New Phytologist Supporting Information Methods S1 for the Article

Relaxation of Tyrosine Pathway Regulation Underlies the Evolution of Betalain Pigmentation in Caryophyllales

Samuel Lopez-Nieves, Ya Yang, Alfonso Timoneda, Minmin Wang, Tao Feng, Stephen A. Smith, Samuel F. Brockington, and Hiroshi A. Maeda

Article acceptance date: August 25, 2017

ADH activity from plant tissue extracts

Spinach oleracea seeds (HighMowing, Wolcott, VT) and pink *Dianthus barbatus* (BloomIQ, Lansing, MI) seedlings were purchased from a nursery and were grown together with *Arabidopsis thaliana* (ecotype Columbia) in 22˚C, 60% humidity, and 12/12 h light cycle growth chamber. Leaves of spinach and Arabidopsis seedlings were harvested at 3-week-old, and *Dianthus barbatus* leaves were harvested at 6-week-old. The crude extracts of Arabidopsis or *Dianthus barbatus* were prepared from ~1 g leaf tissues according to Aryal *et al.* (2014). For spinach, ~10 g leaf tissues were used to isolate the plastids according to Aryal *et al.* (2014) in order to avoid the undesired cytosolic polyphenol oxidase activity. Crude or plastid fractions were desalted by Sephadex G50 column to obtain protein extracts, and protein concentration of all biological replicates were adjusted to 0.06, 0.85, and 0.6 mg/mL for spinach, *Dianthus barbatus*, and Arabidopsis extracts, respectively. Time course ADH activity assays at 0, 1, 2, and 3 hr were performed in the presence and absence of 500 µM Tyr analog, 3-fluoro-Tyr, in 10 µL reaction containing 50 mM sodium phosphate (pH 8.0), 1 mM arogenate, 1 mM NADP$^+$, 10 µg/mL tetracycline (to inhibit prokaryotic-type protein synthesis of plastids or bacterial contamination), and 0.3, 4.25, and 3 µg of spinach, Dianthus, and Arabidopsis protein, respectively. The reaction was stopped by adding 20 µL methanol containing 10 µM norvaline as an internal standard. Respective boiled protein extracts were used as negative controls. ADH activity was quantified by the formation of tyrosine according to (Schenck *et al.*, 2015), except that tyrosine was detected as *o*-phthalaldehyde derivative with excitation/emission wavelength of 360/455 nm by fluorescence detector, and *o*-phthalaldehyde derivative of the norvaline internal standard was quantified at 336 nm by DAD detector.

Analysis of Tyr contents from Caryophyllales tissues

Metabolite extracts of thirteen Caryophyllales species were prepared from ~70 mg of youngest leaves, except for flowers of a Cactaceae species to avoid succulent tissues. All plants were grown and harvested at Botany Greenhouse of the University of Wisconsin-Madison. Young leaf tissues of ~4 weeks-old Arabidopsis Columbia ecotype were used as a control. Harvested tissues were extracted by adding 400 µL extraction buffer containing methanol: chloroform (2:1, v/v) and 100 µM 4-chlorobenzoic acid (an internal standard). After adding 300 µL water and 125 µL chloroform, the mixture was vigorously mixed by a vortex mixer for 5 min and centrifuged at 20,000*g* for 5 min for phase separation. The upper polar phase of 400 µL was transferred to a new centrifuge tube and dried down in a benchtop speed vacuum (Labconco, Kansas City, MO, USA). The dried polar phase was resuspended in 200 µL methanol. After centrifugation at 20,000*g* for 5 min, 20 µL was injected into the Agilent 1260 HPLC equipped with Atlantis T3 C-18 column (3 µm, 2.1x150 mm, Waters, Milford, MA), and separated by the following gradient of acetonitrile (B) in 0.1% formic acid (A): 1% B for the first 5 min, followed by a linear increase to 76% B at 10 min, an isocratic elution at 76% B until 16 min, followed by re-equilibration at 1% B. Tyr was monitored with the fluorescence detector at 274 and 303 nm for excitation and emission, respectively. The internal standard was monitored by photodiode array detector at 270 nm. Statistical analyses were conducted by the Statistica Analysis Software (SAS) based on the "mixed" effect model (Pinheiro, 2000) to compare between the two groups having and not-having ADHα and using the "fixed" effect model (Milliken, 2009) to compare individual samples against Arabidopsis control.

