13	4R	-
-	AHp019088	AHp019088.1	Scaffold_12	16532558	16544063	+	3R	541	7	3R	-
-	AHp019259	AHp019259.1	Scaffold_12	20323254	20324899	+	R2R3	313	2	-	-
-	AHp019341	AHp019341.1	Scaffold_12	21674277	21675477	-	R2R3	335	3	-	-
-	AHp019341	AHp019341.2	Scaffold_12	21674277	21675477	-	R2R3	325	3	-	-
-	AHp020198	AHp020198.1	Scaffold_13	17051563	17053916	+	R2R3	259	3	S17	-
-	AHp020524	AHp020524.1	Scaffold_13	20104457	20106612	-	R2R3	284	3	S1	Defense, stress response
-	AHp020968	AHp020968.1	Scaffold_14	12366084	12368376	-	R2R3	285	3	S14	Axillary meristem, root growth
-	AHp021016	AHp021016.1	Scaffold_14	13270874	13277249	+	R2R3	412	3	S25	Embryogenesis, seed maturation
-	AHp021152	AHp021152.1	Scaffold_14	14968799	14969911	-	R2R3	371	1	-	-
-	AHp021566	AHp021566.1	Scaffold_14	19210847	19212496	-	R2R3	381	3	S11	Defense, stress response
-	AHp021567	AHp021567.1	Scaffold_14	19222097	19225148	-	R2R3	381	3	S11	Defense, stress response
-	AHp021873	AHp021873.1	Scaffold_15	1897249	1899908	-	R2R3	278	3	S17	-
-	AHp022482	AHp022482.1	Scaffold_15	10193355	10194334	+	R2R3	260	3	S17	-
-	AHp022746	AHp022746.1	Scaffold_16	532237	533511	-	R2R3	274	3	-	-
AhMYB2	AHp022773	AHp022773.1	Scaffold_16	988734	991447	+	R2R3	236	3	BvMYB1-like	Betalain biosynthesis
AhMYB2	AHp022773	AHp022773.2	Scaffold_16	988734	991447	+	R2R3	240	3	BvMYB1-like	Betalain
-	AHp022995	AHp022995.1	Scaffold_16	3273279	3276051	-	R2R3	422	2	S23	-
-	AHp023085	AHp023085.1	Scaffold_16	4378346	4379815	+	R2R3	243	2	-	-
-	AHp023286	AHp023286.1	Scaffold_16	7455684	7465379	-	3R	1017	11	3R	-

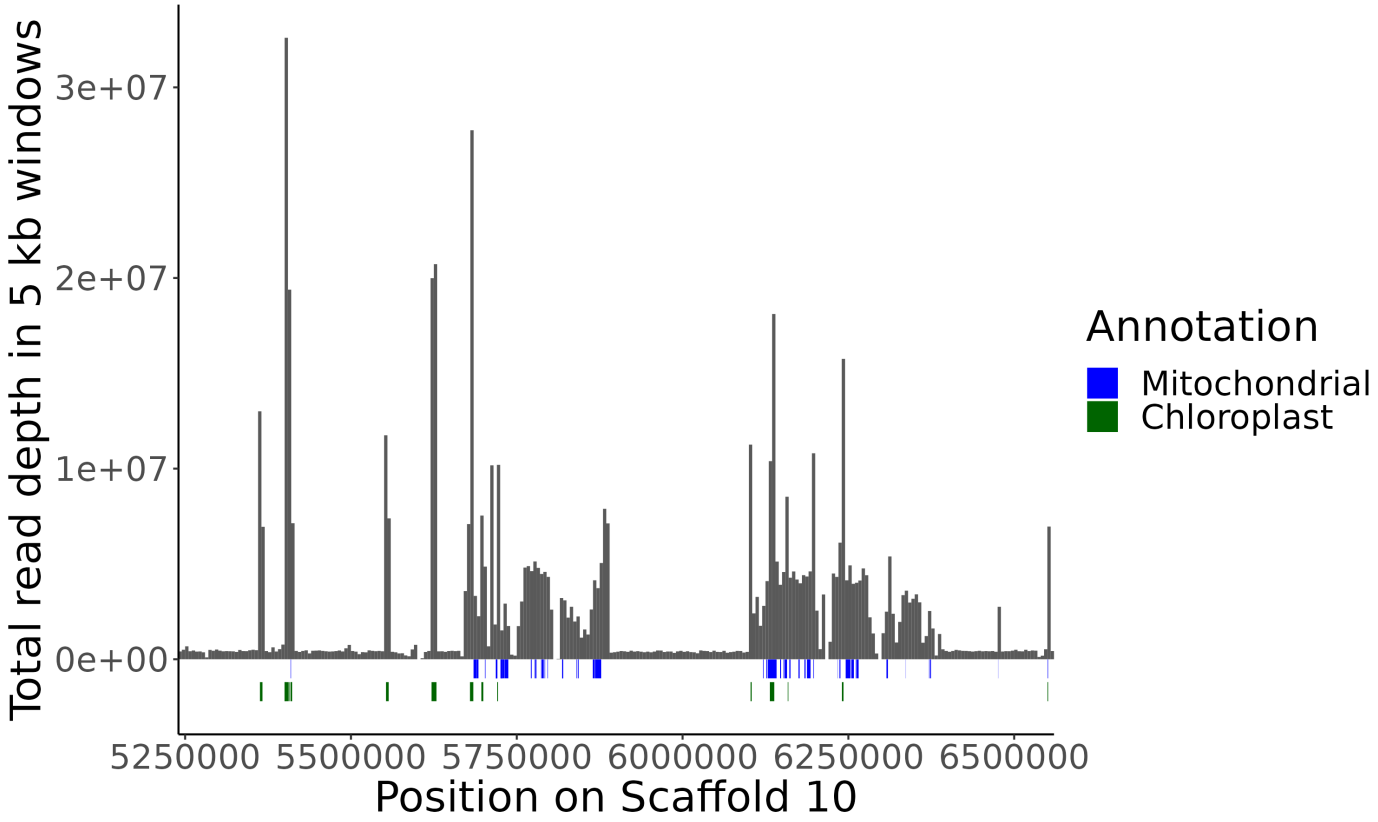

Figure S1 Total read depth of whole genome sequencing data in 5 kb windows along Scaffold 10 region with mapping bias. High read depth regions correspond partially to mapped chloroplast and mitochondrial regions in the genome.

