Lineage-specific gene radiations underlie the evolution of novel betalain pigmentation in Caryophyllales

Samuel F. Brockington[1]*, Ya Yang[2]*, Fernando Gandia-Herrero[3], Sarah Covshoff[1], Julian M. Hibberd[1], Rowan F. Sage[4], Gane K. S. Wong[5,6,7], Michael J. Moore[8] and Stephen A. Smith[2]

[1]Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, UK; [2]Department of Ecology & Evolutionary Biology, University of Michigan, 830 North University Avenue, Ann Arbor, MI 48109-1048, USA; [3]Departamento de Bioquímica y Biología Molecular A, Unidad Docente de Biología, Facultad de Veterinaria, Regional Campus of International Excellence 'Campus Mare Nostrum', Universidad de Murcia, E-30100 Espinardo, Murcia, Spain; [4]Department of Ecology and Evolutionary Biology, University of Toronto, 25 Willcocks Street, Toronto, ON M5S 3B2, Canada; [5]Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada; [6]Department of Medicine, University of Alberta, Edmonton, AB T6G 2E1, Canada; [7]BGI-Shenzhen, Beishan Industrial Zone, Yantian District Shenzhen 518083, China; [8]Department of Biology, Oberlin College, 119 Woodland St, Oberlin, OH 44074-1097, USA

Author for correspondence:
Samuel F. Brockington
Tel: +44 01223 333900
Email: sb771@cam.ac.uk

Received: *16 February 2015*
Accepted: *3 April 2015*

New Phytologist (2015) **207:** 1170–1180
doi: 10.1111/nph.13441

Key words: anthocyanin, betalains, Caryophyllales, lineage-specific genes, pigmentation, taxonomically restricted genes.

Summary

- Betalain pigments are unique to the Caryophyllales and structurally and biosynthetically distinct from anthocyanins. Two key enzymes within the betalain synthesis pathway have been identified: 4,5-dioxygenase (DODA) that catalyzes the formation of betalamic acid and CYP76AD1, a cytochrome P450 gene that catalyzes the formation of cyclo-DOPA.
- We performed phylogenetic analyses to reveal the evolutionary history of the DODA and CYP76AD1 lineages and in the context of an ancestral reconstruction of pigment states we explored the evolution of these genes in relation to the complex evolution of pigments in Caryophylalles.
- Duplications within the CYP76AD1 and DODA lineages arose just before the origin of betalain pigmentation in the core Caryophyllales. The duplications gave rise to DODA-α and CYP76AD1-α isoforms that appear specific to betalain synthesis. Both betalain-specific isoforms were then lost or downregulated in the anthocyanic Molluginaceae and Caryophyllaceae.
- Our findings suggest a single origin of the betalain synthesis pathway, with neofunctionalization following gene duplications in the CYP76AD1 and DODA lineages. Loss of DODA-α and CYP76AD1-α in anthocyanic taxa suggests that betalain pigmentation has been lost twice in Caryophyllales, and exclusion of betalain pigments from anthocyanic taxa is mediated through gene loss or downregulation. [Correction added after online publication 13 May 2015: in the last two paragraphs of the Summary the gene name CYP761A was changed to CYP76AD1.]

Introduction

Comparative genomic analysis has revealed a significant fraction of genes that occur only in defined organismal lineages, and which have been variously termed orphan genes, taxonomically restricted or lineage-specific genes (Khalturin *et al.*, 2009). Several studies have demonstrated that lineage-specific gene radiations can contribute to unique evolutionary changes and novel phenotypic adaptation (Khalturin *et al.*, 2008; Fry *et al.*, 2010; Simola *et al.*, 2013). In a recent phylotranscriptomic analysis of the flowering plant order Caryophyllales sensu APG III (Bremer *et al.*, 2009) we identified gene lineages that exhibit substantially higher rates of gene duplication than in non-Caryophyllales outgroups (Yang *et al.*, 2015). The Caryophyllales are well recognized for extraordinary levels of morphological and physiological adaptation, and in this context, we proposed that these clade-specific, highly duplicated gene lineages might contain important novel genes associated with unusual evolutionary adaptations. Consistent with this hypothesis, we found that two of the most highly duplicated gene lineages among Caryophyllales transcriptomes encode a cytochrome P450 gene and a 4,5-dioxygenase gene, respectively (Yang *et al.*, 2015), that had previously been functionally implicated in the synthesis of betalains, a group of pigments unique to core Caryophyllales.

In the core Caryophyllales, betalains can largely substitute for the otherwise ubiquitous anthocyanins (Bischoff, 1876; Mabry, 1964), which are the dominant form of pigmentation across land plants (Campanella *et al.*, 2014) (see Fig. 1). Betalains are water-soluble, possess high antioxidant and free radical scavenging activities (Escribano *et al.*, 1998; Cai *et al.*, 2003; Wu *et al.*, 2006), exhibit preventative properties with respect to several types of cancer (Lu *et al.*, 2009; Khan *et al.*, 2012; Krajka-Kuźniak

*These authors contributed equally to this work.
The copyright line for this article was changed on 19 August 2015 after original online publication.

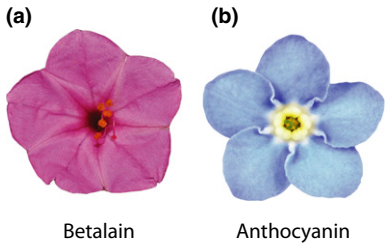

Fig. 1 (a) Pink hue of a betalain-pigmented flower (*Mirabilis jalapa*). (b) Blue anthocyanins, a color that is not obtainable with betalains (*Myosotis* sp.).

et al., 2012). Given the clear health benefits of betalains (Gandía-Herrero *et al.*, 2013), there is interest in expressing the betalain synthesis pathway in the anthocyanic background of food crops (Harris *et al.*, 2012; Hatlestad *et al.*, 2012). Understanding the restricted distribution of betalains within Caryophyllales may therefore have implications for human health and nutrition (Gandía-Herrero *et al.*, 2013). Although some of the genetic components necessary for betalain pigmentation are suggested to be present in anthocyanic plants outside of the Caryophyllales, thereby minimizing the steps needed to express the pathway in heterologous species, other elements appear to be Caryophyllales-specific (Harris *et al.*, 2012). In this context, a clear understanding of the origin and evolution of the genetic components comprising the betalain synthesis pathway is valuable.

Betalains are structurally and biosynthetically distinct from the more common anthocyanin pigments, with the former derived from tyrosine, and the latter from phenylalanine (Clement & Mabry, 1996). Synthesis of betalains is proposed to require three enzyme-mediated steps (Hatlestad *et al.*, 2012) (see Fig. 2). The enzyme or enzymes responsible for the first step are currently unknown but are possibly tyrosinase-like in action (Christinet *et al.*, 2004), converting tyrosine to L-3,4-dihydroxyphenylalanine (L-DOPA). L-DOPA is then a key substrate in the formation of 4,5-seco-DOPA, which spontaneously cyclizes to betalamic acid (Christinet *et al.*, 2004), and is also a substrate in the formation of cyclo-DOPA (Tanaka *et al.*, 2008). Betalamic acid and cyclo-DOPA in turn spontaneously condense to form red betanidin pigments, whereas betalamic acid can condense with additional amine- or amino-groups to form yellow betaxanthin pigments. The genes contained within our highly duplicated gene lineages encode proteins that catalyze two of these key steps: (1) 4,5-dioxygenase (DODA) catalyzes the formation of betalamic acid (Christinet *et al.*, 2004; Sasaki *et al.*, 2009; Gandía-Herrero & García-Carmona, 2012); and (2) CYP76AD1, a cytochrome P450 gene, catalyzes the formation of cyclo-DOPA (Hatlestad *et al.*, 2012).