Reverse Transcription PCR (RT-PCR) analysis

RT-PCR was carried out on five biological replicates for each infiltrated vector (**Fig. S7b**). Two technical replicates were additionally analyzed for one sample each for BvADHα and BvADHβ infiltrations. RNA was extracted and DNAse treated using the RNeasy Plant Mini Kit and the RNAse-free DNAse set (Qiagen, Hilden, Germany). cDNA was prepared using BioScript Reverse Transcriptase (Bioline Reagents, London, UK) and an oligo(dT)$_{18}$ primer according to the manufacturer's recommendations. A control with no reverse transcription was included to test the presence of genomic DNA. RT-PCR was performed on a 1:10 cDNA dilution with the KAPA 2G Fast DNA Polymerase kit (KAPA Biosystems, Wilmington, MA, USA) and an Eppendorf Mastercycler Nexus (Eppendorf, Hamburg, Germany). Amplification conditions were as follow:

initial step of 1 min at 95°C followed by 30 cycles of 10 s at 95°C, 10 s at 60°C and 2 s at 72°C, and a final step of 5 min at 72°C. Amplicons were visualised on 2% agarose gel electrophoresis using ethidium bromide (0.1 μg/ml) and run at 120V for 20 min. The expected size for the reactions is 140, 90 and 111 bp for *BvADHα*, *BvADHβ,* and *tGFP*, respectively. Primers used are described in **Table S1**.

Quantitative real-time PCR (qRT-PCR) analysis

For quantification of endogenous expression of *BvACTIN* (internal control), *BvADHα*, *BvADHβ*, *BvDODA*, *BvMYB1* and *BvCYP76AD1*, red beet (*W357B*) and sugar beet (*Big Buck*) plants were grown in 22˚C, 60% humidity, and 12/12 hr light cycle in a growth chamber. The seedlings were harvested at 7-days after germination and the tissue was divided into cotyledon and hypocotyl. RNA was extracted (Oñate-Sánchez and Vicente-Carbajosa, 2008) and DNAse treated (Ambion, Austin TX, USA) following by cDNA preparation using MLV Reverse Transcriptase (Promega, Madison, WI, USA). qRT-PCR was performed using the GoTaq qPCR Master Mix (Promega, Madison, WI, USA), and the Stratagene Mx3000P qPCR System (Agilent Technologies, Stratagene, La Jolla, CA, USA). Amplification conditions were as follow: an initial step of 1 min at 95°C followed by 45 cycles of 15 s at 95°C, 30 s at 60°C and 30 s at 72°C. The gene expression of BvADH was normalized using *BvACTIN* as an internal control and analyzed by using the relative expression of the genes. The results are shown in % expression relative to the highest sample (**Fig. 1d**). Primers used in all qPCR analysis are listed in **Table S1.**

Phylogenetic analysis

Amino acids from genomes (full open reading frame) and transcriptomes (full or partial open reading frame) of Brockington *et al*. (2015) were used in this analysis with minor modifications in species included (**Table S2**). The final taxon sampling in this study consisted of 95 species, with 91 ingroup species (89 transcriptomes and 2 genomes) representing 26 of the 39 families in Caryophyllales (Hernández-Ledesma *et al.*, 2015) and four outgroup genomes from eudicots and monocots (**Table S2**). Amino acid sequences of the 11 functionally characterized *ADH* genes were used as baits to search against each of the 95 species. To maximize the sensitivity of homology searches in order to identify short and incomplete sequences from *de novo* assembled transcriptomes, we used SWIPE v2.0.11 (Rognes, 2011) with a high E-value cutoff of 10 and low minimal bitscore cutoff of 30. Hits from all 11 query sequences against each species were ranked from high to low by bitscore, and the top 10 hits from each species were pooled and used for the initial phylogenetic analysis.