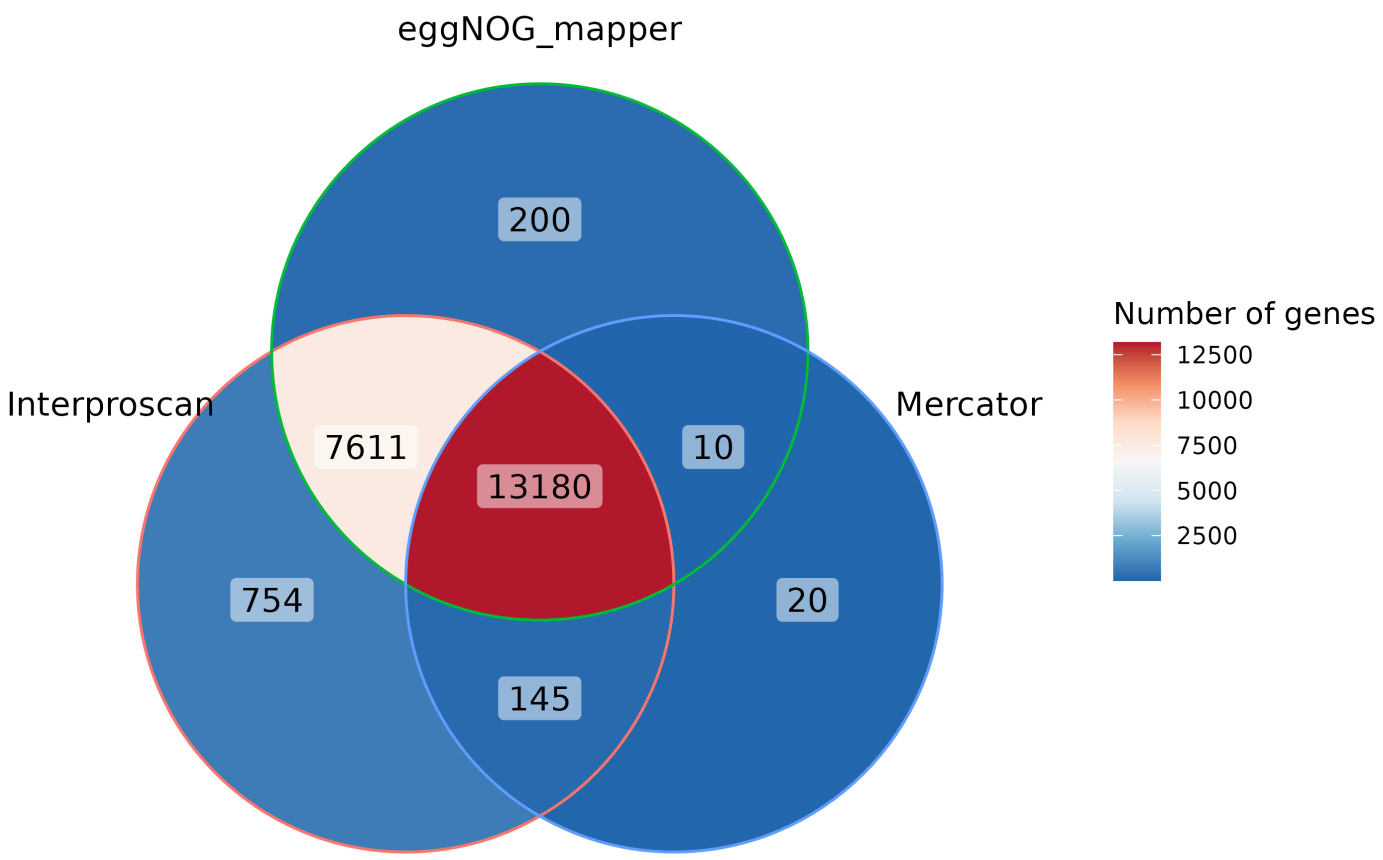

Figure S2 Overlap between functionally annotated genes using the annotation tools Interproscan (considering only genes with identified Pfam, PANTHER or CDD domains as annotated), eggNOG-mapper and Mercator.

