Large-scale Phylogenomics Reveals Systematic Loss of Anthocyanin Biosynthesis Genes at the Family Level in Cucurbitaceae

Nancy Choudhary[1,*], Marie Hagedorn[1], Boas Pucker[1,*]

[1] Plant Biotechnology and Bioinformatics, Institute for Cellular and Molecular Botany (IZMB), University of Bonn, Kirschallee 1, 53115 Bonn, Germany

* corresponding author: pucker@uni-bonn.de

Summary

- Plant pigmentation secrets are among the oldest interests of plant scientists, with pigments such as chlorophyll, carotenoids, flavonoids, and betalains contributing to the diversity of hues in higher plants. Anthocyanins, a class of flavonoids responsible for vibrant shades of pink, red, and blue pigmentation, are almost ubiquitous in angiosperms but are replaced by betalains in some families in the order Caryophyllales.
- We investigated anthocyanin pigmentation in *Cucurbitaceae*, the second-largest fruit and vegetable family, characterised by white and yellow flowers and red, orange, and green fruits predominantly pigmented by carotenoids. Using a comprehensive collection of 258 datasets representing species across all 15 tribes of *Cucurbitaceae* and a phylogenomics approach, we observed a systematic absence of genes involved in anthocyanin and proanthocyanidin biosynthesis. Absence of the structural genes *DFR*, *ANS*, *arGST*, *LAR*, and *ANR*, along with the anthocyanin-related regulatory MYB genes, was consistently confirmed by synteny and phylogenetic analysis.
- These results suggest that anthocyanin loss in angiosperms is more common than previously assumed. In light of this new discovery, we propose a stepwise loss of anthocyanin pigmentation in *Cucurbitaceae*, likely accompanied by a partial functional replacement with carotenoid pigmentation.

Introduction

The colours of plants are one of the primary ways they interact with the environment through attraction of pollinators and seed dispersers (Chittka & Raine, 2006), defence against herbivores (Lev-Yadun *et al.*, 2004; Wolff & Pucker, 2025), and communication through visual signalling (Lev-Yadun, 2016; Garcia *et al.*, 2019). Dawned with complex modifications, anthocyanins, derived from the flavonoid pathway, are one of the key contributors to this chromatic diversity in flowering plants. Widely distributed in seed plants, they are primarily responsible for the diversity of hues seen in most plant tissues, specifically the vibrant red, pink, and blue hues of flowers and fruits (Winkel-Shirley, 2001). Beyond aesthetics, they serve multiple ecological and physiological functions (Grünig *et al.*, 2024), including photoprotection against ultraviolet (UV) radiation (Gould, 2004), attraction of pollinators and seed dispersers (Thompson *et al.*, 1972), and defence responses against herbivory and pathogen attack (Steyn *et al.*, 2002). Moreover, anthocyanins have been implicated in plant adaptation to environmental stress, including drought and cold tolerance (Chalker-Scott, 1999; Pietrini *et al.*, 2002).

The core flavonoid biosynthesis pathway is well understood and begins with the synthesis of naringenin chalcone from one molecule of 4-coumaroyl-CoA and three molecules of malonyl-CoA, catalysed by chalcone synthase (CHS) (**Figure 1a**). Naringenin chalcone is rapidly and stereospecifically isomerised to the colourless naringenin by chalcone isomerase (CHI). From this central intermediate, multiple enzymatic steps branch into distinct flavonoid classes. Hydroxylation reactions mediated by flavanone 3-hydroxylase (F3H), flavonoid 3′-hydroxylase (F3′H), and flavonoid 3′,5′-hydroxylase (F3′5′H) produce diverse dihydroflavonols, which are then converted by flavonol synthase (FLS) into flavonols, converted by the flavone synthase (FNS) into flavones, or channelled towards anthocyanin production via dihydroflavonol reductase (DFR), anthocyanidin synthase (ANS), and anthocyanin-related glutathione S-transferase (arGST). Unstable anthocyanidins are converted to stable anthocyanins by UDP-glucose flavonoid/anthocyanidin 3-O-glycosyltransferase (UGT). Glycosylation, acylation, and methylation further modify the properties of the anthocyanidin-3-glucoside derivatives and generate a variety of anthocyanins. Synthesised on the cytosolic face of the endoplasmic reticulum, anthocyanins must be sequestered into the vacuole for stabilisation and accumulation. Two major mechanisms have been proposed: (i) GST-mediated ligand transport, in which glutathione S-transferases act as carrier proteins to facilitate anthocyanin transport through ABCC or MATE-type transporters (Mueller *et al.*, 2000;

Sun *et al.*, 2012), and (ii) vesicle-mediated trafficking, where anthocyanins are packaged into vesicles and delivered to the vacuole (Poustka *et al.*, 2007) (**Figure 1b**).

In addition to anthocyanins, the flavonoid pathway produces proanthocyanidins (condensed tannins), which play crucial roles in plant defence, particularly through anti-herbivory and anti-pathogen activities, and contribute to seed coat pigmentation and dormancy (Xie *et al.*, 2003; Dixon *et al.*, 2005). The enzymatic steps diverging from the general flavonoid pathway, mediated by leucoanthocyanidin reductase (LAR) and anthocyanidin reductase (ANR), direct metabolic flux towards proanthocyanidin biosynthesis (**Figure 1a**) (Dixon *et al.*, 2005; Lepiniec *et al.*, 2006).

The regulation of anthocyanidin and proanthocyanidin biosynthesis is tightly controlled by the MBW transcriptional complex, which is composed of MYB, bHLH, and WD40 proteins. Well-characterised examples include the TT2-TT8-TTG1 complex in Arabidopsis, which regulates proanthocyanidin biosynthesis (Gonzalez *et al.*, 2009). and the AN2-AN1-AN11 complex in *Petunia* regulating anthocyanin production (Spelt *et al.*, 2000). Certain R2-R3 MYB proteins generally act as the primary activators of key branch-specific genes (DFR, ANS, ANR) and determine specificity, while bHLH and WD40 proteins provide stabilisation and combinatorial control (Gonzalez *et al.*, 2008; Hichri *et al.*, 2011). The WD40 and bHLH proteins are pleiotropic and involved in multiple processes beyond anthocyanin synthesis (Walker *et al.*, 1999; Spelt *et al.*, 2002; Bernhardt *et al.*, 2003).

Pigmentation in plants is not limited to flavonoids. Chlorophylls, the most abundant pigments, confer green colouration and are essential for photosynthesis. Carotenoids, typically yellow, orange, or red, are plastid-derived isoprenoids produced via the methylerythritol phosphate (MEP) pathway (**Figure 1c**). Phytoene synthase (PSY) catalyses the first committed step, condensing geranylgeranyl diphosphate (GGPP) molecules into phytoene. Sequential desaturation and isomerisation steps mediated by phytoene desaturase (PDS), ζ-carotene desaturase (ZDS), and carotenoid isomerases yield lycopene, the central branching point. Cyclisation reactions then give rise to the α- and β-carotene branches, which are further modified by hydroxylases and epoxidases to synthesise lutein, zeaxanthin, violaxanthin, and neoxanthin (Cazzonelli & Pogson, 2010). Unlike anthocyanins, carotenoids are primary metabolites and are essential for plant life. They are involved in photoprotection and light harvesting (Triantaphylidès & Havaux, 2009), act as precursors of signalling molecules like abscisic acid and strigolactones, and accumulate in chromoplasts (**Figure 1d**) to attract pollinators and seed dispersers (Sun *et al.*, 2022).

In plants, another class of pigments, betalains, occur exclusively in the order Caryophyllales (Mabry & Mabry, 1964). Derived from tyrosine, they produce red, yellow, and violet colours and completely replaced anthocyanins in several lineages (Stafford, 1994; Brockington *et al.*, 2011; Sheehan *et al.*, 2020). This mutual exclusivity, potentially resulting from the loss of arGST and MBW complex members, demonstrates dynamic shifts in pigmentation pathways during evolution (Timoneda *et al.*, 2019; Pucker *et al.*, 2024). By contrast, anthocyanins and carotenoids often coexist within the same species, organs, or even the same tissues, resulting in an increase in colour variety (Van Eijk *et al.*, 1987; Ooe *et al.*, 2016; Clemente *et al.*, 2023).

While loss of anthocyanin pigmentation has been reported, such cases are typically restricted to individual species or even cultivar-level variants, often caused by single-gene mutations or regulatory changes (Smith & Goldberg, 2015; Marin-Recinos & Pucker, 2024). Large-scale, lineage-wide loss is exceedingly rare. One notable exception is the genus *Brachypodium*, where all examined species lack key anthocyanin biosynthetic genes, including MYB regulators, *ANS*, and *arGST* (Khatun *et al.*, 2025). Similarly, in *Poaceae*, a recent study revealed a striking rewiring of the anthocyanin pathway, with canonical MYBs and arGSTs (TT19-type) lost and functionally replaced by non-canonical MYBs and arGST (BZ2-type) (Khatun *et al.*, 2025).

The *Cucurbitaceae* family, which consists mainly of tropical and subtropical distribution, herbaceous annual or perennial climbers and woody lianas, were among the first cultivated plants by humans (Bisognin, 2002; Schaefer *et al.*, 2008). Harboring economically important fruit and vegetable crops like pumpkins, gourds, melons, cucumbers, and squashes, the family contains ~115 genera and ~1000 species. Their fruit and flower colour generally ranges from green, yellow, and orange to various shades of red (Schaefer & Renner, 2011a). Historically, carotenoids have been associated with yellow and orange pigmentation in cucurbits (Azevedo-Meleiro & Rodriguez-Amaya, 2007), while chlorophylls explain green tissues. To the best of our knowledge, no reports of blue or purple pigmentation in *Cucurbitaceae* exist, and comprehensive characterisation of the pigment repertoire within this family remains lacking.

In this study, we performed a comprehensive phylogenomic analysis to elucidate the genetic basis of pigmentation in the *Cucurbitaceae* family. Leveraging publicly available genomic and transcriptomic resources, we systematically surveyed over 200 species, representing all 15 tribes of *Cucurbitaceae*, for the presence of key structural and regulatory genes involved in flavonoid biosynthesis. Our results reveal a consistent absence of essential anthocyanin biosynthetic genes, including the structural and regulatory genes. Additionally, we identified parallel losses in the proanthocyanidin branch genes. This widespread loss of core flavonoid biosynthesis components strongly suggests an evolutionary abandonment of anthocyanin and proanthocyanidin pigmentation pathways within *Cucurbitaceae* and highlights carotenoids as the primary alternative. Our findings provide key insights into the evolutionary remodelling of pigmentation systems and lay a foundation for understanding the molecular determinants of colour diversity in cucurbits.