In the betalain-producing species of the Caryophyllales, anthocyanins have never been detected (Bate-Smith & Lerner, 1954; Mabry, 1964) and, conversely, the anthocyanic lineages within Caryophyllales are not known to produce betalains (Clement & Mabry, 1996). On the basis of these data, it has been proposed

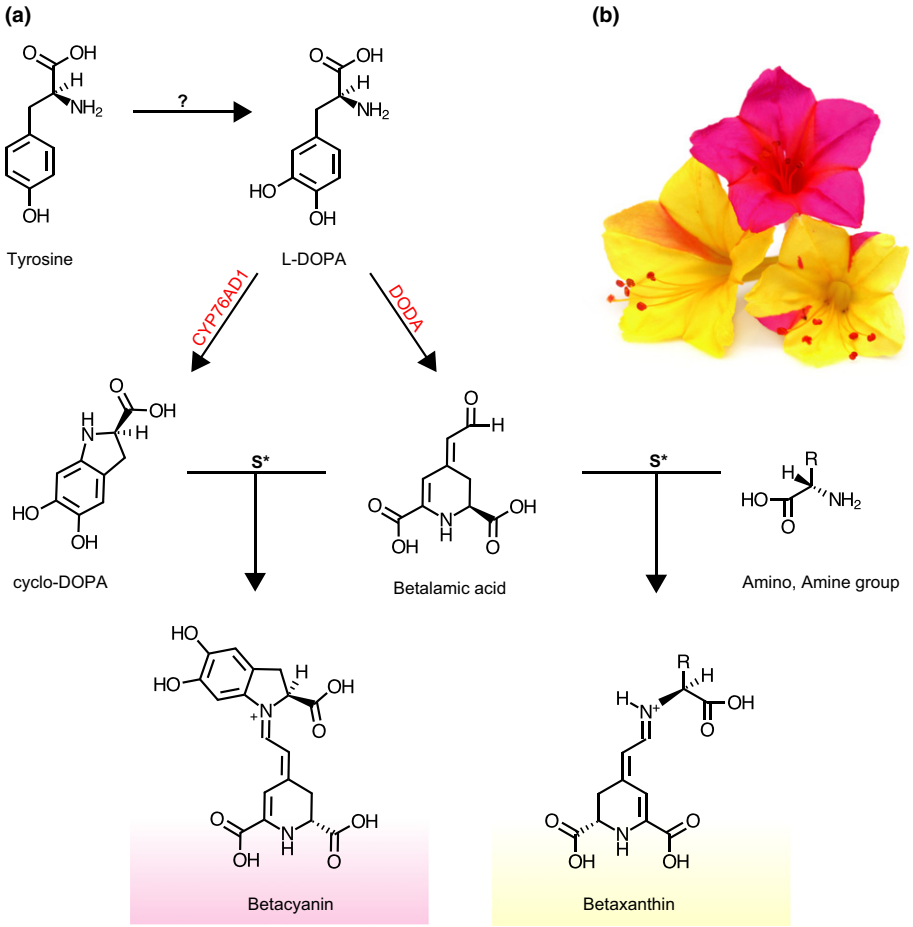

Fig. 2 (a) Outline of the betalain biosynthetic pathway with the key enzymes CYP76AD1 and 4,5-dioxygenase (DODA) marked in red. CYP76AD1 catalyses the conversion of L-3,4-dihydroxyphenylalanine (L-DOPA) into cyclo-DOPA, whereas DODA catalyzes the conversion of L-DOPA to betalamic acid. Betalamic acid spontaneously cyclizes with cyclo-DOPA to give red-pigmented betanidins, whereas betalamic acid can conjugate with amino and amine groups to generate yellow betaxanthins. (b) Variegated forms of *Mirabilis jalapa*, with yellow forms in which betaxanthins are the dominant pigment and pink forms in which betacyanins are dominant.

that anthocyanins and betalains are mutually exclusive (Stafford, 1994; Clement & Mabry, 1996). Furthermore, the occurrence of anthocyanins and betalains exhibit interesting patterns of homoplasy such that anthocyanic lineages are either nested within (e.g. Molluginaceae s.s. and some of its recent segregates such as *Kewa* Kewaceae; previously included in *Hypertelis*; Christenhusz *et al.*, 2014) or sister to (e.g. Caryophyllaceae) betalain-pigmented lineages (Brockington *et al.*, 2011). Therefore, in understanding the evolution of betalain pigmentation, it is necessary to explain these three observations; that is, the unique origin of betalains, their apparent mutual exclusivity with anthocyanins, and the homoplastic distribution of the two pigment types (Brockington *et al.*, 2011).

Previous analyses have determined that three key structural enzymes within the anthocyanin synthesis pathway – chalcone synthase (CHS), dihydroflavonol-4-reductase (DFR) and anthocyanin synthase (ANS) – are maintained in betalain-pigmented taxa, but in the case of ANS are downregulated except in seeds (Shimada *et al.*, 2004, 2005). The maintenance of this structural pathway is thought to be due to the continued requirement of proanthocyanidins in seeds, and the lack of anthocyanins in other tissues in betalain-pigmented taxa is attributed to downregulation or suppression at the regulatory level (Shimada *et al.*, 2004, 2005). By contrast, the evolutionary fate of betalain synthesis genes in anthocyanic Caryophyllales is undetermined, and little is known about the presence or absence of specific betalain synthesis genes in anthocyanic species outside of the Caryophyllales. Putative homologs of DODA have been isolated from a range of anthocyanic taxa across angiosperms (Christinet *et al.*, 2004) but their precise relationship to betalain-specific isoforms is unknown. Similarly, the cytochrome P450 gene family that contains CYP76AD1 has radiated across the land plants (Mizutani & Ohta, 2010), but the evolutionary origin of CYP76AD1 within this large gene family is unknown.

Here, we explore the highly duplicated CYP76AD1 and DODA lineages in Caryophyllales, in order to understand the unique origin of the betalain synthesis pathway and its fate in anthocyanic taxa within Caryophyllales. We demonstrate that the earliest duplications in these gene lineages were apparently coincidental and gave rise to isoforms specific to betalain synthesis immediately before the earliest inferred origin of betalain pigmentation, and that these betalain-specific duplicate loci were then mostly lost or downregulated in the anthocyanic Molluginaceae and Caryophyllaceae. This pattern of duplication and loss is consistent with a single origin of betalain-specific function in core Caryophyllales. We further demonstrate that the betalain-specific isoforms of CYP76AD1 and DODA in *Beta vulgaris* are in close physical proximity on chromosome 2, suggesting the possibility of a metabolic operon and supporting the idea of their coincident duplication. We also analyze patterns of positive selection within and among these radiating DODA and CYP76AD1 lineages and implicate key residues in the neofunctionalization of the betalain-specific loci. Finally, we report the asymmetric diversification of the betalain-specific DODA genes that may be implicated in the evolution of color within the betalain pigment system.

Materials and Methods

Genome and transcriptome data from a total of 100 species were included in our analysis, including both amino acid and coding sequence (see Supporting Information Table S1). Among them, 95 were ingroup species representing 26 of the 34 families in Caryophyllales (Bremer *et al.*, 2009). Of these, two were from annotated genomes and the remaining 93 were from transcriptomes. Fifty-nine transcriptomes were obtained from the One Thousand Plants (1KP) Consortium, including 58 that were previously published (Finn *et al.*, 2014; Matasci *et al.*, 2014; Wickett *et al.*, 2014) and one as yet unpublished (*Portulacca amilis*, Table S1). Raw reads for another 20 transcriptomes were downloaded from the NCBI Sequence Read Archive (SRA; Table S1). Fourteen transcriptomes were newly generated for this study (Table S2). RNA isolation was carried out using the Bio-Rad Aurum Total RNA Mini Kit (Bio-Rad Life Science Research, Hercules, CA, USA) following the manufacturer's instructions, or using TRIzol® Reagent (Life Technologies, Thermo Fisher Scientific, Waltham, MA, USA) followed by a DNase treatment using the TURBO DNA-free™ Kit (Life Technologies, Thermo Fisher Scientific). Total RNA was quantified on an Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Santa Clara, CA, USA). Libraries were prepared using the TruSeq Stranded mRNA Sample Prep Kit (Illumina Inc., San Diego, CA, USA), and were quantified using an Agilent 2100 Bioanalyzer. Six or eight libraries were multiplexed per lane on an Illumina HiSeq2000 sequencer at the University of Michigan DNA Sequencing Core. Newly generated reads were deposited in SRA (BioProject: PRJNA280277).