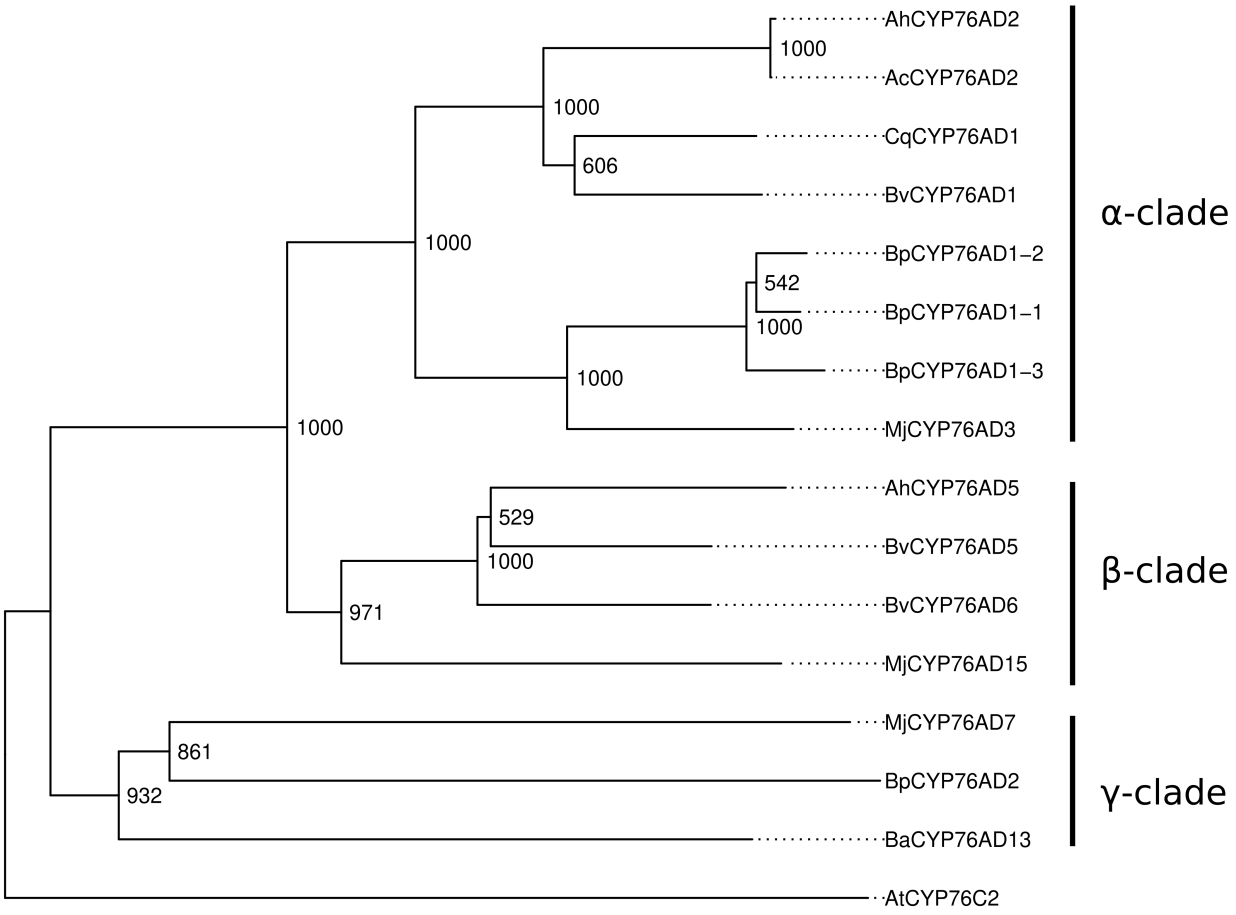

Figure S3 Phylogenetic analysis of amaranth CYP76AD proteins with CYP protein sequences from other species. CYP76AD α-, β- and γ-clades are annotated in the alignment. Bootstrap values from 1000 replicates are depicted on the tree.

