

Multi-Tissue Transcriptome Analysis of Two *Begonia* Species Reveals Dynamic Patterns of Expression Evolution In The Chalcone Synthase Gene Family

Katie Emelianova (✉ katie.emelianova@ed.ac.uk)

University of Edinburgh

Andrea Martínez Martínez

Royal Botanic Gardens Edinburgh

Lucia Campos-Dominguez

Royal Botanic Gardens Edinburgh

Catherine Kidner

Royal Botanic Gardens Edinburgh

Research Article

Keywords: Begonia, Cucumis sativus, Arabidopsis thaliana

Posted Date: April 14th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-403106/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Begonia is an important horticultural plant, as well as one of the most speciose Angiosperm genera, with over 2000 described species. Genus wide studies of genome size have shown that *Begonia* has a highly variable genome size, and analysis of paralog pairs has previously suggested that *Begonia* underwent a whole genome duplication. We address the contribution of gene duplication to the generation of diversity in *Begonia* using a multi-tissue RNA-seq approach. We chose to focus on the chalcone synthase (CHS) gene family due to its role in biotic and abiotic stress response, and in particular its importance in maximising the use of variable light levels in tropical plants.

We used RNA-seq to sample six tissues across two closely related but ecologically and morphologically divergent species, *Begonia conchifolia* and *B. plebeja*, yielding 17,012 and 19,969 annotated unigenes respectively. We identified the chalcone synthase gene family members in our *Begonia* study species, as well as in *Hillebrandia sandwicensis*, the monotypic sister genus to *Begonia*, *Cucumis sativus*, *Arabidopsis thaliana*, and *Zea mays*. Phylogenetic and expression analysis revealed the recent origin of CHS duplicates in *Begonia*, which showed both conserved and divergent expression profiles between duplicates. We conclude that there is evidence for a role of gene duplication in generating diversity through expression divergence in *Begonia*.

Introduction

Begonia is one of the most diverse Angiosperm genera, with more than 2000 species described to date ¹. The genus is thought to have originated in Africa between 24–45 MYA and since then diversified across South America and Asia ², where it occupies a wide range of niches due to the diversity of vegetative forms across species ³.

Strong population structure, high levels of drift, and genetic divergence at local scales are thought to contribute to the high species diversity in *Begonia* ⁴. Endemism is very common ⁵, and strong population structure is known to coincide with high variation in morphological characteristics such as leaf shape and size ⁶.

Begonia has also been shown to have highly variable genome sizes ⁷, and evidence of whole genome duplications has been identified from paralog kS peaks ⁸. The contribution of gene and genome duplication has long been associated with the evolution of phenotypic novelty ⁹. This study used multi tissue RNA-seq to study diversification in duplicated genes in two closely related but morphologically divergent species *B. conchifolia* and *B. plebeja* (Fig. 1). *B. conchifolia* is a small plant with long-lived fleshy peltate leaves and small white flowers. It has a restricted distribution in wet rainforests across Southern Mexico and Central America. *B. plebeja* is more widespread, occupying seasonally dry forests in Northern Mexico. It has larger, thinner leaves which are deciduous in some populations and often blotched. Flowers are larger and sometimes tinged with pink ¹⁰.

The use of RNA-seq enables the interrogation of divergence at the level of gene sequence and expression pattern, both of which are important drivers of biochemical and phenotypic novelty ¹¹. Here we focus on the evolutionary and gene duplication patterns in the key anthocyanin biosynthetic gene chalcone synthase (CHS, EC 2.3.1.47).

Flavonoid pigments are important in the attenuation of high UV exposure, preventing PSII inhibition and the reduction of carbon intake ^{12 13}. Of particular relevance to shade tolerant species, as found in *Begonia*, flavonoid distribution across a variety of tissues helps low light dwelling plants make the most use of intermittent high intensity sunflecks ^{14 15} while avoiding photodamage and attenuating stress response through ROS scavenging ¹⁶.

CHS is the first committed step of the flavonoid biosynthesis pathway ¹⁷, and is a crucial enzyme in the production of compounds used in biotic and abiotic stress responses ^{18 19 20}. Previous studies in other plant species have found the CHS gene family to have dynamic patterns of duplicate evolution ^{21 22}.