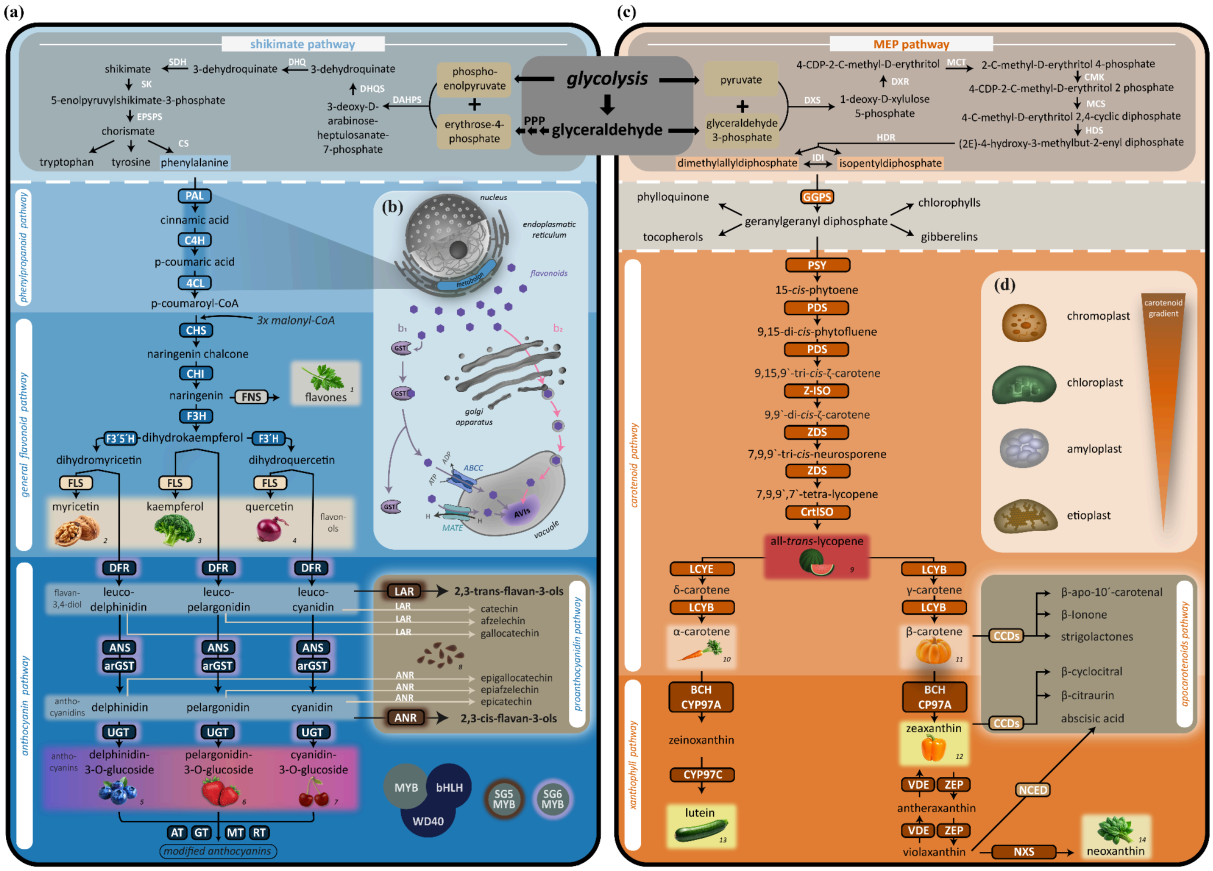

(a)
shikimate pathway
shikimate
SDH
3-dehydroquinate
DHQ
3-dehydroquinate
SK
5-enolpyruvylshikimate-3-phosphate
EPSPS
chorismate
CS
tryptophan
tyrosine
phenylalanine
DHQS
DAHPS
3-deoxy-D-arabinose-heptulosanate-7-phosphate
phospho-enolpyruvate
+
erythrose-4-phosphate
glycolysis
PPP
glyceraldehyde
phenylpropanoid pathway
PAL
cinnamic acid
C4H
p-coumaric acid
4CL
p-coumaroyl-CoA
general flavonoid pathway
3x malonyl-CoA
CHS
naringenin chalcone
CHI
naringenin
FNS
flavones
1
F3H
F3′5′H
dihydrokaempferol
F3′H
dihydromyricetin
dihydroquercetin
FLS
FLS
FLS
myricetin
2
kaempferol
3
quercetin
4
flavon-ols
(b)
nucleus
endoplasmatic reticulum
metabolon
flavonoids
b₁
b₂
GST
GST
golgi apparatus
ADP
ATP
ABCC
GST
H
H
MATE
AVIs
vacuole
anthocyanin pathway
DFR
DFR
DFR
flavan-3,4-diol
leuco-delphinidin
leuco-pelargonidin
leuco-cyanidin
LAR
2,3-trans-flavan-3-ols
LAR
catechin
LAR
afzelechin
LAR
gallocatechin
8
ANS
arGST
ANS
arGST
ANS
arGST
ANR
epigallocatechin
ANR
epiafzelechin
ANR
epicatechin
antho-cyanidins
delphinidin
pelargonidin
cyanidin
ANR
2,3-cis-flavan-3-ols
proanthocyanidin pathway
UGT
UGT
UGT
antho-cyanins
delphinidin-3-O-glucoside
5
pelargonidin-3-O-glucoside
6
cyanidin-3-O-glucoside
7
MYB
bHLH
WD40
SG5 MYB
SG6 MYB
AT
GT
MT
RT
modified anthocyanins
(c)
MEP pathway
pyruvate
+
glyceraldehyde 3-phosphate
DXS
1-deoxy-D-xylulose 5-phosphate
DXR
4-CDP-2-C-methyl-D-erythritol
MCT
2-C-methyl-D-erythritol 4-phosphate
CMK
4-CDP-2-C-methyl-D-erythritol 2 phosphate
MCS
4-C-methyl-D-erythritol 2,4-cyclic diphosphate
HDS
(2E)-4-hydroxy-3-methylbut-2-enyl diphosphate
HDR
dimethylallyldiphosphate
IDI
isopentyldiphosphate
phylloquinone
GGPS
geranylgeranyl diphosphate
chlorophylls
tocopherols
gibberelins
carotenoid pathway
PSY
15-cis-phytoene
PDS
9,15-di-cis-phytofluene
PDS
9,15,9`-tri-cis-ζ-carotene
Z-ISO
9,9`-di-cis-ζ-carotene
ZDS
7,9,9`-tri-cis-neurosporene
ZDS
7,9,9`,7`-tetra-lycopene
CrtISO
all-trans-lycopene
9
LCYE
δ-carotene
LCYB
α-carotene
10
LCYB
γ-carotene
LCYB
β-carotene
11
(d)
chromoplast
chloroplast
amyloplast
etioplast
carotenoid gradient
CCDs
β-apo-10′-carotenal
β-Ionone
strigolactones
CCDs
β-cyclocitral
β-citraurin
abscisic acid
apocarotenoids pathway
xanthophyll pathway
BCH
CYP97A
zeinoxanthin
CYP97C
lutein
13
BCH
CP97A
zeaxanthin
12
VDE
ZEP
antheraxanthin
VDE
ZEP
violaxanthin
NCED
NXS
neoxanthin
14

Figure 1: Key enzymes of the two major pigmentation pathways in plants: flavonoid biosynthesis pathway (a) and carotenoid biosynthesis pathway (c).
Flavonoid biosynthesis begins with phenylalanine, which is synthesised via the shikimate pathway (a). Branches of this pathway, including the anthocyanin and proanthocyanidin pathways, contribute strongly to visible pigmentation in higher plants. DAHPS, 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase; DHQS, 3-dehydroquinate synthase; DHQ, 3-dehydroquinate dehydratase; SDH, shikimate 5-dehydrogenase; SK, shikimate kinase; EPSPS, 5-enolpyruvylshikimate 3-phosphate synthase; CS, chorismate synthase; PAL, phenylalanine ammonia-lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumarate-CoA ligase;

CHS, chalcone synthase; CHI, chalcone isomerase; FNS, flavone synthase; F3H, flavanone-3-hydroxylase; F3′H, flavonoid 3′-hydroxylase; F3′5′H, flavonoid 3′,5′-hydroxylase; FLS, flavonol synthase; DFR, dihydroflavonol 4-reductase; ANS, anthocyanidin synthase; arGST, anthocyanin-related glutathione S-transferase; UGT, UDP-glycosyltransferase; AT, acyltransferase; GT, glycosyltransferase; MT, methyltransferase; RT, rhamnosyltransferase; LAR, leucoanthocyanidin reductase; ANR, anthocyanidin reductase. Flavonoids are synthesised at the endoplasmic reticulum and subsequently transported into vacuoles for further modifications and storage (b). Current evidence supports two major transport mechanisms: ligand-based transport (b1) and vesicle-mediated transport (b2). GST, glutathione S-transferase; ABCC, ATP-binding cassette subfamily C; MATE, multidrug and toxic compound extrusion; AVI, anthocyanin vacuolar inclusions.

Carotenoid biosynthesis is initiated by the condensation of isopentenyl diphosphate and dimethylallyl diphosphate, derived from the MEP pathway (c). The first committed step is catalysed by PSY, which converts geranylgeranyl diphosphate to 15-cis-phytoene. Following sequential desaturation and isomerisation steps, lycopene is formed, which serves as the branching point into the α-carotene and β-carotene lineages. Carotenoids are further modified to generate diverse xanthophylls, and cleavage by CCDs or NCEDs produces apocarotenoids. MEP, 2-C-methyl-D-erythritol 4-phosphate; DXS, 1-deoxy-D-xylulose 5-phosphate synthase; DXR, DXP reductoisomerase; MCT, MEP cytidyltransferase; CMK, 4-(cytidine 5′-diphospho)-2-C-methyl-D-erythritol kinase; MCS, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; HDS, 1-hydroxy-2-methyl-2-butenyl 4-diphosphate synthase; HDR, HMBPP reductase; IDI, isopentenyl diphosphate isomerase; GGPS, geranylgeranyl diphosphate synthase; PSY, phytoene synthase; PDS, phytoene desaturase; Z-ISO, ζ-carotene isomerase; ZDS, ζ-carotene desaturase; CrtISO, carotenoid isomerase; LCYE, lycopene ε-cyclase; LCYB, lycopene β-cyclase; CYP97A/C, cytochrome P450-type monooxygenases 97A and 97C; BCH, β-carotene hydroxylase; VDE, violaxanthin de-epoxidase; ZEP, zeaxanthin epoxidase; NXS, neoxanthin synthase; CCDs, carotenoid cleavage dioxygenases; NCED, 9-cis-epoxycarotenoid dioxygenase. Carotenoid biosynthesis and storage occur in plastids. The accumulation of carotenoids in different plastid types is shown (d).

The fruit and vegetable icons illustrate examples of known plants enriched in specific pigments (1 parsley, 2 walnut, 3 broccoli, 4 red onion, 5 blueberry, 6 strawberry, 7 cherry, 8 apple seeds, 9 watermelon, 10 carrot, 11 pumpkin, 12 orange bell pepper, 13 zucchini, 14 spinach).

Material and Methods

Data collection, annotation, and preparation of final datasets

The species analysed here included members of Cucurbitales, primarily from *Cucurbitaceae*, with outgroups from Datiscaceae, Begoniaceae, and Coriariaceae, as well as representatives from the orders Rosales and Fagales. Sequence datasets were retrieved from the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/, last accessed July 2025), Cucurbit Genomics Database (CucGenDBv2; http://cucurbitgenomics.org/v2/), Watermelon Genomics and Mutation Database (WaGMDB; http://www.watermelondb.cn), Melonomics v4 (http://melonomics.net), GigaDB, Figshare, Genome Warehouse (GWH; https://ngdc.cncb.ac.cn/gwh/), and the China National GeneBank Sequence Archive (CNSA; https://db.cngb.org/cnsa). In total, 282 genomic and transcriptomic datasets were obtained, comprising 258 *Cucurbitaceae* datasets and additional outgroups (**Additional File 1**). Eighty-two genome assemblies from NCBI lacked corresponding annotations and were annotated *de novo* using GeMoMa v1.9 (Keilwagen *et al.*, 2016). This was performed with reference to high-quality (96.7-99.8 % BUSCO conservation) annotations of ten species: *Alnus glutinosa* (Christenhusz *et al.*, 2024), *Juglans regia* (Marrano *et al.*, 2020), *Quercus robur* (GCA_932294415.1), *Sicyos edulis* (Fu *et al.*, 2021a), *Benincasa hispida* (Xie *et al.*, 2019), *Cucumis sativus*, *Cucumis melo*, *Cucurbita maxima*, *Cucurbita pepo*, and *Momordica charantia*. For each annotated species, the ten GeMoMa outputs were merged and filtered using the parameters f="start=='M' and stop=='*' and (isNaN(score) or score/aa>='0.75')" and atf="sumWeight>5". To assess the completeness of available annotations, an additional seven species were re-annotated using the same parameters described above using GeMoMa v1.9. All annotation details can be found in **Additional File 2**. For all genomic datasets, genome sequence (.fasta) and annotation (.gff3) files were retrieved, and coding sequences (CDS) were extracted using agat_sp_extract_sequences.pl from the AGAT toolkit (Dainat *et al.*, 2025). All CDS were retained for flavonoid and carotenoid biosynthesis gene identification and expression analysis. To streamline downstream analyses and reduce redundancy, only the longest transcript per gene was selected for use in gene trees, species phylogeny, synteny analysis, and whole-genome duplication (WGD) detection. For this, agat_sp_keep_longest_isoform.pl was used. CDS were translated to protein sequences using a custom Python script 'transeq.py' v0.3 available at https://github.com/bpucker/PBBtools.

The completeness of all protein datasets was assessed using BUSCO v5.8.1 (Tegenfeldt *et al.*, 2025) in protein mode with the eudicotyledons_odb12 reference dataset. Since some *Cucurbitaceae* datasets were later found to be contaminated, to systematically address such cases and ensure dataset homogeneity, all *Cucurbitaceae* CDS were taxonomically classified using Kraken v2.1.3 (Wood *et al.*, 2019) with the 'core_nt' database (December 2024 release, containing GenBank, RefSeq, TPA and PDB entries). Bracken v2.9 (Lu *et al.*, 2017) was applied to the Kraken outputs to estimate species-level abundance. Any dataset with >10% of classified sequences assigned to a non-*Cucurbitaceae* taxon was flagged as "contaminated" (**Additional File 3**). These contaminated datasets were excluded from phylogenetic tree construction and presence-absence analyses.

Species phylogeny and flower and fruit colour mapping

All non-contaminated datasets were used to construct the species phylogeny. Single-copy orthologs were identified using BUSCO_phylogenomics.py (https://github.com/jamiemcg/BUSCO_phylogenomics/) across 2805 unique BUSCO-defined single-copy marker genes (*eudicotyledons_odb12*). Peptide sequences of each ortholog group were aligned with MUSCLE v5.3 (Edgar, 2022), and poorly aligned regions were trimmed with trimAl v1.5.rev0 (Capella-Gutiérrez *et al.*, 2009) using the -automated1 option. For the first tree, a multi-species coalescent approach was used, where individual gene trees were inferred for all 2805 alignments using FastTree v2.1.11 (Price *et al.*, 2010), followed by species tree estimation using wASTRAL v1.23.3.7 (-R -u 2) (Zhang & Mirarab, 2022). Based on this initial tree, five datasets were removed due to their suspicious phylogenetic placement (see results section for details). For the final species phylogeny, both multi-species coalescent and concatenation (supermatrix) approaches were used to infer the *Cucurbitaceae* phylogeny. For the coalescent-based phylogeny, 20 of the 2805 alignments were randomly selected and individually tested with ModelFinder (Kalyaanamoorthy *et al.*, 2017), which most frequently identified the Q.plant+I+G4 model. This model was therefore applied to all gene trees for consistency. Gene trees were inferred using IQ-TREE v2.3.6 (Minh *et al.*, 2020) with the parameters --bb 1000 --alrt 1000 -m Q.plant+I+G4. The species tree was estimated from the gene trees using wASTRAL v1.23.3.7 (Zhang & Mirarab, 2022) using -R and -u 2 parameters. For the concatenation-based species phylogeny, only BUSCO orthogroups present in at least 80% of species were used for the supermatrix construction, leading to the use of 516 BUSCO genes. All trimmed alignments for these orthogroups were concatenated into a supermatrix within IQ-TREE v2.3.6 (Minh *et al.*, 2020), followed by model selection with ModelFinder

(Kalyaanamoorthy *et al.*, 2017) and tree inference. Fruit and flower colour information was obtained from published literature, floras, and species descriptions.

Identification of flavonoid and carotenoid biosynthesis genes

KIPEs3 v0.38 (Rempel *et al.*, 2023) was used to identify the structural genes in the carotenoid and flavonoid biosynthesis pathways, utilising the polypeptide sequences of all the analysed datasets. The provided baits and residues datasets (FlavonoidBioSynBaits_v3.4 and CarotenoidBioSynBaits_v2.0, respectively) were used to identify functional sequences. For transcriptional regulators, members of the MBW complex were identified as follows: MYB candidates using MYB annotator v0.2 (Pucker, 2022), bHLH candidates using bHLH annotator v0.2 (Thoben & Pucker, 2023), and WD40 candidates by constructing phylogenetic trees including previously characterised anthocyanin-regulating WD40 sequences from the literature (see **Additional File 5** for sequence IDs). Phylogenetic trees were also generated as an additional line of evidence for confirming the presence/absence of all identified genes. Structural gene candidates predicted by KIPEs3 that retained at least 75% of residues known to be critical for function were retained to ensure that even potential pseudogenes are captured when inferring gene loss. Previously characterised functional sequences from the literature were also included. Multiple sequence alignments were generated separately for the following groups: CHS, CHI, CYP450 (including FNS II, F3′H, and F3′5′H), 2ODDs (including FNS I, F3H, FLS, and ANS), SDRs (including DFR, LAR, and ANR), arGST, UGTs (including F3GT and A3GT), MYBs (SG5, SG6, and SG7 MYBs), anthocyanin-associated bHLHs, and anthocyanin-associated WD40 proteins.

For the regulatory gene candidates, the script collect_best_blast_hits.py (available at https://github.com/bpucker/ApiaceaeFNS1) (Pucker & Iorizzo, 2023) was used to collect similar sequences to known MBW genes from *Malus domestica* and *Arabidopsis thaliana*. These sequences, together with reference MBW sequences from the literature, were used for phylogenetic analysis. Alignments were generated using MAFFT v7.490 (G-INS-i algorithm) (Katoh *et al.*, 2019) and translated back to codon alignments with pxaa2cdn (Brown *et al.*, 2017). Alignments were filtered with pxclsq from the PHYX toolkit (-p 0.1) (Brown *et al.*, 2017). Maximum-likelihood phylogenies were inferred using IQ-TREE v2.3.6 (Minh *et al.*, 2020) with the parameters: --alrt 1000 -B 1000 -m MFP, where the best-fit evolutionary model was determined by ModelFinder (Kalyaanamoorthy *et al.*, 2017).

Synteny analysis and WGD dating

Genome-wide syntenic and collinear blocks were identified across 19 representative genome sequences included in this study; *Fragaria vesca*, *Ulmus minor*, *Quercus robur*, *Coriaria nepalensis*, *Begonia loranthoides*, *Begonia peltatifolia*, *Bayabusua clarkie*, *Gynostemma pentaphyllum*2, *Thladiantha cordifolia*1, *Momordica charantia*2, *Herpetospermum pedunculosum*1, *Luffa aegyptiaca*1, *Sicyos edulis*1, *Trichosanthes truncata*1, *Cucurbita moschata*1, *Cucurbita pepo*1, *Citrullus lanatus*2, *Cucumis melo*1, and *Cucumis sativus*1 (**Additional File 1**). These genome sequences were selected based on completeness (as measured by the BUSCO score), continuity, and taxonomic diversity. Synteny analysis was performed using JCVI/MCscan v1.4.14 (Tang *et al.*, 2015). Coding sequence (CDS) and GFF annotation files were used as input. Pairwise orthology and synteny searches were conducted using *F. vesca* as the reference, with default parameters and a c-score cutoff of 0.99. For each species, synteny depth was plotted against *F. vesca* to determine the appropriate number of syntenic block iterations, which is critical for accurately representing synteny in species with whole-genome duplications (WGDs). Based on this approach, the inferred synteny depth ratios against one region of *F. vesca* were as follows: *U. minor*:1, *Q. robur*:1; *C. nepalensis*:4, *B. loranthoides*:4, *B. peltatifolia*:4, *B. clarkei*:2, *G. pentaphyllum*:3, *T. cordifolia*:2, *M. charantia*:2, *H. pedunculosum*:2, *L. aegyptiaca*:2, *S. edulis*:4, *T. truncata*:2, *C. moschata*:4, *C. pepo*:4, *C. lanatus*:2, *C. melo*:2, *C. sativus*:2.