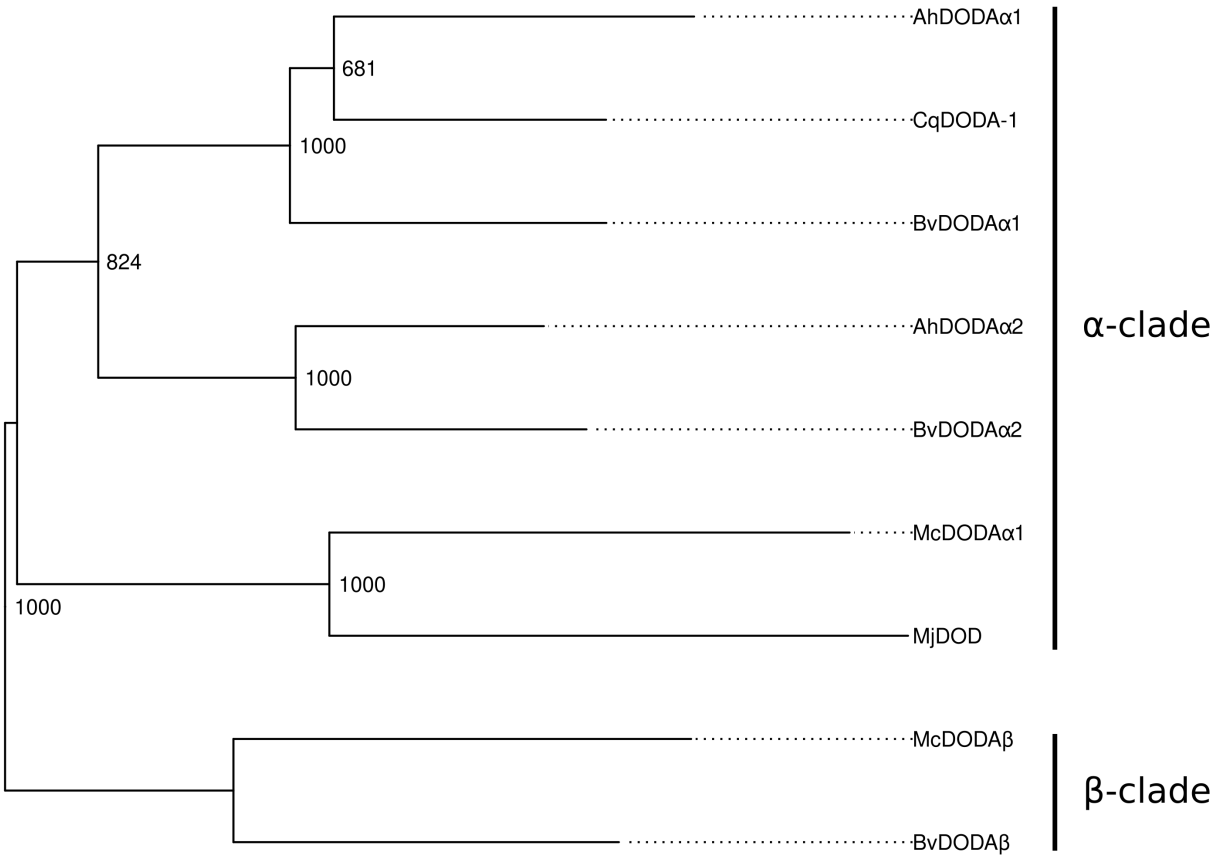

Figure S4 Phylogenetic analysis of amaranth DODA proteins with DODA protein sequences from other species. DODA α- and β-clades are annotated in the alignment. Bootstrap values from 1000 replicates are depicted on the tree.

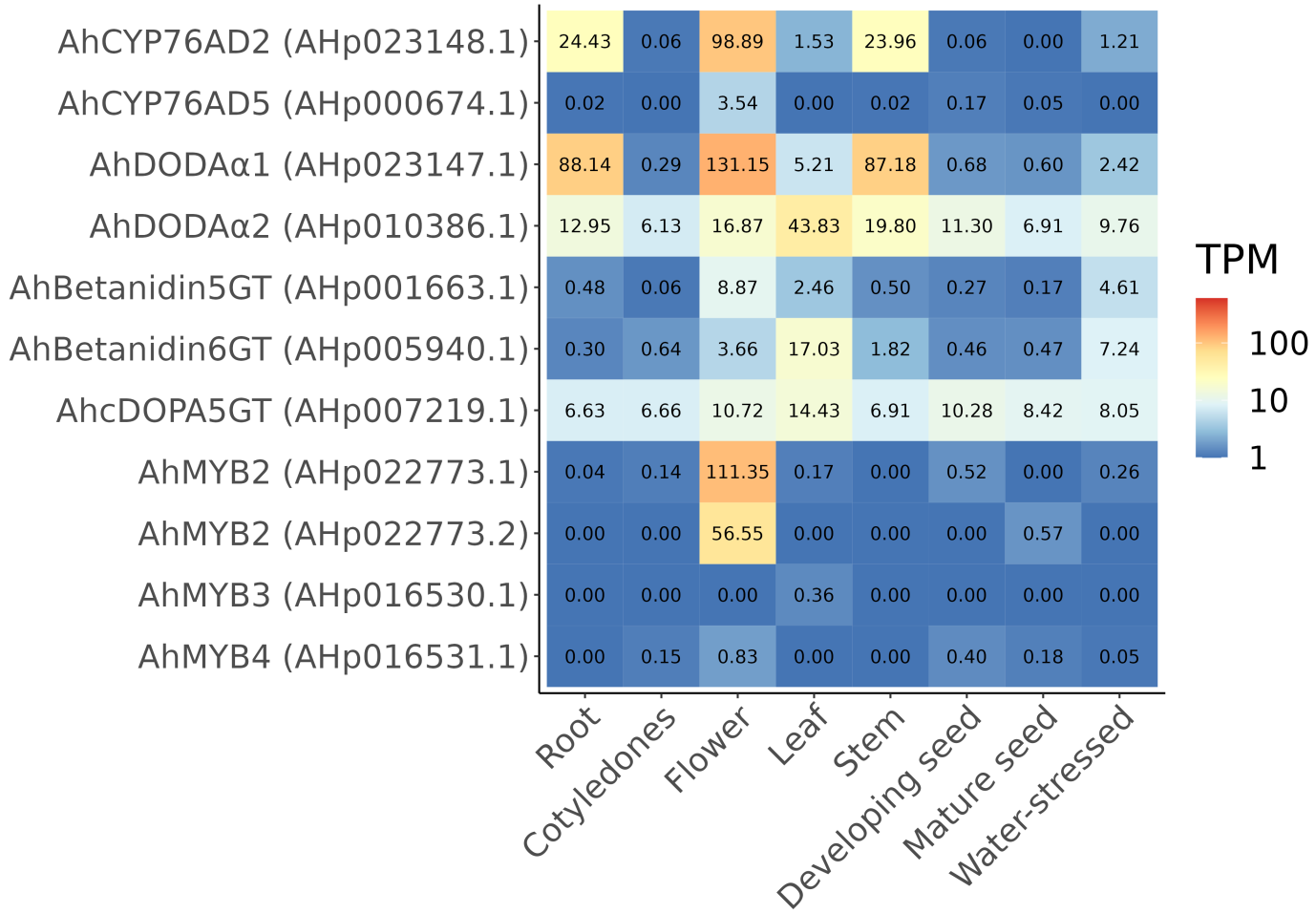

Figure S5 Relative gene expression in TPM of identified betalain pathway gene expression in different tissues. Quantification was done using kallisto from short-read RNA sequencing data.