Multi-tissue RNA-seq in *B. conchifolia* and *B. plebeja* aims to capture as many genes as possible by sampling across the main plant tissues, generating a near complete transcriptome for each species, significantly increasing the genetic resources available for investigating the origins of diversity in this important genus.

Methods

Tissue Sampling and RNA Extraction

Plant tissue was donated from plants in the living collection at Royal Botanic Gardens Edinburgh. The tissues chosen for study, mature leaf, mature petiole, vegetative bud, female flower, male flower and root, were harvested between 9am and 10am between January and May 2015 from *B. conchifolia* (Accession Number: 20042082) and *B. plebeja* (Accession Number: 20051406). Leaves were the first fully expanded leaf on the axis, petioles were from these leaves, flowers were staged between tepals just opening and tepals fully expanded, roots were young white roots within 5–10 cm of the apex. RNA from three biological replicates per tissue was extracted using the phenol chloroform protocol ²³ and quantitated using Qubit (Thermo Fisher). Sample purity was estimated using a NanoVue Spectrophotometer.

Sequencing

Library preparation and sequencing, carried out by Edinburgh Genomics, consisted of preparation of TruSeq mRNA-seq libraries, and generation of c. 240 million 150 base pair paired-end reads on one lane of a HiSeq rapid v1 machine. Raw reads are stored in the European Nucleotide Archive under the study accession PRJEB26711.

Removal of contaminants from RNA-seq reads

BlobTools ²⁴ was used to screen for and remove contaminants from assemblies. Reads were first adapter trimmed using Trimmomatic ²⁵ using a 4 base sliding window and a minimum mean quality of 15. Leading and trailing bases lower than quality score 3 were trimmed. Total adapter trimmed reads were assembled using Trinity v2.5.1 ²⁶ using default parameters. Coverage was estimated by mapping reads back to their corresponding species assembly with STAR v2.5.3a ²⁷ using default parameters. Finally, contigs from each assembly were used as a query to search against the NCBI nucleotide database (nt) with BLAST v2.2.28 for taxonomy assignment.

Using the assembly, taxonomy and coverage files, BlobPlots were created to visualise contigs partitioned by taxon, GC content, and coverage. Contigs which were annotated as Streptophyta were used to extract associated reads belonging to this taxon using the BlobTools bamfilter functionality.

Assembly and quality control

Decontaminated total reads were assembled using Trinity v2.6.4 using default parameters. The longest isoform per gene was obtained with Trinity utility scripts to obtain a set of unigenes for each assembly.

Transcriptome assembly quality was assessed with Transrate v1.0.3 ²⁸. Transrate reports basic metrics for a transcriptome assembly and provides quality information for assembled contigs using coverage and accuracy information by mapping reads to assembled contigs.

Transcriptome assembly completeness was estimated using BUSCO v4.0.0 ²⁹, using transcriptome mode.

Annotation

The Trinotate v3.2.1 pipeline ³⁰ was used to functionally annotate unigenes for each species. Unigenes were searched against the Swissprot database with blastx v2.2.28 using an E value cutoff of $1e^{-3}$ and setting maximum target sequences to 1.

Most likely longest ORF peptide sequences were predicted from unigenes with Transdecoder v5.5.0. The resulting predicted peptides were used to search against the Swissprot database with blastp v2.2.28, using an E-value cutoff of $1e^{-3}$ and setting maximum target sequences to 1. Protein domains were identified by searching predicted longest ORF peptides against the Pfam database using hmmscan v3.1b1. Blast homologies from blastp and blastx results and Pfam domains from hmmscan results were loaded into the Trinotate provided SQLite database, and an annotation report was generated.

Coverage

Decontaminated reads per tissue and replicate were mapped to unigenes for each species with STAR v2.5.3a using default parameters. Read counts were summarized across features using Subread's FeatureCounts v1.5.2 ³¹, not including read pairs

which map to different contigs.

Expression normalization

EdgeR³² was used to normalise counts generated by FeatureCounts. Library size and composition was accounted for using TMM (trimmed mean of M-values) normalisation, and average FPKM values were calculated for replicates of tissue groups per species.