To investigate *DFR*, *ANS*, and *arGST* synteny, candidate sequences were identified in *F. vesca* and used to extract conserved syntenic regions across all selected species. For *arGST*, the local synteny was less conserved than for *DFR* and *ANS*, possibly due to genome rearrangements associated with speciation and WGDs. To ensure robust orthology assignment, we used a phylogenomics approach, consisting of (i) extraction of ±35 flanking genes around the *F. vesca arGST* locus, identification of candidate homologs using collect_best_blast_hits.py (available at https://github.com/bpucker/ApiaceaeFNS1) (Pucker & Iorizzo, 2023), (iii) multiple sequence alignment using MAFFT v7.490 (--auto strategy) (Katoh *et al.*, 2019), (iv) alignment cleaning with pxclsq (Brown *et al.*, 2017), and (v) phylogenetic inference using FastTree v2.1.11 (Price *et al.*, 2010), under default parameters. Only the syntelogs (orthologs in a macrosyntenic context) were accepted as syntenic orthologs and used for downstream visualisation.

Since synteny depth suggested four syntenic regions of *Coriaria* against one region of *F. vesca*, this suggested a previously unreported WGD event in *C. nepalensis*. To test this, WGD analysis was performed using wgd v2.0.38 (Chen & Zwaenepoel, 2023; Chen *et al.*, 2024) based on coding sequences of high-quality genome sequences (*Q. robur*, *C. nepalensis*, *B. loranthoides*, *G. pentaphyllum*2, and *C. sativus*1). Default workflows were followed to infer orthogroups, compute synonymous substitutions per site of duplicated genes (Ks distributions), detect intra- and interspecies synteny, and estimate absolute WGD ages via Bayesian MCMC dating. The species phylogeny used for dating was reconstructed under a multi-species coalescent framework as described above. The following lower and upper-bound divergence times were used as secondary calibration points from timetree.org (Kumar *et al.*, 2022): *Q. robur* vs. Cucurbitales (93.0-109.4 mya); *C. nepalensis* vs. *B. loranthoides* (55.8-76.0 mya); *B. loranthoides* vs. *Cucurbitaceae* (56.6-91.0 mya); and *G. pentaphyllum* vs. *C. sativus* (26.5-68.7 mya).

Expression analysis

To compare the expression of remaining flavonoid biosynthesis genes in *Cucurbitaceae* with those of close outgroups, we collected and analyzed publicly available RNA-seq datasets from NCBI SRA following a previously established approach (Pucker *et al.*, 2024). Species were selected based on three criteria: (i) broad taxonomic representation, (ii) high completeness of gene annotation (protein BUSCO), and (iii) availability of >10 RNA-seq datasets to ensure robust comparisons. Species used in this analysis are shown in Figure 5.

All paired-end RNA-seq FASTQ files available at NCBI SRA were collected for all selected species using the NCBI SRA toolkit v3.1.0 (SRA IDs in **Additional File 6**). Transcript abundance was quantified using Kallisto v0.50.1 (Bray *et al.*, 2016) based on the coding sequences. The individual count tables were then merged and filtered with customised Python scripts (Pucker & Iorizzo, 2023; Choudhary & Pucker, 2024).

Because expression values must be comparable across species and pathways, we implemented a reference-gene-based normalisation strategy:

1. Identification of candidate reference genes: Multiple reference genes showing stable gene expression from *Arabidopsis thaliana* were curated from the literature.
2. Ortholog detection: Using the phylogenomics approach described above, we identified syntelogs of these reference genes in all candidate species. Reference genes with missing orthologs in any species were excluded.
3. Stability check: For each gene, relative expression was computed across samples as

$$R_{i,j} = \frac{E_{i,j}}{\mu_j}$$

where $E_{i,j}$ is the expression of gene *i* in sample *j*, and μ_j is the mean expression of all samples for that gene. Reference genes showing $R_{i,j}$≈1 with low variance across species and samples were considered stable.

4. Final reference set: Ten genes, which were the most stable across all species (**Additional File 7**), were selected as internal controls.

For per-sample normalisation, the mean expression of the reference genes (internal controls) in each sample was computed, and all gene counts in that sample were divided by this mean, yielding reference-normalised expression values. Alternative isoforms and paralogs contributing to the same function were summed to derive total functional gene expression.

Finally, we compared the expression of anthocyanin and proanthocyanidin biosynthesis genes with that of carotenoid biosynthesis genes in both *Cucurbitaceae* and outgroups to test whether carotenoids functionally compensate for the absence of anthocyanins in cucurbits.

Results

A highly supported nuclear phylogeny of *Cucurbitaceae*

In this study, a total of 258 datasets of *Cucurbitaceae* species were analysed, representing all 15 tribes and 68 genera (out of an estimated 95-115 genera; (Schaefer *et al.*, 2008; Schaefer & Renner, 2011b)). An additional 24 outgroup species were included, eight each from Cucurbitales (excluding *Cucurbitaceae*), Fagales, and Rosales, reflecting their phylogenetic closeness to *Cucurbitaceae* (The Angiosperm Phylogeny Group *et al.*, 2016). Among the *Cucurbitaceae* datasets, 136 were genomic from various sources, and the remaining were transcriptomic datasets from a previous study (Guo *et al.*, 2020). Of the genomic datasets, 82 were newly annotated in this study.

Given the known risk of contamination in all transcriptome assemblies, all 258 *Cucurbitaceae* datasets were screened for non-cucurbit sequences. Twenty-seven contaminated datasets were identified and excluded from downstream analysis (**Additional File 3**). Using the remaining 255 datasets (including multiple accessions per species), we extracted 2805 single-copy orthologs from the eudicotyledons_odb12 BUSCO dataset as phylogenetic markers.

An initial coalescent-based species phylogeny was constructed using FastTree and wASTRAL (**Figure S1 in Additional File 4**). Based on this preliminary tree, five datasets were excluded due to low completeness or phylogenetic incongruence: Xerosicyos perrieri1 and Ampelosicyos leandrii (0.1 % and 0% BUSCO, respectively) lacked sufficient gene recovery; Neoalsomitra pilosa (9% BUSCO) and Momordica subangulata (96.4% BUSCO) clustered in unexpected tribes, likely due to low completeness or misidentified source material; and Diplocyclos schliebenii grouped with Peponium; whereas Diplocyclos palmatus and related taxa clustered with Coccinia both in our study and previous studies (Chomicki *et al.*, 2020), suggesting possible misidentification or sample contamination.

To reconstruct a robust species phylogeny, we adopted a two-method-based approach. First, we generated 2805 gene trees using IQ-TREE under maximum-likelihood (ML) and inferred the species tree using the coalescent method implemented in weighted ASTRAL (wASTRAL) v1.23.3.7 (**Figure 2a**; **Figure S2 in Additional File 4**). Second, we performed concatenated ML analyses using a 516-gene supermatrix (**Figure S3 in Additional File 4**), as the concatenation method with a relatively small number of loci can reduce systemic errors, whereas very large supermatrices may yield spuriously high support for incorrect topologies (Seo, 2008). Both methods were evaluated for congruence.

The coalescent-based tree provided a more resolved topology and stronger support for monophyletic tribes compared to the concatenated analysis. Branch support was evaluated using local posterior probability (LPP) values. *Cucurbitaceae* was recovered as a strongly supported monophyletic clade, subdivided into eight major clades (Clade I to Clade VIII). Clade I consists of early diverging lineages Triceratieae, Zanonieae, Actinostemmateae, and Gomphogyneae (LPP=0.62). Clade II consists of Indofevilleae (LPP=1), Clade III-VI containing individual tribes Thladiantheae, Siraitieae, Momordiceae, and Jollifieae (all LPP=1). Clade VII consists of sister tribes Bryonieae and Scizopeponeae, clustered together as sister to Sicyoeae (LPP=1). Lastly, clade VIII consists of sister tribes Benincaseae and Coniandreae, with *Cucurbitaceae* sister to this group (LPP=1).

The concatenation-based tree largely mirrored this topology, except that the early-diverging lineages formed three individual clades rather than a single combined clade: Triceratie and Zanonieae (Clade I), Gomphogyneae (Clade II), and Actinostemmateae (Clade III). However, the concatenated tree failed to recover monophyly for the *Bryonieae* tribe and has nested phylogenies for *Trichosanthes/Hodgsonia/Cyclanthera* and *Coccinia/Diplocyclos* genera. Therefore, the coalescent-based phylogeny was selected as the most reliable species tree and used for all downstream analyses (**Figure 2a**).

Cucurbitaceae comprise a wide diversity of fruit and vegetable crops (**Figure 2b**). Mapping of fruit and flower pigmentation traits onto the phylogeny (**Figure 2a**) revealed that green, orange and red were the most frequent fruit colours, while white, yellow, and green were predominant among flowers in *Cucurbitaceae*.

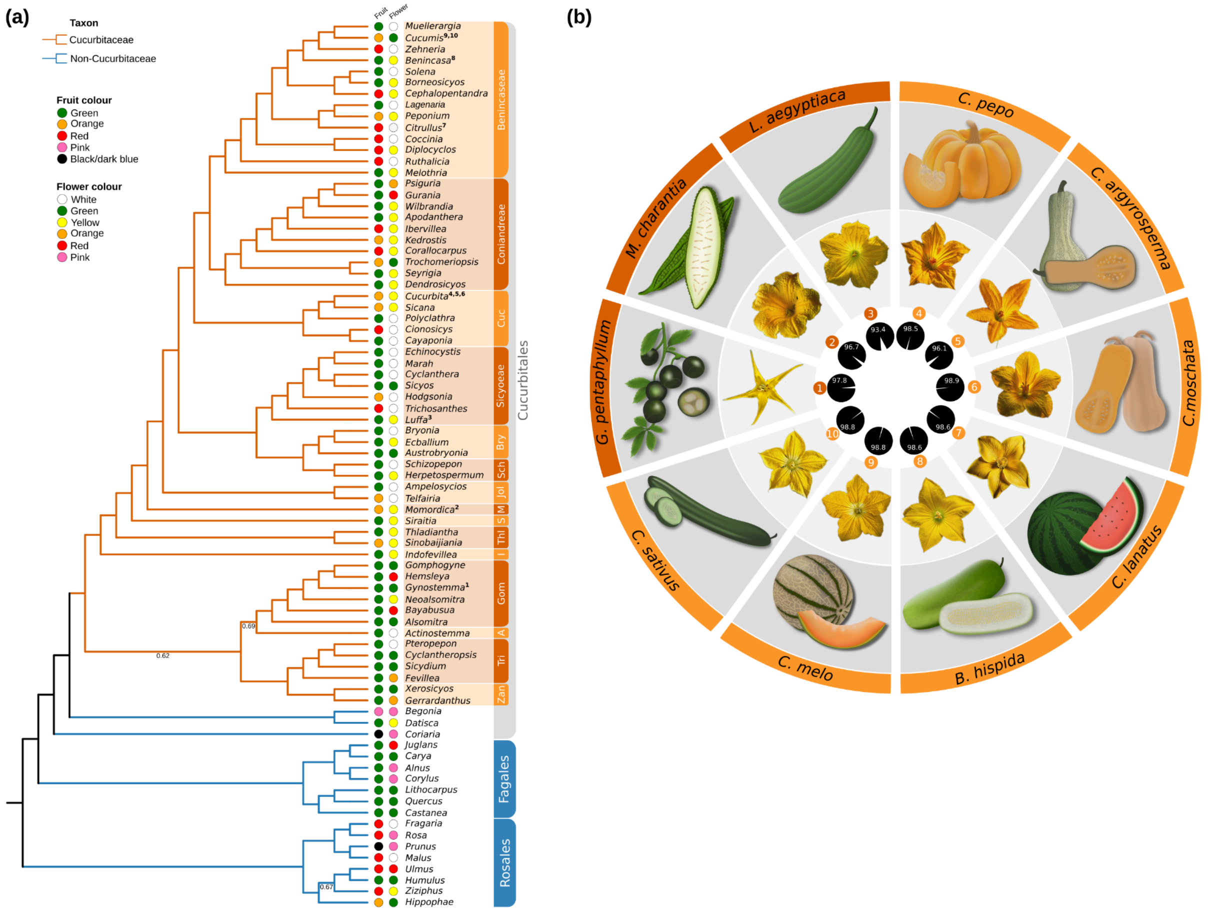

(a)
Taxon
Cucurbitaceae
Non-Cucurbitaceae
Fruit colour
Green
Orange
Red
Pink
Black/dark blue
Flower colour
White
Green
Yellow
Orange
Red
Pink
Fruit
Flower
Muellerargia
Cucumis9,10
Zehneria
Benincasa8
Solena
Borneosicyos
Cephalopentandra
Lagenaria
Peponium
Citrullus7
Coccinia
Diplocyclos
Ruthalicia
Melothria
Psiguria
Gurania
Wilbrandia
Apodanthera
Ibervillea
Kedrostis
Corallocarpus
Trochomeriopsis
Seyrigia
Dendrosicyos
Cucurbita4,5,6
Sicana
Polyclathra
Cionosicys
Cayaponia
Echinocystis
Marah
Cyclanthera
Sicyos
Hodgsonia
Trichosanthes
Luffa3
Bryonia
Ecballium
Austrobryonia
Schizopepon
Herpetospermum
Ampelosycios
Telfairia
Momordica2
Siraitia
Thladiantha
Sinobaijiania
Indofevillea
Gomphogyne
Hemsleya
Gynostemma1
Neoalsomitra
Bayabusua
Alsomitra
Actinostemma
Pteropepon
Cyclantheropsis
Sicydium
Fevillea
Xerosicyos
Gerrardanthus
Begonia
Datisca
Coriaria
Juglans
Carya
Alnus
Corylus
Lithocarpus
Quercus
Castanea
Fragaria
Rosa
Prunus
Malus
Ulmus
Humulus
Ziziphus
Hippophae
Benincaseae
Coniandreae
Cuc
Sicyoeae
Bry
Sch
Jol
M
S
Thl
I
Gom
A
Tri
Zan
Cucurbitales
Fagales
Rosales
0.69
0.62
0.67
(b)
L. aegyptiaca
C. pepo
C. argyrosperma
C.moschata
C. lanatus
B. hispida
C. melo
C. sativus
G. pentaphyllum
M. charantia
1
2
3
4
5
6
7
8
9
10
97.8
96.7
93.4
98.5
96.1
98.9
98.6
98.6
98.8
98.8

Figure 2: A summary of phylogenetic relationships and pigmentation traits in *Cucurbitaceae*. (a) Species phylogeny inferred from 2805 BUCO single copy loci, aggregated at the genus level and representing all 15 tribes of the *Cucurbitaceae* family. Numbers on branches represent local posterior probability if below 1. Branches corresponding to *Cucurbitaceae* are shown in orange, while non-*Cucurbitaceae* outgroups are shown in blue. Fruit and flower colours are mapped onto the terminal tips at the genus level. When multiple colours were observed within a genus, the most intense colour (e.g., red over orange, orange over yellow) was selected. For fruit colour, preference is given to the mature fruit stage and the cut-open fruit section when available. Full species phylogenies are provided in the supporting information in Additional File 4; Figures S1, S2 and S3. Tribe abbreviations: Zan, Zanonieae; Tri, Triceratieae; Act, Actinostemmateae; Gom, Gomphogyneae; I, Indofevilleae; T, Thladiantheae; S, Siraitieae; M, Momordiceae; Jol, Joliffieae; Sch, Schizopeponeae; Bry, Bryonieae; Cuc, Cucurbiteae (b) Representative *Cucurbitaceae* species included in this study, illustrating variation in flowers and fruits. The innermost pie chart indicates BUSCO protein completeness scores for each species, followed by concentric circles showing flower morphology (middle) and fruit morphology (outer).