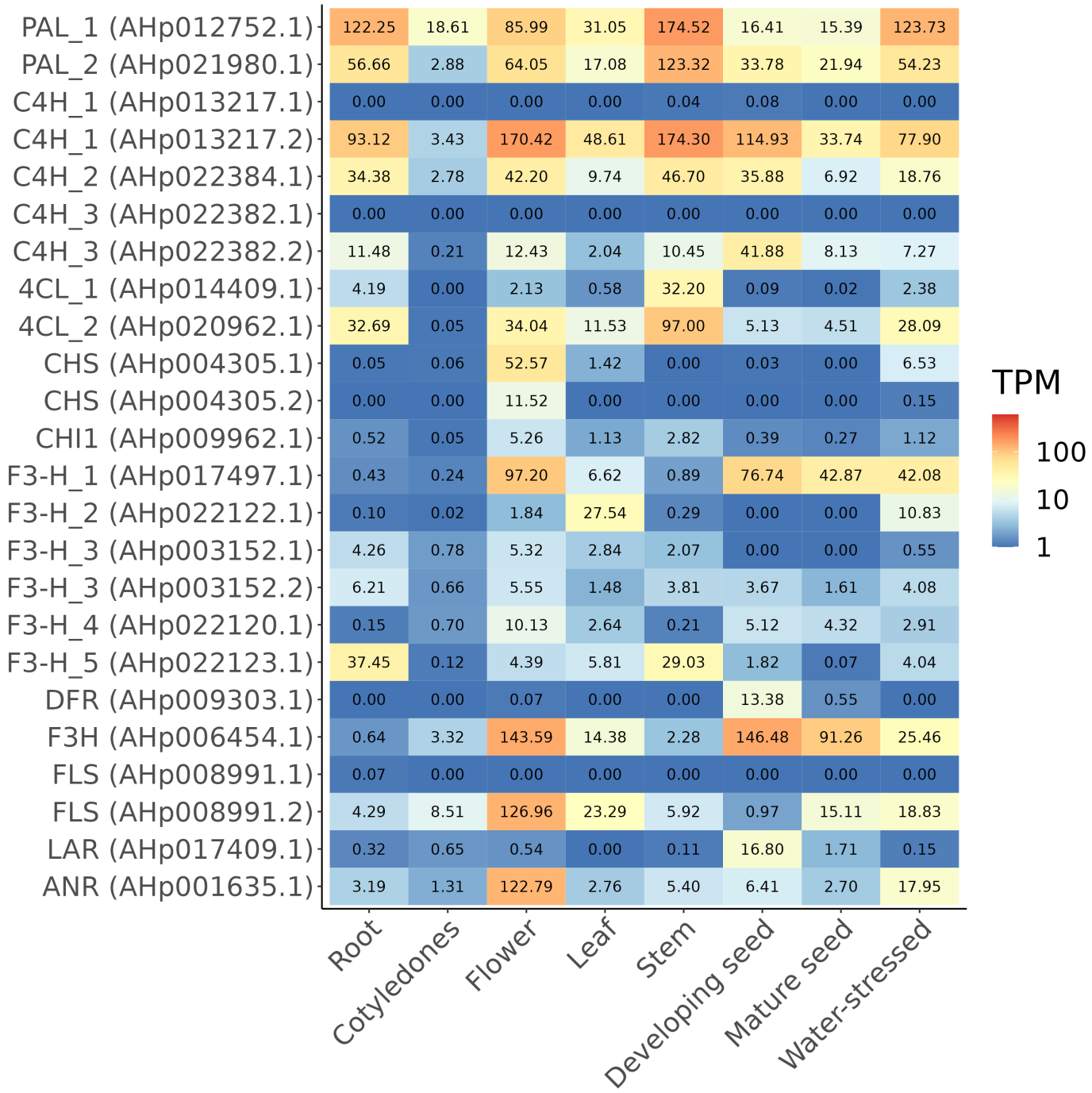

Figure S6 Relative gene expression in TPM of identified flavonoid pathway gene expression in different tissues. Quantification was done using kallisto from short-read RNA sequencing data.