Characterization of chalcone synthase

The *A. thaliana* protein sequence (AT5G13930) was used to search the *B. conchifolia* and *B. plebeja* nucleotide databases of longest assembled isoforms with tblastn, using an E value threshold of 1e-20 and a percent identity threshold of 50%.

The same strategy was used to find homologs of CHS in the draft *Hillebrandia sandwicensis* genome³³, tblastn coordinates were used to extract CHS coding sequences from contigs containing hits. Nucleotide sequences of positive hits in *B. conchifolia*, *B. plebeja* and *H. sandwicensis* were aligned with the *A. thaliana* CHS cDNA sequence and a *Zea mays* homolog of CHS (C2, gene symbol LOC100274415) with Geneious³⁴ using the Geneious aligner, specifying global alignment with free end gaps, a similarity threshold of 65%, and a cost matrix of 5/-4 for matches and mismatches respectively. Sequence alignment was manually inspected and any sequences showing poor homology were discarded.

Sequences with an overlap shorter than 200bp with all other sequences were not included in further analysis. Intronic sequence introduced by genomic sequences was excised, and conserved sequence composed of the first and second exon was extracted. The alignment was manually checked and corrected prior to further analysis.

The final alignment of CHS sequences was used to perform a model selection procedure using Model Generator³⁵ based on the Akaike Information Criterion (AIC).

RAxML³⁶ was used to construct a gene tree using the GTR + R substitution model with 1000 bootstrap replicates.

Results

To remove sequences sampled from other taxa during RNA extraction, BlobTools was used to classify assembled transcript sequences, and only reads contributing to sequences classified as Streptophyta were used for all downstream analysis.

BlobTools infers taxonomic annotation from a similarity search of the input sequences against a public sequence collection (e.g. NCBI nt), and determines the taxonomy of each sequence using a taxrule algorithm. Coverage and GC content of the annotated sequences are plotted in order to visualise the partitioned sequences and perform downstream contaminant screening.

Screening for contaminants revealed the majority of taxonomically assigned transcripts belonged to Streptophyta, with 62,082 and 68,696 transcripts from *B. conchifolia* and *B. plebeja* respectively assigned to the taxon (supplementary Figs. 1 and 2). The next most frequent taxon represented by annotated transcripts in *B. conchifolia* is Arthropoda with 3,629 transcripts, and Ascomycota in *B. plebeja* with 7,309 transcripts, representing plausible sources of contamination from a greenhouse setting.

The sequence length weighted span of coverage in both species' BlobPlots shows Streptophyta having the second highest peak of coverage after no-hit sequences, and the widest span, reflecting the range of expression levels of the transcripts screened.

Using the BlobPlot information, reads from both annotated contaminant and no-hit transcripts were removed from further analysis totalling around 80 million reads in *B. conchifolia* and 64 million in *B. plebeja* (supplementary table 1).

Comparison of assemblies prior to and after decontamination showed that removal of contaminant and no-hit reads predominantly affected the number of short transcripts (Fig. 2), median transcript length increasing to 1,641bp and 1,304bp in the decontaminated assembly from 595bp and 505bp in the contaminated assembly for *B. conchifolia* and *B. plebeja* respectively. The maximum contig length also increased in assemblies after decontamination, increasing from 15,079bp to 15,923bp in *B. conchifolia* and 15,948bp to 16,037bp in *B. plebeja*, suggesting an improvement in the Trinity assembly process when faced with fewer reads and reduced K-mer graph complexity.

Decontaminated assemblies had fewer shorter and incomplete sequences, only 39% and 35% of sequences in the contaminated *B. conchifolia* and *B. plebeja* assemblies had a detected ORF, compared to 77% and 73% for each species after removal of contaminants.

B. conchifolia showed less evidence of fragmentation, with fewer total transcripts, and a greater N50 and N90 than *B. plebeja* (Table 1). Sequence length distribution corroborates this pattern, showing *B. plebeja* to have a greater bulk of shorter transcripts than *B. conchifolia*. BUSCO assessment of transcriptome completeness (Table 2) showed that *B. plebeja* had marginally poorer scores for transcript completeness and fragmentation, however both transcriptomes showed over 80% completeness, indicating that using multi-tissue RNA-seq to sample as much of the plant as possible has successfully captured a large proportion of genes.