All raw reads that were newly generated or downloaded from SRA were filtered following the same procedures as Yang *et al.* (2015). *De novo* assembly was carried out using Trinity v20140413p1 with default settings (Grabherr *et al.*, 2011) for datasets from SRA, and with the stranded 'RF' setting for the newly generated sequences. All assembled transcripts were translated using TransDecoder v16JAN2014 (Haas *et al.*, 2013) taking Pfam domain information into account. Amino acid and coding sequences of seven species were obtained from genome annotations in Phytozome v9 (Goodstein *et al.*, 2012) or respective publications (The Arabidopsis Genome Initiative, 2000; Tomato Genome Consortium, 2012; Huang *et al.*, 2013; Ibarra-Laclette *et al.*, 2013; Dohm *et al.*, 2014; Yagi *et al.*, 2014).

Amino acid and mRNA sequences from the *Beta vulgaris DODA* gene, as well as from the *B. vulgaris CYP76AD1* gene and its orthologs in *Amaranthus cruentus* and *Mirabilis jalapa*, were obtained from Hatlestad *et al.* (2012). The *B. vulgaris DODA* and *CYP76AD1* amino acid sequences were used to search against each of the 101 amino acid datasets using SWIPE v2.0.9 (Rognes, 2011) with an *E*-value cutoff of 10. We used a high *E*-value cutoff to maximize the sensitivity of searches in order to identify short and incomplete sequences. Positive hits with at least 40% amino acid identity to the *B. vulgaris DODA* gene or the *B. vulgaris CYP76AD1* gene were included in downstream analyses, except for the search for *CYP76AD1* homologs in *Arabidopsis thaliana*, for which hits with at least 20% amino acid

identity were included. All utilized DODA and CYP76AD1 homologs are listed in Tables S3 and S4, respectively. Sequences were deposited with GenBank; DODA sequences KR376141–KR376346, and CYP76AD1 sequences KR376350–KR376500.

In order to estimate the gene trees for *DODA* and *CYP76AD1*, an iterative set of alignment and phylogenetic estimation steps was conducted. An initial alignment was carried out for each of the two genes using MAFFT v7.154b using the default settings (Katoh & Standley, 2013), and low occupancy columns were trimmed using Phyutility v2.2.6 (-clean 0.01) (Smith & Dunn, 2008). A phylogeny was estimated with FastTree v2.1.7 (-wag) (Price *et al.*, 2010) and the alignment was then refined using SATé v2.2.7 (–iter-limit = 3) (Liu *et al.*, 2012). The resulting refined alignment was again trimmed (phyutility -clean 0.05) and a tree was then estimated again in FastTree (-wag). Branches longer than 1.5 were assumed to be due to distantly related paralogs and/or to assembly artifacts and were pruned. For *CYP76AD1*, which is part of a very large gene family, we extracted the clade of sequences containing the *B. vulgaris CYP76AD1* sequence together with the closest outgroup sequences for subsequent analysis. Sequences were realigned using PRANK v140110 (Löytynoja, 2014), poorly aligned sequences were removed from the alignment, trimmed (phyutility -clean 0.1), and the phylogeny was estimated by RAxML v8.0.0 (-m PROTCATWAG) (Stamatakis, 2014). Tips with branches longer than 1.0 were removed from the alignment, and the phylogeny re-estimated with 200 rapid bootstrap replicates. In addition to PRANK, we also used SATé and MAFFT for alignments and the resulting trees were highly similar. MAFFT and SATé tend to overalign (force regions to align even when they are highly divergent) whereas PRANK tends to do the opposite (introduce lots of gaps in highly divergent regions). For both DODA and CYP76AD1 lineages, the core Caryophyllales clade containing the known *B. vulgaris* gene was extracted and realigned, followed by removal of poorly aligned sequences and tree inference using RAxML with 100 bootstrap replicates. Analyses were repeated on the alignment using codon-aligned coding DNA sequence (CDS) data by RAxML and were found to be congruent with the amino acid-derived topologies. Phylogenies were rooted with non-Caryophyllales eudicot representatives of these orthogroups for which whole genome sequences could be obtained (e.g. *Solanum*, *Arabidopsis*).

All functionally characterized CYP761AD1 loci were retrieved from the literature and their location mapped onto the phylogeny, including: *MjCYP76AD3* (GenBank accession HQ656026) from *M. jalapa* (Hatlestad *et al.*, 2012; Suzuki *et al.*, 2014); *AcCYP76AD2* (HQ656025) from *A. cruentus* (Hatlestad *et al.*, 2012); and *BvCYP76AD1* from *B. vulgaris* (Hatlestad *et al.*, 2012). Similarly, all functionally characterized DODA loci were retrieved from the literature and mapped onto the phylogeny, including: *PgDODA* from *Portulaca grandiflora* (Christinet *et al.*, 2004); *MjDODA* from *M. jalapa* (Sasaki *et al.*, 2009); and *BvDODA* from *B. vulgaris* (Hatlestad *et al.*, 2012). The loci from *B. vulgaris* belonging to the CYP76AD1 and DODA lineages were blasted against the genome of *B. vulgaris* (RefBeet-1.1) in order to reveal their relative genomic location (Dohm *et al.*, 2014).

Codon models were used to test for the presence of positive selection during the evolutionary history of both the CYP76AD1 and DODA lineages. To avoid missing data, only sequences that were complete between codons 39 and 436 in CYP76AD1 (codons numbered on the basis of *B. vulgaris* CYP76AD1-α) and between 65 and 275 in DODA (codons numbered on the basis of *B. vulgaris* DODA-α) were included in analyses. Topologies were then re-estimated from the reduced alignment using RAxML (GTR + I + G). After confirming that the resulting tree topologies were congruent with trees derived from the original, larger alignments, they were subsequently used as reference topologies for the codon model analyses. CodeML as implemented in PAML v4.4 (Yang, 2007) was used to optimize codon models and estimate dN : dS values (ω) among branches of the RAxML-derived trees and at sites within the sequence alignments. CodeML estimation was performed using the clean data function that ignores all codons with ambiguous or missing data in the sequence alignment.

We employed branch-site model A (Yang & Nielsen, 2002) to test the occurrence of positive selection on individual codons along specific branch groups (model = 2 and NS sites = 2). Model A assumes that the branches in the phylogeny are divided *a priori* into foreground and background clades, where only the former may have experienced positive selection. For these analyses, all branches within the betalain-specific clades CYP76AD1-α and DODA-α were selected as the foreground and remaining branches of the topology outside of these clades served as background. Using CodeML, we analyzed the occurrence of positive selection along codon sites by comparing model M1a with model M2a (Yang, 2007). Codons evolving under positive selection in foreground branches were identified with a posterior probability > 0.95, as estimated by the Bayes Empirical Bayes procedure (Yang, 2007).

Ancestral state reconstruction analyses were performed in Mesquite (version 3.03, build 564) (Maddison *et al.*, 2015) using a likelihood-based approach and an Mk1 model that allows forward and reverse rates to be the same (estimated rate 3.12701442, −log L.: 15.35774177). The tree topology was derived from a concatenated *matK*/*rbcL* dataset from Brockington *et al.* (2009) and Brockington *et al.* (2011), and was reanalyzed with GARLI 2.0 under a GTR + I + G model (Zwickl, 2000) . Data on pigment content of the operational taxonomic units (OTUs) were obtained from a literature survey (primarily: Mabry, 1964; Clement & Mabry, 1996) and are the same as those contained in the dataset previously published by Brockington *et al.* (2011). OTUs were coded as anthocyanic, betalainic or missing data.