Absence of major anthocyanin and proanthocyanidin biosynthesis genes in the *Cucurbitaceae*

Orthologous genes typically retain conserved functions across species and are instrumental for predicting gene functions (Eisen, 1998; Pucker, 2024). Conversely, the inability to identify orthologs in certain lineages can signal lineage-specific gene loss or gene gain in others. To examine this in the context of pigmentation in *Cucurbitaceae*, we performed a comprehensive ortholog-based survey of flavonoid biosynthesis genes (FBGs) across all *Cucurbitaceae* datasets included in this study. The analysis incorporated both sequence similarity to functionally characterized reference genes and conservation of key amino acid residues essential for enzymatic activity. In addition, phylogenetic analyses were performed for each major structural and regulatory gene family to validate orthology and infer *Cucurbitaceae* pigmentation traits.

Consistent with expectations, the core flavonoid and phenylpropanoid pathway genes (e.g., *PAL*, *C4H*, *4CL*, *CHS*, *CHI*, *FNS II*, *F3H*, *F3'H*, and *FLS*) were readily identified across all *Cucurbitaceae* datasets, displaying clear orthology with known functional genes in other eudicots. However, the analyses revealed a striking absence of key anthocyanin and proanthocyanidin biosynthetic genes, *DFR*, *ANS*, and *arGST* for anthocyanin biosynthesis and *ANR* and *LAR* for proanthocyanidin biosynthesis (Figure 3).

A few exceptions were observed among the early diverging species in the *Cucurbitaceae*, where *ANR* was detected in *Bayabusua clarkie*, *LAR* was detected in *Alsomitra macrocarpa* and *B. clarkie*, and *DFR* or DFR-like sequences in *B. clarkie*, *A. macrocarpa*, *Gynostemma* and a few related taxa. In contrast, the vast majority of *Cucurbitaceae* lacked all essential structural genes required for anthocyanidin and proanthocyanidin biosynthesis, suggesting a family-wide irreversible loss of both branches of the flavonoid pathway.

The status of *U3GT* (UDP-glucose:flavonoid 3-O-glucosyltransferase) could not be conclusively resolved. Although *Cucurbitaceae* sequences clustering with canonical *U3GT* gene sequences were identified (Figure 3g), some glycosyltransferases are known to exhibit substrate specificity, often catalyzing glycosylation of both flavonols and anthocyanidins (Lee *et al.*, 2005). The functional role of *U3GT*-like sequences remains uncertain in *Cucurbitaceae* requiring enzymatic validation.

To investigate the regulatory control, the MBW complex was examined. Orthologs of Subgroup 6 (SG6) MYBs, which regulate anthocyanin biosynthesis, and PA1 MYB and PA2 MYB (SG5) controlling proanthocyanidin biosynthesis, were entirely absent from *Cucurbitaceae*. Subgroup 7

(SG7) MYBs, which regulate flavonol biosynthesis, were present and phylogenetically conserved. The other two components of the MBW complex, bHLH and WD40 proteins, were detected in almost all datasets, consistent with their pleiotropic roles in diverse biological processes beyond flavonoid biosynthesis.

The identified presence-absence patterns were mapped onto the Cucurbitaceae species phylogeny (**Figure 3**). While interpreting gene absence as evolutionary loss is vulnerable to technical artefacts like incomplete assemblies; low sequence coverage or extreme sequence divergence, several lines of evidence support the inferred evolutionary gene loss. First, consistent absence across multiple independent, high-quality genome assemblies and annotations; second, concordant absence in both genomic and transcriptomic datasets from diverse lineages; and lastly, the identification of *DFR*, *LAR*, and *ANR* in a few early-diverging taxa using the same approach.

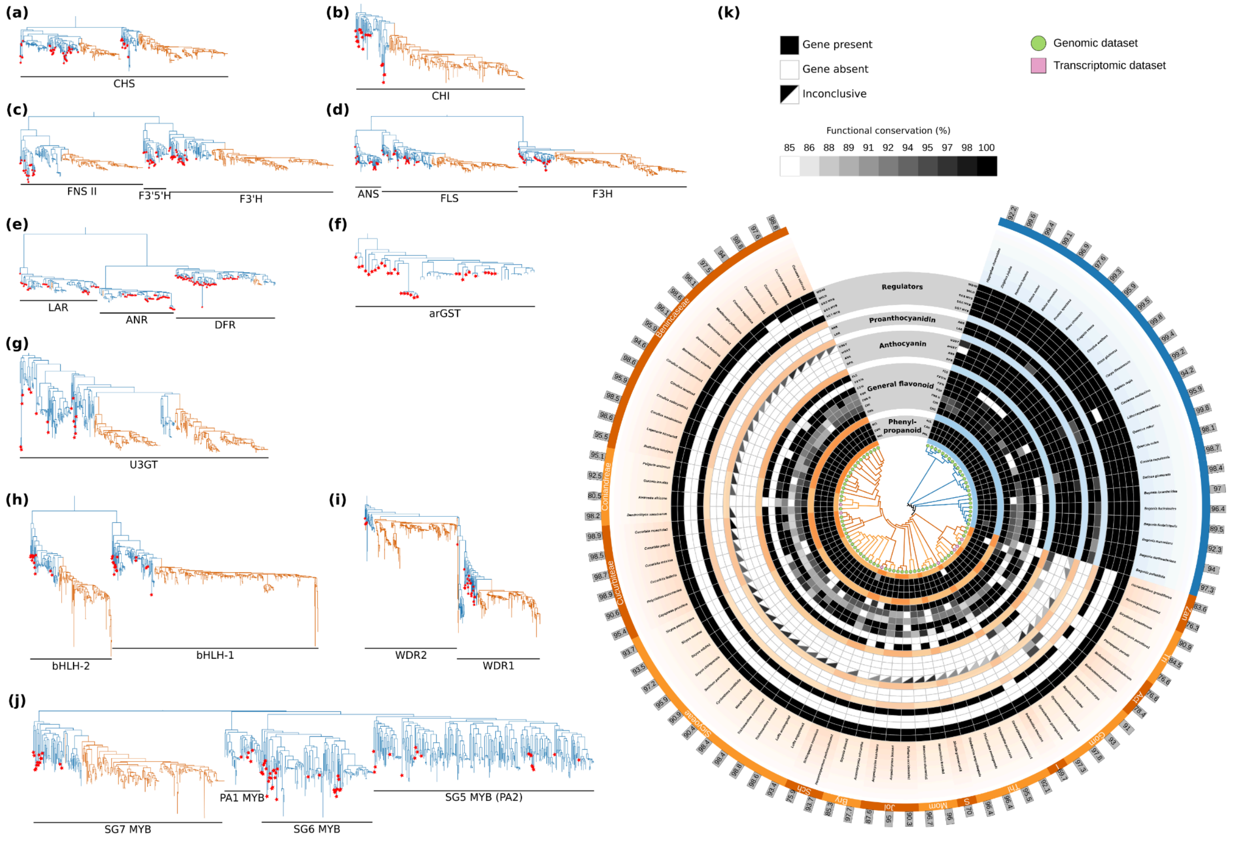


Figure 3: The phylogenetic pattern of major flavonoid pathway genes in *Cucurbitaceae*. Maximum-likelihood phylogenetic trees of major flavonoid pathway genes are shown on the left.

Branches corresponding to *Cucurbitaceae* sequences are shown in orange, while non-*Cucurbitaceae* outgroups are shown in blue. Red stars at the start of terminal branches indicate functional enzymatic sequences identified in previous studies. Gene trees for (a) CHS, (b) CHI, (c) CYP450 family (FNS II, F3'H, and F3'5'H), (d) 2-ODD family (F3H, FLS, and ANS), (e) SDR family (LAR, ANR, and DFR), (f) arGST, (g) U3GT, (h) anthocyanin-related bHLHs, (i) anthocyanin-related WD40, and (j) flavonoid MYBs (SG7 MYB, SG6 MYB, SG5 MYB, and PA1 MYB) are shown. Full phylogenetic trees are provided in **Additional File 5**. On the right, the presence-absence summary matrix of flavonoid biosynthesis genes mapped to the *Cucurbitaceae* phylogeny. Outgroups (blue) and Cucurbitacae (orange) species are annotated with genomic (green tips) and transcriptomic (pink tips) data sources. Gene presence, absence, or putative non-functional sequences are shown as black, white and grey boxes, respectively. The outermost bars represent the percentage of conserved BUSCO genes identified in the polypeptide sequences of each dataset shown. Tribe abbreviations: Zan, Zanonieae; Tri, Triceratieae; Act, Actinostemmateae; Gom, Gomphogyneae; I, Indofevilleae; Thl, Thladiantheae; S, Siraitieae; Mom, Momordiceae; Jol, Joliffieae; Bry, Bryonieae; Sch, Schizopeponeae.

Microsynteny supports the loss of anthocyanin biosynthesis genes

The evolutionary path of anthocyanin and proanthocyanidin genes might help in understanding the inability to find anthocyanidin and proanthocyanidin biosynthesis genes in *Cucurbitaceae*. To elucidate the basis for the absence of anthocyanin and proanthocyanidin biosynthesis genes in *Cucurbitaceae*, we conducted a comparative microsynteny analysis focused on key anthocyanin biosynthetic loci.

Before examining fine-scale microsynteny, it was necessary to understand the large-scale genomic context, particularly the history of whole-genome duplication (WGD) events in the analysed species relative to the reference genome (*Fragaria vesca*). All eudicot share an ancient core eudicot common hexaploidization (ECH) event, estimated to have occurred approximately 115-130 million years ago (Mya) (Jaillon *et al.*, 2007). Following this, the Cucurbitales common tetraploidization (CCT) event, shared across all members of the order, took place around 93-105 Mya (Wang *et al.*, 2022). Begoniaceae experienced a WGD around 35 Mya (Li *et al.*, 2022). Within *Cucurbitaceae*, additional lineage-specific WGDs have been reported. A major event, the Cucurbita-specific tetraploidization (CST), occurred around 30 Mya, resulting from an ancient allotetraploidization followed by extensive gene loss and rediploization (Montero-Pau *et al.*, 2018). Furthermore, a distinct Sicyoeae-specific WGD, dated to approximately 25 Mya (Fu *et al.*, 2021b), was previously identified and is shared among *Sicyos* and closely related taxa (Guo *et al.*, 2020).

Given this complex history of polyploidization, accurately inferring orthology required representing one-to-many syntenic relationships in comparisons with the *F. vesca* reference genome sequence. Synteny depth analyses confirmed the expected orthology ratios that reflect these duplication events: 1:1 between *F. vesca* and *Quercus robur*, 1:2 between *F. vesca* and most *Cucurbitaceae*; and 1:4 between *F. vesca* and the highly polyploid lineages *Begonia*, *Cucurbita*, and *Sicyos* (**Additional File 8**). Unexpectedly, we observed a 1:4 orthology ratio between *F. vesca* and *C. nepalensis*. Because *Coriaria* is known to share only the CCT event, a 1:2 ratio would have been expected. This pattern instead suggests an additional, lineage specific WGD in *Coriaria*. To test this hypothesis, we applied the wgd v2.0.38 pipeline and detected a previously unreported, independent duplication event at 18 Mya, providing strong support for Coriaria-specific WGD (**Additional File 9**).

With the polyploidization history clarified, we next investigated the conservation of genomic regions harbouring key anthocyanin biosynthesis genes. Comparative analyses across 19 high-quality genome sequences (including both outgroups and *Cucurbitaceae*) revealed well-conserved syntenic blocks surrounding the *DFR*, *ANS*, and *arGST* loci (**Figure 4**), as well as for proanthocyanidin biosynthesis genes *ANR* and *LAR* (data not shown). To validate the reliability of the identified syntenic relationships, we also generated phylogenetic trees to confirm identified homologs represented true orthologous relationships.

The microsynteny pattern clearly demonstrates a loss of the anthocyanin biosynthesis genes *DFR*, *ANS*, and *arGST* across *Cucurbitaceae*. A DFR homolog was retained in *Bayabusua clarkie,* and a DFR-like sequence was identified in *Gynostemma pentaphyllum* (early diverging *Cucurbitaceae*). Similarly, the proanthocyanidin genes LAR and ANR were absent from all *Cucurbitaceae* examined except for a few early-diverging taxa. This suggests that the anthocyanin biosynthesis gene loss occurred at the base of *Cucurbitaceae* family while the proanthocyanidin gene loss occurred after the divergence of basal cucurbit lineages.

Together, the phylogenetic trees, biosynthetic gene identification, and microsynteny provides robust genomic evidence for the complete loss of both anthocyanidin and proanthocyanidin biosynthetic genes in the majority of *Cucurbitaceae*.

