

DR HESTER SHEEHAN (Orcid ID : 0000-0002-2169-5206)

MR BOAS PUCKER (Orcid ID : 0000-0002-3321-7471)

DR SAMUEL FRASER BROCKINGTON (Orcid ID : 0000-0003-1216-219X)

Article type : MS - Regular Manuscript

**Evolution of *L*-DOPA 4,5-dioxygenase activity allows for recurrent specialisation to betalain pigmentation in Caryophyllales**

**Authors**

Hester Sheehan<sup>\*1</sup>, Tao Feng<sup>\*1,2</sup>, Nathanael Walker-Hale<sup>\*1</sup>, Samuel Lopez-Nieves<sup>1</sup>, Boas Pucker<sup>1,3</sup>, Rui Guo<sup>1</sup>, Won C. Yim<sup>4</sup>, Roshani Badgami<sup>1</sup>, Alfonso Timoneda<sup>1</sup>, Lijun Zhao<sup>5</sup>, Helene Tiley<sup>6</sup>, Dario Copetti<sup>7</sup>, Michael J. Sanderson<sup>7</sup>, John C. Cushman<sup>4</sup>, Michael J. Moore<sup>6</sup>, Stephen A. Smith<sup>5</sup>, Samuel F. Brockington<sup>1</sup>

\*These authors contributed equally to this work.

Author for correspondence: Samuel F Brockington; [sb771@cam.ac.uk](mailto:sb771@cam.ac.uk); 01223336268

**Author affiliations**

1. Department of Plant Sciences, University of Cambridge, Tennis Court Road, Cambridge, CB2 3EA, United Kingdom
2. CAS Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, 430074, China.
3. CeBiTec & Faculty of Biology, Bielefeld University, Universitaetsstrasse, Bielefeld, 33615, Germany
4. Department of Biochemistry and Molecular Biology, University of Nevada, Reno, NV, 89577, USA
5. Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, 48109, USA
6. Department of Biology, Oberlin College, Science Center K111, Oberlin, OH, 44074, USA

**This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/NPH.16089](https://doi.org/10.1111/NPH.16089)**

This article is protected by copyright. All rights reserved

7. Department of Ecology and Evolutionary Biology, University of Arizona, 1041 E. Lowell St,  
Tucson, AZ, 85721, USA

Received: 30 May 2019

Accepted: 22 July 2019

**MAIN TEXT WORD COUNT: 6412**

**SUMMARY: 199**

**INTRODUCTION: 897**

**MATERIALS AND METHODS: 1663**

**RESULTS: 1686**

**DISCUSSION: 2052**

**ACKNOWLEDGMENTS: 114**

**FIGURES:** 8 total (all colour)

**SUPPORTING INFORMATION:** One main document (DOCX) with one PDF and one XLS file.

#### **ORCID ID's**

Hester Sheehan: 0000-0002-2169-5206

Tao Feng: 0000-0002-0489-2021

Nathanael Walker-Hale: 0000-0003-1105-5069

Samuel Lopez-Nieves: 0000-0002-3583-0392

Boas Pucker: 0000-0002-3321-7471

Rui Guo: 0000-0002-5165-7905

Won C. Yim: 0000-0002-7489-0435

Roshani Badgami: 0000-0002-9290-3420

Alfonso Timoneda: 0000-0002-7024-8947

Lijun Zhao: 0000-0001-7317-830X

Helene Tiley: 0000-0002-4227-1824

Dario Copetti: 0000-0002-2680-2568

Michael J. Sanderson: 0000-0002-0855-9648

John C. Cushman: 0000-0002-5561-1752

Michael J. Moore: 0000-0003-2222-8332

Stephen A. Smith: 0000-0003-2035-9531

Samuel F. Brockington: 0000-0003-1216-219X

## SUMMARY

- The evolution of *L*-DOPA 4,5-dioxygenase activity, encoded by the gene *DODA*, was a key step in the origin of betalain biosynthesis in Caryophyllales. We previously proposed that *L*-DOPA 4,5-dioxygenase activity evolved via a single Caryophyllales-specific neofunctionalisation event within the *DODA* gene lineage. However, this neofunctionalisation event has not been confirmed and the *DODA* gene lineage exhibits numerous gene duplication events, whose evolutionary significance is unclear.
- To address this, we functionally characterised 23 distinct *DODA* proteins for *L*-DOPA 4,5-dioxygenase activity, from four betalain-pigmented and five anthocyanin-pigmented species, representing key evolutionary transitions across Caryophyllales. By mapping these functional data to an updated *DODA* phylogeny, we then explored the evolution of *L*-DOPA 4,5-dioxygenase activity.
- We find that low *L*-DOPA 4,5-dioxygenase activity is distributed across the *DODA* gene lineage. In this context, repeated gene duplication events within the *DODA* gene lineage give rise to polyphyletic occurrences of elevated *L*-DOPA 4,5-dioxygenase activity, accompanied by convergent shifts in key functional residues and distinct genomic patterns of micro-synteny.
- In the context of an updated organismal phylogeny and newly inferred pigment reconstructions, we argue that repeated convergent acquisition of elevated *L*-DOPA 4,5-dioxygenase activity is consistent with recurrent specialisation to betalain synthesis in Caryophyllales.

**Keywords:** Anthocyanins, Betalains, Caryophyllales, Convergent Evolution, *L*-DOPA 4,5-dioxygenase (*DODA*), Metabolic Operon, Specialised Metabolism, Plant Pigments

## INTRODUCTION

As sessile organisms, plants have exploited metabolic systems to produce a plethora of diverse specialised metabolites (Weng, 2014). Specialised metabolites are critical for survival in particular ecological niches and are often taxonomically-restricted (Weng, 2014). In flowering plants, one remarkable example of a taxonomically-restricted specialised metabolite occurs in the angiosperm order Caryophyllales (Brockington *et al.* 2011; Timoneda *et al.*, 2019). Here, tyrosine-derived betalain

pigments have evolved to replace anthocyanins, which are otherwise ubiquitous across flowering plants (Bischoff, 1876; Clement & Mabry, 1996). Betalains are also present in the fungal lineage, Basidiomycota (Musso, 1979), and the bacterial species, *Gluconacetobacter diazotrophicus* (Contreras-Llano *et al.*, 2019).

The phylogenetic distribution of betalain pigmentation within Caryophyllales is complex (Brockington *et al.* 2011). In betalain-producing families within the core Caryophyllales, anthocyanins have never been discovered (Bate-Smith, 1962; Clement & Mabry, 1996), although earlier substrates and associated enzymes in the flavonoid pathway have been detected (Shimada *et al.*, 2004; 2005; Polturak *et al.*, 2018). However, anthocyanins have been reported in six core Caryophyllales families, including Caryophyllaceae, Molluginaceae, Kewaceae, Limeaceae, Macarthuraceae and Simmondsiaceae (Clement & Mabry, 1996; Thulin *et al.*, 2016). In these six anthocyanic lineages, betalains have not been detected, indicating that anthocyanins and betalains are mutually exclusive (Stafford, 1994; Clement & Mabry, 1996). The six anthocyanic lineages are intercalated with betalain-pigmented lineages resulting in a homoplastic distribution of these two pigments (Brockington *et al.*, 2011). The distribution of anthocyanin and betalain-pigmented lineages is consistent with multiple origins of betalain pigmentation, a single origin of betalain pigmentation with multiple reversals to anthocyanin, or a combination of these scenarios (Brockington *et al.*, 2011).

In contrast to the anthocyanin pathway, the betalain biosynthetic pathway is relatively simple, involving as few as four enzymatic steps to proceed from tyrosine to stable betalain pigments: yellow betaxanthins and violet betacyanins (Fig. 1). The core genes encoding betalain synthesis enzymes have been elucidated, primarily through heterologous assay in *Saccharomyces cerevisiae* and *Nicotiana benthamiana*, which has emerged as an essential tool for betalain research *in planta* (Polturak *et al.*, 2016; Timoneda *et al.* 2018). The key enzymatic step in betalain biosynthesis involves conversion of *L*-3,4-dihydroxyphenylalanine (*L*-DOPA) to betalamic acid, the central chromophore of betalain pigments. *L*-DOPA 4,5-dioxygenase is encoded by the gene *DODA*, a member of the *LigB* gene family (Christinet *et al.*, 2004). Within Caryophyllales, a gene duplication in the *LigB/DODA* gene lineage gave rise to the *DODA* $\alpha$  and *DODA* $\beta$  clades, with *L*-DOPA 4,5-dioxygenase activity previously inferred to have evolved at the base of the *DODA* $\alpha$  lineage (Brockington *et al.* 2015). On the basis of this Caryophyllales-specific *DODA* $\alpha$ /*DODA* $\beta$  duplication, and subsequent losses of *DODA* $\alpha$  loci in anthocyanic lineages, we previously argued for a single origin of betalain pigmentation, with multiple reversals to anthocyanin pigmentation (Brockington *et al.* 2015).

However, evolutionary patterns within the *DODA* $\alpha$  lineage are complex (Brockington *et al.*, 2015). In addition to the *DODA* $\alpha$ /*DODA* $\beta$  duplication, there have been at least nine duplications in the *DODA* $\alpha$  lineage resulting in all betalain-pigmented lineages of Caryophyllales containing at least three *DODA* genes – at least one homologue from the *DODA* $\beta$  lineage and at least two paralogues from the *DODA* $\alpha$  lineage, with *Beta vulgaris* containing five copies of *DODA* $\alpha$ . In the case of *B. vulgaris*, only two *DODA* $\alpha$  paralogues have been studied (Sasaki *et al.*, 2009; Gandía-Herrero & García-Carmona, 2012; Hatlestad *et al.*, 2012; Chung *et al.*, 2015; Bean *et al.*, 2018), with one paralogue, BvDODA1 (hereafter

termed BvDODA $\alpha$ 1), found to exhibit high levels of *L*-DOPA 4,5-dioxygenase activity and the other paralogue, BvDODA2 (hereafter termed BvDODA $\alpha$ 2), exhibiting no or marginal *L*-DOPA 4,5-dioxygenase activity. Bean *et al.* (2018) compared BvDODA $\alpha$ 1 and BvDODA $\alpha$ 2 and identified seven divergent residues that, when altered in BvDODA $\alpha$ 2, were sufficient to allow BvDODA $\alpha$ 2 to convert *L*-DOPA to betalamic acid in yeast. Like BvDODA $\alpha$ 2, two DODA $\alpha$  paralogues from other species (*Parakeelya mirabilis* and *Ptilotus hybrid*) have also been shown to have limited or no capacity to produce betalamic acid (Chung *et al.*, 2015). Extensive gene duplication within the DODA $\alpha$  lineage and the conserved presence of paralogues exhibiting no or marginal *L*-DOPA 4,5-dioxygenase activity is suggestive of further sub- and/or neofunctionalisation events occurring within the DODA $\alpha$  clade, but the evolutionary significance of this is unclear.

In the current study, we explore the evolution of *L*-DOPA 4,5-dioxygenase activity in Caryophyllales, focusing on paralogy within the DODA lineage. In the context of an updated organismal phylogeny and new pigment data, we select species representing key inferred transitions in pigment gain and loss. We then assess their DODA paralogues for levels of *L*-DOPA 4,5-dioxygenase activity using an established heterologous assay in *N. benthamiana*. We use the production of betacyanin in this heterologous assay as a proxy for *L*-DOPA 4,5-dioxygenase activity, with the relative strength of betacyanin production indicating the relative strength of *L*-DOPA 4,5-dioxygenase activity between paralogues. By mapping activity to a comprehensive DODA phylogeny, we reveal that multiple acquisitions of high *L*-DOPA 4,5-dioxygenase activity are linked with repeated gene duplication events within the DODA $\alpha$  lineage. Furthermore, we find that marginal levels of *L*-DOPA 4,5-dioxygenase activity are distributed across the DODA $\alpha$  lineage, implying that marginal levels of activity were present prior to multiple origins of high activity. Recurrent origins of high activity are accompanied by convergent shifts in residues known to be sufficient to confer *L*-DOPA 4,5-dioxygenase activity. We reconcile these data with inferred patterns of pigment evolution and argue that recurrent acquisition of high *L*-DOPA 4,5-dioxygenase activity underlies polyphyletic patterns of betalain pigmentation in Caryophyllales.