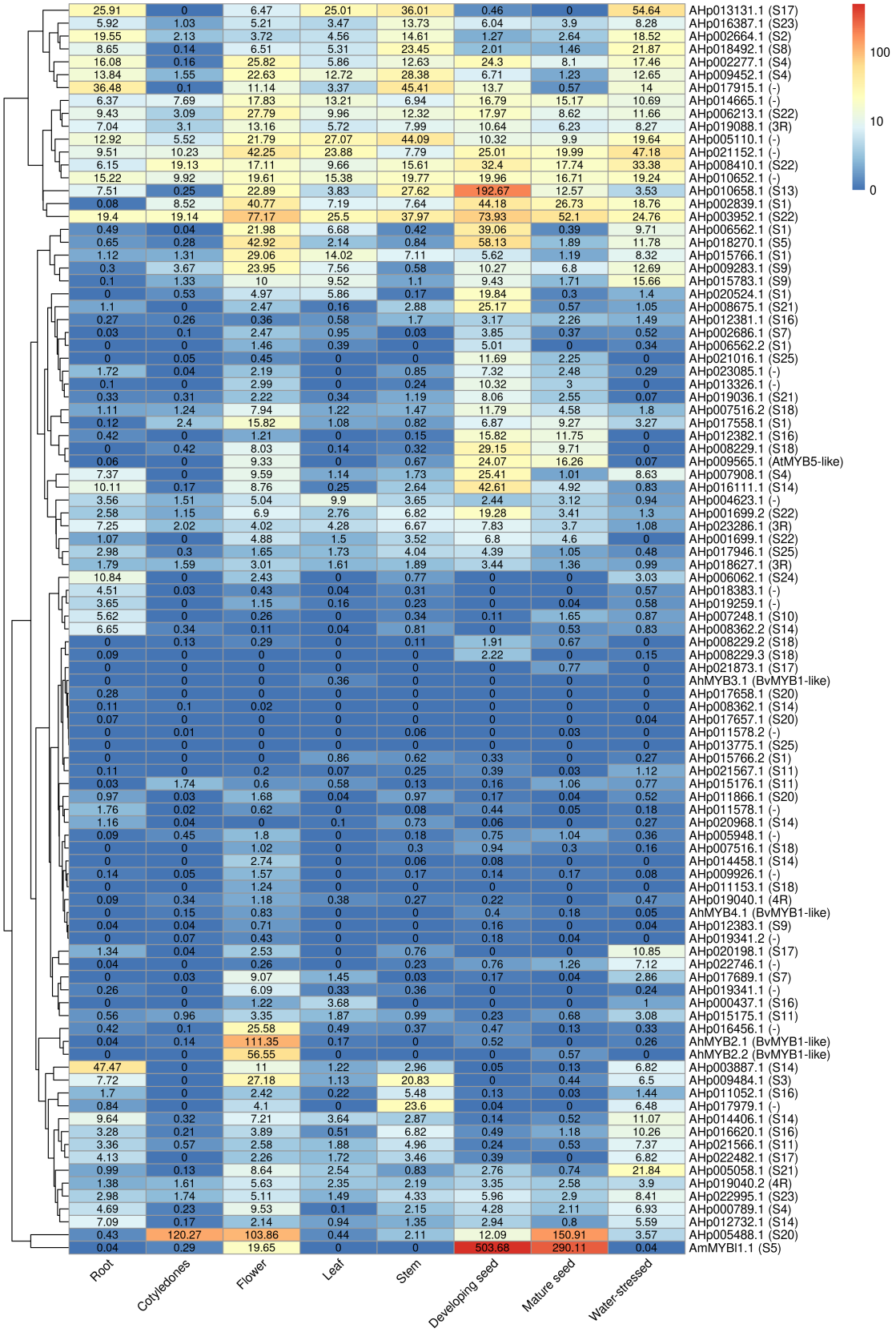

Figure S7 Relative gene expression in TPM of identified MYB transcription factor genes in different tissues. Quantification was done using kallisto from short-read RNA sequencing data.

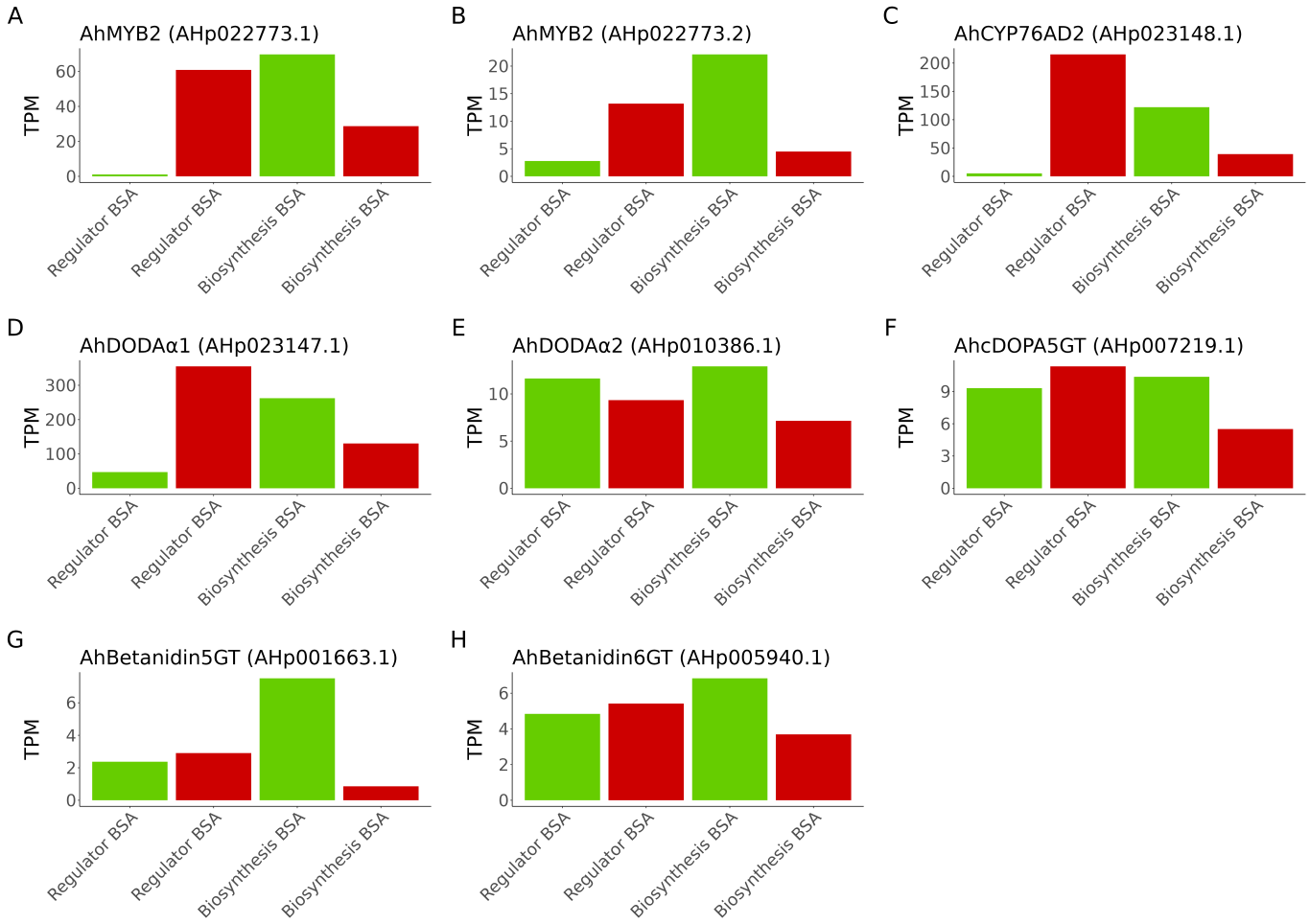

Figure S8 Gene expression of betalain pathway genes in flower tissue in green and red bulks of regulator and biosynthesis BSAs. Relative gene expression in TPM is displayed for betalain pathway genes and candidate regulators. *AhCYP76AD5*, *AhMYB3* and *AhMYB4* had negligible expression in flower tissue in all samples and were excluded.