The pooled top hits from each of the 95 species, together with the 11 baits were used as the starting sequence file (948 sequences). An initial phylogenetic analysis was conducted using MAFFT v7.215 with "--genafpair --maxiterate 1000" (Katoh & Standley, 2013). Columns with more than 90% missing data in the resulting alignment were trimmed using Phyutility v2.2.6 with "-clean 0.1"(Smith & Dunn, 2008) and a phylogeny was estimated using RAxML v8.1.5 with the model "PROTCATWAG" (Stamatakis, 2014). After visually examining the alignment and tree, tips with branch lengths that were outliers were removed (any terminal branches that had on average more than two substitutions for each amino acid site; or more than ten times longer than its sister group and on average had more than one substitution per site; Yang and Smith, 2014). Monophyletic or paraphyletic tips that belonged to the same species from transcriptome data most often resulted from isoforms produced during *de novo* assembly. These were masked, leaving only the tip with the highest number of aligned characters (Yang and Smith, 2014). Internal branches with molecular branch lengths longer than 1 were likely due to distantly related paralogs or assembly artifacts and were pruned. A large number of distantly related genes, isoforms, and assembly errors were removed during the tip trimming and long branch removing process, with 251 sequences left. A new fasta file was written from remaining tips, and this alignment, tree building, and tree trimming procedure was repeated once, with 229 sequences left. While visually examining alignment and tree we found the sequence Cham@c36044_g1_i2_242_1480_minus that belonged to *Chenopodium giganteum*, but was placed in between ADHα and ADHβ, outside of Chenopodiaceae. Further examination of the alignment showed that half of the sequence was closely related to ADHα, and the other half closely related to ADHβ. This is most likely due to an assembly error. Indeed, *Chenopodium giganteum* had additional, correctly assembled ADHα and ADHβ copies nested in respective Chenopodiaceae clades. Therefore the putative chimeric sequence was removed.

Remaining sequences were aligned with MAFFT with "--genafpair --maxiterate 1000" and trimmed by Phyutility with "-clean 0.3". An alternative alignment was constructed with PRANK v140603 using default settings (Löytynoja & Goldman, 2008; 2010), poorly aligned sequences were manually removed, and trimmed by Phyutility with "-clean 0.1". We used two alternative alignment methods because MAFFT tends to force regions to align even when they are highly divergent whereas PRANK tends to introduce lots of gaps in highly divergent regions. On the other hand, PRANK is an iterative alignment, tree building, and refinement pipeline that we run five iterations before obtaining the final alignment. For both trimmed alignments, a phylogenetic tree was constructed using RAxML with "-m PROTCATAUTO" and 200 rapid bootstrap replicates to evaluate support. Given that the resulting tree topologies and support values using both alignments were very similar we are presenting the results from MAFFT in **Fig. S1e**. The code used in the phylogenetic analysis is available from https://bitbucket.org/yangya/adh_2016, and final alignment is **Notes S1.**

Testing for relaxed selection in Caryophyllaceae

To test for shift in selection pressure in ADHα associated with loss of betalain, we carried out selection analysis on a reduced data set that included representative sequences across *ADHα* that were either verified by Sanger sequencing or by mapping reads back to the *de novo* assembled contigs and carefully examining read coverages visually.

Within the family Caryophyllaceae, ADHα expression was detected in the transcriptome of only the subfamily Paronychioideae. Those *ADHα* transcripts from *Corrigiola litoralis* and *Telephium imperati* were both confirmed by PCR and Sanger sequencing. Two *Spergularia media* fragments from transcriptome assembly were both belonged to *ADHα* and are non-overlapping in the alignment. These two fragments could be from two loci or from a single locus. To distinguish between these two scenarios, we first extended the two fragments separately using Assembly by Reduced Complexity (Hunter *et al.*, 2015, ARC v.1.1.3) with maximum 10 cycles, Bowtie 2 v2.2.8 (Langmead & Salzberg, 2012) for read mapping and Newbler v2.9 (454 Life Sciences, downloaded March 17, 2015) for assembly. After extending the original assembly and aligning it with other *ADHα* sequences, the two extended fragments were still 22 base pairs apart. To evaluate whether these two fragments were supported by raw reads we concatenated the two fragments by fixing the direction and adding 22 Ns to the middle, and mapped raw reads to the concatenated reference using Bowtie 2 with the setting "--phred64 --very-fast-local". The 22 bp gap was highly supported by read pairs and the joined read were kept for subsequent dN/dS analysis. We carried out the same procedure for *Polycarpaea repens* but were unable to join the reads nor confirm they are paralogs due to low read coverage and a longer gap between the two fragments. Therefore, the two fragments were kept in the alignments for phylogenetic analysis but were removed for dN/dS analysis.