## MATERIALS AND METHODS

### Plant materials and growth conditions

*Beta vulgaris* ssp. *vulgaris* ‘Bolivar’ (referred to as *B. vulgaris*) was obtained from Thompson & Morgan (Suffolk, UK), and *B. vulgaris* ssp. *vulgaris* ‘YTiBv’ was obtained from Syngenta (Basel, Switzerland). Beet plants were grown at the Cambridge University Botanic Garden under natural light and temperature conditions. The seeds of *Limeum aethiopicum* Burm. f., *Kewa bowkeriana* (Sond.) Christenh., *Macarthuria australis* Hügel ex Endl., *Spergularia marina* (L.) Besser, *Cardionema ramosissimum* Weinm. A. Nelson & J.F. Macbr., *Telephium imperati* L., *Pollichia campestris* Aiton, *Corrigiola litoralis* L., *Spergula arvensis* L., and *Simmondsia chinensis* Link C.K. were grown at the Cambridge University Botanic Garden under natural conditions. Fresh tissue of *Polycarpon tetraphyllum* (L.) L. was collected in Florida (Lake Wauburg Recreation Area, Alachua County, FL,

US). *N. benthamiana* is a standard laboratory line that is maintained by selfing and plants were grown in controlled growth rooms with the following conditions: long-day (16 h light: 8 h dark), 20 °C, 60% humidity.

#### **DNA/RNA extraction and cDNA synthesis**

Tissue sampled for extraction was snap frozen and ground frozen using a Tissue Lyser II homogeniser (QIAGEN, Hilden, Germany). DNA was extracted using the QIAGEN DNeasy Plant Mini Kit (Hilden, Germany) and RNA was removed by the QIAGEN DNase-Free RNase Set. RNA extraction was carried out using the PureLink Plant RNA Reagent (Invitrogen, Carlsbad, CA, USA) and the Ambion TURBO DNA-free kit (Carlsbad, CA, USA). Both DNA and RNA quantity and quality were assessed by NanoDrop (Thermo Fisher Scientific, Waltham, MA, USA) and agarose gel electrophoresis. A Bioanalyzer (Agilent Technologies, USA) was used to assess quantity and quality of RNA for transcriptome sequencing. First strand cDNA synthesis was performed using BioScript Reverse Transcriptase (Bioline Reagents, London, UK) and an oligo dT primer. All protocols were carried out according to the manufacturer's specifications unless otherwise specified.

#### **Transcriptome and genome sequencing and assembly**

Transcriptomes of fresh young leaves of *S. chinensis* and fresh young leaves and flowers of *K. bowkeriana* were sequenced at BGI using BGI-seq (Hong Kong, China). Downstream processing and assembly optimisation was performed following Haak *et al.* (2018). Genome sequencing of *C. littoralis*, *L. aethiopicum*, *S. chinensis* and *S. arvensis* was performed on a HiSeq X-Ten with one sample per lane and assembled following Pucker *et al.* (2019). Full details can be found in Methods S1.

#### **Isolating DODA genes for functional analysis**

For species with sequenced genomes (*B. vulgaris*, *Mesembryanthemum crystallinum*, *Carnegiea gigantea*, *K. caespitosa*), DODA sequences were identified by BLAST searching genomes and annotated gene files. Annotated gene sequences were checked manually and adjusted if necessary (see Methods S2). For species for which no published genome was available, diverse strategies were used to obtain the full-length coding sequences. For *Stegnosperma halimifolium*, *M. australis*, *P. tetraphyllum* and *L. aethiopicum*, full-length coding sequences were recovered from transcriptomes. For *C. ramosissimum*, a partial coding sequence was recovered from RNAseq, and then RACE PCR was performed using the RACE System for Rapid Amplification of cDNA Ends Kit (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's instructions in order to amplify the 3' end of *CrDODa*. For *P. campestris*, *S. marina* and *T. imperati*, degenerate PCR primers were designed based on known *DODa* sequences in order to amplify a partial sequence, then inverse PCR was used to obtain the full-length coding sequences, following the protocol described in Ren *et al.* (2005). The full-length coding sequences for *DODA* genes were isolated from cDNA or gDNA by PCR using Phusion High-Fidelity DNA polymerase (Thermo Fisher Scientific, Waltham, MA, USA), and then cloned into pBlueScript SK (New England Biolabs, Hitchin, UK) and verified by Sanger sequencing (Source BioScience, Nottingham, UK); except for *M. crystallinum*, *C. gigantea* and *S.*

*halimifolium* for which *DODA* genes were synthesised by BioMatik (Cambridge, Canada), Twist Bioscience (San Francisco, CA, USA) and Integrated DNA Technologies (Iowa, IA, USA), respectively. Oligonucleotides are listed in Table S1 and the sequences have been deposited in GenBank (Table S2).

### Species phylogeny and pigment reconstruction

To enable trait reconstruction across the order Caryophyllales, we generated a comprehensive genus-level species tree using publicly-available sequence data compiled by PyPHLAWD (Smith & Walker, 2019), with constraints of the backbone topology from Walker *et al.* (2018) and Thulin *et al.* (2016, 2018). We calibrated branch lengths to time using treePL (Smith & O'Meara, 2012), which implements the penalised likelihood approach of Sanderson (2002). Full details can be found in Methods S3.

Pigment data at genus resolution was used to reconstruct the evolution of betalain pigmentation on the time-calibrated genus-level phylogeny of Caryophyllales. We surveyed the literature for pigment data and determined the pigmentation status of 174 genera, classifying them as anthocyanin-pigmented, betalain-pigmented, or unknown (Table S3). We reconstructed ancestral states using maximum likelihood (Pagel, 1994, 1999) and Bayesian inference via stochastic mapping (Huelsenbeck *et al.*, 2003; Bollback, 2006), using the R packages ape v5.0 and phytools v0.6-70, respectively in R v3.6.0 (Revell, 2012; Paradis & Schliep, 2019; R Core Team, 2019) under an equal rate and an asymmetric rate model. For stochastic mapping, we enforced a prior that the root of Caryophyllales was anthocyanin-pigmented. Full details can be found in Methods S4.

### DODA gene phylogeny and ancestral sequence reconstruction

We compiled a dataset of publicly available and early release genome and transcriptome assemblies (Table S4) and used a baited search approach with iterative refinement (Lopez-Nieves *et al.*, 2018) to infer a gene tree of *DODA* sequences in Caryophyllales. Full details can be found in Methods S5. To create a sequence dataset computationally and numerically tractable for ancestral sequence reconstruction, we used a custom python script to subsample the *DODAA* gene tree (Fig. S1), using a strategy designed to maintain within-paralogue diversity. We created a final dataset of 198 sequences (indicated in Table S5), ensuring that a representative of all functionally characterised *DODAA* loci was included. Ancestral sequence reconstructions were conducted for codons and amino acids in IQ-TREE v1.6.10 (Nguyen *et al.*, 2015; Kalyaanamoorthy *et al.*, 2017). All scripts, alignments and trees are available on GitHub (<https://github.com/NatJWalker-Hale/DODA>). Full details can be found in Methods S6.

### Vector generation and transient expression assay

Construction of the multi-gene vectors containing the genes of the betalain biosynthetic pathway (*DODA*, *BvCYP76AD1*, *MjcDOPA-5GT*) was carried out using MoClo GoldenGate cloning following the protocol described (Engler *et al.*, 2014; Timoneda *et al.*, 2018) in order to produce level 2 binary vectors (Fig. S2). *DODA* genes were cloned into Level 0 vectors using their coding sequence, except *PtDODAA* for which the gDNA sequence was used. Level 1 vectors were verified by sequencing and

level 2 vectors were verified by restriction digests. Transient expression using agroinfiltration of *N. benthamiana* was performed as described previously (Timoneda *et al.*, 2018). The following were used as controls in every experiment: positive, pBC-BvDODA $\alpha$ 1; negative, pLUC (Fig. S2). Upon transient transformation in *N. benthamiana*, *DODA* genes encoding enzymes that carry out the *L*-DOPA 4,5-dioxygenase reaction necessary for betalamic acid production will produce betalains. For instance, the previously characterised *B. vulgaris* *DODA*, BvDODA $\alpha$ 1 (Hatlestad *et al.*, 2012), exhibits high levels of *L*-DOPA 4,5-dioxygenase activity as inferred by a high level of betacyanin production under heterologous expression (Polturak *et al.*, 2016; Timoneda *et al.*, 2018). Accordingly, we use the production of betacyanin in this heterologous assay as a proxy for *L*-DOPA 4,5-dioxygenase activity, with the relative strength of betacyanin production indicating the relative strength of *L*-DOPA 4,5-dioxygenase activity between loci. The assay is designed to give a clear comparative measure of biologically relevant levels of *L*-DOPA 4,5-dioxygenase activity *in planta*. We categorise the levels as high *L*-DOPA 4,5-dioxygenase activity (hereafter also referred to as ‘high activity’), low or marginal *L*-DOPA 4,5-dioxygenase activity (hereafter also referred to simply as ‘marginal activity’), or no *L*-DOPA 4,5-dioxygenase activity (hereafter also referred to as ‘no activity’)

#### **Betalain quantification using high performance liquid chromatography**

A single sample was taken from each infiltration spot four days post-infiltration, snap frozen in liquid nitrogen and stored at  $-80^{\circ}\text{C}$  until needed. Samples were homogenised frozen using a single 5 mm glass bead in a Tissue Lyser II homogeniser (QIAGEN, Hilden, Germany). Betalains were extracted overnight at  $4^{\circ}\text{C}$  in 80% aqueous methanol containing 50 mM ascorbic acid with a volume of 1 mL extraction buffer per 50 mg fresh weight leaf tissue. After extraction, the samples were clarified twice by centrifugation at 21,130 g for 10 min. HPLC analysis was performed using a Thermo Fisher Scientific Accela HPLC autosampler and pump system incorporating a photodiode array detector (Waltham, MA, USA). Betalains were separated using a Luna Omega column (100 Å, 5  $\mu\text{m}$ , 4.6 mm X 150 mm) from Phenomenex (Torrance, CA, USA) under the following conditions: 3 min, 0% B; 3–19 min, 0–75% B; 7 min, 0% B where mobile phase A was 0.1% formic acid in 1% acetonitrile and solvent B was 100% acetonitrile, and at a flow rate of 500  $\mu\text{L min}^{-1}$ . We quantified the betacyanin compound, betanin, because it has been shown to be the predominant pigment arising from the transient expression assay (Timoneda *et al.*, 2018). Betanin was detected by UV/VIS absorbance at a wavelength of 540 nm. Identification and quantification of betanin was carried out using a commercially available *B. vulgaris* extract (Tokyo Chemical Industry UK Ltd, Oxford, UK) and a pure betanin standard (provided by Fernando Gandía-Herrero, Universidad de Murcia, Spain).

Negative controls (uninfiltrated tissue or pLUC) were set as background and removed from all other samples.

### Synteny analysis of gene cluster

Synteny from sequenced genomes was evaluated to explore the conservation of clustering of *BvDODAA1* and *BvCYP76AD1* as observed in *B. vulgaris* (Brockington *et al.*, 2015). *BvDODAA1* (Hatlestad *et al.*, 2012) and *BvCYP76AD1* (DeLoache *et al.*, 2015; Polturak *et al.*, 2016; Sunnadaniya *et al.*, 2016) were identified from the *B. vulgaris* genome and the related microsynteny was visualised by MCscanX (Wang *et al.*, 2012). Pairs of homologous genes from the genomes of *Amaranthus hypochondriacus*, *Chenopodium quinoa*, *B. vulgaris* and *M. crystallinum* were identified using LAST with default parameters (Kielbasa *et al.*, 2011). Only restricted syntenic regions containing collinear genes along with their neighboring genes were evaluated.

## RESULTS

### *Ancestral state reconstruction of pigmentation in Caryophyllales suggests at least four origins of betalain pigmentation*

We inferred a time-calibrated, genus-level maximum likelihood species tree for 640 genera of Caryophyllales, constraining our inference to match the most recent phylogenomic hypotheses (Walker *et al.*, 2018). Our inferred topology and divergence times agree well with the current understanding of Caryophyllales (Walker *et al.*, 2018), with some minor or unsupported incongruences (Figs. 2, S3). Our updated pigmentation dataset contains data for 174 genera or 27% of genera represented in the tree topology. Maximum likelihood reconstruction under a symmetric (ER) and an asymmetric (ARD) model produced the same inferences with four transitions from anthocyanins to betalains predicted – in Stegnospermataceae, Amaranthaceae, the raphide clade (*sensu* Stevens, 2017), and the Portulacineae – and one reversal from betalains to anthocyanins in Kewaceae (Fig. S4). The ER model generated slightly more equivocal reconstructions along the backbone of Caryophyllales than the ARD model but statistical support for each model is nearly equivalent ( $\Delta\text{AIC}_{\text{ARD-ER}} = 0.23$ ). Posterior probabilities of node states from Bayesian reconstruction under both models similarly suggested four transitions from anthocyanins to betalains along the backbone of the tree and one reversal (Figs. 2, S4).