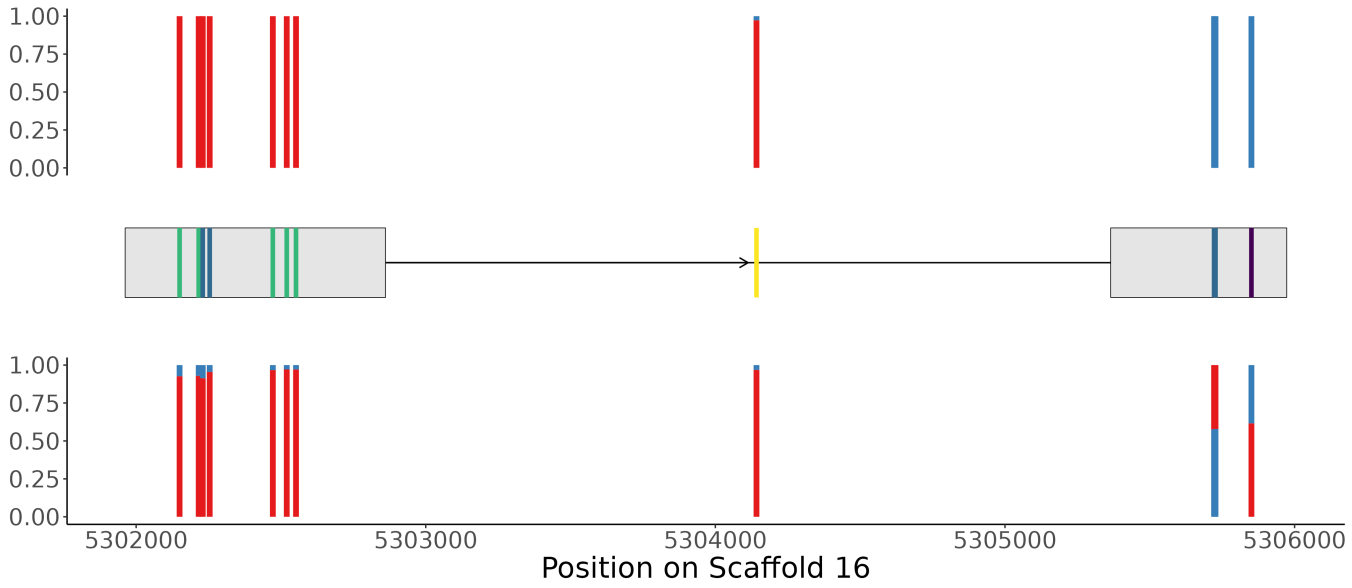

Figure S9 Segregating variants called from whole-genome sequencing data in the biosynthesis BSA and their relative allele depth for the *AhCYP76AD2* gene. Top row shows the relative depth of reference (blue) and alternative allele (red) for the red bulk, the bottom row the relative depth for the green bulk. Middle part shows the exon structure of the gene and all variants with their predicted effect compared to the reference genome (yellow: intron variant, green: synonymous variant, dark blue: missense variant, purple: stop-gain variant).