To obtain *ADHα* sequences from additional species of Caryophyllaceae, primers were designed to the conserved portion of the *Spergularia media* contig, and were used to amplify *ADHα* sequences from the closely related *Spergularia marina*. Inverse PCR was used to obtain *ADHα* sequences from *Spergularia marina*, *Paronychia polygonifolia* and *Herniaria latifolia.* For inverse PCR, genomic DNA was digested with restriction enzymes *EcoRI* and *MfeI*, and fragments were circularised with T4 ligase (Biolabs, New England). Nested primers were used to amplify the fragment containing the ADH*α* ortholog. Amplified products were sanger sequenced to acquire the 5' and 3' terminals of the locus. In summary, a total of six well-supported *ADHα* sequences were then taken forward for the dN/dS selection analyses.

Our final alignment for selection analysis included eight ADHα sequences in Caryophyllaceae and six additional sequences from representative betalain-producing species across rest of the ADHα lineage (**Notes S2**). We first trimmed the alignment to remove signal peptide and poorly aligned ends, leaving the region from *BvADHα* amino acid no. 79 to 354 that covered the enzyme active domain (**Notes S3** and **Notes S4**). We then carried out phylogenetic analyses for both alignments in RAxML, with the model "GTRCAT" for the codon alignment and "PROTCATAUTO" for the amino acids alignment, and 200 rapid bootstrap replicates to evaluate node support (**Fig. S9a,b**). To quantify the rate shift, we carried out RELAX analysis (Wertheim *et al.*, 2014) as implemented in the online portal Datamonkey (Kosakovsky Pond & Frost, 2005, accessed March 19, 2016), using the trimmed CDS matrix with *Polycarpaea repens* removed. RELAX has the advantage of distinguishing between increased positive selection vs. reduced purifying selection, both of which would result in accelerated average dN/dS values. We designated all crown branches in Caryophyllaceae as the testing branches and the rest branches as the background. We fitted the partitioned MG94xREV model that assumes all sites having unified dN and dS value, allowing the rate to vary between the test and background branches. We also fitted the RELAX model that takes site heterogeneity into account. The RELAX null model assumes all background and test branches share the same rate in each rate category, whereas the RELAX alternative model allows substitution rate to vary between the test and background branches in each rate category, and sites can move among rate categories.

Supplemental References

Aryal UK, Xiong Y, McBride Z, Kihara D, Xie J, Hall MC, Szymanski DB. **2014**. A proteomic strategy for global analysis of plant protein complexes. *The Plant Cell* **26**: 3867–3882.

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.

Hernandez-Ledesma P, Berendsohn WG, Borsch T, Mering SV, Akhani H, Arias S, Castañeda-Noa I, Eggli U, Eriksson R, Flores-Olvera H, Fuentes-Bazán S, Kadereit G, Klak C, Korotkova N, Nyffeler R, Ocampo G, Ochoterena H, Oxelman B, Rabeler RK, Sanchez A, Schlumpberger BO & Uotila P. 2015. A taxonomic backbone for the global synthesis of species diversity in the angiosperm order Caryophyllales. *Willdenowia* **45**: 281-383.

Hunter SS, Lyon RT, Sarver BAJ, Hardwick K, Forney LJ, Settles ML. **2015**. Assembly by Reduced Complexity (ARC): a hybrid approach for targeted assembly of homologous sequences. *bioRxiv*. doi:10.1101/014662.

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

Kosakovsky Pond SL, Frost SDW. **2005**. Datamonkey: Rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* **21**: 2531–2533.

Langmead B, Salzberg SL. **2012**. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.

Löytynoja A, Goldman N. **2008**. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science (New York, N.Y.)* **320**: 1632–1635.

Löytynoja A, Goldman N. **2010**. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics* **11**: 579.

Milliken, GA. 2009. Analysis of messy data. Boca Raton :CRC Press.

Oñate-Sánchez L, and Vicente-Carbajosa J. 2008. DNA-free RNA isolation protocols for *Arabidopsis thaliana*, including seeds and siliques. *BMC Research Note* **1**: 93.

Pinheiro, JC. 2000. Mixed-effects models in S and S-PLUS. New York :Springer.

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

Schenck CA, Chen S, Siehl DL, Maeda HA. **2015**. Non-plastidic, tyrosine-insensitive prephenate dehydrogenases from legumes. *Nature Chemical Biology* **11**: 52–57.

Smith SA, Dunn CW. **2008**. Phyutility: A phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* **24**: 715–716.

Stamatakis A. **2014**. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. **2014**. RELAX: Detecting relaxed selection in a phylogenetic framework. *Molecular Biology and Evolution* **32**: 1–13.

Yang, Y. and S.A. Smith. 2014. Orthology inference in non-model organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution* **31**: 3081–3092.