Table 1
Assembly statistics for *B. conchifolia* and *B. plebeja* before and after contaminant removal.

	Before BlobTools		After BlobTools	
	<i>B. conchifolia</i>	<i>B. plebeja</i>	<i>B. conchifolia</i>	<i>B. plebeja</i>
Number of seqs	200,849	251,473	42,614	59,106
Number of unigenes	95,553	124,116	17,012	19,969
Smallest seq	201	201	201	201
Largest seq	15,079	15,948	15,923	16,037
N over 1kb	70,738	71,283	31,876	37,865
N over 10kb	44	13	28	16
N with ORF	77,736	86,926	32,848	43,119
N90	400	328	1,090	842
N50	1,808	1,348	2,381	1905

Table 2
BUSCO assessment of transcriptome completeness for *B. conchifolia* and *B. plebeja*

BUSCO category	<i>B. conchifolia</i>	<i>B. plebeja</i>
Complete	362 (85.2%)	351 (82.5%)
Complete single copy	352 (82.8%)	344 (80.9%)
Complete duplicated	10 (2.4%)	7 (1.6%)
Fragmented	23 (5.4%)	36 (8.5%)
Missing	40 (9.4%)	38 (9%)
Total	425	425

The Trinotate pipeline was used to annotate 17,012 *B. conchifolia* and 19,969 *B. plebeja* unigenes (supplementary files 3 and 4). More unigenes were annotated in *B. plebeja* than in *B. conchifolia* within each annotation source (Fig. 3), likely due to the larger number of input transcripts. Concordantly, *B. plebeja* also has more transcripts without any annotation, 16.65% and 17.48% of unigenes are unannotated in *B. conchifolia* and *B. plebeja* respectively. Roughly equal proportions of unigenes from both species were annotated to Streptophyta (70.87% in *B. conchifolia* and 70.51% in *B. plebeja*).

Unigenes which shared sources of annotation were identified using UpSet plots, showing that, due to very sparse EggNOG annotation (Fig. 3), the largest group of unigenes sharing annotation sources were annotated with all sources except for EggNOG

(supplementary Figs. 5 and 6), *B. conchifolia* having 7,217 unigenes in this group and *B. plebeja* 7,787.

Unigene presence or absence across tissues was compared for *B. conchifolia* and *B. plebeja*, using a cutoff of 1 FPKM for whether a transcript is present or absent in a tissue. UpSet plots were used to visualise the shared and tissue specific distribution of unigene expression (supplementary Figs. 3 and 4). The majority of unigenes are expressed in all tissues (11,495 unigenes in *B. conchifolia* and 13,609 unigenes in *B. plebeja*).

Both species had a high frequency of unigenes expressed uniquely in root (528 and 502 unigenes in *B. conchifolia* and *B. plebeja* respectively) and in and male flower (223 and 359 unigenes in *B. conchifolia* and *B. plebeja* respectively).

To identify unigenes which are uniquely expressed in each tissue, we used a 1 FPKM cutoff for unigene expression, identifying unigenes which are expressed in a single tissue and not expressed in any other tissues. For example, a unigene which is uniquely expressed in root has an expression of more than 1 FPKM in root, and less than 1 FPKM in all other tissues. To identify the functional categories of genes uniquely expressed per tissue, we mapped GO terms to tissue specific unigenes (plotted in supplementary Figs. 7 and 8). The total number of GO terms mapped to these uniquely expressed unigenes, here referred to as unique GO terms (UGT), is shown in Table 3. *B. conchifolia* and *B. plebeja* had a comparable number of UGTs across tissues. Both species had fewer UGTs in female flower compared to male flower (199 and 131 in female flower and 772 and 924 in male flower in *B. conchifolia* and *B. plebeja* respectively). Of all tissues, leaf had the least UGTs in both species (69 and 0 in leaf in *B. conchifolia* and *B. plebeja* respectively), and root tissue had the greatest number (1183 and 1820 in leaf in *B. conchifolia* and *B. plebeja* respectively). Tissues which had the biggest difference in number of UGTs between species were petiole (a difference of 225 UGTs) and vegetative bud (a difference of 341 UGTs).