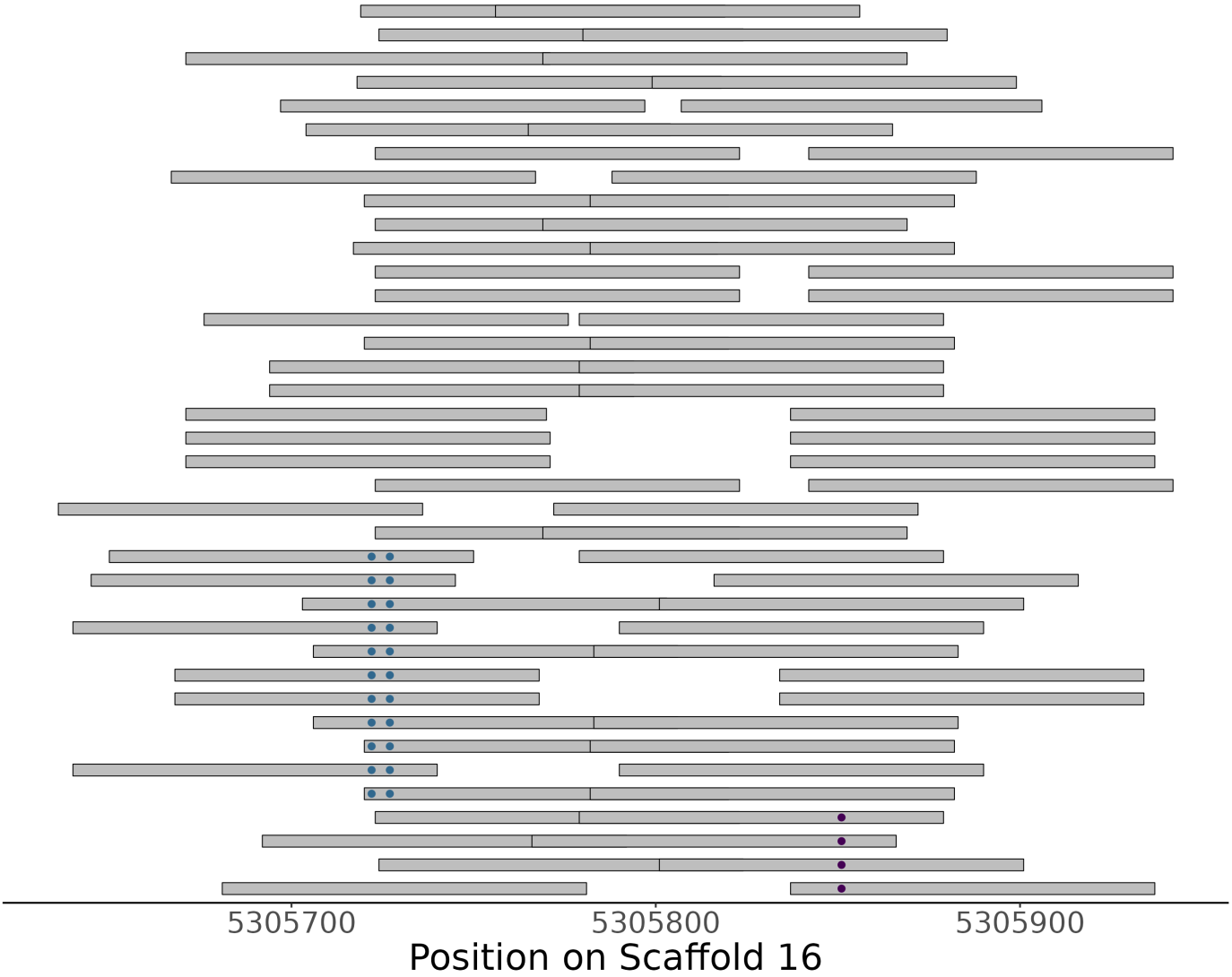

Figure S10 RNA sequencing read pairs from flower tissue of the red bulk of the biosynthesis BSA covering at least one of the non-synonymous allele T1258A and A1263T positions and the stop allele C1387T position on the *AhCYP76AD2* gene. Variant positions were colored based on the assigned impact of the allele on a read on the translated protein.

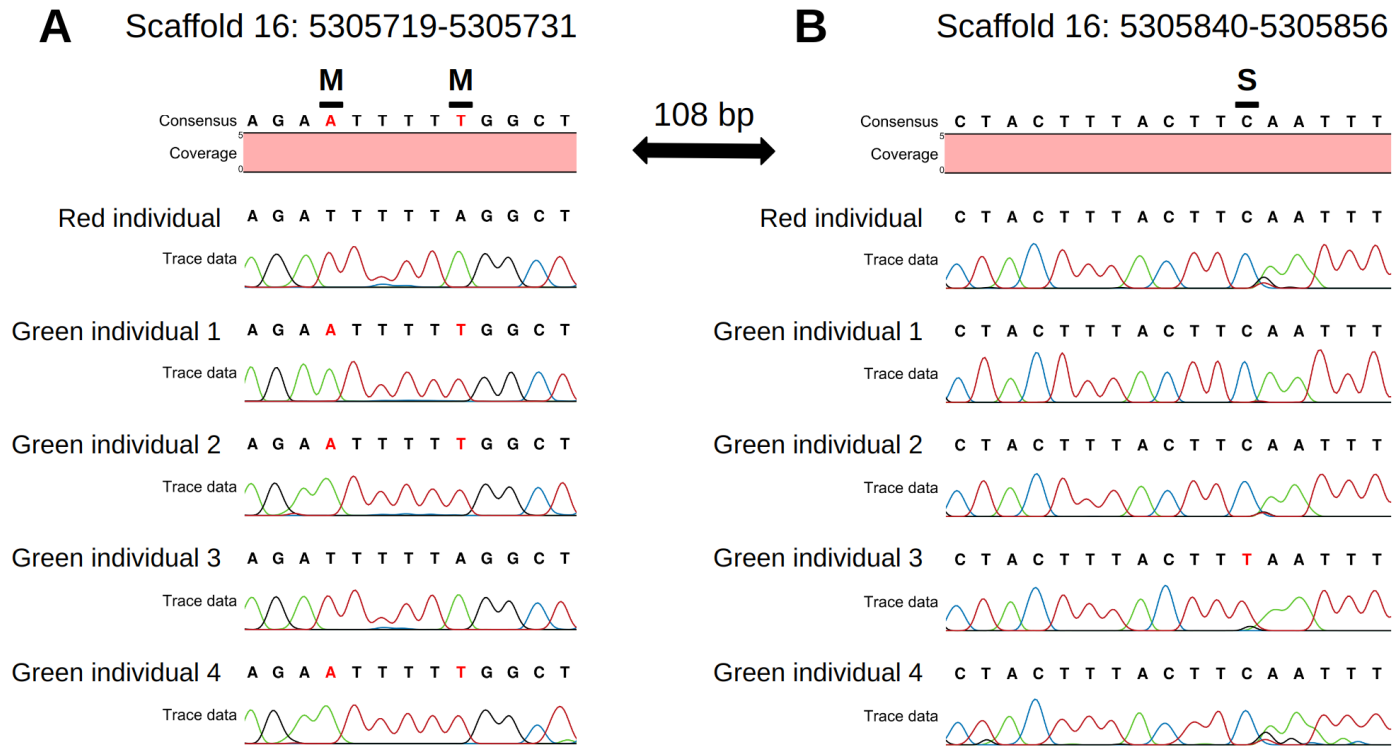

Figure S11 Sanger sequencing of *AhCYP76AD2* (primer SV4) from individuals from biosynthesis BSA were mapped to the *AhCYP76AD2* genomic sequence. One individual was chosen from the red bulk (top), the other four individuals were from the green bulk. Positions differing from the reference genome were marked in red. The positions of the two missense alleles T1258A and A1263T were marked with M (A), the stop allele C1387T was marked with S (B).