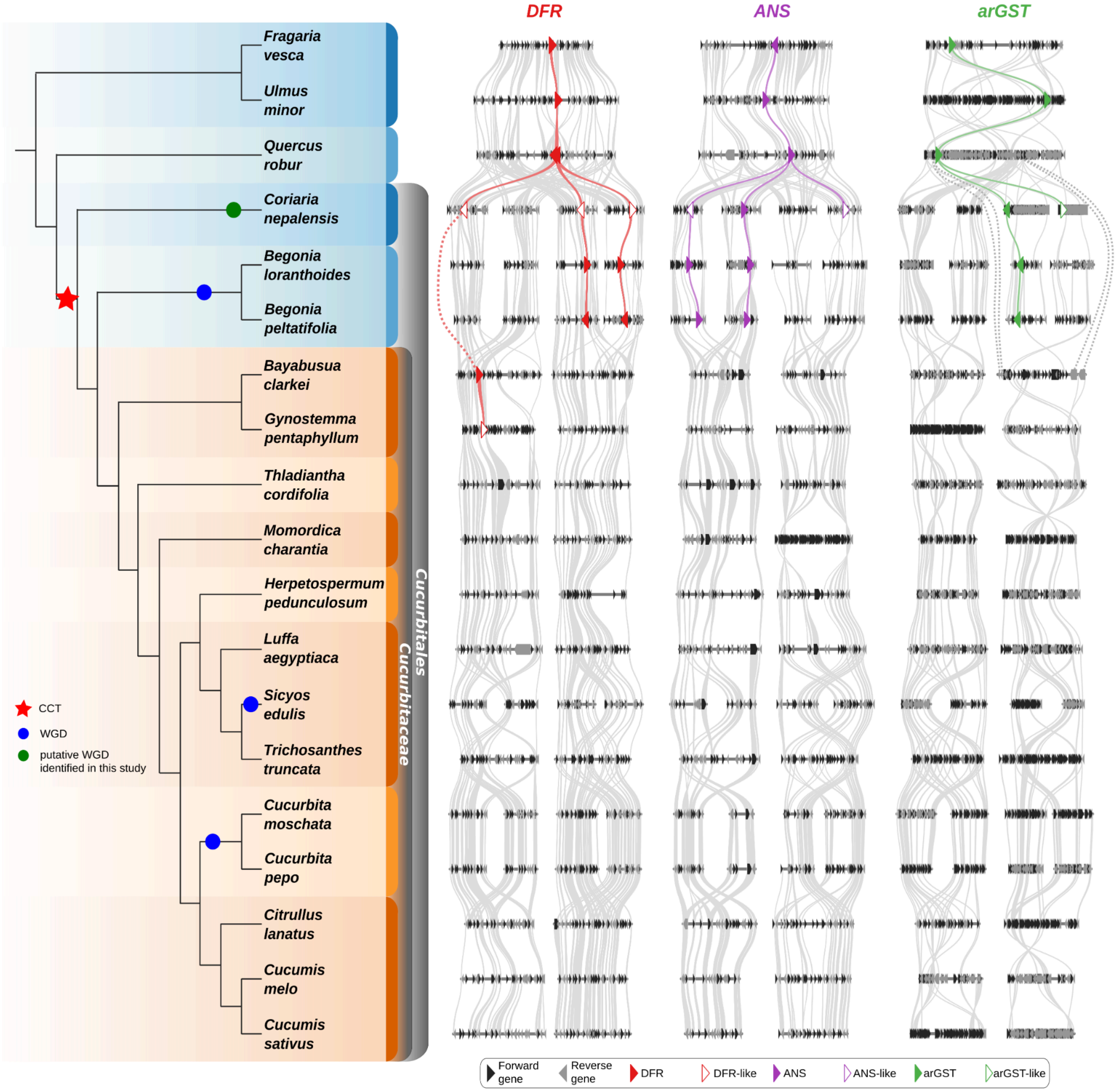

DFR
ANS
arGST
Fragaria vesca
Ulmus minor
Quercus robur
Coriaria nepalensis
Begonia loranthoides
Begonia peltatifolia
Bayabusua clarkei
Gynostemma pentaphyllum
Thladiantha cordifolia
Momordica charantia
Herpetospermum pedunculosum
Luffa aegyptiaca
Sicyos edulis
Trichosanthes truncata
Cucurbita moschata
Cucurbita pepo
Citrullus lanatus
Cucumis melo
Cucumis sativus
Cucurbitales
Cucurbitaceae
CCT
WGD
putative WGD identified in this study
Forward gene
Reverse gene
DFR
DFR-like
ANS
ANS-like
arGST
arGST-like

Figure 4: On the left, the species phylogeny of all taxa used for synteny analysis is shown. A red star denotes the Cucurbitales common tetraploidization (CCT) event, while blue circles indicate lineage-specific WGDs: the Begoniaceae common tetraploidization (BCT), the Sicyoeae WGD, and the Cucurbita-specific polyploidization (CST). A green circle marks the putative Coriaria-specific WGD identified in this study. On the right, conserved syntenic regions for the anthocyanin biosynthesis genes, *DFR* (red), *ANS* (purple), and *arGST* (green) are shown. Genes outlined with unfilled coloured borders represent putatively non-functional copies, based on conserved amino acid residue analysis. Other genes are shown in black (forward strand) and grey (reverse strand). The number of syntenic blocks displayed per species reflects

the degree of collinearity relative to *F. vesca* and corresponds to the number of polyploidization events in each lineage's evolutionary history.

Carotenoids as functional replacements for anthocyanins

The complete loss of the anthocyanin and proanthocyanidin biosynthetic genes in *Cucurbitaceae* raises an important question regarding the evolutionary fate of the upstream flavonoid biosynthetic pathway. To assess whether these upstream genes remain transcriptionally active, we performed comparative expression analysis using RNA-seq data from eight representative *Cucurbitaceae* species and eight phylogenetically close outgroup species from *Fagales* and *Rosales*.

Across all *Cucurbitaceae* species, we observed a marked reduction in transcript abundance for the remaining flavonoid biosynthesis genes (**Figure 5a**). Particularly, *CHS*, which encodes the enzyme catalyzing the first committed step of flavonoid biosynthesis, exhibited near-negligible expression levels in *Cucurbitaceae* species, suggesting that the entire flux into the flavonoid pathway is effectively constrained at its entry point. While *CHI*, *F3'H* and *FLS* retained residual expression, their overall transcript levels were significantly lower than those observed in taxa with functional anthocyanin biosynthesis. Quantitative comparisons confirmed that this expression difference was statistically significant ($p<0.001$) between *Cucurbitaceae* and outgroups (**Figure 5b**). These results might imply that the upstream flavonoid biosynthesis genes in *Cucurbitaceae* are undergoing relaxed or negative purifying selection, consistent with functional redundancy, and gradual pathway erosion following the loss of anthocyanin and proanthocyanidin branches. However, this requires further work before making any definitive conclusions.

Given the multifunctional roles of anthocyanins in pigmentation, photoprotection, and pollinator attraction, we next investigated whether carotenoids, the other major class of plant pigments, may have functionally compensated for the anthocyanin loss. To test this hypothesis, we compared the expression of core carotenoid biosynthetic genes between *Cucurbitaceae* and the same outgroup taxa (**Figure 5d**). Although carotenoid biosynthesis represents a primary metabolic pathway that is constitutively active in photosynthetic tissues, *Cucurbitaceae* exhibited a significantly higher overall transcript abundance across nearly all carotenoid biosynthetic genes compared to outgroups ($p < 0.01$; **Figure 5c**). This elevated expression may suggest adaptive enhancement of carotenoid production, both to compensate for the visual signalling roles and to maintain photoprotection in the absence of anthocyanins and possibly flavonoids. The observed transcript abundance is consistent with the high carotenoid content reported in cucurbit fruits and flowers.

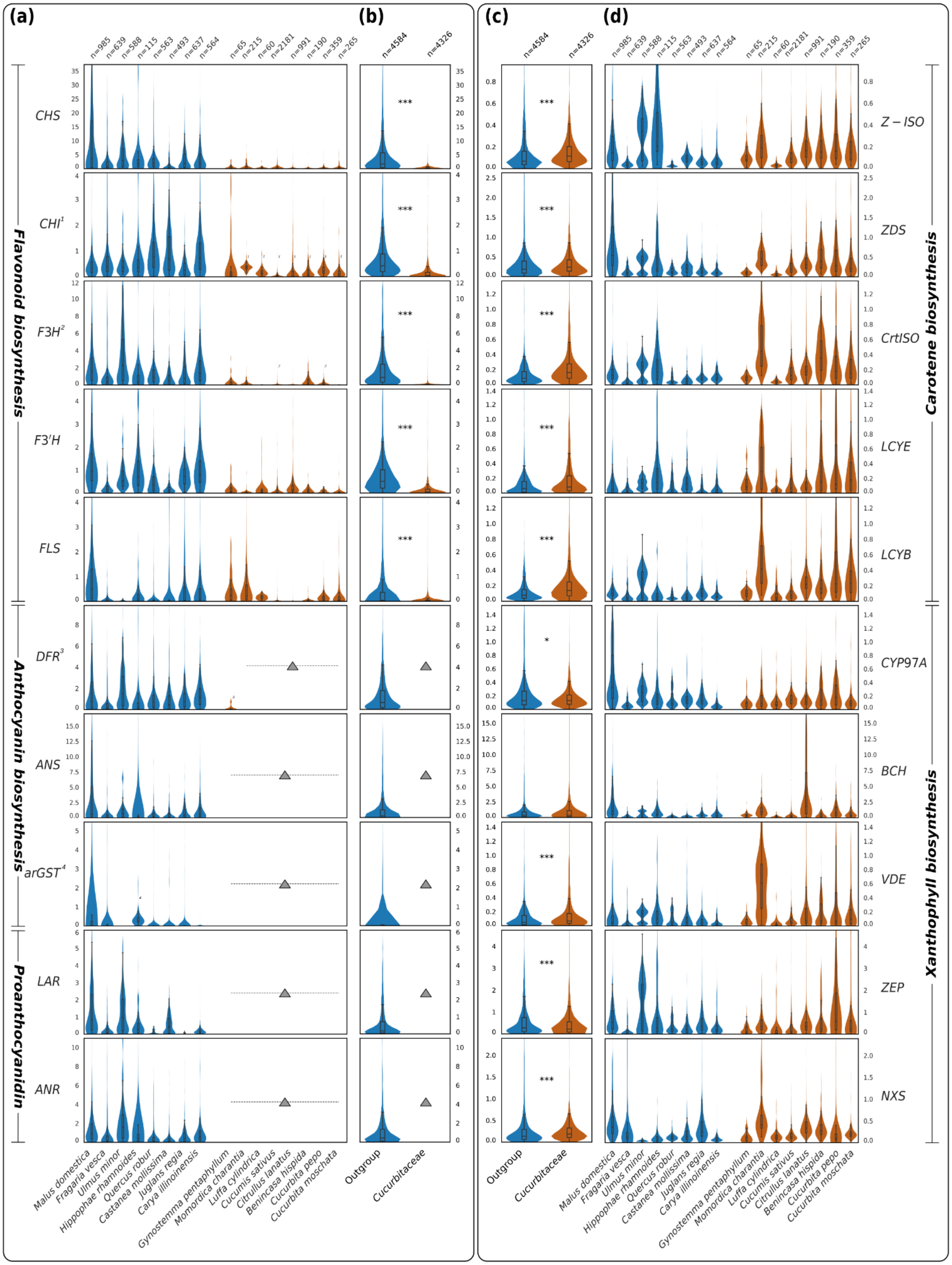

(a)
(b)
(c)
(d)
Flavonoid biosynthesis
Anthocyanin biosynthesis
Proanthocyanidin
Carotene biosynthesis
Xanthophyll biosynthesis
CHS
CHI¹
F3H²
F3′H
FLS
DFR³
ANS
arGST⁴
LAR
ANR
Z − ISO
ZDS
CrtISO
LCYE
LCYB
CYP97A
BCH
VDE
ZEP
NXS
n=985
n=639
n=588
n=115
n=563
n=493
n=637
n=564
n=65
n=215
n=60
n=2181
n=991
n=190
n=359
n=265
n=4584
n=4326
***
*
Malus domestica
Fragaria vesca
Ulmus minor
Hippophae rhamnoides
Quercus robur
Castanea mollissima
Juglans regia
Carya illinoinensis
Gynostemma pentaphyllum
Momordica charantia
Luffa cylindrica
Cucumis sativus
Citrullus lanatus
Benincasa hispida
Cucurbita pepo
Cucurbita moschata
Outgroup
Cucurbitaceae

Figure 5: (a) Expression analysis of core flavonoid biosynthetic genes in *Cucurbitaceae* (orange) and outgroups (blue). Sample size (*n*) for each group is indicated above the plots. Superscript numerals denote putative non-functional gene copies inferred from conserved amino acid residue analyses. Grey triangles indicate gene sequence not detected (loss inferred). (b) Combined expression analysis of flavonoid biosynthesis genes between all *Cucurbitaceae* and outgroups. Statistical significance is denoted as '***' (p <=0.001),'**' (p<=0.01), and '*' (p<=0.05).
(d) Expression analysis of the core carotenoid biosynthetic genes in outgroups (blue) and *Cucurbitaceae* (orange). (c) Combined expression analysis of carotenoid biosynthesis genes between all *Cucurbitaceae* and outgroups. Boxplots within violin plots indicate the median, interquartile range, and outliers.

Discussion

Gene loss is a major evolutionary force shaping morphological diversity, ecological adaptations and lineage-specific innovations across organisms (Smith & Rausher, 2011; Albalat & Cañestro, 2016). Although often associated with pseudogenization and loss of function, gene loss can also drive evolutionary innovation, where the removal of redundant or dispensable genes allows reconfiguration of metabolic or developmental networks, promoting novel adaptive functions (Olson, 1999; Sharma *et al.*, 2018).

In this study, we report an exceptional case of large-scale pathway gene loss defining the pigmentation history of the *Cucurbitaceae* family. Our analysis revealed complete absences of key structural genes required for anthocyanin and proanthocyanidin biosynthesis, as well as regulatory SG6 and SG5 MYB transcription factors responsible for activating these pathways. Such comprehensive loss of both structural and regulatory components across an entire plant family is exceedingly rare and defines a fundamental evolutionary shift in the pigmentation and defense strategies of *Cucurbitaceae*.

Anthocyanins serve a central ecological role by producing diverse colours that attract pollinators and seed dispersers, while proanthocyanidin contribute to defense, seed coat pigmentation, and dormancy. The complete absence of these pathways in *Cucurbitaceae* is therefore remarkable, especially since many species remain insect and bird pollinated (Schaefer & Renner, 2011c). Interestingly, several early-diverging *Cucurbitaceae* lineages, such as *Bayabusua*, *Alsomitra*, *Pteropepon*, and *Siolmatra*, are wind pollinated. Species such as *Alsomitra macrocarpa* and

Bayabusua clarkei produce wing-shaped seed capable of gliding hundreds of meters (Certini *et al.*, 2022) reducing reliance on visual signal for pollination and seed dispersal. This exceptionally adapted pollination strategy may have rendered anthocyanin pigmentation functionally irrelevant in early *Cucurbitaceae*, facilitating relaxed selection pressure and eventual gene loss. Gene loss is an irreversible event that prevents the lineage from acquiring the trait through the same genetic basis again thus opening possibilities for independent evolution or convergent evolution (Pucker *et al.*, 2024; Khatun *et al.*, 2025).

Following the initial loss, other adaptations likely compensated for the lack of anthocyanin-mediated pigmentation and defense. Carotenoids, which produce yellow, red, and orange pigmentation in *Cucurbitaceae* fruits and flowers, likely assumed the visual signaling functions of anthocyanins. The predominance of carotenoid-based coloration across the family, combined with elevated carotenoid biosynthetic gene expression, supports this hypothesis. Similarly, morphological innovations such as fringed petal lobes in genera like Trichosanthes may have evolved to enhance pollinator attraction through visual or tactile cues, partially compensating for the absence of anthocyanin pigments (de Boer *et al.*, 2012). Many genera like *Luffa*, *Cayaponia*, *Hodsgonia*, and *Sicyos* are adapted to water dispersal bypassing the need for anthocyanin pigmentation (Henry Nicholas Ridley, 1930).

Beyond visual attraction, seed pigmentation and defense functions, previously attributed to proanthocyanidins, may have been replaced by other biochemical mechanisms. One potential candidate is melanin, a lesser-studied plant pigment already known to color the black seeds of *Citrullus lanatus* (Glagoleva *et al.*, 2020).

While our findings reveal a clear pattern of gene loss, the temporal and phylogenetic context of these events remains to be fully resolved. Our species phylogeny indicates that the early-diverging tribes, Triceratieae, Zanonieae, Actinostemmateae, and Gomphogyneae, form a monophyletic basal clade (Clade I), consistent with previous reconstructions (Schaefer *et al.*, 2008; Lin *et al.*, 2025). However, alternative topologies reported for these early-diverging taxa (Chomicki *et al.*, 2020; Guo *et al.*, 2020) highlight the need for denser taxon sampling and higher-quality genome sequences. Improved genomic resolution in these early lineages will be essential for pinpointing the exact timing and sequence of gene loss events.