Results

The CYP76AD1 lineage underwent a minimum of three duplication events early on in the history of the core Caryophyllales. Following the divergence of *Physena*, gene duplication events near the base of core Caryophyllales gave rise to three clades within the CYP76AD1 lineage, here termed CYP76AD1-α, CYP76AD1-β and CYP76AD1-γ. All three clades are strongly supported with 100, 80 and 100% bootstrap support (BS),

respectively. The CYP76AD1-α contains both CYP76AD1 genes that are functionally implicated in the betalain synthesis pathway (as mapped in Fig. 3a). The CYP76AD1-α, CYP76AD1-β, and CYP76AD1-γ exhibit asymmetric patterns of gene silencing and loss, with betalainic taxa typically possessing orthologs for all three loci but anthocyanic taxa possessing only one (Figs S3, S4). CYP76AD1-β and CYP76AD1-γ were present in a mutually exclusive fashion in six of the 15 anthocyanic transcriptomes. For example, in the anthocyanic Molluginaceae, CYP76AD1-γ was recovered in *Mollugo pentaphylla* and *M. verticillata*, whereas

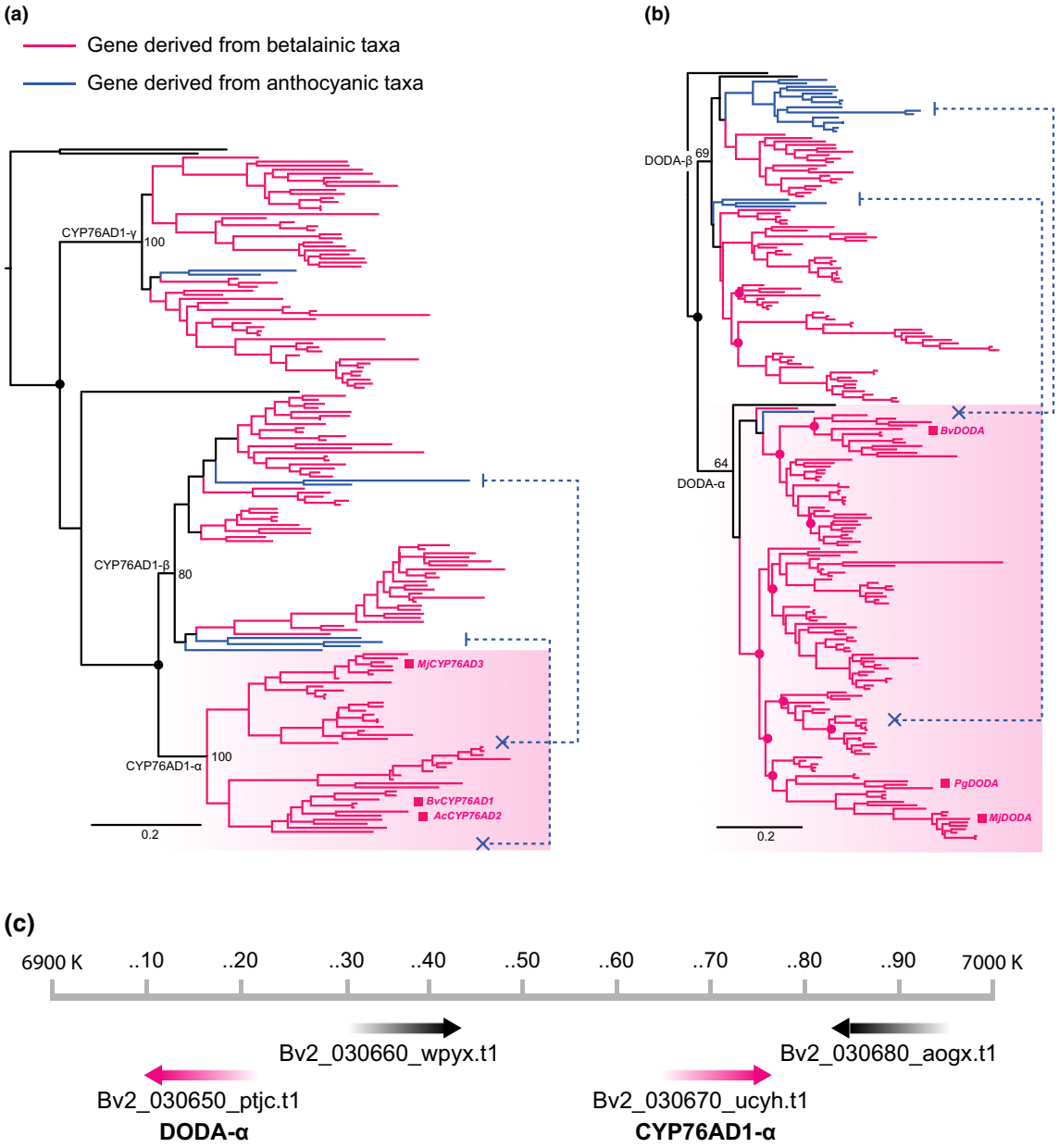

Fig. 3 (a) Phylogeny of the CYP76AD1 lineage, delimiting the CYP76AD1-α, -β and -γ clades, indicating the branching patterns amongst the three clades and their bootstrap support values. Black lines are early diverging Caryophyllales lineages (Microteaceae or Physenaceae). Duplication events are marked with closed circles. CYP76AD1 homologs derived from betalain-producing taxa are marked pink and homologs derived from anthocyanic taxa are marked blue. Blue dashed lines indicate that no CYP76AD1-α orthologs were recovered from anthocyanic transcriptomes or genomes, in contrast to CYP76AD1-β and -γ. The phylogenetic locations of the three functionally characterized betalain-specific CYP76AD1 genes (*MjCYP76AD3*, *AcCYP76AD2*, and *BvCYP76AD1*) are marked within the CYP76AD1-α clade. (b) Phylogeny of the 4,5-dioxygenase (DODA) lineage, containing DODA-α and -β clades, and their bootstrap support values. Grey lines are outgroups. Duplication events are marked with closed circles. Blue dashed lines indicate that in contrast to DODA-β, almost no DODA-α orthologs were recovered from anthocyanic transcriptomes or genomes. In contrast to DODA-β, no DODA-α orthologs were recovered from anthocyanic transcriptomes or genomes. The phylogenetic locations of the three functionally characterized betalain-specific DODA orthologs (*MjDODA*, *PgDODA* and *BvDODA*) are marked within the DODA-α clade. (c) Map of the region of *Beta vulgaris* chromosome 2 that contains the two functionally related but nonhomologous genes DODA-α (Bv2_030650_ptjc.t1) and CYP76AD1-α (Bv2_030670_ucyh.t1).

CYP76AD1-β was recovered in *M. cerviana*. CYP76AD1-β was also recovered in three Caryophyllaceae transcriptomes (Figs S1, S2). CYP76AD1-α was not detected in any anthocyanic transcriptomes (Figs 3a, S1, S2). In support of these results, BLAT and BLAST searches of the genomic scaffolds of the anthocyanic *Dianthus caryophyllus* (Caryophyllaceae) (Yagi *et al.*, 2014) recovered the CYP76AD1-β isoform but were unable to detect the CYP76AD1-α isoform.

The DODA gene lineage underwent a minimum of 11 duplication events specific to the core Caryophyllales (Fig. 3b). The first deep duplication event occurred before the divergence of *Microtea*, generating two major clades within the Caryophyllales, here termed DODA-α and DODA-β. Both clades were weakly supported with 64% and 69% BS, respectively. DODA-α contains all three characterized DODA orthologs (i.e. from *B. vulgaris*, *M. jalapa* and *P. grandiflora*) that have been functionally implicated in the betalain synthesis pathway (as mapped in Fig. 3b). As with the CYP76AD1 lineage, the DODA-α and DODA-β clades also exhibited asymmetric patterns of gene silencing and loss, such that the DODA-α lineage has undergone a minimum of nine separate gene duplication events, but only two duplication events were inferred in the DODA-β lineage (Fig. 3b). Whereas isoforms of DODA-β were present in all 15 anthocyanic Caryophyllales transcriptomes, the DODA-α isoform was detected in only one, in *Spergularia media* (Caryophyllaceae); that is, this species possessed both isoforms of DODA (Figs S3, S4). BLAT and BLAST searches of the genomic scaffolds of the anthocyanic *D. caryophyllus* recovered the DODA-β isoform but were unable to detect the DODA-α isoform (Yagi *et al.*, 2014).

In the *B. vulgaris* genome (Dohm *et al.*, 2014) CYP76AD1-β and CYP76AD1-γ are located on chromosomes 1 and 9, respectively, whereas the betalain-specific CYP76AD1-α is located on chromosome 2 (Fig. 4). The *B. vulgaris* DODA-β locus is located on chromosome 4, as are four of the DODA-α homologs (data not shown). The remaining DODA-α homolog (Bv2_030650_ptjc.t1), which is functionally implicated in betalain pigmentation, resides on chromosome 2 and is located *c.* 50 kb from the betalain-specific CYP76AD1-α homolog (Bv2_030670_ucyh.t1). These two genes are separated by a single functionally uncharacterized locus (Bv2_030660_wpyx.t1) (Fig. 3c).