Table 3
number of GO terms mapped to unigenes
expressed uniquely in each tissue in *B.*
conchifolia and *B. plebeja*

Tissue	<i>B. conchifolia</i>	<i>B. plebeja</i>
Female flower	199	131
Leaf	69	0
Male flower	772	924
Petiole	338	113
Root	1183	1820
Vegetative bud	102	443

After alignment of the CHS copies identified, any CHS sequences which had an overlap of less than 200bp within the conserved coding sequence with any other sequence were discarded (supplementary file 1). A maximum likelihood tree was inferred from the remaining subset of CHS copies identified in *B. conchifolia* and *B. plebeja*, using *Z. mays* as an outgroup, and including the CHS sequence of *C. sativus*, the closest relative to *Begonia* with a publicly available genome sequence³⁷, and *H. sandiwcensis*, the monotypic sister genus to *Begonia* (supplementary file 2). Four copies of CHS were identified in *B. conchifolia* and five copies in *B. plebeja*.

At least one more copy exists in both species (supplementary table 2), but due to incomplete sequence reconstruction, it is not possible to assign a copy number with any certainty. Of the nine CHS sequences included for *B. conchifolia* and *B. plebeja*, four pairs of closely related orthologs were identified in *B. conchifolia* and *B. plebeja*, as well as one additional *B. plebeja* duplicate.

The phylogenetic tree reconstructed for CHS copies from *Begonia*, *Hillebrandia*, *C. sativus*, *A. thaliana* and *Z. mays* revealed duplicates from *B. conchifolia* and *B. plebeja* arose after the divergence of *Begonia* and its closest sequenced neighbour in the Cucurbitales, *C. sativus*. *Begonia* sequences were obtained from RNA-seq, therefore it is not possible to confirm that no older duplicates exist in *B. conchifolia* and *B. plebeja*, however the absence of older duplicates in the genome of *H. sandiwcensis*, the monotypic species of *Begonia*'s sister genus *Hillebrandia* supports a pattern of high duplicate turnover in the *Begoniaceae*.

The nine copies identified in *B. conchifolia* and *B. plebeja* are four putative orthologs and one single *B. plebeja* duplicate, and are colour coded for ease of comparison (Fig. 4 and Fig. 5).

The oldest duplication identified in *Begonia* gives rise to group 4 orthologs (in green, Fig. 4), and is placed after the divergence of *Begonia* and *Cucumis*.

The intra-ortholog expression similarity (e.g. group 1 *B. conchifolia* ortholog vs group 1 *B. plebeja* ortholog) is reasonably high (Fig. 5), and reflects the recent speciation of the two study species. The single *B. plebeja* ortholog in group 5 (red) also shows high similarity in expression profile to group 2 orthologs (pink), mirrored by the phylogenetic proximity of the two groups (Fig. 4). Any changes between each ortholog pair are therefore the result of expression changes since divergence of *B. conchifolia* and *B. plebeja*.

Inter-ortholog expression is more variable; ortholog group 1 (blue) has the highest expression (*B. conchifolia* FPKM min = 530, max = 2350, *B. plebeja* FPKM min = 408, max = 2903), and group 3 (yellow) has the lowest expression (*B. conchifolia* FPKM min = 0.11, max = 2.42, *B. plebeja* FPKM min = 2.06, max = 14.91). Ortholog groups 2 (*B. conchifolia* FPKM min = 27.57, max = 576.92, *B. plebeja* FPKM min = 11.91, max = 635.85 and 4 (*B. conchifolia* FPKM min = 13.65, max = 365.58, *B. plebeja* FPKM min = 12.26, max = 606.76) have much more comparable expression, indicating that phylogenetic proximity is not correlated with expression similarity in this case. The only exception is the group 5 single *B. plebeja* duplicate (FPKM min = 27.28, max = 871.03), which is the product of a duplication shortly before the group 2 orthologs, and shares similarity in expression due to phylogenetic proximity.

The group 3 CHS orthologs appear to have considerably decreased expression levels in both species, the *B. conchifolia* ortholog showing less than 1 FPKM expression across all tissues except for root. While the *B. plebeja* ortholog has slightly higher expression, this ortholog pair may represent an example of a gene duplicate in the process of pseudogenisation. This possibility is supported by the considerably longer branch length leading to the group 3 orthologs, suggesting an increased rate of mutation accumulation allowed by the reduced selective constraints during pseudogenisation ³⁸.