The detection of *DFR*, *LAR*, and *ANR* orthologs in *Bayabusua* and related early-diverging taxa suggests that the proanthocyanidin branch may have persisted longer than the anthocyanin branch. This pattern supports a scenario in which anthocyanin biosynthesis was lost first, possibly following the loss of its specific regulatory MYB activators, while the proanthocyanidin pathway remained partially functional before being completely lost in later diverging

Cucurbitaceae. Testing this hypothesis will require multiple high-quality genome sequences from early branching lineages to confirm whether the remaining flavonoid biosynthesis genes are functional or represent degenerated relics.

In summary, our findings reveal a systematic loss of anthocyanin and proanthocyanidin genes in *Cucurbitacae* and compensatory enhancement by carotenoid pigmentation. By uncovering the genomic and phylogenetic foundation of this transition, our study provides a framework for understanding the evolution of pigmentation systems in plants and suggests that anthocyanin pigmentation is not as conserved as previously believed.

Declarations

Data availability statement

All data sets underlying this study are publicly available. Customised Python scripts for the analyses in this study are available through GitHub (https://github.com/NancyChoudhary28/Cucurbits).

Author contributions

BP and NC designed the project. NC conducted the analyses and BP supervised the work. MH designed illustrations. NC wrote the manuscript. NC, MH, and BP revised the manuscript, and all authors agreed to its submission.

Acknowledgements

This work was supported by the de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) and ELIXIR-DE (Forschungszentrum Jülich and W-de.NBI-001, W-de.NBI-004, W-de.NBI-008, W-de.NBI-010, W-de.NBI-013, W-de.NBI-014, W-de.NBI-016, W-de.NBI-022). We thank all members of the research group Plant Biotechnology and

Bioinformatics for their discussion and support. We acknowledge support from Project DEAL and the University of Bonn for open access publication.

Supporting information

Description:

Additional_File1: Sources of genomic and transcriptomic datasets used in this study
Additional_File2: Gene annotation and re-annotation statistics
Additional_File3: Detailed results of Kraken and Bracken classification of all datasets.
Additional_File4: **Figure S1**: Coalescence species tree constructed from 2805 BUSCO orthogroups across 255 non-contaminated datasets, **Figure S2**: Coalescence species tree constructed from 2805 BUSCO orthogroups across 250 non-contaminated datasets, **Figure S3**: The ML tree reconstructed from a concatenated supermatrix of 516 BUSCO orthogroups across 250 non-contaminated datasets
Additional_File5: **Figure S4**: Codon-based maximum-likelihood phylogenetic tree of CHS sequences constructed with **(a)** MAFFT and IQ-TREE, and **(b)** MUSCLE and IQ-TREE
Figure S5: Codon-based maximum-likelihood phylogenetic tree of CHI sequences constructed with **(a)** MAFFT and IQ-TREE, and **(b)** MUSCLE and IQ-TREE
Figure S6: Codon-based maximum-likelihood phylogenetic tree of CYP450 sequences constructed with **(a)** MAFFT and IQ-TREE, and **(b)** MUSCLE and IQ-TREE
Figure S7: Codon-based maximum-likelihood phylogenetic tree of 2ODD sequences constructed with **(a)** MAFFT and IQ-TREE, and **(b)** MUSCLE and IQ-TREE
Figure S8: Codon-based maximum-likelihood phylogenetic tree of SDR sequences constructed with **(a)** MAFFT and IQ-TREE, and **(b)** MUSCLE and IQ-TREE
Figure S9: Codon-based maximum-likelihood phylogenetic tree of arGST sequences constructed with **(a)** MAFFT and IQ-TREE, and **(b)** MUSCLE and IQ-TREE
Figure S10: Codon-based maximum-likelihood phylogenetic tree of U3GT sequences constructed with **(a)** MAFFT and IQ-TREE, and **(b)** MUSCLE and IQ-TREE
Figure S11: Codon-based maximum-likelihood phylogenetic tree of MYBs sequences constructed with **(a)** MAFFT and IQ-TREE, and **(b)** MUSCLE and IQ-TREE
Figure S12: Codon-based maximum-likelihood phylogenetic tree of bHLH sequences constructed with **(a)** MAFFT and IQ-TREE, and **(b)** MUSCLE and IQ-TREE
Figure S13: Codon-based maximum-likelihood phylogenetic tree of WD40 sequences constructed with **(a)** MAFFT and IQ-TREE, and **(b)** MUSCLE and IQ-TREE
Additional_File6: SRA IDs of species used for expression analysis
Additional_File7: Selected internal controls for expression analysis
Additional_File8: Synteny depth plots of all reference vs target datasets used in the microsynteny analysis,
Additional_File9: (b) WGD detection and dating of *Coriaria nepalensis*.

References

Albalat R, Cañestro C. **2016**. Evolution by gene loss. *Nature Reviews Genetics* **17**: 379–391.

Azevedo-Meleiro CH, Rodriguez-Amaya DB. **2007**. Qualitative and Quantitative Differences in Carotenoid Composition among Cucurbita moschata, Cucurbita maxima, and Cucurbita pepo. *Journal of Agricultural and Food Chemistry* **55**: 4027–4033.

Bernhardt C, Lee MM, Gonzalez A, Zhang F, Lloyd A, Schiefelbein J. **2003**. The bHLH genes GLABRA3 (GL3) andENHANCER OF GLABRA3(EGL3) specify epidermal cell fate in the Arabidopsis root. *Development* **130**: 6431–6439.

Bisognin DA. **2002**. Origin and evolution of cultivated cucurbits. *Ciência Rural* **32**: 715–723.

de Boer HJ, Schaefer H, Thulin M, Renner SS. **2012**. Evolution and loss of long-fringed petals: a case study using a dated phylogeny of the snake gourds, Trichosanthes(Cucurbitaceae). *BMC Evolutionary Biology* **12**: 108.

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

Brockington SF, Walker RH, Glover BJ, Soltis PS, Soltis DE. **2011**. Complex pigment evolution in the Caryophyllales. *New Phytologist* **190**: 854–864.

Brown JW, Walker JF, Smith SA. **2017**. Phyx: phylogenetic tools for unix. *Bioinformatics* **33**: 1886–1888.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. **2009**. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Cazzonelli CI, Pogson BJ. **2010**. Source to sink: regulation of carotenoid biosynthesis in plants. *Trends in Plant Science* **15**: 266–274.

Certini D, Cummins C, Giorgio Serchi F, Yang Y, Nakayama N. **2022**. *The flight of Alsomitra macrocarpa*.

Chalker-Scott L. **1999**. Environmental Significance of Anthocyanins in Plant Stress Responses. *Photochemistry and Photobiology* **70**: 1–9.

Chen H, Zwaenepoel A. **2023**. Inference of Ancient Polyploidy from Genomic Data. In: Van de Peer Y, ed. Polyploidy: Methods and Protocols. New York, NY: Springer US, 3–18.

Chen H, Zwaenepoel A, Van de Peer Y. **2024**. wgd v2: a suite of tools to uncover and date ancient polyploidy and whole-genome duplication. *Bioinformatics* **40**: btae272.

Chittka L, Raine NE. **2006**. Recognition of flowers by pollinators. *Current Opinion in Plant Biology* **9**: 428–435.

Chomicki G, Schaefer H, Renner SS. **2020**. Origin and domestication of Cucurbitaceae crops: insights from phylogenies, genomics and archaeology. *New Phytologist* **226**: 1240–1255.

Choudhary N, Pucker B. **2024**. Conserved amino acid residues and gene expression patterns associated with the substrate preferences of the competing enzymes FLS and DFR. *PLOS ONE* **19**: e0305837.

Christenhusz MJM, Goodwin Z, Bell DG, Martin CA, Royal Botanic Gardens Kew Genome Acquisition Lab, Royal Botanic Garden Edinburgh Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Plant Genome Sizing collective, Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team, Wellcome Sanger Institute Scientific Operations: Sequencing Operations, *et al.* 2024. The genome sequence of the common alder, Alnus glutinosa (L.) Gaertn. (Betulaceae). *Wellcome Open Research* **9**: 570.

Clemente AA, Maciel GM, Siquieroli ACS, Gallis RB de A, Luz JMQ, Sala FC, Pereira LM, Yada RY. **2023**. Nutritional Characterization Based on Vegetation Indices to Detect Anthocyanins, Carotenoids, and Chlorophylls in Mini-Lettuce. *Agronomy* **13**: 1403.

Dainat J, Cannoodt R, Soares A, Ruano DG, Hereñú D, Murray DKD, Davis E, Ugrin I, Crouch K, Soler L, *et al.* 2025. NBISweden/AGAT: AGAT v1.5.1.

Dixon RA, Xie D-Y, Sharma SB. **2005**. Proanthocyanidins – a final frontier in flavonoid research? *New Phytologist* **165**: 9–28.

Edgar RC. **2022**. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications* **13**: 6968.

Eisen JA. **1998**. Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. *Genome Research* **8**: 163–167.

Fu A, Wang Q, Mu J, Ma L, Wen C, Zhao X, Gao L, Li J, Shi K, Wang Y, *et al.* 2021a. Combined genomic, transcriptomic, and metabolomic analyses provide insights into chayote (Sechium edule) evolution and fruit development. *Horticulture Research* **8**: 35.

Fu A, Wang Q, Mu J, Ma L, Wen C, Zhao X, Gao L, Li J, Shi K, Wang Y, *et al.* 2021b. Combined genomic, transcriptomic, and metabolomic analyses provide insights into chayote (Sechium edule) evolution and fruit development. *Horticulture Research* **8**: 35.

Garcia JE, Shrestha M, Howard SR, Petersen P, Dyer AG. **2019**. Signal or cue: the role of structural colors in flower pollination. *Current Zoology* **65**: 467–481.

Glagoleva AY, Shoeva OY, Khlestkina EK. **2020**. Melanin Pigment in Plants: Current Knowledge and Future Perspectives. *Frontiers in Plant Science* **11**.

Gonzalez A, Mendenhall J, Huo Y, Lloyd A. **2009**. TTG1 complex MYBs, MYB5 and TT2, control outer seed coat differentiation. *Developmental Biology* **325**: 412–421.

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

Gould KS. **2004**. Nature's Swiss Army Knife: The Diverse Protective Roles of Anthocyanins in Leaves. *Journal of Biomedicine and Biotechnology* **2004**: 314–320.

Grünig N, Horz JM, Pucker B. **2024**. Diversity and Ecological Functions of Anthocyanins.

Guo J, Xu W, Hu Y, Huang J, Zhao Y, Zhang L, Huang C-H, Ma H. **2020**. Phylotranscriptomics in Cucurbitaceae Reveal Multiple Whole-Genome Duplications and Key Morphological and Molecular Innovations. *Molecular Plant* **13**: 1117–1133.

Henry Nicholas Ridley. **1930**. *The Dispersal Of Plants Throughout The World*.

Hichri I, Barrieu F, Bogs J, Kappel C, Delrot S, Lauvergeat V. **2011**. Recent advances in the transcriptional regulation of the flavonoid biosynthetic pathway. *Journal of Experimental Botany* **62**: 2465–2483.

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, *et al.* 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–467.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. **2017**. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**: 587–589.

Katoh K, Rozewicki J, Yamada KD. **2019**. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* **20**: 1160–1166.

Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. **2016**. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research* **44**: e89.

Khatun N, Jones A, Rahe A, Choudhary N, Pucker B. **2025**. Evolutionary Rewiring of Anthocyanin Biosynthesis Pathway in Poaceae. : 2025.09.21.677584.

Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. **2022**. TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution* **39**: msac174.

Lee Y, Yoon HR, Paik YS, Liu JR, Chung W, Choi G. **2005**. Reciprocal regulation ofArabidopsis UGT78D2 and BANYULS is critical for regulation of the metabolic flux of anthocyanidins to condensed tannins in developing seed coats. *Journal of Plant Biology* **48**: 356–370.

Lepiniec L, Debeaujon I, Routaboul J-M, Baudry A, Pourcel L, Nesi N, Caboche M. **2006**. GENETICS AND BIOCHEMISTRY OF SEED FLAVONOIDS. *Annual Review of Plant Biology* **57**: 405–430.

Lev-Yadun S. **2016**. Visual Signaling by Plants to Animals via Color. In: Lev-Yadun S, ed. Defensive (anti-herbivory) Coloration in Land Plants. Cham: Springer International Publishing, 25–26.

Lev-Yadun S, Dafni A, Flaishman MA, Inbar M, Izhaki I, Katzir G, Ne'eman G. **2004**. Plant coloration undermines herbivorous insect camouflage. *BioEssays* **26**: 1126–1130.

Li L, Chen X, Fang D, Dong S, Guo X, Li N, Campos-Dominguez L, Wang W, Liu Y, Lang X, *et al.* 2022. Genomes shed light on the evolution of Begonia, a mega-diverse genus. *New Phytologist* **234**: 295–310.

Lin D-L, Wen Y-Y, Wang Z-Q, Li J-W, Wang C-W, Hidayat A, Yu L-Y, Huang L-Q, Jin X-H. **2025**. Elucidation of the phylogeny of Cucurbitaceae, particularly Trichosanthes, based on plastome data and nuclear single-copy genes. *BMC Plant Biology* **25**: 929.

Lu J, Breitwieser FP, Thielen P, Salzberg SL. **2017**. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science* **3**: e104.

Mabry TJ, Mabry TJ. **1964**. The beta-cyanins, a new class of red-violet pigments, and their phylogenetic significance.

Marin-Recinos MF, Pucker B. **2024**. Genetic factors explaining anthocyanin pigmentation differences. *BMC Plant Biology* **24**: 627.

Marrano A, Britton M, Zaini PA, Zimin AV, Workman RE, Puiu D, Bianco L, Pierro EAD, Allen BJ, Chakraborty S, *et al.* 2020. High-quality chromosome-scale assembly of the walnut (Juglans regia L.) reference genome. *GigaScience* **9**: giaa050.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. **2020**. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**: 1530–1534.

Montero-Pau J, Blanca J, Bombarely A, Ziarsolo P, Esteras C, Martí-Gómez C, Ferriol M, Gómez P, Jamilena M, Mueller L, *et al.* 2018. De novo assembly of the zucchini genome reveals a whole-genome duplication associated with the origin of the Cucurbita genus. *Plant Biotechnology Journal* **16**: 1161–1171.

Mueller LA, Goodman CD, Silady RA, Walbot V. **2000**. AN9, a Petunia Glutathione S-Transferase Required for Anthocyanin Sequestration, Is a Flavonoid-Binding Protein1. *Plant Physiology* **123**: 1561–1570.

Olson MV. **1999**. When Less Is More: Gene Loss as an Engine of Evolutionary Change. *The American Journal of Human Genetics* **64**: 18–23.

Ooe E, Ogawa K, Horiuchi T, Tada H, Murase H, Tsuruma K, Shimazawa M, Hara H. **2016**. Analysis and characterization of anthocyanins and carotenoids in Japanese blue tomato. *Bioscience, Biotechnology, and Biochemistry* **80**: 341–349.

Pietrini F, Iannelli MA, Massacci A. **2002**. Anthocyanin accumulation in the illuminated surface of maize leaves enhances protection from photo-inhibitory risks at low temperature, without further limitation to photosynthesis. *Plant, Cell & Environment* **25**: 1251–1259.

Poustka F, Irani NG, Feller A, Lu Y, Pourcel L, Frame K, Grotewold E. **2007**. A Trafficking Pathway for Anthocyanins Overlaps with the Endoplasmic Reticulum-to-Vacuole Protein-Sorting Route in Arabidopsis and Contributes to the Formation of Vacuolar Inclusions. *Plant Physiology* **145**: 1323–1335.