Likelihood-based ancestral state reconstructions of pigment evolution under the Mk1 model recovered anthocyanins as the most likely dominant pigment at the node uniting *Macarthuria* and remaining core Caryophyllales (betalain: 0.29, anthocyanin: 0.71; Fig. 4), but the probability of betalain pigmentation being dominant was higher at the following node, uniting Caryophyllaceae and remaining core Caryophyllales (betalain: 0.53, anthocyanin: 0.47). Three reversions to anthocyanic pigmentation were inferred, in the ancestors of Kewaceae, Molluginaceae and Caryophyllaceae (Fig. 4), although the node uniting Caryophyllaceae with *Phaulothamnus* and Amaranthaceae s.l. had approximately equal probabilioties of betalain (0.55) and anthocyanin (0.45) pigmentation. The inferred phylogenetic positions of the key duplications leading to the DODA-α/β, and CYP76AD1-α/β lineages are mapped onto the topology in Fig. 4.

Manual inspection of the DODA alignment identified two residues that are diagnostic for the DODA-α clade, including proline at site 183 and trytophan at site 227 (Fig. 5a; residues are numbered on the basis of *B. vulgaris* DODA-α). Positive selection analyses implemented in PAML revealed three additional sites (82, 102 and 115) that are under positive selection ($P = 0.99$). These three sites were variable among the in-paralog clades within the DODA-α lineage. For CYP76AD1, several residues were also found to be diagnostic for the CYP76AD1-α clade, including histidine at site 111, threonine at sites 114 and 131, leucine at site 186, isoleucine at site 207, aspartic acid at site 241 and valine at site 275 (Fig. 5b; residues are numbered on the basis of *B. vulgaris* CYP76AD1-α). The positive selection analyses implemented in PAML identified sites 111, 114, 131, 186, 241 and 275 as being under positive selection ($P = 0.99$) in CYP76AD1-α.

Discussion

Our analyses demonstrate that the key betalain synthesis genes CYP76AD1 and DODA arose via gene duplications that are specific to the core Caryophyllales. Furthermore, the gene lineages containing CYP76AD1 and DODA have been subject to numerous and successive gene duplication events in core Caryophyllales. Nonetheless, within these broad Caryophyllales-specific radiations, all DODA and CYP76AD1 homologs that have been experimentally verified as functionally involved in the betalain synthesis pathway fall into single clades of related orthologs, despite being functionally characterized in phylogenetically diverse taxa. For example *MjCYP76AD2* (characterized in *Mirabilis jalapa*), *BvCYP76AD1* (characterized in *Beta vulgaris*), and *AcCYP76AD3* (characterized in *Amaranthus cruentus*) fall into the CYP76AD1-α clade, and all experimentally verified, functionally annotated DODA genes (*MjDODA* from *M. jalapa*, *PgDODA* from *P. grandiflora*, and *BvDODA* from *B. vulgaris*) fall into the DODA-α clade. Both the CYP76AD1-α and DODA-α clades arose via gene duplication events just before the divergence of *Microtea*, after which our reconstruction analyses suggest the probability of betalain pigmentation increases (Fig. 4; Brockington *et al.*, 2011). The timing of the duplication events close to the inferred origin of betalain pigmentation, together with the observation that the CYP76AD1-α and DODA-α clades harbor all homologs functionally implicated in the betalain pathway, is significant. These observations imply that gene duplication and subsequent neofunctionalization were key events in the evolutionary origin of betalain pigmentation. However, we recognize that full confirmation of betalain-specific neofunctionalization will require future analysis of betalain-associated enzymatic activity between CYP76AD1-α and DODA-α and their sister clades, within single species.

Asymmetric patterns of gene loss in the CYP76AD1-α and DODA-α clades relative to their paralogous sister clades are informative with respect to the evolution of betalain-specific gene function. In contrast to the CYP76AD1-β and CYP76AD1-γ clades, no homologs of the CYP76AD1-α clade were identified in transcriptomes from anthocyanic taxa. We confirmed the

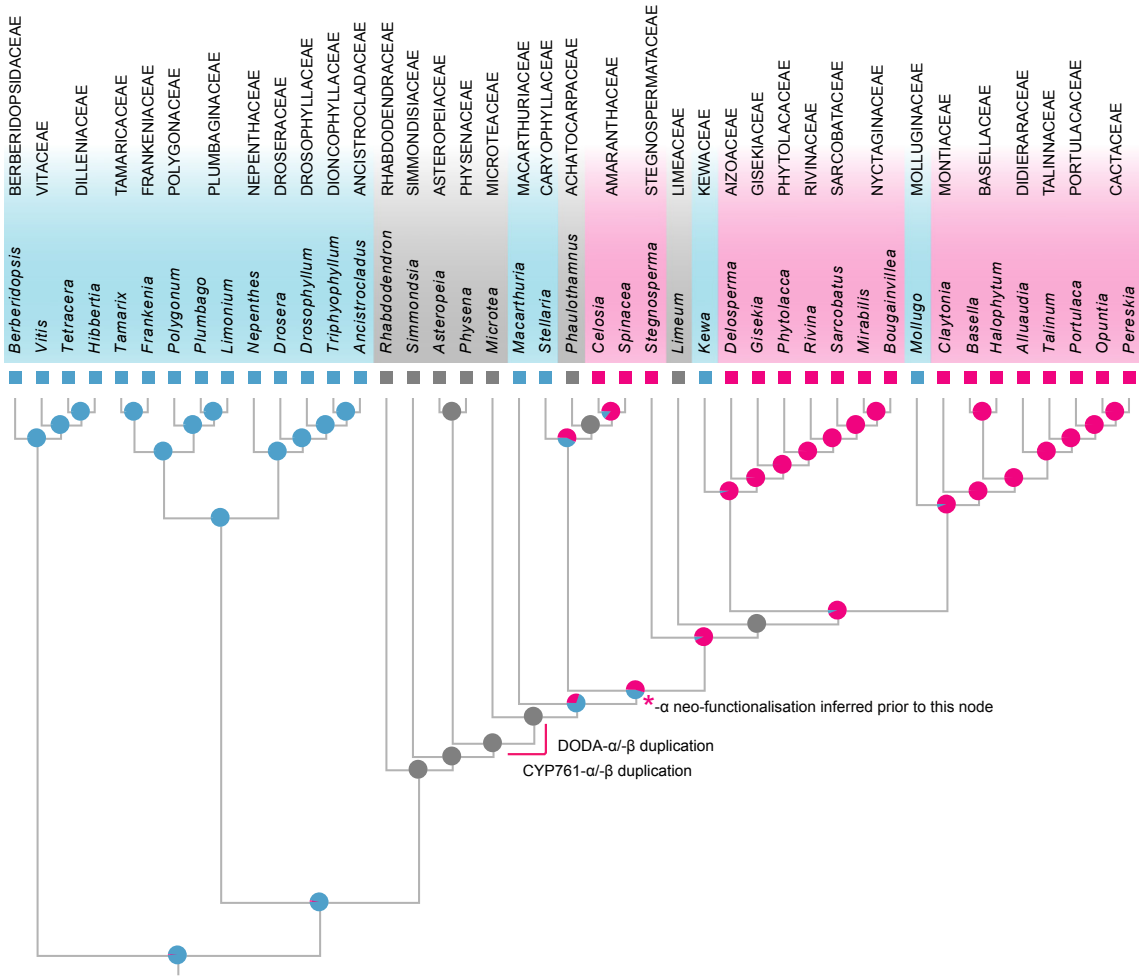

Fig. 4 Topology derived from a maximum-likelihood (ML) analysis of a concatenated *matK*/*rbcL* dataset, with terminals coded blue for anthocyanin, pink for betalain and grey for missing data. Results of the ML reconstruction of dominant pigment type are also depicted; probabilities for character states at internal nodes are reported as pie charts. The solid pink line indicates the approximate phylogenetic location of the duplication event in the CYP76AD1 lineage that gave rise to CYP76AD1-α, -β, and -γ and the duplication event in the 4,5-dioxygenase (DODA) lineage that gave rise to DODA-α and β. The pink asterisk marks the node at which we infer neofunctionalisation of the CYP76AD1-α and DODA-α must have taken place on the basis of duplication and subsequent losses in the CYP76AD1 and DODA lineages.