Discussion

The diversity of form seen across the Angiosperms is the topic of a wide scope of research, including conservation ³⁹, plant breeding ⁴⁰ and evolution ⁴¹.

In this study, we address the role of gene and genome duplication in the generation of phenotypic and ecological diversity in *Begonia* using two closely related but ecologically distinct species, *B. conchifolia* and *B. plebeja*. We use multi-tissue RNA-seq to sample six tissues across both species, thereby also producing valuable transcriptomic sequence data to add to the growing genetic resources for *Begonia*.

We used contamination screening to find and remove contaminants. We identified a sizable number of sequences of non-plant origin representing environmental contamination during tissue collection and RNA extraction. Contamination of genomic and transcriptomic datasets is a widespread problem ⁴², and while some contaminants are easy to spot, such as odorant binding proteins and chemosensory proteins unique to insects ⁴³, contaminants which are plausible homologs can pose a danger to the conclusions drawn from a study ⁴⁴. Our results have indicated that contaminants have lower coverage, presumably due to contaminant taxon tissue being more sparsely sampled, and therefore contribute to a large proportion of fragmented and incomplete transcripts. More transcript fragments may act to increase the perceived transcriptome complexity, and thus reduce the transcriptome assembly quality ⁴⁵.

After contaminant screening, 137,985,258 and 91,321,126 total reads were assembled into reference transcriptomes composed of 17,012 and 19,969 unigenes for *B. conchifolia* and *B. plebeja* respectively. Both transcriptomes had good metrics for completeness, 85.2% and 82.5% of BUSCOs were recovered completely and less than 10% of BUSCOs were missing in both species. Quality metrics suggest a higher rate of fragmented transcripts in *B. plebeja*; more total transcripts and a lower N50 and N90 (1,905 and 842 and 2,381 and 1,090 for *B. conchifolia* and *B. plebeja* respectively).

More than 80% of both species' transcriptomes were annotated using at least one source, providing valuable context to the reference transcriptomes as well as to tissue specific expression. Analysis of GO terms revealed leaves to be the most conserved in gene expression profiles between the two species and vegetative bud the most distinct. This may reflect the different developmental decisions during development of the leaf and meristems as the different leaf shapes and plant architectures are laid down, compared to very similar functional expression patterns in the mature leaf.

Multi-tissue RNA-seq allows for greater spatial resolution when investigating the fates of duplicated genes in isolation as well as within coexpression networks^{46 47}. Expression divergence in gene duplicates is a key process that allows for tissue specialization and morphological diversification⁴⁸, however these changes are also dependent on mode of duplication⁴⁹. Further development of genomic data in *Begonia* may help to distinguish between the products of tandem and whole genome duplication and changes in their expression profiles since duplication.

We used our multi-tissue RNA-seq dataset to investigate the flavonoid biosynthesis gene chalcone synthase, due to its role in interaction with and response to the environment and therefore of ecological relevance to *B. conchifolia* and *B. plebeja*.

Due to incomplete assembly of all CHS copies, four copies of CHS were used in *B. conchifolia* and five in *B. plebeja*, however both species have at least one additional copy of CHS.

CHS copies used in phylogenetic and expression analysis revealed they were all derived from duplications after the divergence of *Begonia* and *Cucumis*. Detection of expression in only recent *Begonia* duplicates, as well as shared and lineage specific duplications in *Begonia* and *Hillebrandia* CHS may suggest high rate of duplicate turnover in the Begoniaceae.

A *B. plebeja* specific duplicate was identified (Fig. 5, group 5), which had a similar expression profile to group 2 orthologs, to which it was the closest related. The absence of a *B. conchifolia* ortholog may reflect a pseudogenization event, causing no detectable expression of an ortholog in any tissue, potentially further supporting a high rate of duplicate gain and loss in the *Begoniaceae*, however further work is needed to confirm this.