Price MN, Dehal PS, Arkin AP. **2010**. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**: e9490.

Pucker B. **2022**. Automatic identification and annotation of MYB gene family members in plants. *BMC Genomics* **23**: 220.

Pucker B. **2024**. Functional Annotation – How to Tackle the Bottleneck in Plant Genomics.

Pucker B, Iorizzo M. **2023**. Apiaceae FNS I originated from F3H through tandem gene duplication. *PLOS ONE* **18**: e0280155.

Pucker B, Walker-Hale N, Dzurlic J, Yim WC, Cushman JC, Crum A, Yang Y, Brockington SF. **2024**. Multiple mechanisms explain loss of anthocyanins from betalain-pigmented Caryophyllales, including repeated wholesale loss of a key anthocyanidin synthesis enzyme. *New Phytologist* **241**: 471–489.

Rempel A, Choudhary N, Pucker B. **2023**. KIPEs3: Automatic annotation of biosynthesis pathways. *PLOS ONE* **18**: e0294342.

Schaefer H, Heibl C, Renner SS. **2008**. Gourds afloat: a dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. *Proceedings of the Royal Society B: Biological Sciences* **276**: 843–851.

Schaefer H, Renner SS. **2011a**. Cucurbitaceae. In: Kubitzki K, ed. Flowering Plants. Eudicots: Sapindales, Cucurbitales, Myrtaceae. Berlin, Heidelberg: Springer, 112–174.

Schaefer H, Renner SS. **2011b**. Phylogenetic relationships in the order Cucurbitales and a new classification of the gourd family (Cucurbitaceae). *TAXON* **60**: 122–138.

Schaefer H, Renner SS. **2011c**. Cucurbitaceae. In: Kubitzki K, ed. Flowering Plants. Eudicots: Sapindales, Cucurbitales, Myrtaceae. Berlin, Heidelberg: Springer, 112–174.

Seo T-K. **2008**. Calculating Bootstrap Probabilities of Phylogeny Using Multilocus Sequence Data. *Molecular Biology and Evolution* **25**: 960–971.

Sharma V, Hecker N, Roscito JG, Foerster L, Langer BE, Hiller M. **2018**. A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nature Communications* **9**: 1215.

Sheehan H, Feng T, Walker-Hale N, Lopez-Nieves S, Pucker B, Guo R, Yim WC, Badgami R, Timoneda A, Zhao L, *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.

Smith SD, Goldberg EE. **2015**. Tempo and mode of flower color evolution. *American Journal of Botany* **102**: 1014–1025.

Smith SD, Rausher MD. **2011**. Gene Loss and Parallel Evolution Contribute to Species Difference in Flower Color. *Molecular Biology and Evolution* **28**: 2799–2810.

Spelt C, Quattrocchio F, Mol JNM, Koes R. **2000**. anthocyanin1 of Petunia Encodes a Basic Helix-Loop-Helix Protein That Directly Activates Transcription of Structural Anthocyanin Genes. *The Plant Cell* **12**: 1619–1631.

Spelt C, Quattrocchio F, Mol J, Koes R. **2002**. ANTHOCYANIN1 of Petunia Controls Pigment Synthesis, Vacuolar pH, and Seed Coat Development by Genetically Distinct Mechanisms. *The Plant Cell* **14**: 2121–2135.

Stafford HA. **1994**. Anthocyanins and betalains: evolution of the mutually exclusive pathways. *Plant Science* **101**: 91–98.

Steyn WJ, Wand SJE, Holcroft DM, Jacobs G. **2002**. Anthocyanins in vegetative tissues: a proposed unified function in photoprotection. *New Phytologist* **155**: 349–361.

Sun Y, Li H, Huang J-R. **2012**. Arabidopsis TT19 Functions as a Carrier to Transport Anthocyanin from the Cytosol to Tonoplasts. *Molecular Plant* **5**: 387–400.

Sun T, Rao S, Zhou X, Li L. **2022**. Plant carotenoids: recent advances and future perspectives. *Molecular Horticulture* **2**: 3.

Tang H, Krishnakumar V, Li J. **2015**. jcvi: JCVI utility libraries.

Tegenfeldt F, Kuznetsov D, Manni M, Berkeley M, Zdobnov EM, Kriventseva EV. **2025**. OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Research* **53**: D516–D522.

The Angiosperm Phylogeny Group, Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS, *et al.* 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* **181**: 1–20.

Thoben C, Pucker B. **2023**. Automatic annotation of the bHLH gene family in plants. *BMC Genomics* **24**: 780.

Thompson WR, Meinwald J, Aneshansley D, Eisner T. **1972**. Flavonols: Pigments Responsible for Ultraviolet Absorption in Nectar Guide of Flower. *Science* **177**: 528–530.

Timoneda A, Feng T, Sheehan H, Walker-Hale N, Pucker B, Lopez-Nieves S, Guo R, Brockington S. **2019**. The evolution of betalain biosynthesis in Caryophyllales. *New Phytologist* **224**: 71–85.

Triantaphylidès C, Havaux M. **2009**. Singlet oxygen in plants: production, detoxification and signaling. *Trends in Plant Science* **14**: 219–228.

Van Eijk JP, Nieuwhof M, Van Keulen HA, Keijzer P. **1987**. Flower colour analyses in tulip (Tulipa L.). The occurrence of carotenoids and flavonoids in tulip tepals. *Euphytica* **36**: 855–862.

Walker AR, Davison PA, Bolognesi-Winfield AC, James CM, Srinivasan N, Blundell TL, Esch JJ, Marks MD, Gray JC. **1999**. The TRANSPARENT TESTA GLABRA1 Locus, Which Regulates Trichome Differentiation and Anthocyanin Biosynthesis in Arabidopsis, Encodes a WD40 Repeat Protein. *The Plant Cell* **11**: 1337–1349.

Wang J, Yuan M, Feng Y, Zhang Y, Bao S, Hao Y, Ding Y, Gao X, Yu Z, Xu Q, *et al.* 2022. A common whole-genome paleotetraploidization in Cucurbitales. *Plant Physiology* **190**: 2430–2448.

Winkel-Shirley B. **2001**. It Takes a Garden. How Work on Diverse Plant Species Has Contributed to an Understanding of Flavonoid Metabolism. *Plant Physiology* **127**: 1399–1404.

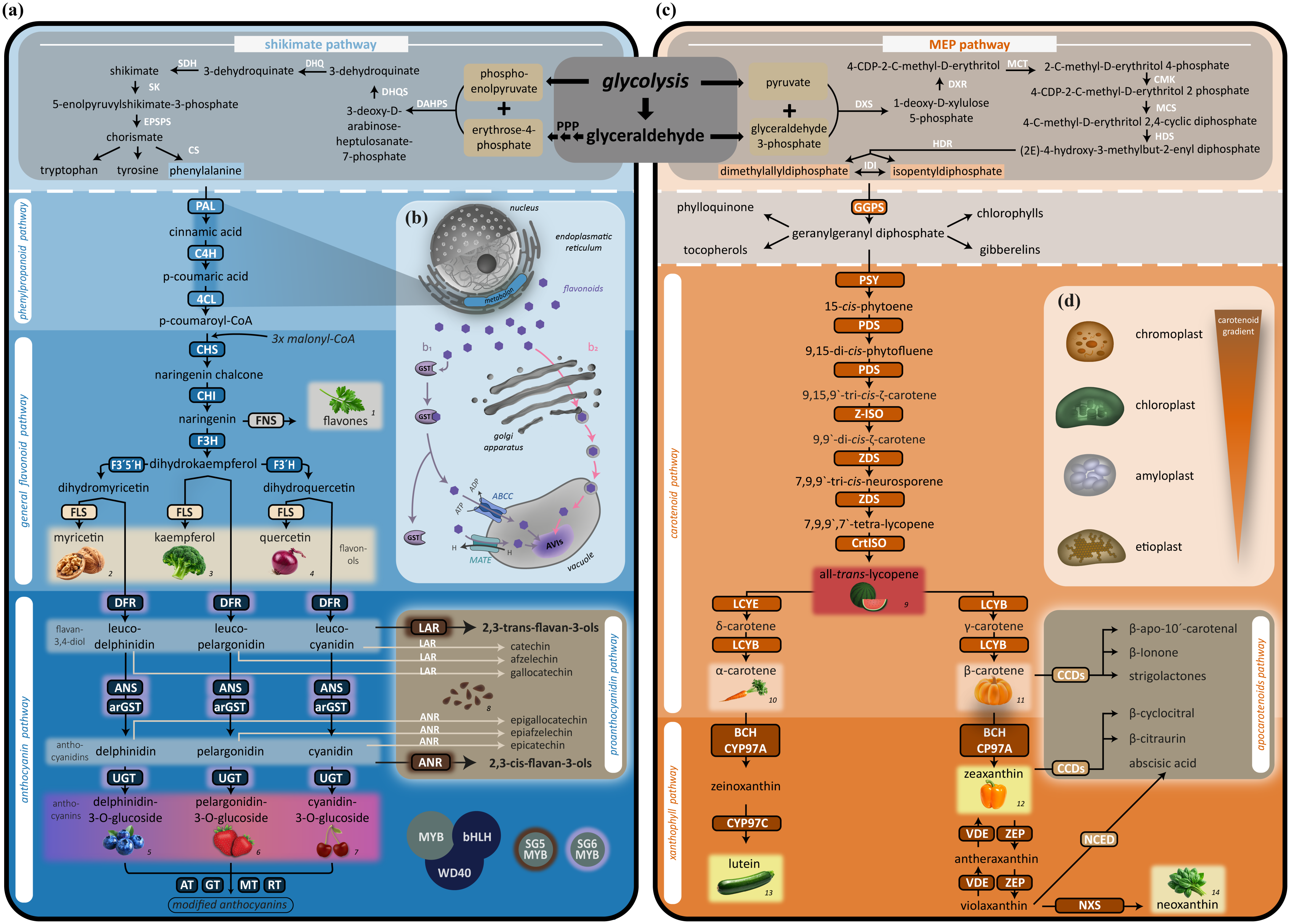
Wolff K, Pucker B. **2025**. Dark side of anthocyanin pigmentation. *Plant Biology* **27**: 935–947.

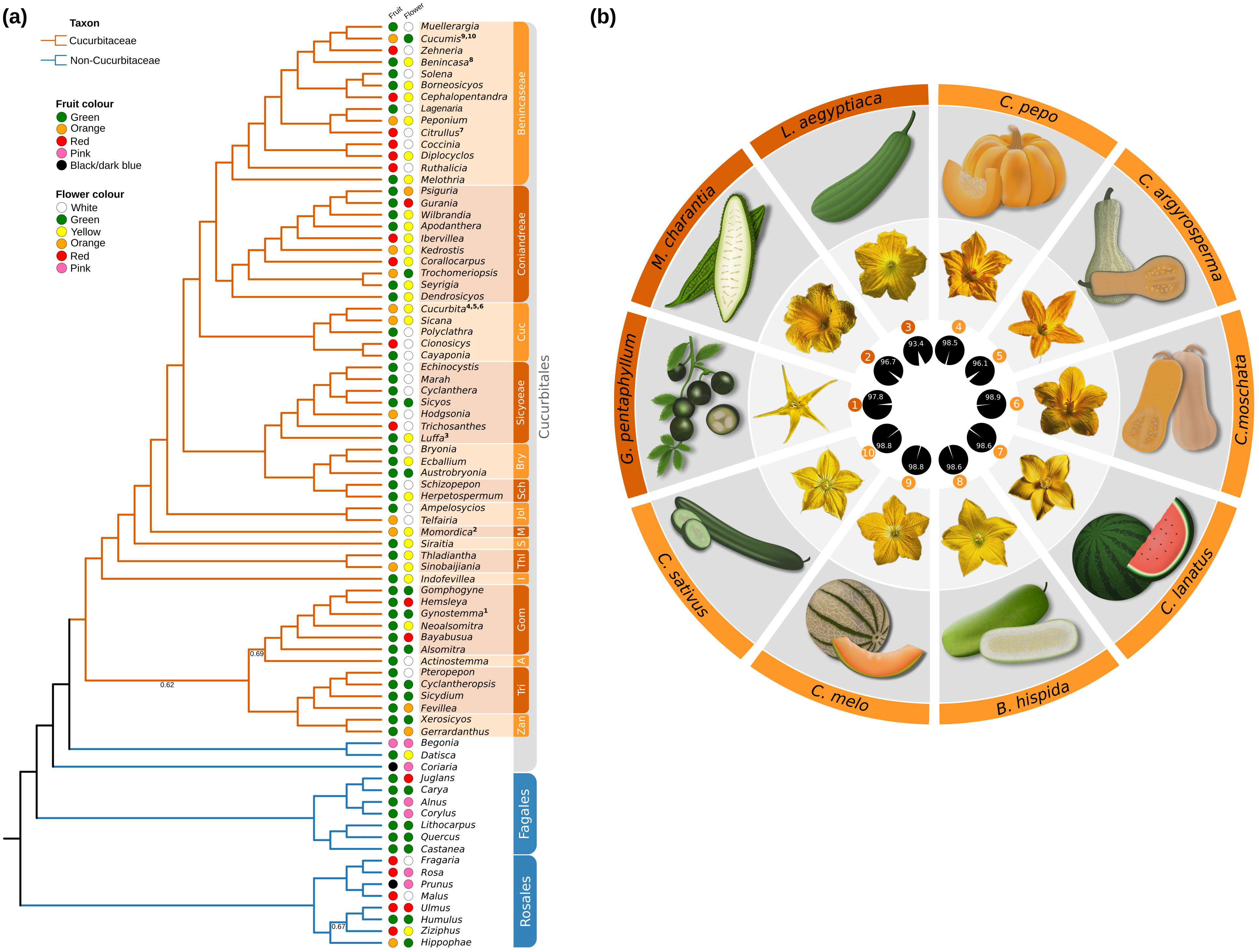
Wood DE, Lu J, Langmead B. **2019**. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**: 257.