absence of the CYP76AD1-α homolog from the genome of the anthocyanic *Dianthus caryophyllus*, suggesting that gene loss may underlie the general absence of CYP6AD1-α homologs from at least some anthocyanic transcriptomes. Exactly the same patterns of asymmetric gene loss occur in DODA-α vs DODA-β. Isoforms of DODA-β are present in 15 anthocyanic transcriptomes plus the genome of *D. caryophyllus*, but the DODA-α variant was only detectable from the transcriptome of a single putatively anthocyanic species, *Spergularia media*. The DODA-α sequence of *S. media* occupies the typical phylogenetic position of Caryophyllaceae, suggesting that this locus has been inherited and retained in *S. media* rather than acquired via horizontal gene transfer (HGT). Loss or downregulation of these loci in taxa with anthocyanin pigmentation also strongly implies that these clades have a betalain-specific function. Furthermore, it indicates that the betalain-specific function of the CYP76AD1-α and DODA-α homologs arose before the origin of these anthocyanic lineages. Consequently, we infer that the common ancestors to both the Caryophyllaceae and Molluginaceae each had a functional betalain synthesis pathway, and that the betalain-specific genes were then lost in the anthocyanic taxa because betalain pigmentation is no longer maintained. Loss of both the CYP76AD1-α and DODA-α isoforms is confirmed in only one anthocyanic species with a fully sequenced genome, *D. caryophyllus*. The bulk of our data were derived from transcriptomes and we therefore cannot distinguish between gene loss and downregulation. Although we appreciate that absence from the transcriptomes is not proof of absence, the common expression of the DODA-β in 15 different anthocyanic transcriptomes compared with the presence of the DODA-α in a single anthocyanic transcriptome is compelling.

The asymmetric loss or downregulation of the DODA-α and CYP76AD1-α homologs in anthocyanic taxa suggests an interesting mechanistic difference underlying the mutual exclusion of these two pigment types. Whereas the suppression of anthocyanin pigmentation in betalain taxa appears to be achieved at the regulatory level (Shimada *et al.*, 2004, 2005, 2007; Hatlestad *et al.*, 2014), the long-term exclusion of betalain pigments from anthocyanic taxa is likely mediated through gene loss. These

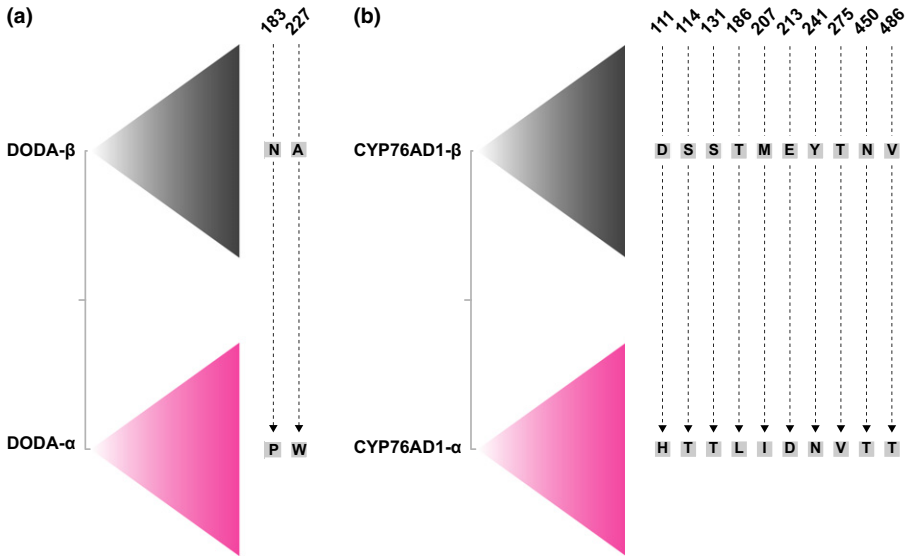

Fig. 5 Diagnostic invariant residues (denoted by grey squares with black text) that distinguish the (a) 4,5-dioxygenase (DODA)-α and (b) CYP76AD1-α lineages, from their sister -β lineages. Denoted amino acid residues are invariant in the respective clades. CYP76AD1 residues are numbered on the basis of *Beta vulgaris* CYP76AD1-α and DODA residues are numbered on the basis of *Beta vulgaris* DODA-α.

patterns of gene loss or downregulation in betalain-pigmented taxa also inform our understanding of the homoplastic distribution of anthocyanic lineages and betalain-pigmented clades. Given the intercalation of anthocyanic and betalain pigmented lineages, previous analyses concluded that derived instances of anthocyanin pigmentation are reversals from an ancestral betalain condition. However, in a previous study we showed that the phylogenetic disintegration of the polyphyletic Molluginaceae s.l. broadens the distribution of anthocyanin pigmentation such that each major clade of betalain taxa is now subtended at, or towards its base by an anthocyanic lineage (Brockington *et al.*, 2011). The anthocyanic Molluginaceae s.s. are sister to the betalain-pigmented Portulacineae, the anthocyanic *Kewa* (formerly included in *Hypertelis*; Christenhusz *et al.*, 2014) are sister to the betalain-pigmented Aizoaceae, Gisekiaceae, Phytolaccaceae and Nyctaginaceae, and the anthocyanic Caryophyllaceae are sister to the betalain-pigmented Amaranthaceae s.l. We previously argued that this pattern was consistent with multiple origins of betalain pigmentation from an ancestral anthocyanic condition (Brockington *et al.*, 2011). However, the loss or downregulation of CYP76AD1-α and DODA-α homologs from anthocyanic lineages argues against multiple origins of betalain pigmentation, and instead suggests that a betalain-pigment pathway was fully functioning in the common ancestor to the Caryophyllaceae, Molluginaceae and *Kewa*. A single origin of betalain pigmentation seems more likely in light of these data.

The phylogenetic timing and location of the duplication events giving rise to CYP76AD1-α and DODA-α do not coincide with the previously identified hotspots of gene duplication within Caryophyllales that are putatively associated with whole genome duplication events (Dohm *et al.*, 2012, 2014; Yang *et al.*, 2015). We therefore reasoned that the DODA and CYP76AD1 genes might be located physically close together and thus would be subject to more localized duplication events. When we examined the physical location of the *B. vulgaris* DODA and CYP76AD1 homologs we found that the betalain-specific DODA-α homolog (Bv2_030650_ptjc.t1) and betalain-specific CYP76AD1-α homolog (Bv2_030670_ucyh.t1) are in close proximity on the same scaffold of chromosome 2 in the *B. vulgaris* genome. The fact that these genes are < 50 kb apart greatly increases the probability of both loci being subject to the same localized segmental duplication(s), which may explain the phylogenetic proximity of the duplication events that led to the betalain-specific DODA-α and CYP76AD1-α clades. Alternatively, this clustering of functionally related but nonhomologous loci is suggestive of an operon-like cluster, albeit of small gene number, such as has been described for other plant secondary metabolic pathways including triterpenes, steroidal and isoquinoline alkaloids, and cyanogenic glycosides (Boycheva *et al.*, 2014). A MYB gene (the R locus) that regulates both DODA-α and CYP76AD1-α in *B. vulgaris* (Hatlestad *et al.*, 2014), is also located on chromosome 2 and is linked to the CYP76AD1-α (Y) locus (Keller, 1936; Goldman & Austin, 2000), further supporting the concept of a metabolic operon.

We explored the role for neofunctionalization in the evolution of the betalain-specific DODA-α and CYP76AD1-α loci by examining diagnostic molecular substitutions and patterns of selection. Christinet *et al.* (2004) proposed that paralogous copies of DODA possess diagnostic residues that are positioned close to the putative catalytic site of the DODA protein. In our taxon-dense dataset, two of these residues hold up as invariant sites that distinguish DODA-α and DODA-β. Immediately following the histidine at site 182, which is universally conserved across all DODA genes (Christinet *et al.*, 2004), a proline at site 183 is invariantly present in the DODA-α clade but is almost invariantly asparagine and never proline in the DODA-β clade. Similarly, at site 227 a tryptophan is invariably present in DODA-α whereas alanine is invariably present at the same site in DODA-β. Analyses of selection patterns in the DODA lineage indicate that DODA-α underwent positive selection following the duplication event that gave rise to DODA-α and its sister lineage DODA-β. Furthermore, positive selection analyses revealed a number of variable sites that are under positive selection ($P = 0.99$) within DODA-α, including sites 82, 102 and 115. Interestingly, the

residues at these sites sometimes vary between paralogous clades within DODA-α suggesting possible adaptive evolution after duplication events within the DODA-α lineage.