### *Betalain-pigmented species are inferred to contain a minimum of three DODA genes with at least two DODAA and one DODAB*

We used a baited search approach to populate an expanded *DODA* gene tree containing 318 Caryophyllales species from 34 families (increased from the 95 species and 26 families previously analysed; Brockington *et al.*, 2015). The phylogeny includes denser sampling from the anthocyanin-pigmented lineages Caryophyllaceae and Molluginaceae, and additionally samples the anthocyanin-

pigmented lineages Macarthuriaceae, Kewaceae and Limeaceae, and the betalain-pigmented lineage Stegnospermataceae. The phylogeny is mostly congruent with earlier analyses, revealing a gene duplication post the divergence of Physenaceae, resulting in two well-supported paralogues corresponding to DODAA $\alpha$  and DODAA $\beta$  clades in Brockington *et al.* (2015) (Figs. S5, S6). Further duplications have occurred, particularly in the DODAA $\alpha$  lineage, so that betalain-pigmented species are inferred to contain at least three genes predicted to encode a full-length DODA protein, with at least two DODAA $\alpha$  and one DODAA $\beta$  (Figs. S5, S6).

#### ***DODAA* homologues are present in a wide range of anthocyanic lineages across the Caryophyllales**

We performed a search for *DODA* loci across all anthocyanic lineages (Macarthuriaceae, Limeaceae, Kewaceae, Caryophyllaceae and Molluginaceae). Here, we recovered single *DODAA* genes from seven species representing four early diverging lineages of Caryophyllaceae: *P. tetraphyllum* (Polycarpaeae), *C. ramosissimum* and *P. campestris* (Paronychieae), *T. imperati* and *C. litoralis* (Corrigioleae), and *S. marina* and *S. arvensis* (Sperguleae; Figs. S5, S6). The *DODAA* genes from two of these species, *C. litoralis* and *P. campestris*, were found to have mutations causing premature stop codons (Fig. S7). We then searched for *DODA* genes in anthocyanic *M. australis* (Macarthuriaceae), *L. aethiopicum* (Limeaceae), *K. caespitosa* (Kewaceae), and *P. exiguum* (Molluginaceae) using transcriptome and genome sequencing. We detected a *DODAA* gene in all four species (Figs. S5, S6). Single *DODAA* genes were recovered from *M. australis* and *L. aethiopicum*, and two *DODAA* homologues were recovered from *K. caespitosa*. A *DODAA* gene could not be detected in *P. exiguum* despite ~206x short read coverage based on its estimated genome size (Pucker *et al.*, 2019).

#### ***DODAA* homologues exhibit no or negligible amounts of L-DOPA 4,5-dioxygenase activity**

We selected four betalain-pigmented species that represent each of the inferred origins of betalain pigmentation, three of which are represented by complete annotated genomes: *S. halimifolium* (Stegnospertaceae), *B. vulgaris* (Amaranthaceae; Dohm *et al.*, 2014), *M. crystallinum* (Aizoaceae; W. C. Yim, unpublished) and *C. gigantea* (Cactaceae; Copetti *et al.*, 2017). *S. halimifolium*, *M. crystallinum* and *C. gigantea* each contain one *DODAA* and two *DODAA*, and *B. vulgaris* contains one *DODAA* and five *DODAA* (Figs. S5, S6). The *DODA* genes for all four species were separately cloned into multi-gene vectors containing all necessary genes for betalain biosynthesis and these vectors were transiently transformed into *N. benthamiana* to test their heterologous expression (Fig. S1; Timoneda *et al.* 2018). We used the production of betacyanin in this heterologous assay as a proxy for L-DOPA 4,5-dioxygenase activity, with the relative strength of betacyanin production indicating the relative strength of L-DOPA 4,5-dioxygenase activity between loci. Upon heterologous expression, none of the *DODAA* genes from *S. halimifolium*, *B. vulgaris*, *M. crystallinum* and *C. gigantea* produced visible betacyanin pigmentation (Fig. 3a; Timoneda *et al.*, 2018). Traces of betanin were detected for all loci by HPLC, but amounts were extremely low compared to the positive control, BvDODAA1 (<0.1%; Fig. 3b).

***A single DODAA gene in each betalain-pigmented species exhibits high levels of L-DOPA 4,5-dioxygenase activity***

Using the same procedure as that used to test the DODA $\beta$  enzymes, we found that only a single DODA $\alpha$  from each study species showed a strong production of betalain pigmentation including BvDODA $\alpha$ 1 (Figs. 3, S8-S11). All paralogues exhibiting a high production of betanin in the heterologous assay (which we term as having high levels of L-DOPA 4,5-dioxygenase activity) are named henceforth “ $\alpha$ 1”. The amount of betanin produced by the different orthologues of DODA $\alpha$ 1 varies relative to BvDODA $\alpha$ 1, indicating that there may be differences in the effectiveness of these DODA $\alpha$  enzymes to convert L-DOPA to betalamic acid (Fig. 3b). For example, ShDODA $\alpha$ 1 produced a comparable amount of pigment to BvDODA $\alpha$ 1, CgDODA $\alpha$ 1 produced approximately 60% as much as BvDODA $\alpha$ 1, and McDODA $\alpha$ 1 produced approximately 80% more than BvDODA $\alpha$ 1.

***Numerous DODAA homologues in both betalain-pigmented and anthocyanin-pigmented lineages exhibit marginal levels of L-DOPA 4,5-dioxygenase activity***

We found that in each betalain-pigmented species there is at least one DODA $\alpha$  paralogue that exhibits marginal activity (consistent with Chung *et al.*, 2015 and Bean *et al.*, 2018). For example, in *M. crystallinum*, pigmentation was also observed for McDODA $\alpha$ 2, albeit at a much lower level than the DODA $\alpha$ 1 gene from *M. crystallinum* or any other species (Figs. 3a,b, S8-S11). Depending on the strength of transient expression in particular leaves, faint pigment was also sometimes observed for BvDODA $\alpha$ 2 and CgDODA $\alpha$ 2, from *B. vulgaris* and *C. gigantea* respectively (Fig. 3a). Quantification of betanin in these infiltration spots showed that McDODA $\alpha$ 2 produced approximately 5% the amount of betanin as BvDODA $\alpha$ 1, whereas BvDODA $\alpha$ 2 and CgDODA $\alpha$ 2 showed below 3% the amount of BvDODA $\alpha$ 1 (Figs. 3b, S9-11). Despite being visually undetectable, betanin could also be detected by HPLC in ShDODA $\alpha$ 2 and BvDODA $\alpha$ 4 (Figs. 3b, S8-9). For two *B. vulgaris* paralogues, BvDODA $\alpha$ 3 and BvDODA $\alpha$ 5, betanin was undetectable (Figs. 3b, S9). DODA $\alpha$  enzymes from anthocyanic species *M. australis* and *K. caespitosa* were also found to produce a small amount of betanin, as detected by HPLC (Figs. 4, S12). MaDODA $\alpha$  produces approximately 2.5% the amount of BvDODA $\alpha$ 1, and KcDODA $\alpha$ 1 and KcDODA $\alpha$ 2 produce approximately 12% and 2% of BvDODA $\alpha$ 1, respectively. Within Caryophyllaceae, DODA $\alpha$  from *C. ramosissimum* and *T. imperati* produced a small amount of betanin (6% and 2.5% of BvDODA $\alpha$ 1, respectively; Figs. 4b,c, S12).

***Betalain biosynthetic genes, DODAA1 and CYP76AD1, are colocalised in Amaranthaceae but not in the representative Aizoaceae genome***

We previously showed that BvDODA $\alpha$ 1 and BvCYP76AD1 are part of a putative gene cluster, being located close to one another (<50 kb) and also in close linkage with the MYB that regulates both of these genes (Keller, 1936; Goldman and Austin, 2000; Hatlestad *et al.*, 2014; Brockington *et al.*, 2015). In *C. quinoa* and *A. hypochondriacus*, other Amaranthaceae species for which there is a genome

sequence available (Yasui *et al.*, 2016; Lightfoot *et al.*, 2017), these genes are also colocalised (Fig. 5). However, in *M. crystallinum* in the family Aizoaceae, the high activity DODAA, *McDODAA1*, is located on a different chromosome from the *McCYP76AD1* orthologue, and the genes are not colocalised despite considerable conservation of synteny between putatively homologous chromosomes (Fig. 5). This analysis is limited to two of the four betalain-pigmented study species because they are currently the only species for which genome assemblies are of sufficient quality to allow syntenic analysis.

### ***Homologues exhibiting high levels of L-DOPA 4,5-dioxygenase activity are not monophyletic***

In order to understand the evolution of *L*-DOPA 4,5-dioxygenase activity, we mapped our functional data to the *DODA* gene tree (Fig. 6), and also mapped functional data from studies for which *DODA* activity has been tested *in planta*, either in heterologous transient assays or stable transgenics, or using recombinant expression in *E. coli* or yeast (Figs. 3, 4; Table S6). Mapping the functional data to the tree reveals that DODAA homologues exhibiting *L*-DOPA 4,5-dioxygenase activity have a homoplastic distribution across the *DODA* phylogeny (Fig. 6). In total, there are three polyphyletic gene lineages containing high levels of *L*-DOPA 4,5-dioxygenase activity (DODAA1): a lineage specific to *Stegnosperma* singly represented by *ShDODAA1*; a clade arising by duplication within Amaranthaceae containing *BvDODA1* and *CqDODA-1*, and a clade representing the remaining betalain-pigmented lineages, and containing *CgDODA1*, *MjDODAA1*, *McDODAA1*, *PmDOD* and *PgDODA* (Table S6). Each clade containing high-activity DODAA paralogues is sister to clades containing marginal activity DODAA paralogues, and in the case of Amaranthaceae, the DODAA1 clade is nested within clades exhibiting no or marginal levels of *L*-DOPA 4,5-dioxygenase activity (Fig. 6).

### ***Ancestral sequence reconstruction indicates that high L-DOPA 4,5-dioxygenase activity is a derived state within the DODAA lineage***

A recent study characterised seven residues that are important for *L*-DOPA 4,5-dioxygenase activity (Bean *et al.*, 2018). We carried out ancestral sequence reconstruction using coding sequences on a reduced *DODAA* gene tree (Figs. S13, S14) and observed that the three putative origins of highly active DODAA are marked by three convergent shifts in these seven residues (Figs. 6, 7), which occur post-gene duplication. For the residues inferred for the clade containing the *M. crystallinum* and *C. gigantea* highly active DODAA (DDFNDDI; Fig 7), 4/7 of the predicted residues are identical to those inferred for the Amaranthaceae highly active DODAA clade (DDYNDET). For *Stegnosperma*, the motif is more divergent, with three residues identical to the Amaranthaceae betalain-producing DODAA clade and two residues identical to the clade containing the *M. crystallinum* and *C. gigantea* highly active DODAA. The highly active DODAA are derived from ancestral nodes with inferred motifs that share more similarity with the marginal or no activity DODAA lineages (XGFNN[N/D]T), and this motif is highly conserved across the backbone and represented at almost all ancestral nodes (Figs. 6, 7). Amino acid reconstruction gave similar results (Fig. S15, S16).

## DISCUSSION

### *Polyphyletic patterns of elevated L-DOPA 4,5-dioxygenase activity support the recurrent specialisation to betalain pigmentation*

Since our earlier ancestral state pigment reconstructions (Brockington *et al.*, 2011, 2015), phylogenomic data have advanced new hypotheses for relationships within Caryophyllales (Walker *et al.*, 2018) and the previously uncharacterised Limeaceae and Simmondsiaceae have been shown to be anthocyanic (Thulin *et al.*, 2016). Here, we explicitly account for the large number of taxa for which pigmentation status is unknown and constrain our phylogenetic hypothesis to match the latest phylogenomic study of Walker *et al.* (2018), which places betalain-pigmented Stegnospermataceae as sister to anthocyanic Macarthuriaceae. In the context of these new data (Fig. 2), our reconstructions do not infer a single evolution of betalains, as previously suggested (Brockington *et al.* 2015), but rather imply up to four separate origins of betalain pigmentation in Caryophyllales.

To explore the hypothesis of multiple origins of betalain pigmentation suggested by our reconstructions, we then selected four species that represent each of the putative origins: *S. halimifolium*, *B. vulgaris*, *M. crystallinum* and *C. gigantea*. Using heterologous transient assays, we inferred relative levels of L-DOPA 4,5-dioxygenase activity based on the proxy of betanin production (Fig. 3). Our data show that activity is barely detectable in *DODAβ* loci from betalain-pigmented species, supporting our original hypothesis that the high levels of L-DOPA 4,5-dioxygenase activity evolved exclusively within the *DODAα* lineage (Brockington *et al.*, 2015). However, although betalain-pigmented species have multiple paralogues of *DODAα* genes, *only one* of these *DODAα* in each species encodes a protein which exhibits high levels of L-DOPA 4,5-dioxygenase activity. All betalain-pigmented species also exhibit *DODAα* paralogues with no or marginal L-DOPA 4,5-dioxygenase activity and we also detected marginal activity in several anthocyanic taxa (Fig. 4). Thus, marginal activity is widespread among Caryophyllales *DODAα* enzymes, suggesting broader underlying catalytic promiscuity in the *DODAα* lineage. Catalytic promiscuity is well documented in the broader protocatechuate dioxygenase gene family of which the LigB/*DODA* lineage is a member (Burroughs *et al.*, 2019). Such promiscuity is an important feature of metabolic evolution, potentially conferring evolvability (Weng & Noel, 2012; Leong & Last, 2017), and may have significant implications for the recurrent evolution of betalain pigmentation, as discussed below.