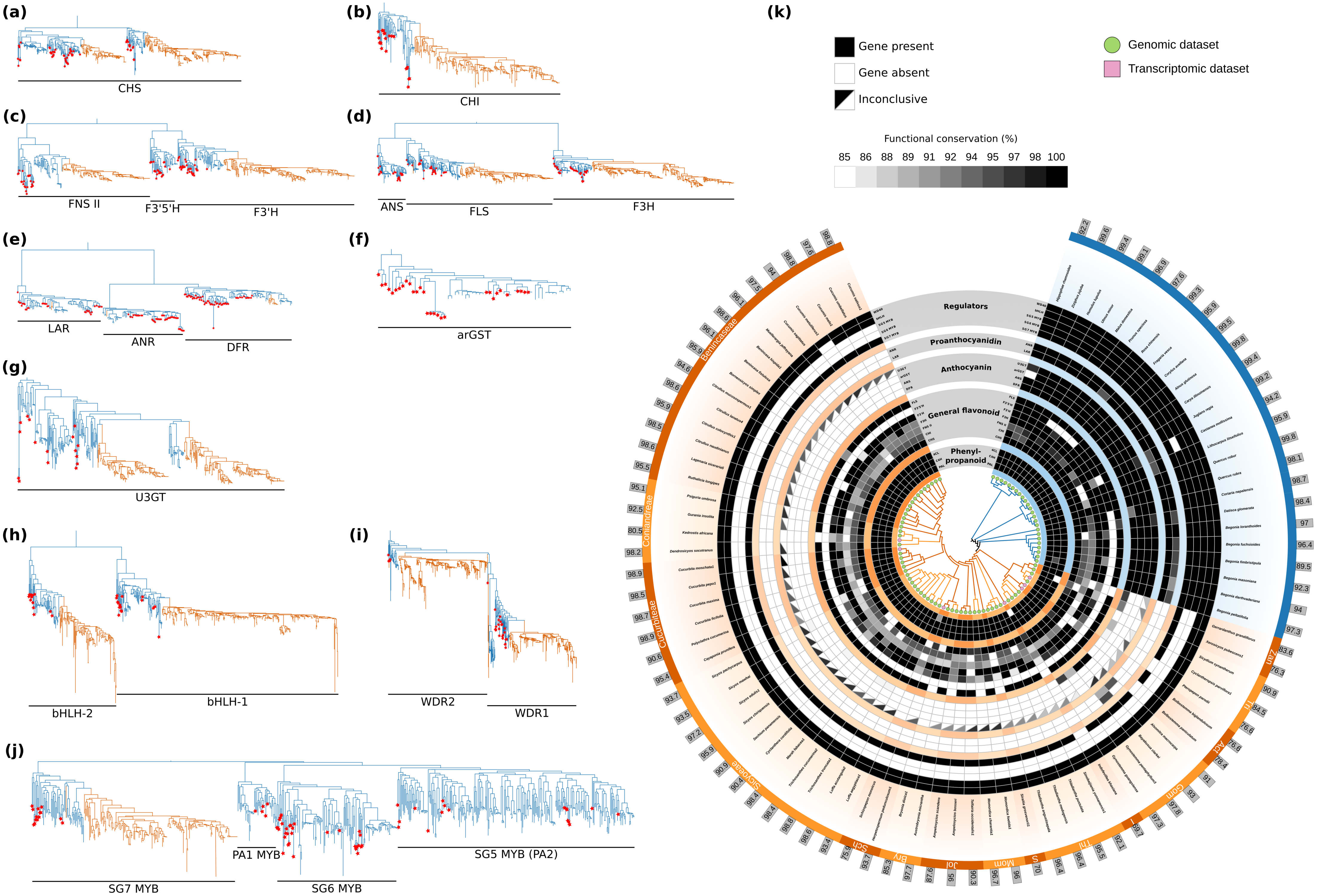
Xie D-Y, Sharma SB, Paiva NL, Ferreira D, Dixon RA. **2003**. Role of Anthocyanidin Reductase, Encoded by BANYULS in Plant Flavonoid Biosynthesis. *Science* **299**: 396–399.

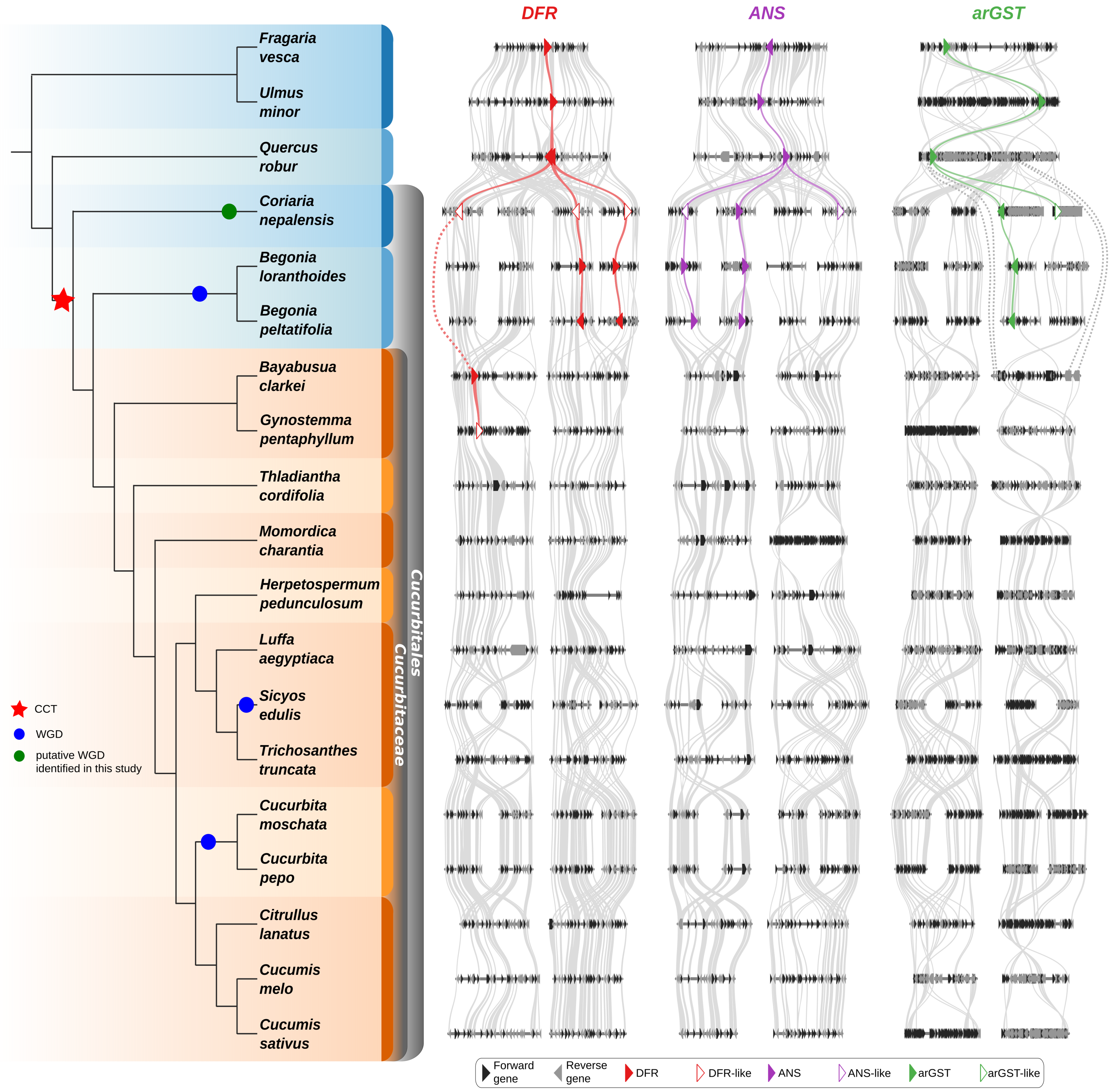
Xie D, Xu Y, Wang J, Liu W, Zhou Q, Luo S, Huang W, He X, Li Q, Peng Q, *et al.* 2019. The wax gourd genomes offer insights into the genetic diversity and ancestral cucurbit karyotype. *Nature Communications* **10**: 5158.

Zhang C, Mirarab S. **2022**. Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees. *Molecular Biology and Evolution* **39**: msac215.


(a)
shikimate pathway
shikimate
SDH
3-dehydroquinate
DHQ
3-dehydroquinate
DHQS
SK
5-enolpyruvylshikimate-3-phosphate
EPSPS
chorismate
CS
tryptophan
tyrosine
phenylalanine
DAHPS
3-deoxy-D-arabinose-heptulosanate-7-phosphate
phospho-enolpyruvate
erythrose-4-phosphate
glycolysis
PPP
glyceraldehyde
phenylpropanoid pathway
PAL
cinnamic acid
C4H
p-coumaric acid
4CL
p-coumaroyl-CoA
3x malonyl-CoA
general flavonoid pathway
CHS
naringenin chalcone
CHI
naringenin
FNS
flavones
1
F3H
F3′5′H
dihydrokaempferol
F3′H
dihydromyricetin
dihydroquercetin
FLS
FLS
FLS
myricetin
kaempferol
quercetin
flavon-ols
2
3
4
anthocyanin pathway
DFR
DFR
DFR
flavan-3,4-diol
leuco-delphinidin
leuco-pelargonidin
leuco-cyanidin
LAR
2,3-trans-flavan-3-ols
LAR
LAR
LAR
catechin
afzelechin
gallocatechin
8
ANS
arGST
ANS
arGST
ANS
arGST
ANR
ANR
ANR
epigallocatechin
epiafzelechin
epicatechin
antho-cyanidins
delphinidin
pelargonidin
cyanidin
ANR
2,3-cis-flavan-3-ols
proanthocyanidin pathway
UGT
UGT
UGT
antho-cyanins
delphinidin-3-O-glucoside
pelargonidin-3-O-glucoside
cyanidin-3-O-glucoside
5
6
7
MYB
bHLH
WD40
SG5 MYB
SG6 MYB
AT
GT
MT
RT
modified anthocyanins
(b)
nucleus
endoplasmatic reticulum
metabolon
flavonoids
b₁
b₂
GST
GST
golgi apparatus
ADP
ATP
ABCC
GST
H
H
MATE
AVIs
vacuole
(c)
MEP pathway
pyruvate
glyceraldehyde 3-phosphate
DXS
1-deoxy-D-xylulose 5-phosphate
DXR
4-CDP-2-C-methyl-D-erythritol
MCT
2-C-methyl-D-erythritol 4-phosphate
CMK
4-CDP-2-C-methyl-D-erythritol 2 phosphate
MCS
4-C-methyl-D-erythritol 2,4-cyclic diphosphate
HDS
(2E)-4-hydroxy-3-methylbut-2-enyl diphosphate
HDR
dimethylallyldiphosphate
IDI
isopentyldiphosphate
GGPS
phylloquinone
tocopherols
geranylgeranyl diphosphate
chlorophylls
gibberelins
carotenoid pathway
PSY
15-cis-phytoene
PDS
9,15-di-cis-phytofluene
PDS
9,15,9`-tri-cis-ζ-carotene
Z-ISO
9,9`-di-cis-ζ-carotene
ZDS
7,9,9`-tri-cis-neurosporene
ZDS
7,9,9`,7`-tetra-lycopene
CrtISO
all-trans-lycopene
9
LCYE
δ-carotene
LCYB
α-carotene
10
LCYB
γ-carotene
LCYB
β-carotene
11
CCDs
β-apo-10´-carotenal
β-Ionone
strigolactones
β-cyclocitral
β-citraurin
abscisic acid
CCDs
apocarotenoids pathway
xanthophyll pathway
BCH CYP97A
zeinoxanthin
CYP97C
lutein
13
BCH CP97A
zeaxanthin
12
VDE
ZEP
antheraxanthin
VDE
ZEP
violaxanthin
NCED
NXS
neoxanthin
14
(d)
chromoplast
chloroplast
amyloplast
etioplast
carotenoid gradient


(a)
Taxon
Cucurbitaceae
Non-Cucurbitaceae
Fruit colour
Green
Orange
Red
Pink
Black/dark blue
Flower colour
White
Green
Yellow
Orange
Red
Pink
Fruit
Flower
Muellerargia
Cucumis[9,10]
Zehneria
Benincasa[8]
Solena
Borneosicyos
Cephalopentandra
Lagenaria
Peponium
Citrullus[7]
Coccinia
Diplocyclos
Ruthalicia
Melothria
Psiguria
Gurania
Wilbrandia
Apodanthera
Ibervillea
Kedrostis
Corallocarpus
Trochomeriopsis
Seyrigia
Dendrosicyos
Cucurbita[4,5,6]
Sicana
Polyclathra
Cionosicys
Cayaponia
Echinocystis
Marah
Cyclanthera
Sicyos
Hodgsonia
Trichosanthes
Luffa[3]
Bryonia
Ecballium
Austrobryonia
Schizopepon
Herpetospermum
Ampelosycios
Telfairia
Momordica[2]
Siraitia
Thladiantha
Sinobaijiania
Indofevillea
Gomphogyne
Hemsleya
Gynostemma[1]
Neoalsomitra
Bayabusua
Alsomitra
Actinostemma
Pteropepon
Cyclantheropsis
Sicydium
Fevillea
Xerosicyos
Gerrardanthus
Begonia
Datisca
Coriaria
Juglans
Carya
Alnus
Corylus
Lithocarpus
Quercus
Castanea
Fragaria
Rosa
Prunus
Malus
Ulmus
Humulus
Ziziphus
Hippophae
Benincaseae
Coniandreae
Cuc
Sicyoeae
Bry
Sch
Jol
M
S
Thl
I
Gom
A
Tri
Zan
Cucurbitales
Fagales
Rosales
0.69
0.62
0.67
(b)
L. aegyptiaca
C. pepo
C. argyrosperma
C.moschata
C. lanatus
B. hispida
C. melo
C. sativus
G. pentaphyllum
M. charantia
1
97.8
2
96.7
3
93.4
4
98.5
5
96.1
6
98.9
7
98.6
8
98.6
9
98.8
10
98.8


(a)
CHS
(b)
CHI
(c)
FNS II
F3'5'H
F3'H
(d)
ANS
FLS
F3H
(e)
LAR
ANR
DFR
(f)
arGST
(g)
U3GT
(h)
bHLH-2
bHLH-1
(i)
WDR2
WDR1
(j)
SG7 MYB
PA1 MYB
SG6 MYB
SG5 MYB (PA2)
(k)
Gene present
Gene absent
Inconclusive
Genomic dataset
Transcriptomic dataset
Functional conservation (%)
85 86 88 89 91 92 94 95 97 98 100
Regulators
Proanthocyanidin
Anthocyanin
General flavonoid
Phenyl-propanoid
WD40
bHLH
SG5 MYB
SG6 MYB
SG7 MYB
ANR
LAR
U3GT
arGST
ANS
DFR
FLS
F3'5'H
F3'H
F3H
FNS II
CHI
CHS
4CL
C4H
PAL
Benincaseae
Coniandreae
Cucurbiteae
Sicyoeae
Sch
Bry
Jol
Mom
S
Thl
Gom
Acy
Tri
Zan


DFR
ANS
arGST
Fragaria vesca
Ulmus minor
Quercus robur
Coriaria nepalensis
Begonia loranthoides
Begonia peltatifolia
Bayabusua clarkei
Gynostemma pentaphyllum
Thladiantha cordifolia
Momordica charantia
Herpetospermum pedunculosum
Luffa aegyptiaca
Sicyos edulis
Trichosanthes truncata
Cucurbita moschata
Cucurbita pepo
Citrullus lanatus
Cucumis melo
Cucumis sativus
Cucurbitales
Cucurbitaceae
CCT
WGD
putative WGD identified in this study
Forward gene
Reverse gene
DFR
DFR-like
ANS
ANS-like
arGST
arGST-like

(a)
(b)
(c)
(d)
Flavonoid biosynthesis
Anthocyanin biosynthesis
Proanthocyanidin
Carotene biosynthesis
Xanthophyll biosynthesis
CHS
CHI¹
F3H²
F3′H
FLS
DFR³
ANS
arGST⁴
LAR
ANR
Z – ISO
ZDS
CrtISO
LCYE
LCYB
CYP97A
BCH
VDE
ZEP
NXS
n=985
n=639
n=588
n=115
n=563
n=493
n=637
n=564
n=65
n=215
n=60
n=2181
n=991
n=190
n=359
n=265
n=4584
n=4326
***
*
Malus domestica
Fragaria vesca
Ulmus minor
Hippophae rhamnoides
Quercus robur
Castanea mollissima
Juglans regia
Carya illinoinensis
Gynostemma pentaphyllum
Momordica charantia
Luffa cylindrica
Cucumis sativus
Citrullus lanatus
Benincasa hispida
Cucurbita pepo
Cucurbita moschata
Outgroup
Cucurbitaceae