Analyses of selection patterns also indicate that the CYP76AD1-α lineage underwent positive selection following the duplication event that gave rise to CYP76AD1-α and its sister lineage CYP76AD1-β, at sites 111, 114, 131, 186, 241 and 275. However, these putatively functional significant changes are spread across the CYP76AD1 protein and are hence less obviously associated with one active site. These results are in line with analyses of mammalian cytochrome P450 proteins, which have documented that substrate recognition is mediated by several substrate recognition sites (SRSs) that are broadly distributed across the protein (Gotoh, 1992). Point mutations have been shown to significantly affect substrate recognition (Gotoh, 1992). It is likely that the betalain-specific function of CYP76AD1-α arose in a switch of substrate specificity towards L-DOPA, and hence these identified residues are an important starting point in understanding the evolution of a putative specificity change.

Comparison of the asymmetric radiation in DODA-α and DODA-β is intriguing. Only two gene duplications are inferred (one within Nyctaginaceae, one within Portulacaceae) in DODA-β, whereas a minimum of nine duplications are inferred at the familial level or above in DODA-α, including four at the base of the clade containing Aizoaceae, Nyctaginaceae, Phytolaccaceae and Sarcobataceae, three within the Portulacineae, and three within the Amaranthaceae. These duplications largely coincide with hotspots of gene duplication that are likely the product of genome duplication (Yang *et al.*, 2015). Therefore, we suggest that the asymmetric radiation of DODA-α genes is the result of differential retention of DODA-α homologs vs DODA-β, rather than the product of localized segmental duplications of chromosomal regions that contain the DODA-α homologs. It is tempting to speculate that this radiation of DODA-α homologs is connected with the radiation of the betalain color tool-kit and that different paralogs of DODA-α may contribute to differential coloring in some way. With this in mind, it is interesting that the paralogous clades within DODA-α differ in diagnostic residues at sites that are suggested to be under positive selection, suggesting adaptive evolution following gene duplication within the DODA-α clade. Alternatively these paralogs could be differentially expressed and enable finer control of differential coloring patterns through the development of betalain-pigmented taxa. Functional analysis of these additional paralogs will ultimately be necessary to determine their *in planta* significance.

Conclusion

In summary, we propose that the betalain synthesis pathway arose in part through coincidental duplication events in the DODA and CYP76AD1 lineages. We infer that these duplication events gave rise to the betalain-specific DODA-α and CYP76AD1-α clades. Several lines of evidence suggest that this betalain-specific neofunctionalization occurred in the DODA-α and CYP76AD1-α clades, including: the origin of these clades just before the probable origin of betalain pigmentation; the asymmetric loss or downregulation of DODA-α and CYP76AD1-α homologs in anthocyanic lineages; the restricted distribution of isoforms functionally implicated in betalain synthesis to the DODA-α and CYP76AD1-α clades; and the asymmetric radiation of the DODA-α lineage in betalain taxa vs the low copy number of DODA-β detectable from transcriptomes in equivalent species. These patterns imply that betalain pigmentation arose once, early in the evolution of the Caryophyllales, and then was lost in anthocyanic lineages. Furthermore our data imply that the long-term exclusion of betalain pigments from anthocyanic taxa is likely mediated through gene loss, rather than at the regulatory level, as is the case with the suppression of anthocyanin in betalain species. We identify numerous diagnostic residues and sites under positive selection that are good candidates with which to explore the evolution of this specificity change in these putatively neofunctionalized clades. Finally we identify that the functionally related but otherwise nonhomologous DODA-α and CYP76AD1-α genes are physically located close together in the *B. vulgaris* genome, suggestive of a possible operon cluster in relation to the betalain synthesis pathway. Future experiments should aim to assess the betalain-specific activity of the CYP76AD1-α and DODA-α lineages relative to their paralogous sister clades, explore the functional significance of the asymmetric radiation in the DODA-α lineages, and seek to understand how specific amino acid changes have contributed to the evolution of betalain-specific activity. Finally, understanding the evolutionary forces that led to multiple losses of betalain pigmentation and reversions to the anthocyanic condition remains a key unanswered question.

Acknowledgements

We thank Beverley Glover for critical reading of the manuscript and members of the Glover Lab for useful discussion. We thank Pascal-Antoine Christin for advice on use of PAML. We thank the Royal Botanic Gardens Kew for supplying material of *Pisonia umbellifera* and *Phytolacca americana*, and Edith Kapinos for help with tissue sampling. We thank M. Raquel Marchán Rivadeneira and Venkata Shiva Mandala for help with lab work. The molecular work was conducted in the Genomic Diversity Laboratory of the Department of Ecology and Evolutionary Biology, University of Michigan. S.C. was supported by a grant to IRRI from the Bill and Melinda Gates Foundation and UKAID. This work was supported by a National Science Foundation award (grant numbers DEB 1354048 and DEB 1352907) to S.F.B., M.J.M. and S.A.S., and a NERC Independent Research Fellowship to S.F.B. The 1000 Plants (1KP) initiative, led by G.K.S.W., is funded by the Alberta Ministry of Enterprise and Advanced Education, Alberta Innovates Technology Futures (AITF), Innovates Centre of Research Excellence (iCORE), Musea Ventures and BGI-Shenzhen.

References

Bate-Smith EC, Lerner NH. 1954. Leuco-anthocyanins. 2. Systematic distribution of leuco-anthocyanins in leaves. *Biochemical Journal* **58**: 126–132.

Bischoff H. 1876. *Das Caryophyllinenroth*. Inaugural dissertation, University of Tubingen, Tubingen, Germany.

Boycheva S, Daviet L, Wolfender J-L, Fitzpatrick TB. 2014. The rise of operon-like gene clusters in plants. *Trends in Plant Science* **19**: 447–459.

Bremer B, Bremer K, Chase M, Fay M, Reveal J, Soltis D, Soltis P, Stevens P. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* **161**: 105–121.

Brockington SF, Alexandre R, Ramdial J, Moore M, Crawley S, Dhingra A, Hilu K, Soltis P, Soltis DE. 2009. Phylogeny of the Caryophyllales *sensu lato*: Revisiting hypotheses on pollination biology and perianth differentiation in the core Caryophyllales. *International Journal of Plant Science* **170**: 627–643.

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

Cai Y, Sun M, Corke H. 2003. Antioxidant activity of betalains from plants of the amaranthaceae. *Journal of Agriculture and Food Chemistry* **51**: 2288–2294.

Campanella JJ, Smalley JV, Dempsey ME. 2014. A phylogenetic examination of the primary anthocyanin production pathway of the Plantae. *Botanical Studies* **55**: 10.

Christenhusz MJM, Brockington SF, Christin P-A, Sage RF. 2014. On the disintegration of Molluginaceae: a new genus and family (*Kewa*, Kewaceae) segregated from *Hypertelis*, and placement of *Macarthuria* in Macarthuriaceae. *Phytotaxa* **181**: 238–242.

Christinet L, Burdet FX, Zaiko M, Hinz U, Zrÿd J-P. 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.

Clement J, Mabry T. 1996. Pigment evolution in the Caryophyllales: a systematic overview. *Botanica Acta* **109**: 360–367.

Dohm JC, Lange C, Holtgräwe D, Sörensen TR, Borchardt D, Schulz B, Lehrach H, Weisshaar B, Himmelbauer H. 2012. Palaeohexaploid ancestry for Caryophyllales inferred from extensive gene-based physical and genetic mapping of the sugar beet genome (*Beta vulgaris*). *Plant Journal* **70**: 528–540.

Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutíerrez S, Zakrzewski F, Tafer H, Rupp O, Sörensen TR, Stracke R, Reinhardt R *et al.* 2014. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature* **505**: 546–549.

Escribano J, Pedreño MA, García-Carmona F, Muñoz R. 1998. Characterization of the antiradical activity of betalains from *Beta vulgaris* L. roots. *Phytochemical Analysis* **9**: 124–127.

Finn RD, Miller BL, Clements J, Bateman A. 2014. iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic acids research* **42**: D364–D373.

Fry BG, Roelants K, Winter K, Hodgson WC, Griesman L, Kwok HF, Scanlon D, Karas J, Shaw C, Wong L *et al.* 2010. Novel venom proteins produced by differential domain-expression strategies in beaded lizards and gila monsters (genus *Heloderma*). *Molecular Biology and Evolution* **27**: 395–407.