Previous analysis of the genome of *B. vulgaris* revealed a putative “gene cluster” (*sensu* Osbourne, 2010) in which *DODAα* and *CYP76AD1* are colocalised in chromosome 2 (Brockington *et al.*, 2015). The *B. vulgaris* locus with high levels of L-DOPA 4,5-dioxygenase activity falls within this operon, while the paralogues with no or marginal L-DOPA 4,5-dioxygenase activity occur outside of the gene cluster, supporting the concept of a betalain gene cluster in *B. vulgaris*. Our analysis, which describes the relative synteny of homologous loci between genomes, also shows that the betalain gene cluster appears to be conserved in the genomes of *C. quinoa* and *A. hypochondriacus* (Fig. 5). These divergent species all belong to the family Amaranthaceae and represent one putative origin of betalain

pigmentation (Fig. 2, **origin #2**). However, *CYP76AD1* and *DODAA1* are not clustered in the genome of *M. crystallinum*, which represents a different putative origin (Fig. 2, **origin #3**). Therefore, the absence of colocalisation in *M. crystallinum* highlights interesting structural genomic differences among putative betalain origins.

Mapping loci encoding functionally characterised DODA proteins to a comprehensively taxon-sampled *DODA* gene tree, we found that within the *DODAA* lineage, loci encoding proteins with high levels of *L*-DOPA 4,5-dioxygenase activity are not monophyletic (Fig. 6). Specifically, each clade containing high-activity *DODAA* paralogues is sister to clades containing marginal activity *DODAA*, suggesting polyphyletic origins of high activity, associated with gene duplication *within* the *DODAA* lineage. Previous research identified residues in seven sites which were necessary and sufficient to confer higher levels of *L*-DOPA 4,5-dioxygenase activity (Bean *et al.*, 2018). Phylogenetic reconstruction of these seven residues across the *DODAA* clade showed that polyphyletic clades containing proteins with high activity have distinctive motifs for these seven residues (Figs. 6, 7). Furthermore, motifs associated with high activity evolved at least three times from a background of motifs more similar to those proteins with no or marginal *L*-DOPA 4,5-dioxygenase activity. The diversity we recognise at these motifs in high activity *DODAA* sequences (e.g. between BvDODAA1 and ShDODAA1) indicates that high activity may arise from divergent sequence motifs and represent molecular convergence at key functional residues.

Intriguingly, the origins of high *L*-DOPA 4,5-dioxygenase activity following gene duplication, and associated residue shifts, are congruent with at least three of the four origins of betalain pigmentation inferred from our pigment reconstructions (Fig. 8). On the basis of these integrated observations we argue that betalain biosynthesis evolved multiple times in concert with recurrent gene duplication and neofunctionalisation *within* the *DODAA* clade, rather than as a single event at the base of the *DODAA* clade. Specifically, data in support of this model include: **1**) a background of marginal levels of *L*-DOPA 4,5-dioxygenase activity implying inherent evolvability of the ancestral enzyme (see below); **2**) polyphyletic origins of high *L*-DOPA 4,5-dioxygenase activity coincident with multiple inferred origins of betalain pigmentation; **3**) derived and convergent shifts in key residues underpinning high *L*-DOPA 4,5-dioxygenase activity. Together this model explains and conceptually links the homoplastic distribution of betalain lineages and the high levels of gene duplication observed in the *DODAA* clade.

Polyphyletic origins of betalain pigmentation occur only within Caryophyllales (Fig. 2), while polyphyletic origins of high *L*-DOPA 4,5-dioxygenase activity are constrained to the Caryophyllales-specific *DODAA* clade (Fig. 6). Both patterns speak to the concept of evolutionary precursors in which underlying evolutionary state(s) potentiate recurrent evolution of subsequent complex traits (*sensu* Marazzi *et al.* 2012). In this scenario, we propose that an initial precursor step was the evolution of tyrosine hydroxylase activity by duplication in the *CYP76AD* lineage, leading to an abundance of *L*-DOPA (DeLoache *et al.*, 2015; Polturak *et al.*, 2016; Sunnadaniya *et al.*, 2016), the necessary substrate for betalamic acid biosynthesis. Given an abundance of *L*-DOPA, a subsequent precursor step was duplication in the *DODA* lineage that gave rise to a clade of *DODAA* enzymes, whose ancestral function is currently unknown, but presumably with some promiscuous ability to act on *L*-DOPA to

produce trace betalamic acid. Subsequently, further repeated duplication within the *DODAA* lineage led to recurrent neofunctionalisation towards high levels of *L*-DOPA 4,5-dioxygenase activity and the production of betalamic acid.

Such a model, in which similar enzymatic function has arisen repeatedly from homologous but non-orthologous enzymes, is not unprecedented. Many studies indicate that this form of convergent evolution is widespread in specialised plant metabolism (Pichersky and Lewinsohn, 2011). For example: the enzymes that methylate purine intermediates in caffeine biosynthesis in *Coffea* versus *Thea* evolved from different branches of the SABATH carboxyl methyltransferase family (Yoneyama *et al.*, 2006); and, disparate origins of pyrrolizidine alkaloids have arisen through recurrent evolution of homospermidine synthase from the ubiquitous enzyme deoxyhypusine synthase (Reimann *et al.*, 2004). However, the detection of this same phenomenon in the evolution of betalains is perhaps surprising, because the retention of enzymes for anthocyanin biosynthesis in betalain-pigmented species (Shimada *et al.*, 2004; 2005) has encouraged the assumption that reversals to anthocyanin are more likely than multiple shifts to betalain pigmentation (Brockington *et al.* 2011, 2015).

#### ***Reconciling polyphyletic origins of high L-DOPA 4,5-dioxygenase activity with DODAA gene loss***

It has always been striking that each major betalain pigmented clade is subtended at, or towards its base, by an anthocyanic lineage (Brockington *et al.*, 2011). Given this essential pattern, our trait reconstructions suggest multiple origins of betalain pigmentation. In turn, this implies that many lineages are anthocyanic through retention of an ancestral state, rather than through reversal from betalain pigmentation (Brockington *et al.*, 2011; Fig. 2). Yet, we earlier reported that *DODAA* are lost or down-regulated in anthocyanic Caryophyllaceae and Molluginaceae, and previously argued that loss of *DODAA* in these anthocyanic lineages is consistent with reversals from betalain pigmentation to anthocyanin pigmentation (Brockington *et al.*, 2015). However, the emerging picture is more complex, and with new data presented here, we find: **1)** evidence of *DODAA* loci in all but one of the anthocyanic lineages in Caryophyllales; **2)** evidence of *DODAA* gene loss in Caryophyllaceae and Molluginaceae, but no evidence of *DODAA* gene loss in the anthocyanin-pigmented lineages, *K. caespitosa*, *M. australis*, and *L. aethiopicum*, and; **3)** evidence of *DODAA* loci with marginal *L*-DOPA 4,5-dioxygenase activity within anthocyanic *M. australis*, *K. caespitosa*, *C. ramosissimum* and *T. imperati*. Clearly, evolutionary disparate anthocyanic lineages show different patterns of molecular evolution with respect to *DODAA*. Given this complex evolutionary milieu, and in the face of compelling evidence for polyphyletic origins of elevated *L*-DOPA 4,5-dioxygenase activity, below we propose two alternative hypotheses to explain loss of *DODAA* loci in anthocyanic lineages.

- 1) The patterns we detect may evidence lability in early stages of betalain evolution. Anthocyanins and betalains have never been found to co-occur, however, it is possible that the two classes of pigment did co-occur in early evolutionary stages. In this scenario, repeated evolution of increased *L*-DOPA 4,5-dioxygenase activity allowed for betalain pigmentation, initially co-occurring with anthocyanins. But evolution of an integrated betalain pathway requires more than biosynthetic enzymes (e.g. the recruitment of the MYB transcriptional

- regulators; Hatlestad *et al.*, 2014). Therefore, establishment and enhancement of the betalain pathway was only achieved in certain lineages, those which specialised to betalain pigmentation. In contrast, other lineages arising close to the origins of elevated *L*-DOPA 4,5-dioxygenase activity, specialised to anthocyanins rather than betalains, ultimately losing the *DODA* $\alpha$  paralogues with elevated *L*-DOPA 4,5-dioxygenase activity. In these cases, anthocyanic lineages may have retained anthocyanins from ancestors in which both pigments co-existed, and inferences of reversals to anthocyanins based on *DODA* $\alpha$  loss are misleading. This scenario is appealing because we do not see any anthocyanin lineages that are nested deeply within the betalain-pigmented clades that stem from each inferred origin (Fig. 2), and so shifts to anthocyanin pigmentation seem less likely with evolutionary distance from inferred origins of betalain pigmentation.
- 2) Previously, Lopez-Nieves *et al.* (2018) speculated that the evolution of tyrosine-derived betalains occurred in a metabolic environment enriched for tyrosine. The arogenate dehydrogenase enzyme (*ADH* $\alpha$ ) responsible for increased tyrosine availability is also lost and/or down-regulated in the anthocyanic Caryophyllaceae and Molluginaceae lineages (Lopez-Nieves *et al.*, 2018). Therefore, the shift to anthocyanins in these taxa may be indicative of deeper shifts away from a tyrosine-rich metabolism towards a metabolism in which phenylalanine plays a canonical role. The primary function of the *DODA* $\alpha$  paralogues with marginal *L*-DOPA 4,5-dioxygenase activity is unknown, but may catalyse production of tyrosine-derived metabolites, other than betalains. In this scenario, the duplication that gave rise to the *DODA* $\beta$  and *DODA* $\alpha$  lineages led to neofunctionalisation in the *DODA* $\alpha$  lineage towards an unknown, but tyrosine-derived enzymatic activity, with marginal *L*-DOPA 4,5-dioxygenase activity. Loss of *DODA* $\alpha$  in Caryophyllaceae and Molluginaceae could instead reflect loss of the unknown tyrosine-derived enzymatic activity, in the context of shifts towards more phenylalanine-biased metabolism, independent of origins of high *L*-DOPA 4,5-dioxygenase activity.

## CONCLUSION

The evolutionary origin of betalain pigments in Caryophyllales and the processes that led to their homoplastic distribution have been the subject of much debate. Our new data and analyses offer compelling evidence for recurrent specialisation to betalain biosynthesis. Specifically: 1) a background of marginal levels of *L*-DOPA 4,5-dioxygenase activity implying inherent evolvability; 2) polyphyletic origins of high *L*-DOPA 4,5-dioxygenase activity coincident with multiple inferred origins of betalain pigmentation; 3) derived and convergent shifts in key residues underpinning high *L*-DOPA 4,5-dioxygenase activity, and; 4) lack of conservation of the betalain metabolic gene cluster between putative origins of betalain pigmentation. But our hypothesis requires future experimentation. First, it will be important to identify the primary function of those *DODA* $\alpha$  proteins that only exhibit marginal *L*-DOPA 4,5-dioxygenase activity. Elucidation of this unknown function will further inform the inferences made in this study and direct future hypotheses. Further validation could also emerge by considering other aspects of the betalain biosynthesis pathway. For example, as exemplified by our

syntenic analyses, it may be possible to discern the signal of multiple origins at different hierarchical levels of the betalain pathway, including: patterns of gene clustering by genomic colocalisation; patterns of co-option of transcriptional regulators and other genes; and the dissection of the molecular convergence signal at key residues. It is fortunate here that three of the putative origins of betalain biosynthesis are represented by well-resourced experimental systems: *Portulaca grandiflora*, *Mirabilis jalapa* and *B. vulgaris*, which together promise rapid progress in this era of betalain renaissance.

## ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their comments. We thank Hiroshi Maeda, Ya Yang, Fernando Gandía-Herrero and Joseph Walker for critical reading of the manuscript, Fernando Gandía-Herrero for providing the betanin standard, and Syngenta for providing the *B. vulgaris* YTiBv seeds. We thank Cambridge University Botanic Garden for growing various species included in this study, the Desert Botanical Garden for access to Aizoaceae samples, and Kevin Thiele for the provision of *Macarthuria australis* seed. We acknowledge support from the following funding bodies: SFB, BBSRC High Value Chemicals from Plants Network; HS, SNF P2BEP3\_165359 & P300PA\_174333; TF, U1802232 & XDA20050203; NWH, Woolf Fisher Cambridge Scholarship; MJM, NSF DEB 1352907; SS, NSF DEB 1354048; JCC, DOE, GSP DE-SC0008834.

## AUTHOR CONTRIBUTION

SFB, HS, TF, and NWH planned and designed the research; SFB, HS, and NWH wrote the manuscript; HS, TF, NWH, SLN, BP, RG, WCY, AT, RB, MJS and SAS performed experiments and analysed the data; MJM, HT, LZ, SAS, WCY, MJS, DC, and JCC provided sequence data. All authors read and approved the manuscript.

## REFERENCES

- Bate-Smith EC. 1962.** The phenolic constituents of plants and their taxonomic significance. *Botanical Journal of the Linnean Society* **60**: 325–356.
- Bean A, Sunnadeniya R, Akhavan N, Campbell A, Brown M, Lloyd A, Lloyd A. 2018.** Gain-of-function mutations in beet DODA2 identify key residues for betalain pigment evolution. *New Phytologist* **219**: 287-296.
- Bischoff H. 1876.** *Das Caryophyllinenroth*. Inaugural dissertation, University of Tübingen, Tübingen, Germany.
- Bollback JP. 2006.** SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* **7**: 88.
- Burroughs AM, Glasner ME, Barry KP, Taylor EA, Aravind L. 2019.** Oxidative opening of the aromatic ring: tracing the natural history of a large superfamily of dioxygenase domains and their relatives. *Journal of Biological Chemistry* **294**: 10211-10235.

- Brockington SF, Walker RH, Glover BJ, Soltis PS, Soltis DE. 2011.** Complex pigment evolution in the Caryophyllales. *New Phytologist* **190**: 854–864.
- Brockington SF, Yang Y, Gandia-Herrero F, Covshoff S, Hibberd JM, Sage RF, Wong GKS, Moore MJ, Smith SA. 2015.** Lineage-specific gene radiations underlie the evolution of novel betalain pigmentation in Caryophyllales. *New Phytologist* **207**: 1170–1180.
- Christinet L, Burdet FX, Zaiko M, Hinz U, Zrýd 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.
- Chung H-H, Schwinn KE, Ngo HM, Lewis DH, Massey B, Calcott KE, Crowhurst R, Joyce DC, Gould KS, Davies KM, et al. 2015.** Characterisation of betalain biosynthesis in *Parakeelya* flowers identifies the key biosynthetic gene DOD as belonging to an expanded LigB gene family that is conserved in betalain-producing species. *Frontiers in Plant Science* **6**: 1–16.
- Clement JS, Mabry TJ. 1996.** Pigment evolution in the Caryophyllales: a systematic overview. *Botanica Acta* **109**: 360–367.
- Copetti D, Búrquez A, Bustamante E, Charboneau JLM, Childs KL, Eguiarte LE, Lee S, Liu TL, McMahon MM, Whiteman NK, et al. 2017.** Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. *Proceedings of the National Academy of Sciences, USA* **114**: 12003–12008.
- DeLoache WC, Russ ZN, Narcross L, Gonzales AM, Martin VJJ, Dueber JE. 2015.** An enzyme-coupled biosensor enables (S)-reticuline production in yeast from glucose. *Nature Chemical Biology* **11**: 465–471.
- Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez 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**: 546–9.
- Engler C, Youles M, Gruetzner R, Ehnert TM, Werner S, Jones JDG, Patron NJ, Marillonnet S. 2014.** A Golden Gate modular cloning toolbox for plants. *ACS Synthetic Biology* **3**: 839–843.
- 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 and Applied Genetics* **100**: 337–343.
- Haak M, Vinke S, Keller W, Droste J, Rückert C, Kalinowski J, Pucker B. 2018.** High quality *de novo* transcriptome assembly of *Croton tiglium*. *Frontiers in Molecular Bioscience* **5**: 62.
- Hatlestad GJ, Akhavan N a, 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
- Huelsenbeck JP, Nielsen R, Bollback JP. 2003.** Stochastic mapping of morphological characters. *Systematic Biology* **52**: 131–158.

- Keller W. 1936.** Inheritance of some major color types in beets. *Journal of Agricultural Research* **52**: 27–38.
- Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011.** Adaptive seeds tame genomic sequence comparison. *Genome Research* **21**: 487–493.
- 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.
- Leong BJ, Last RL. 2017.** Promiscuity, impersonation and accommodation: evolution of plant specialized metabolism. *Current Opinion in Structural Biology* **47**: 105–112.
- Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ. 2017.** Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. *BMC Biology* **15**: 74.
- Lopez-Nieves S, Yang Y, Timoneda A, Wang M, Feng T, Smith SA, Brockington SF, Maeda HA. 2018.** Relaxation of tyrosine pathway regulation underlies the evolution of betalain pigmentation in Caryophyllales. *New Phytologist* **217**: 896–908.
- Marazzi B, Ane C, Simon MF, Delgado-Salinas A, Luckow M, Sanderson MJ. 2012.** Locating evolutionary precursors on a phylogenetic tree. *Evolution* **66**: 3918–3930
- Musso H. 1979.** The pigments of fly agaric, *Amanita muscaria*. *Tetrahedron* **35**: 2843–2853.
- Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015.** IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**: 268–274.
- Osbourn A. 2010.** Gene clusters for secondary metabolic pathways: an emerging theme in plant biology. *Plant Physiology* **154**: 531–535.
- Pagel M. 1994.** Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society B* **255**: 37–45.
- Pagel M. 1999.** The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology* **48**: 612–622.
- Paradis E, Schliep K. 2019.** Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**: 526–528.
- Pichersky E and Lewinsohn E. 2011.** Convergent evolution in plant specialized metabolism. *Annual Review of Plant Biology* **62**: 549–566.
- Polturak G, Breitel D, Grossman N, Sarrion-Perdigones A, Weithorn E, Pliner M, Orzaez D, Granell A, Rogachev I, Aharoni A. 2016.** Elucidation of the first committed step in betalain biosynthesis enables the heterologous engineering of betalain pigments in plants. *New Phytologist* **210**: 269–283.
- Polturak G, Heinig U, Grossman N, Battat M, Leshkowitz D, Malitsky S, Rogachev I, Aharoni A. 2018.** Transcriptome and metabolic profiling provides insights into betalain biosynthesis and evolution in *Mirabilis jalapa*. *Molecular Plant* **11**: 189–204.
- Pucker B, Feng T, Brockington SF. 2019.** Next generation sequencing to investigate genomic diversity in Caryophyllales. *bioRxiv*. doi: <https://doi.org/10.1101/646133>
- R Core Team. 2019.** R: a language and environment for statistical computing. Version 3.6.1. R

- Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>
- Reimann A, Nurhayati N, Backenköhler A, Ober D. 2004.** Repeated evolution of the pyrrolizidine alkaloid-mediated defense system in separate angiosperm lineages. *Plant Cell* **16**: 2772–84.
- Ren M, Chen Q, Li L, Zhang R, Guo S. 2005.** Successive chromosome walking by compatible ends ligation inverse PCR. *Molecular Biotechnology* **30**: 95–101.
- Revell LJ. 2012.** phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**: 217–223.
- Sanderson MJ. 2002.** Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution* **19**: 101–109.
- 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 Cell Physiology* **50**: 1012–1016.
- Shimada S, Inoue YT, Sakuta M. 2005.** Anthocyanidin synthase in non-anthocyanin-producing Caryophyllales species. *Plant Journal* **44**: 950–959.
- Shimada S, Takahashi K, Sato Y, Sakuta M. 2004.** Dihydroflavonol 4-reductase cDNA from non-anthocyanin-producing species in the Caryophyllales. *Plant & Cell Physiology* **45**: 1290–1298.
- Smith SA, O’Meara BC. 2012.** TreePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**: 2689–2690.
- Smith SA, Walker JF. 2019.** PyPHLAWD: a python tool for phylogenetic dataset construction. *Methods in Ecology and Evolution* **10**: 104–108.
- Stafford HA. 1994.** Anthocyanins and betalains: evolution of the mutually exclusive pathways. *Plant Science* **101**: 91–98.
- Stevens P. 2017.** *Angiosperm Phylogeny Website*. [WWW document] URL <http://www.mobot.org/MOBOT/research/APweb/>. [accessed 20 May 2019]
- Sunnadeniya R, Bean A, Brown M, Akhavan N, Hatlestad G, Gonzalez A, Symonds VV, Lloyd A. 2016.** Tyrosine hydroxylation in betalain pigment biosynthesis is performed by cytochrome P450 enzymes in beets (*Beta vulgaris*). *PLoS One* **11**: e0149417.
- Thulin M, Larsson A, Edwards EJ, Moore AJ. 2018.** Phylogeny and systematics of *Kewa* (Kewaceae). *Systematic Botany* **43**: 689–700.
- Thulin M, Moore AJ, El-seedi H, Larsson A, Christin P, Edwards EJ. 2016.** Phylogeny and generic delimitation in Molluginaceae, new pigment data in Caryophyllales, and the new family Corbichoniaceae. *Taxon* **65**: 775–793.
- Timoneda A, Feng T, Sheehan H, Walker-Hale N, Pucker B, Lopez-Nieves S, Guo R, Brockington SF. 2019.** The evolution of betalain biosynthesis in Caryophyllales. *New Phytologist*. doi: 10.1111/nph.15980
- Timoneda A, Sheehan H, Feng T, Lopez-Nieves S, Maeda HA, Brockington SF. 2018.** Redirecting primary metabolism to boost production of tyrosine-derived specialised metabolites *in planta*. *Scientific Reports* **8**: 17256.
- Walker JF, Yang Y, Feng T, Timoneda A, Mikenas J, Hutchinson V, Edwards C, Wang N, Ahluwalia S, Olivieri J, et al. 2018.** From cacti to carnivores: improved phylotranscriptomic

sampling and hierarchical homology inference provides further insight to the evolution of Caryophyllales. *American Journal of Botany* **105**: 446–462.

- Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, Lee T, Jin H, Marler B, Guo H, et al. 2012.** MCSanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**: e49.
- Weng J-K. 2014.** The evolutionary paths towards complexity : a metabolic perspective. *New Phytologist* **201**: 1141–1149.
- Weng J-K, Noel JP. 2012.** The remarkable pliability and promiscuity of specialized metabolism. *Cold Spring Harbour Symposia on Quantitative Biology* **77**: 309–320.
- Yasui Y, Hirakawa H, Oikawa T, Toyoshima M, Matsuzaki C, Ueno M, Mizuno N, Nagatoshi Y, Imamura T, Miyago M, et al. 2016.** Draft genome sequence of an inbred line of *Chenopodium quinoa*, an allotetraploid crop with great environmental adaptability and outstanding nutritional properties. *DNA Research* **23**: 535-546.
- Yoneyama N, Morimoto H, Ye CX, Ashihara H, Mizuno K, Kato M. 2006.** Substrate specificity of *N*-methyltransferase involved in purine alkaloids synthesis is dependent upon one amino acid residue of the enzyme. *Molecular Genetics and Genomics* **275**:125–35.

## FIGURE LEGENDS

**Fig. 1 The betalain biosynthetic pathway.** A schematic showing the enzymatic and spontaneous reactions that form the specialised metabolites, betalains. The focus of this study is *L*-DOPA 4,5-dioxygenase (DODA; highlighted in red) which catalyses the formation of betalamic acid, the core chromophore of betalain pigments, from *L*-3,4-dihydroxyphenylalanine (*L*-DOPA). This figure was adapted from Timoneda *et al.* (2019).

**Fig. 2 Reconstruction of pigment state across the Caryophyllales supports four origins of betalain pigmentation.** (a) Bayesian ancestral state reconstructions of pigmentation on a time-calibrated, genus-level maximum likelihood species tree of Caryophyllales, inferred from seven nuclear and plastid markers. Tips are coloured according to pigmentation state: anthocyanin (blue), betalain (pink) and unknown (grey). Pie charts at nodes give posterior probabilities at nodes for anthocyanin (blue) and betalain (pink), inferred from  $n = 1000$  stochastic mapping simulations under the asymmetric (ARD) model of character evolution. (b) The reconstruction shown in (a) simplified to show family-level relationships. The numbers represent the four inferred origins of betalain pigmentation. Branches in bold indicate relationships that were constrained during inference.