No compelling differences were observed in CHS expression levels in ortholog pairs across *B. conchifolia* and *B. plebeja*. Rather, high expression variation was seen between different pairs of orthologs; one ortholog pair (Fig. 5, group 1) duplicated before the divergence of *Begonia* and *Hillebrandia* was expressed at very high levels (> 400 FPKM) in all tissues in both species, while another more recently duplicated ortholog pair (Fig. 5, group 3) showed almost no expression in *B. conchifolia* (with the exception of root tissue) and very low expression in *B. plebeja*. The remaining two ortholog pairs and the *B. plebeja* specific duplicate (Fig. 5, groups 2, 4 and 5) showed a similar expression pattern, possibly reflecting an ancestral expression pattern.

Full data from a genome assembly is needed to resolve the likeliest evolutionary scenario, however the data presented here supports the idea that a gene duplication can lead to release of constraint on one of the duplicated copies and generation of novel gene expression patterns whilst ancestral function is retained by the other copy⁵⁰.

Other species studied have CHS gene families comprising very old duplicates^{51 52 53}, whilst other taxa have only retained CHS copies from more recent, clade-specific duplications⁵⁴. Some taxa have almost no evidence of CHS duplication such as the Brassicaceae⁵⁵, suggesting lineage specific patterns in the CHS gene family across plants.

Lineage specific expansions of genes are important mechanisms for adaptive responses to environmental stimuli⁵⁶, and identification of selectively retained duplicate genes can reveal functional biases of ecological importance⁵⁷.

Conclusions

Begonia is a mega-diverse genus, with excellent applicability to research in conservation and plant breeding. We have produced a multi-tissue RNA-seq dataset in two closely related but morphologically and ecologically divergent species of *Begonia*, providing a valuable addition to the growing base of genomic resources in the genus. Recent duplications in an important flavonoid biosynthetic gene, chalcone synthase, have led to diversification of expression of duplicate copies, suggesting duplication patterns in this gene family are dynamic and prone to high turnover rates.

Further transcriptome wide investigations using the RNA-seq dataset generated here may help uncover species specific changes in expression profiles that have contributed to the ecological divergence of *B. conchifolia* and *B. plebeja* and answer wider questions about the role gene duplication plays in the generation of diversity in *Begonia*.

Declarations

Author contributions

KE performed the experiments, KE analysed the data, CK, AMM and LCD helped in interpretation of results and editing of the manuscript, and CK supervised the project.

Corresponding author

Correspondence to Katie Emelianova at katie.emelianova@ed.ac.uk

Additional Information

Competing interests

The authors declare no competing interests.

Statement on plant guidelines

Collection of plant material complies with relevant institutional, national, and/or international guidelines and legislation.

References

1. Hughes, N. M. *et al.* Begonia Resource Centre, Online database available from <http://padme.rbge.org.uk/begonia/>. (2015).
2. Goodall-Copestake, W. P., Harris, D. J. & Hollingsworth, P. M. The origin of a mega-diverse genus: dating Begonia (Begoniaceae) using alternative datasets, calibrations and relaxed clock methods. *Bot. J. Linn. Soc.* **159**, 363–380 (2009).
3. Dewitte, A., Twyford, A. D., Thomas, D. C., Kidner, C. A. & Huylensbroeck, J. V. The Origin of Diversity in Begonia: Genome Dynamism, Population Processes and Phylogenetic Patterns. *Dyn. Process. Biodivers. - Case Stud. Evol. Spat. Distrib.* <https://doi.org/10.5772/23789> (2011).
4. Hughes, M. & Hollingsworth, P. M. Population genetic divergence corresponds with species-level biodiversity patterns in the large genus Begonia. *Mol. Ecol.* **17**, 2643–2651 (2008).
5. Kiew, R. The Limestone Begonias of Sabah, Borneo - Flagship Species for Conservation. *Gard. Bull. Singap.* **53**, 241–286 (2001).
6. McLellan, T. Geographic variation and plasticity of leaf shape and size in Begonia dregei and B. homonyma (Begoniaceae). *Bot. J. Linn. Soc.* **132**, 79–95 (2000).
7. Dewitte, A. D. *et al.* Genome size variation in Begonia. *Genome*. <https://doi.org/10.1139/G09-056> (2009).
8. Brennan, A. C. *et al.* Genomic Resources for Evolutionary Studies in the Large, Diverse, Tropical Genus, Begonia. *Trop. Plant Biol.* **5**, 261–276 (2012).
9. Zhang, L. *et al.* The ancient wave of polyploidization events in flowering plants and their facilitated adaptation to environmental stress. *Plant Cell Environ.* **43**, 2847–2856 (2020).
10. Burt-Utley, K. A revision of Central American species of Begonia section Gireoudia (Begoniaceae). *Tulane Stud. Zool. Bot.* **25**, 1–131 (1985).
11. Li, W. H., Yang, J. & Gu, X. Expression divergence between duplicate genes. *Trends Genet.* **21**, 602–607 (2005).
12. Krause, G. H. *et al.* Photosynthesis, photoprotection, and growth of shade-tolerant tropical tree seedlings under full sunlight. *Photosynth. Res.* **113**, 273–285 (2012).