Gandía-Herrero F, Cabanes J, Escribano J, García-Carmona F, Jiménez-Atiénzar M. 2013. Encapsulation of the most potent antioxidant betalains in edible matrixes as powders of different colors. *Journal of Agriculture and Food Chemistry* **61**: 4294–4302.

Gandía-Herrero F, García-Carmona F. 2012. Characterization of recombinant *Beta vulgaris* 4,5-DOPA-extradiol-dioxygenase active in the biosynthesis of betalains. *Planta* **236**: 91–100.

Goldman IL, Austin D. 2000. Linkage among the *R, Y* and *Bl* loci in table beet. *Theoretical Applied Genetics* **100**: 337–343.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N *et al.* 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **40**: D1178–D1186.

Gotoh O. 1992. Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. *Journal of Biological Chemistry* **267**: 83–90.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**: 644–652.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M *et al.* 2013. *De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature Protocols* **8**: doi: 10.1038/nprot.2013.084.

Harris NN, Javellana J, Davies KM, Lewis DH, Jameson PE, Deroles SC, Calcott KE, Gould KS, Schwinn KE. 2012. Betalain production is possible in anthocyanin-producing plant species given the presence of DOPA-dioxygenase and L-DOPA. *BMC Plant Biology* **12**: 34.

Hatlestad GJ, Akhavan NA, Sunnadeniya RM, Elam L, Cargile S, Hembd A, Gonzalez A, McGrath JM, Lloyd AM. 2014. The beet *Y* locus encodes an anthocyanin MYB-like protein that activates the betalain red pigment pathway. *Nature Genetics* **47**: 92–96.

Hatlestad GJ, Sunnadeniya RM, Akhavan NA, Gonzalez A, Goldman IL, McGrath JM, Lloyd AM. 2012. The beet *R* locus encodes a new cytochrome P450 required for red betalain production. *Nature Genetics* **44**: 816–820.

Huang S, Ding J, Deng D, Tang W, Sun H, Liu D, Zhang L, Niu X, Zhang X, Meng M *et al.* 2013. Draft genome of the kiwifruit *Actinidia chinensis*. *Nature Communications* **4**: 2640.

Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang T-H, Lan T, Welch AJ, Juárez MJA, Simpson J *et al.* 2013. Architecture and evolution of a minute plant genome. *Nature* **498**: 94–98.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Keller W. 1936. Inheritance of some major color types in beets. *Journal of Agricultural Research* **52**: 27–38.

Khalturin K, Anton-Erxleben F, Sassmann S, Wittlieb J, Hemmrich G, Bosch TCG. 2008. A novel gene family controls species-specific morphological traits in hydra. *PLoS Biology* **6**: e278.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends in Genetics* **25**: 404–413.

Khan MI, Sri Harsha PSC, Giridhar P, Ravishankar GA. 2012. Pigment identification, nutritional composition, bioactivity, and *in vitro* cancer cell cytotoxicity of *Rivina humilis* L. berries, potential source of betalains. *LWT–Food Science and Technology* **47**: 315–323.

Krajka-Kuźniak V, Szaefer H, Ignatowicz E, Adamska T, Baer-Dubowska W. 2012. Beetroot juice protects against N-nitrosodiethylamine-induced liver injury in rats. *Food and Chemical Toxicology* **50**: 2027–2033.

Liu K, Warnow TJ, Holder MT, Nelesen SM, Yu J, Stamatakis AP, Linder CR. 2012. SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology* **61**: 90–106.

Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. *Methods in Molecular Biology* **1079**: 155–170.

Lu X, Wang Y, Zhang Z. 2009. Radioprotective activity of betalains from red beets in mice exposed to gamma irradiation. *European Journal of Pharmacology* **615**: 223–227.

Mabry T. 1964. *The betacyanins, a new class of red violet pigments, and their phylogenetic significance*. New York, NY, USA: Roland Press.

Maddison WP, Maddison DR. 2015. *Mesquite: a modular system for evolutionary analysis*. version 3.03. URL http://mesquiteproject.org [accessed 15 March].

Matasci N, Hung L-H, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M *et al.* 2014. Data access for the 1,000 Plants (1KP) project. *GigaScience* **3**: 17.

Mizutani M, Ohta D. 2010. Diversification of P450 genes during land plant evolution. *Annual Review of Plant Biology* **61**: 291–315.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2-approximately maximum-likelihood trees for large alignments. *PLoS One* **5**: e9490.

Rognes T. 2011. Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation. *BMC Bioinformatics* **12**: 221.

Sasaki N, Abe Y, Goda Y, Adachi T, Kasahara K, Ozeki Y. 2009. Detection of DOPA 4,5-dioxygenase (DOD) activity using recombinant protein prepared from *Escherichia coli* cells harboring cDNA encoding DOD from *Mirabilis jalapa*. *Plant and Cell Physiology* **50**: 1012–1016.

Shimada S, Inoue Y, Sakuta M. 2005. Anthocyanidin synthase in non-anthocyanin-producing Caryophyllales species. *Plant Journal* **44**: 950–959.
Shimada S, Otsuki H, Sakuta M. 2007. Transcriptional control of anthocyanin biosynthetic genes in the Caryophyllales. *Journal of Experimental Botany* **58**: 957–967.
Shimada S, Takahashi K, Sato Y, Sakuta M. 2004. Dihydroflavonol 4-reductase cDNA from non-anthocyanin-producing species in the Caryophyllales. *Plant and Cell Physiology* **45**: 1290–1298.
Simola DF, Wissler L, Donahue G, Waterhouse RM, Helmkampf M, Roux J, Nygaard S, Glastad KM, Hagen DE, Viljakainen L *et al.* 2013. Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Research* **23**: 1235–1247.
Smith SA, Dunn CW. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* **24**: 715–716.
Stafford HA. 1994. Anthocyanins and betalains: evolution of the mutually exclusive pathways. *Plant Science* **101**: 91–98.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.
Suzuki M, Miyahara T, Tokumoto H, Hakamatsuka T, Goda Y, Ozeki Y, Sasaki N. 2014. Transposon-mediated mutation of CYP76AD3 affects betalain synthesis and produces variegated flowers in four o'clock (*Mirabilis jalapa*). *Journal of plant physiology* **171**: 1586–1590.
Tanaka Y, Sasaki N, Ohmiya A. 2008. Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids. *Plant Journal* **54**: 733–749.
The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.
Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**: 635–641.
Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA *et al.* 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences, USA* **111**: E4859–E4868.
Wu L-C, Hsu H-W, Chen Y-C, Chiu C-C, Lin Y-I, Ho J-AA. 2006. Antioxidant and antiproliferative activities of red pitaya. *Food Chemistry* **95**: 319–327.
Yagi M, Kosugi S, Hirakawa H, Ohmiya A, Tanase K, Harada T, Kishimoto K, Nakayama M, Ichimura K, Onozaki T *et al.* 2014. Sequence analysis of the genome of carnation (*Dianthus caryophyllus* L.). *DNA Research* **21**: 231–241.
Yang Y, Moore MJ, Brockington SF, Soltis DE, Wong GK-S, Carpenter EJ, Zhang Y, Chen L, Yan Z-X, Xie Y-L *et al.* 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution*. doi:10.1093/molbev/msv081.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**: 1586–1591.
Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* **19**: 908–911.
Zwickl D. 2000. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*, PhD Thesis. University of Texas, Austin, USA.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Taxon-labeled phylogeny of the CYP76AD1 lineages (and CYP76AD1–γ).

Fig. S2 Taxon-labeled phylogeny of the CYP76AD1 lineages (CYP76AD1-α, CYP76AD1-β).

Fig. S3 Taxon-labeled phylogeny of the 4,5-dioxygenase (DODA) lineages (DODA-β).

Fig. S4 Taxon-labeled phylogeny of the 4,5-dioxygenase (DODA) lineage (DODA-α).

Table S1 Sources of the transcriptome data

Table S2 Information for 14 newly sequenced transcriptomes: provenance, herbarium vouchers, type of material

Table S3 List of 4,5-dioxygenase (DODA) gene accessions and corresponding taxa

Table S4 List of CYP76AD1 gene accessions and corresponding taxa

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.