**Fig. 3 Betalain-pigmented species have a single DODA that has high activity when heterologously expressed in *Nicotiana benthamiana* and other DODA that have no or marginal activity.** (a) A representative leaf is shown from the agroinfiltration of *L*-DOPA 4,5-dioxygenase (*DODA*) genes from *Stegnosperma halimifolium*, *Beta vulgaris*, *Carnegiea gigantea* and *Mesembryanthemum crystallinum*.

The *DODA* coding sequences were cloned into multi-gene constructs containing the other structural genes necessary for betacyanin production (*BvCYP76AD1*, *MjcDOPA-5GT*). The infiltration spots are labelled according to the *DODA* variant. For all species, the pBC-BvDODA $\alpha$ 1 multi-gene vector is included as a positive control (P). **(b)** The betanin content of the infiltration spots was measured using HPLC and data were represented relative to the BvDODA $\alpha$ 1 spot present in each biological replicate. The data were combined after calculating relative amounts for each species from individual species-specific experiments (Supporting Information Figs. **S8-S11**). Bars show means  $\pm$  SD;  $n = 5$  for *S. halimifolium*, *B. vulgaris*, *C. gigantea* and *M. crystallinum*, except for BvDODA $\alpha$ 2 ( $n = 4$ ) and BvDODA $\alpha$ 3 ( $n = 4$ ). **(c)** A representative HPLC trace for the *DODAA1* gene from each species is shown. The left peak is betanin and the right peak is its isomer, isobetanim. The traces are offset for presentation. ‘pLUC’ is a negative control plasmid carrying the firefly luciferase gene and ‘Betanin’ is a commercially available extract from beet hypocotyl which was used to validate the retention time of betanin in these samples.

**Fig. 4 *L-DOPA* 4,5-dioxygenase activity is marginal or absent in *DODA* from anthocyanin-producing taxa when heterologously expressed in *Nicotiana benthamiana*.** **(a)** A representative leaf is shown from the agroinfiltration of *L-DOPA* 4,5-dioxygenase (*DODA*) genes from *Macarthuria australis* (*MaDODAA*), *Telephium imperati* (*TiDODAA*), *Polycarpon tetraphyllum* (*PtDODAA*), *Cardionema ramosissimum* (*CrDODAA*), *Spergularia marina* (*SmDODAA*), *Limeum aethiopicum* (*LaDODAA*) and *Kewa caespitosa* (*KcDODAA1*, *KcDODAA2*). The *DODA* coding sequences were cloned into multi-gene constructs with the other structural genes necessary for betacyanin production (*BvCYP76AD1*, *MjcDOPA-5GT*). The pBC-BvDODA $\alpha$ 1 multi-gene vector was included as a positive control (P). **(b)** Betanin content of the infiltration spots was measured using HPLC. Amounts of betanin were calculated relative to the average amount of betanin present in BvDODA $\alpha$ 1 infiltration spots for each species and the data combined into one graph (see Supporting Information Fig. **S12 for species-specific data**). Bars show means  $\pm$  SD;  $n \geq 3$ . **(c)** A representative HPLC trace for each *DODAA* gene from each species is shown. The left peak is betanin and the right peak is its isomer, isobetanim. The traces are offset for presentation. ‘pLUC’ is a negative control plasmid carrying the firefly luciferase gene and ‘Betanin’ is a commercially available extract from beet hypocotyl, which was used to validate the retention time of betanin in these samples.

**Fig. 5 Betalain biosynthesis genes are colocalised in the genomes of three Amaranthaceae species but not in the Aizoaceae species, *Mesembryanthemum crystallinum*.** Shown are the genomic regions containing the betalain biosynthetic genes, *DODAA1* and *CYP76AD1*, from *Amaranthus hypochondriacus*, *Beta vulgaris*, *Chenopodium quinoa* and *Mesembryanthemum crystallinum*. Rectangles show predicted gene models, and blue (+ strand) and green (- strand) colours indicate relative orientations. Matching gene pairs are displayed as grey connections, and purple and red connections indicate *DODAA1* and *CYP76AD1* correspondence, respectively. Whole genome duplication is marked with a dotted line.

**Fig. 6 The DODA gene tree shows a homoplasious distribution of functionally characterised DODA genes that produce a high level of betalain pigments.** The maximum likelihood phylogeny of Caryophyllales *L-DOPA* 4,5-dioxygenase (*DODA*) genes inferred from coding sequences derived from genomes and transcriptomes. Branch lengths are expected number of substitutions per site. Scale bar gives 0.2 expected substitutions per site. Branch labels are support values for major paralogous clades from rapid bootstrapping and the SH-like Approximate Likelihood Ratio Test, respectively, given as RBS/SH-aLRT. Putative major duplication nodes are highlighted with stars. Branches are coloured according to the putative pigmentation state of the taxa (blue: anthocyanin, pink: betalain). Labelled tips show functionally characterised DODAs and shades squares correspond to DODA activity (white, no activity; grey, marginal activity; black, high activity). Annotated at tips is a colour-coded alignment of the seven residues conferring high activity reported in Bean *et al.* (2018). *XxDODA*: *Bv*, *Beta vulgaris*; *Cg*, *Carnegiea gigantea*; *Cr*, *Cardionema ramosissimum*; *Ds*, *Dianthus superbus*; *Kc*, *Kewa caespitosa*; *La*, *Limeum aethiopicum*; *Mc*, *Mesembryanthemum crystallinum*; *Ma*, *Macarthuria australis*; *Mj*, *Mirabilis jalapa*; *Pg*, *Portulaca grandiflora*; *Phyb*, *Ptilotus* hybrid; *Pm*, *Parakeelya mirabilis*; *Pt*, *Polycarpon tetraphyllum*; *Sh*, *Stegnosperma halimifolium*; *Sm*, *Spergularia marina*; *Ti*, *Telephium imperati*.

**Fig. 7 Reconstruction of seven functionally important amino acids on the *L-DOPA* 4,5-dioxygenase (*DODA*) phylogeny show a shift in patterns associated with *DODA* genes of high *L-DOPA* 4,5-dioxygenase activity.** (a) Proportional logo plots of the seven residues identified by Bean *et al.* (2018) across functionally characterised high activity (left) and no or marginal activity (right) paralogues included in this study. 1: ShDODA $\alpha$ 1 vs. ShDODA $\alpha$ 2 (Stegnospermataceae); 2: CqDODA-1, BvDODA $\alpha$ 1 vs. PhybDODA $\alpha$ , BvDODA $\alpha$ 2-5 (Amaranthaceae); 3: McDODA $\alpha$ 1, PmDODA $\alpha$ 1, CgDODA $\alpha$ 1, PgDODA $\alpha$ 1 vs. McDODA $\alpha$ 2, PmDODA $\alpha$ 2, CgDODA $\alpha$ 2 (Portulacineae and Aizoaceae). Residue numbering is according to Bean *et al.*, (2018). (b) Ancestral sequence reconstructions of codons by empirical Bayesian inference on a maximum likelihood phylogeny of a reduced DODA $\alpha$  sequence alignment. Logo plots at nodes give the posterior probability of an amino acid (translated from codons and summed over redundantly translated codons) at each of the seven sites identified by Bean *et al.* (2018). Numbers at nodes match the ancestor of the high-activity paralogues summarised in (a).

**Fig. 8 Summary of the major evolutionary changes inferred with respect to transitions in pigment type and corresponding transitions in *L-DOPA* 4,5-dioxygenase activity.** (a) Simplified pigment reconstruction shows family-level relationships with numbers in purple circles representing the four inferred origins of betalain pigmentation. Tips are coloured according to pigmentation state: anthocyanin (blue), betalain (pink) and unknown (grey). (b) Mirror image of pigment reconstruction topology. A black star marks the initial *DODA* $\alpha$ /*DODA* $\beta$  duplication. Purple stars indicate inferred phylogenetic location of *DODA* $\alpha$  gene duplications giving rise to paralogues with high *L-DOPA* 4,5-dioxygenase activity. Black crosses indicate loss of a *DODA* $\alpha$  paralogue. Tips show inferred presence or absence of *DODA* $\alpha$ /*DODA* $\beta$ : black squares indicate presence of *DODA* $\alpha$ /*DODA* $\beta$ , white squares

with a cross indicate loss of *DODA $\alpha$* , white squares with a dashed boundary indicate missing data, *DODA $\alpha$*  squares coloured purple are inferred to have high *L*-DOPA 4,5-dioxygenase activity.

## SUPPORTING INFORMATION

**Fig. S1** Taxon-labeled *DODA $\alpha$*  gene tree with nucleotide branch lengths and support values

**Fig. S2** Schematic of binary vectors used in this study

**Fig. S3** Maximum likelihood, time-calibrated genus-level phylogeny of Caryophyllales, inferred from seven nuclear and plastid markers

**Fig. S4** Ancestral state reconstruction of pigments by maximum likelihood and Bayesian inference via stochastic mapping.

**Fig. S5** A simplified representation of the *DODA* gene tree

**Fig. S6** Taxon-labeled maximum likelihood gene tree of *DODA* in Caryophyllales with support values (pdf)

**Fig. S7** The *Corrigiola litoralis* and *Pollichia campestris* *DODA $\alpha$*  genes are putative pseudogenes

**Fig. S8** Functional characterisation of the full complement of *DODA* genes from *Stegnosperma halimifolium*

**Fig. S9** Functional characterisation of the full complement of *DODA* genes from *Beta vulgaris*

**Fig. S10** Functional characterisation of the full complement of *DODA* genes from *Mesembryanthemum crystallinum*

**Fig. S11** Functional characterisation of the full complement of *DODA* genes from *Carnegiea gigantea*

**Fig. S12** Functional characterisation of *DODA $\alpha$*  genes from anthocyanic Caryophyllales taxa

**Fig. S13** Taxon-labeled subsampled maximum likelihood *DODA $\alpha$*  gene tree with nucleotide branch lengths and support values

**Fig. S14** Taxon-labeled subsampled maximum likelihood *DODA $\alpha$*  gene tree with codon branch lengths

**Fig. S15** Taxon-labeled subsampled maximum likelihood *DODA $\alpha$*  gene tree with amino acid branch lengths

**Fig. S16** Ancestral amino acid sequence reconstruction for seven residues on subsampled *DODA $\alpha$*  gene tree

**Table S1** Oligonucleotide primers used in this study

**Table S2** Details of *DODA* genes functionally characterised in this study

**Table S3** Pigmentation status of 174 genera of Caryophyllales gathered from the literature

**Table S4** Genome and transcriptome assemblies used in this study

**Table S5** *DODA* sequences used in *DODA* phylogeny

**Table S6** Details of *DODA* genes that have been functionally characterised from Caryophyllales species and mapped to the *DODA* gene tree

**Methods S1** Genome assembly

**Methods S2** Identification of full complements of *DODA* genes from betalain-pigmented species for functional analysis

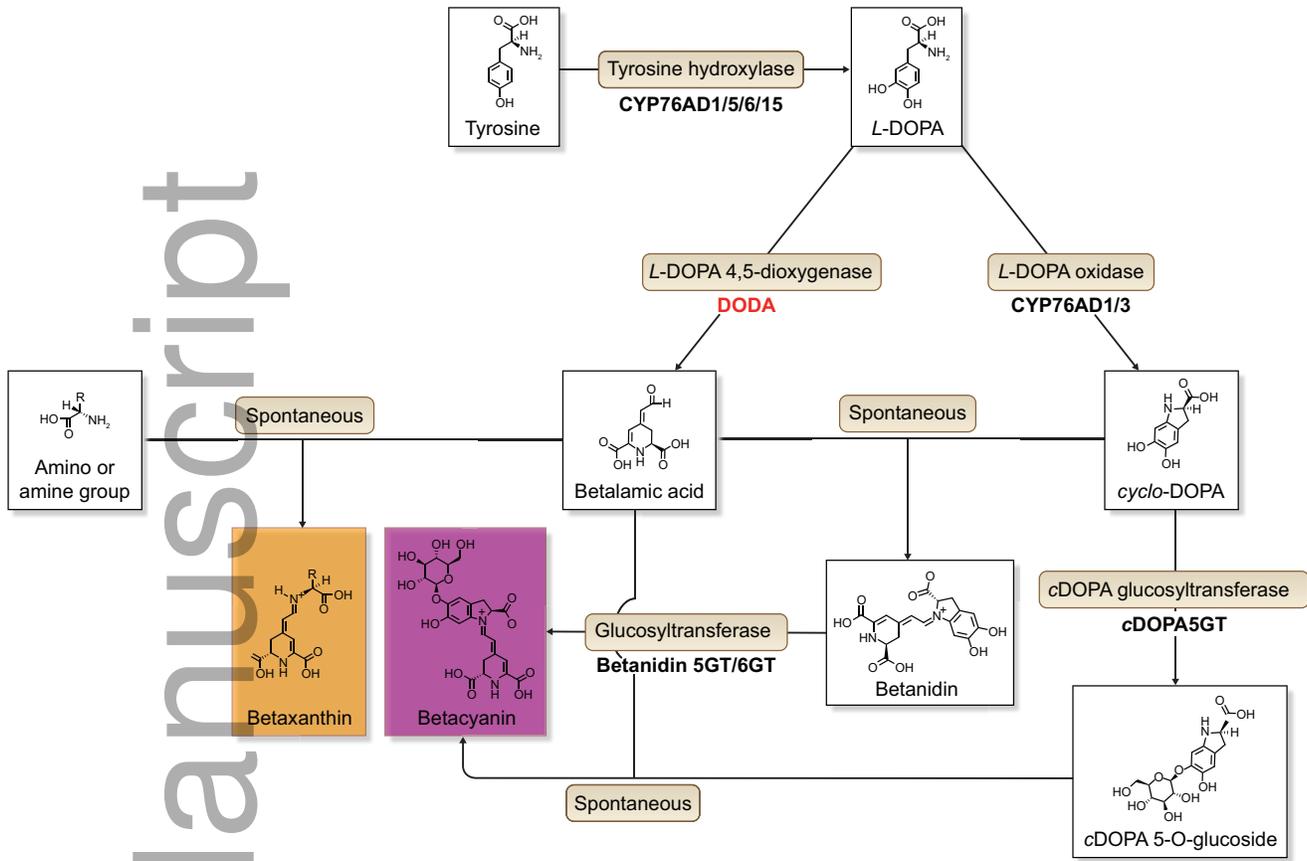
**Methods S3** Phylogeny of Caryophyllales

**Methods S4** Pigment reconstruction

**Methods S5** Gene tree of Caryophyllales *DODA* genes

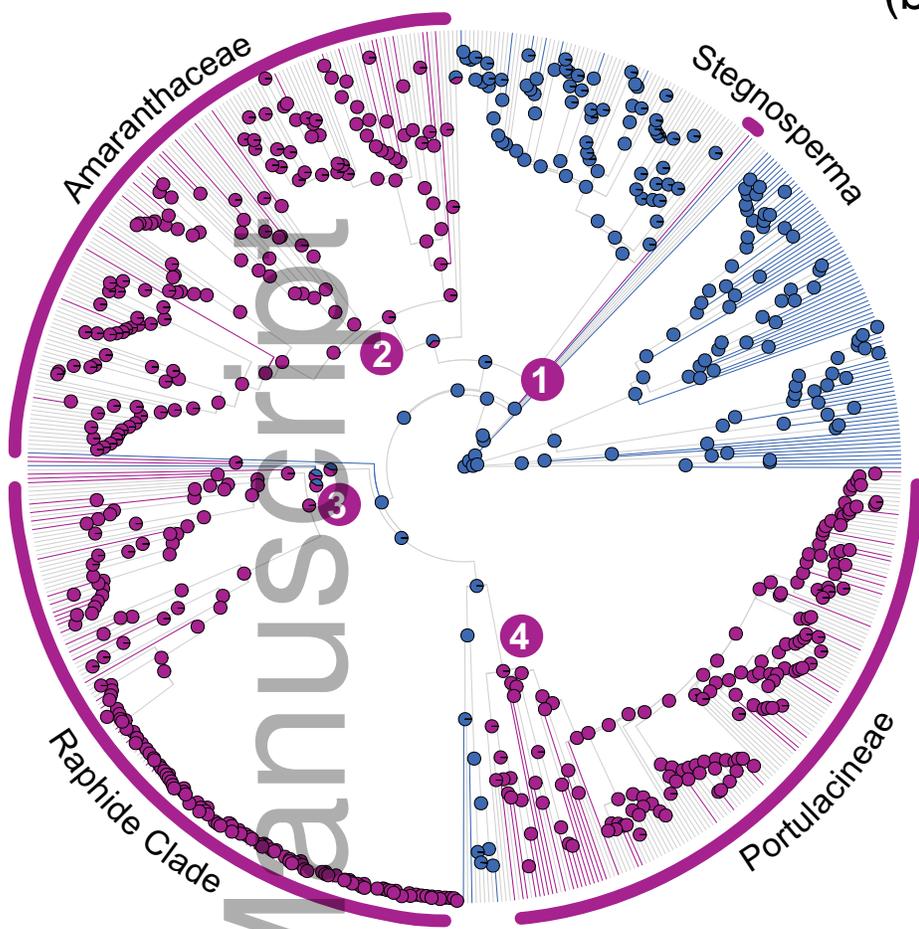
**Methods S6** Ancestral Sequence Reconstruction

Author Manuscript

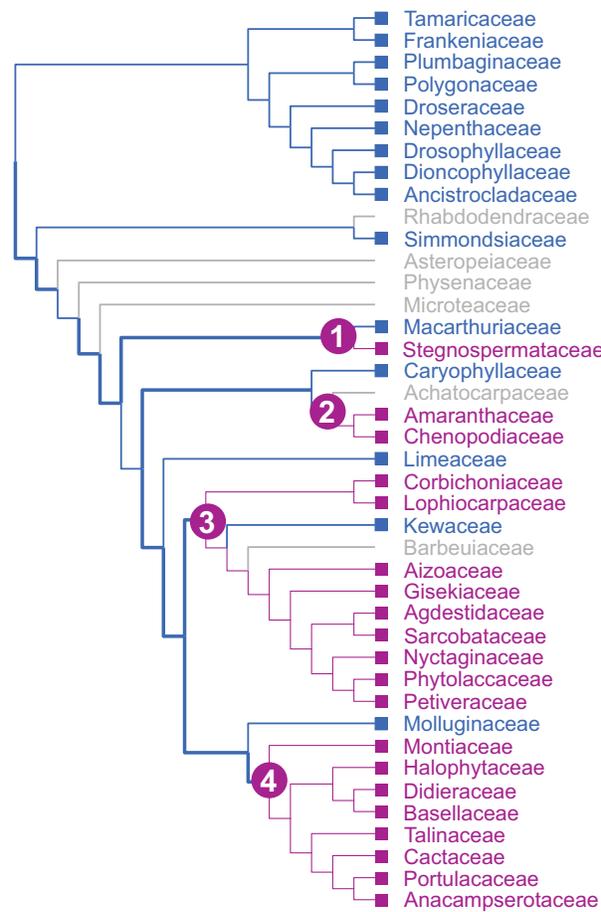


This article is protected by copyright. All rights reserved

(a)

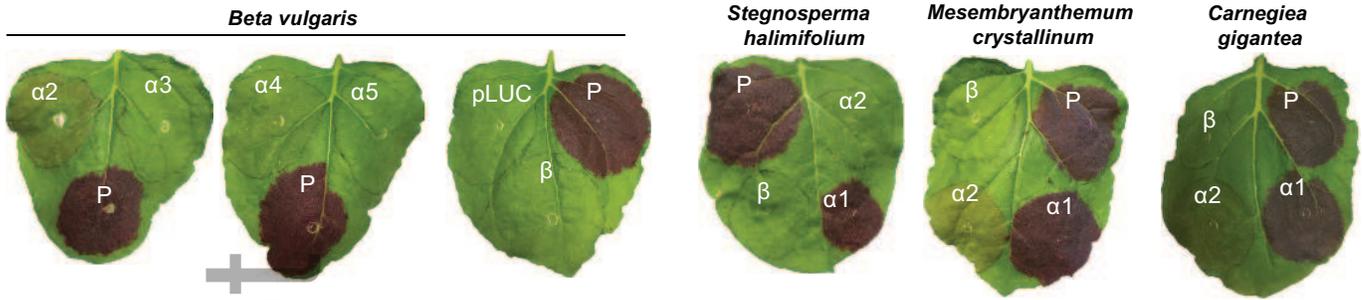


(b)

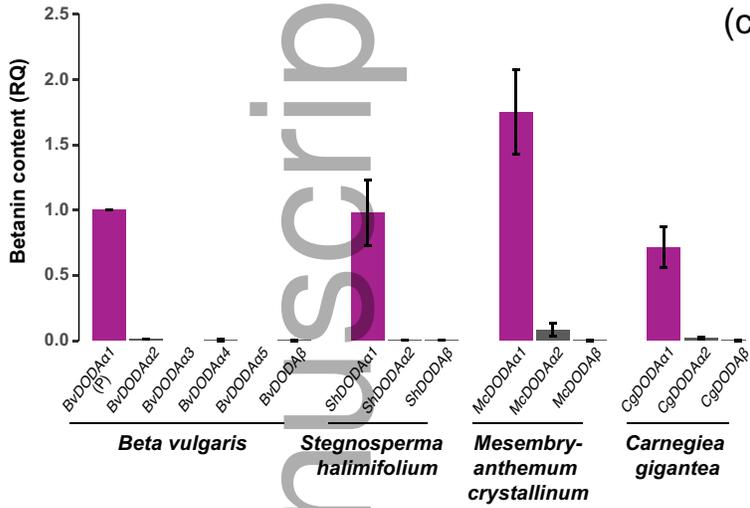


nph\_16089\_f2.eps

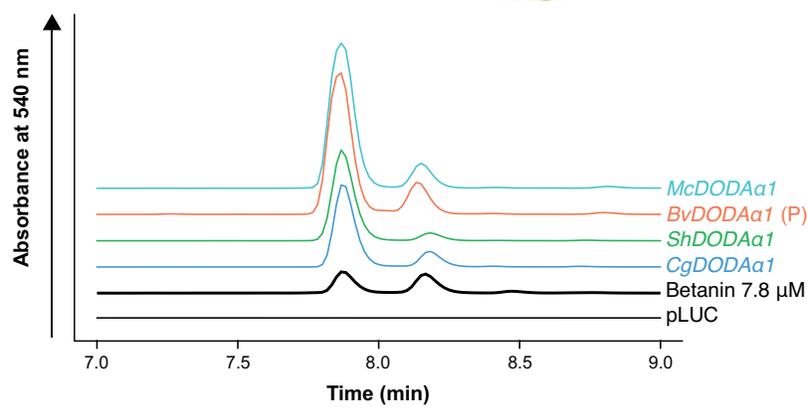
(a)



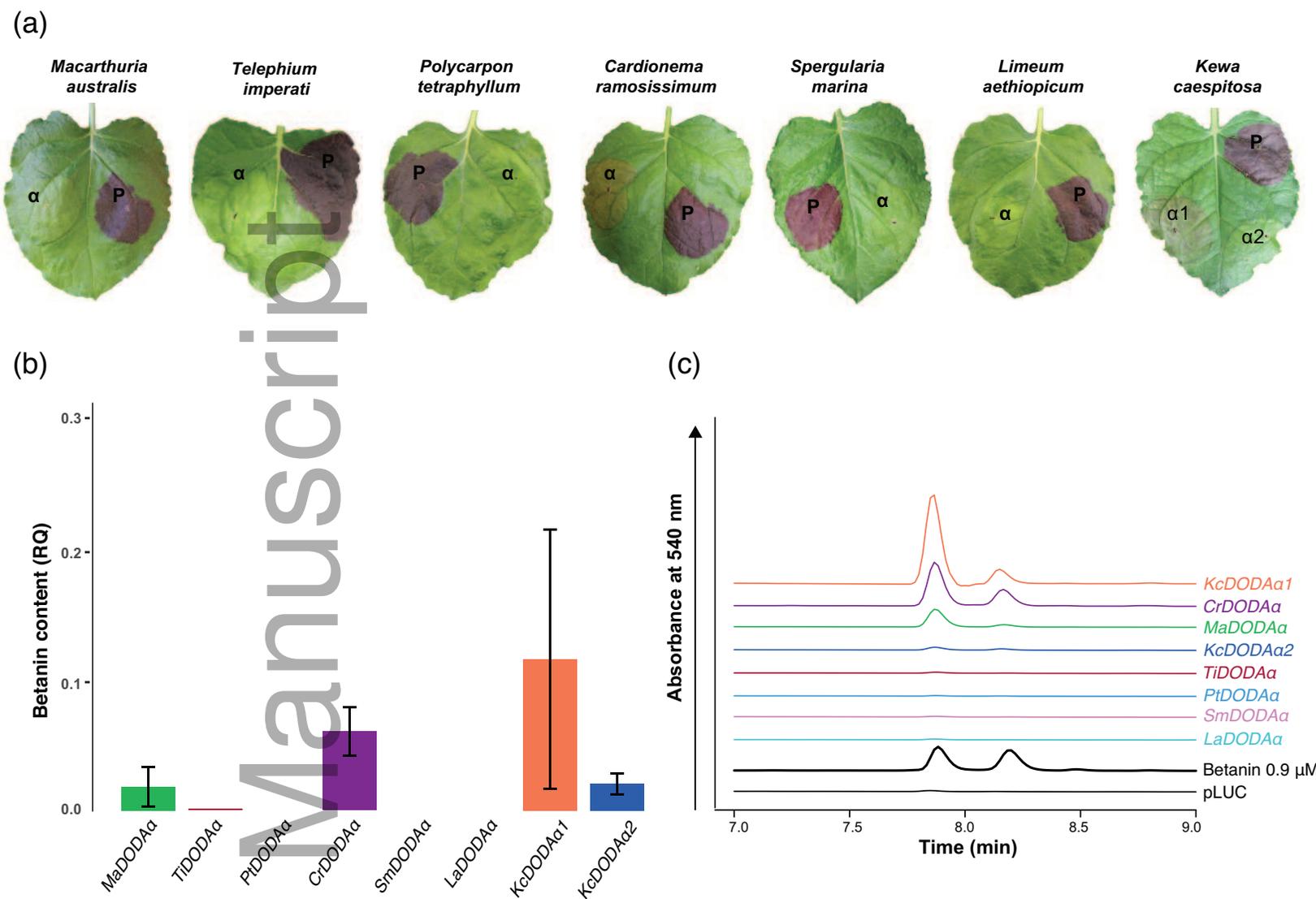
(b)



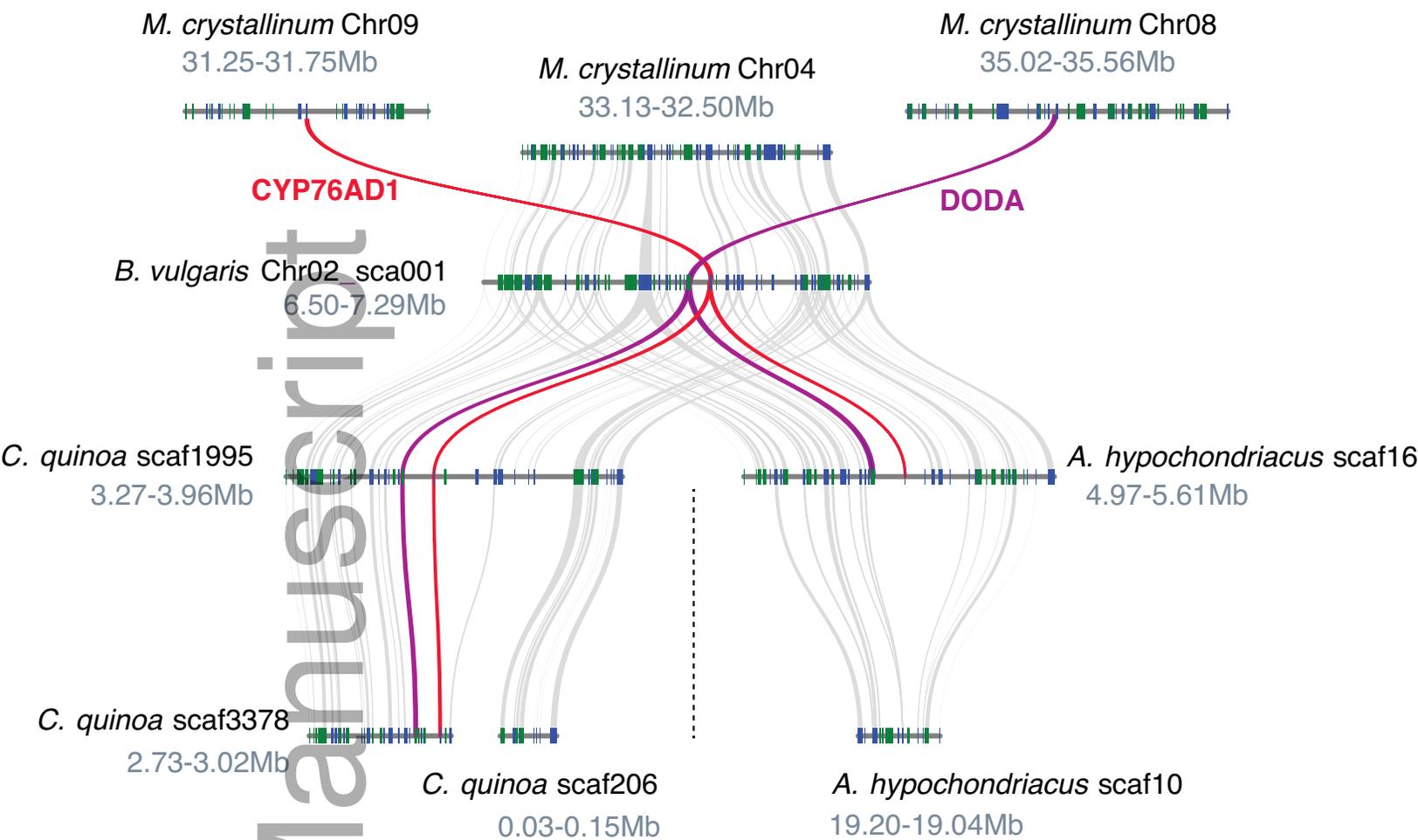
(c)



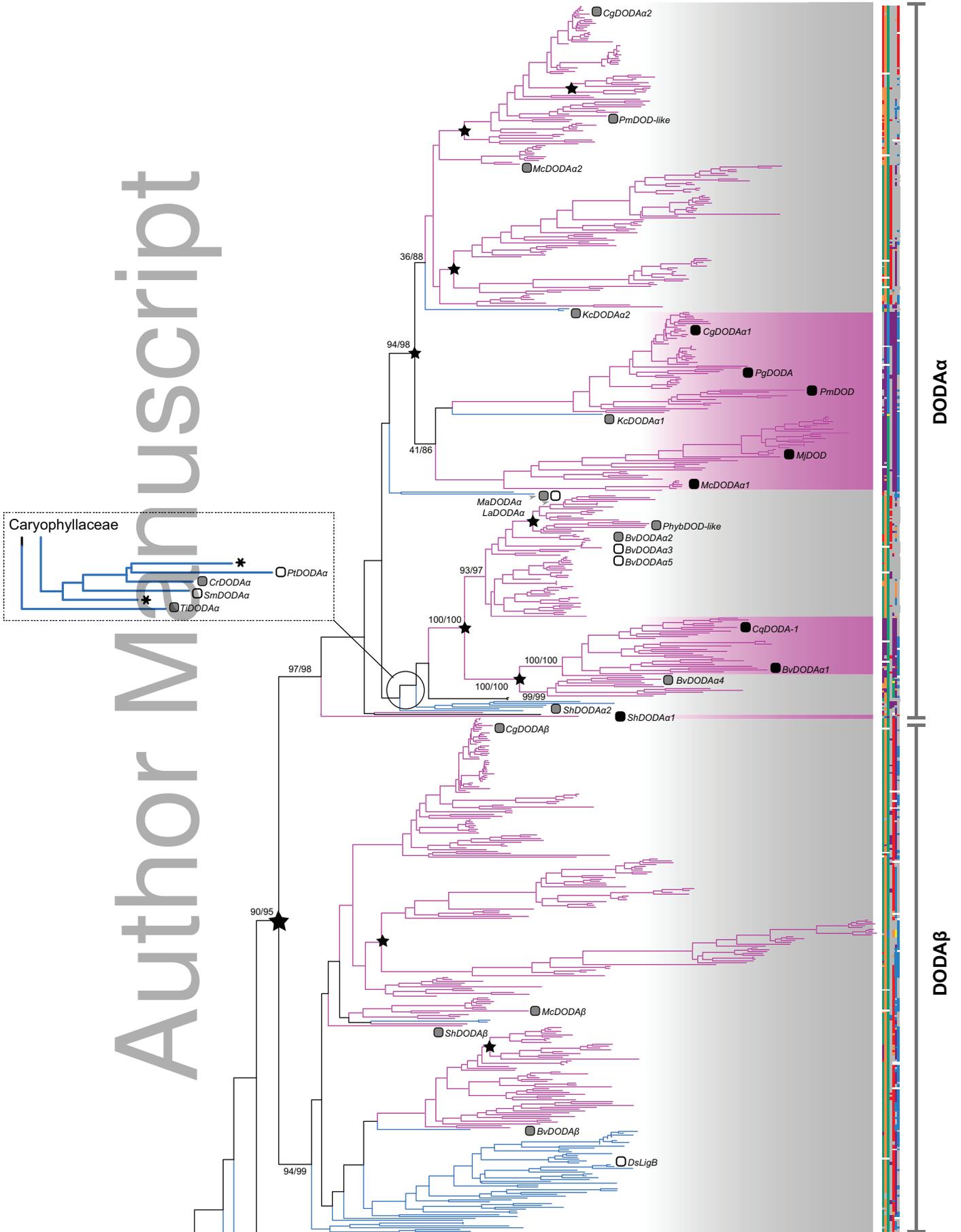
nph\_16089\_f3.eps



nph\_16089\_f4.eps



nph\_16089\_f5.eps



(a)

High Activity DODA      No/Low Activity DODA

74 75 76 152 178 179 203

74 75 76 152 178 179 203

1

TIDNDAT

RGFN-DT

2

DDYNDET

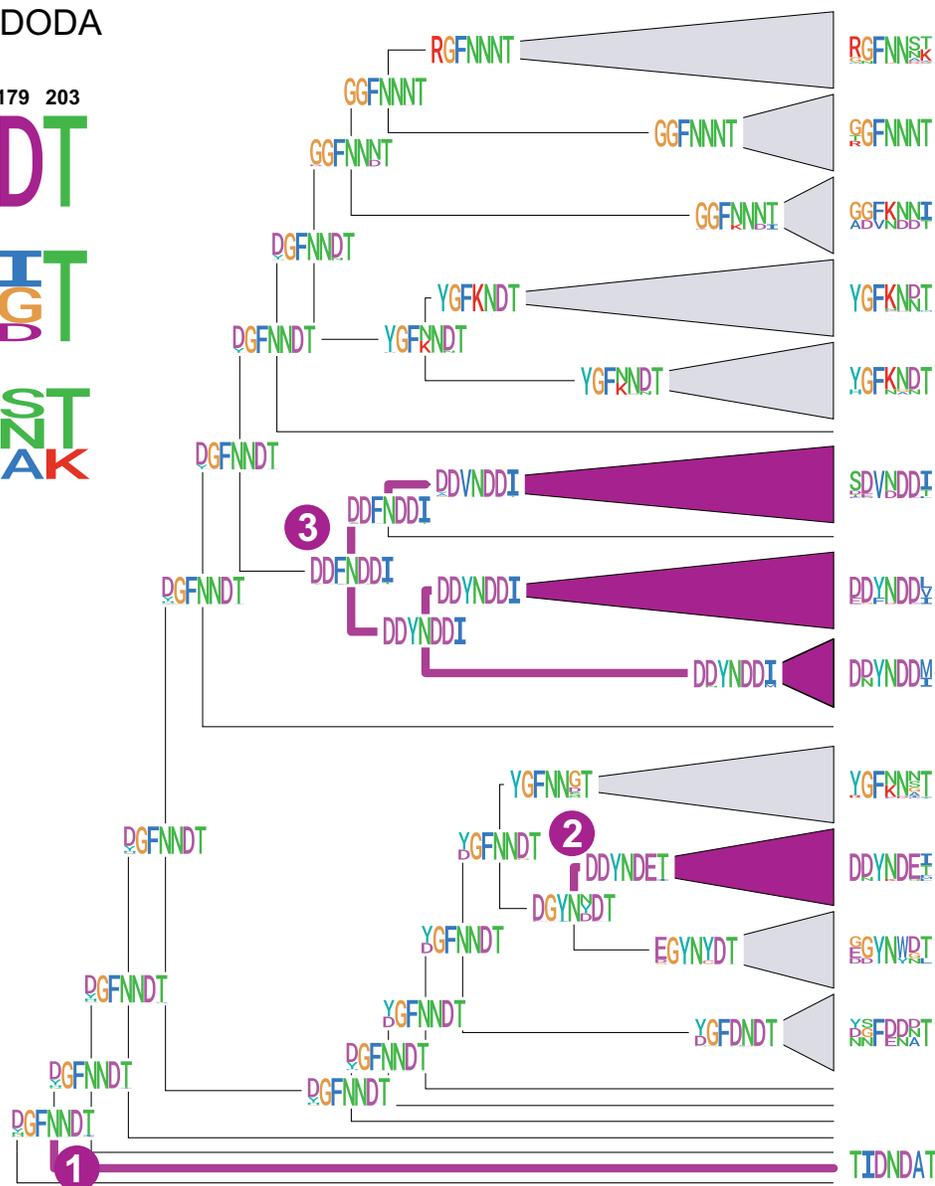
RGEKNET

3

SEVNDDI  
TBNYD

RGFNNST  
GAK

(b)



nph\_16089\_f7.eps