13. Zhang, Q., Liu, M. & Ruan, J. Metabolomics analysis reveals the metabolic and functional roles of flavonoids in light-sensitive tea leaves. *BMC Plant Biol.* **17**, 64 (2017).
14. Chazdon, R. L. & Pearcy, R. W. The Importance of Sunflecks for Forest Understory Plants. *BioScience*. **41**, 760–766 (1991).
15. Hughes, N. M. *et al.* Photosynthetic costs and benefits of abaxial versus adaxial anthocyanins in *Colocasia esculenta* 'Mojito'. *Planta*. **240**, 971–981 (2014).
16. Zhang, T. J. *et al.* A magic red coat on the surface of young leaves: anthocyanins distributed in trichome layer protect *Castanopsis fissa* leaves from photoinhibition. *Tree Physiol.* **36**, 1296–1306 (2016).
17. Ferrer, J. L., Jez, J. M., Bowman, M. E., Dixon, R. A. & Noel, J. P. Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis. *Nat. Struct. Biol.* **6**, 775–784 (1999).
18. Fofana, B. *et al.* Milsana-induced resistance in powdery mildew-infected cucumber plants correlates with the induction of chalcone synthase and chalcone isomerase. *Physiol. Mol. Plant Pathol.* **61**, 121–132 (2002).
19. Banik, N. & Bhattacharjee, S. Complementation of ROS scavenging secondary metabolites with enzymatic antioxidant defense system augments redox-regulation property under salinity stress in rice. *Physiol. Mol. Biol. Plants Int. J. Funct. Plant Biol.* **26**, 1623–1633 (2020).
20. Karageorgou, P., Buschmann, C. & Manetas, Y. Red leaf color as a warning signal against insect herbivory: Honest or mimetic? *Flora - Morphol. Distrib. Funct. Ecol. Plants*. **203**, 648–652 (2008).
21. Durbin, M. L., Learn, G. H., Huttley, G. A. & Clegg, M. T. Evolution of the chalcone synthase gene family in the genus *Ipomoea*. *Proc. Natl. Acad. Sci.* **92**, 3338–3342 (1995).
22. Durbin, M. L., McCaig, B. & Clegg, M. T. Molecular evolution of the chalcone synthase multigene family in the morning glory genome. *Plant Mol. Biol.* **42**, 79–92 (2000).
23. Logemann, J., Schell, J. & Willmitzer, L. Improved method for the isolation of RNA from plant tissues. *Anal. Biochem.* **163**, 16–20 (1987).
24. Laetsch, D. R., Blaxter, M. L. & BlobTools Interrogation of genome assemblies. *F1000Research*. **6**, 1287 (2017).
25. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114–2120 (2014).
26. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
27. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).
28. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.* **26**, 1134–1144 (2016).
29. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. **31**, 3210–3212 (2015).
30. Bryant, D. M. *et al.* A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Rep.* **18**, 762–776 (2017).
31. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. **30**, 923–930 (2014).
32. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. **26**, 139–140 (2010).
33. Martínez, A. M. A draft genome assembly for *Hillebrandia sandwicensis*. (2017).
34. Kears, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. **28**, 1647–1649 (2012).
35. Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & McInerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* **6**, 29 (2006).

36. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. **30**, 1312–1313 (2014).
37. Huang, S. *et al.* The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **41**, 1275–1281 (2009).
38. Xie, J. *et al.* Evolutionary Origins of Pseudogenes and Their Association with Regulatory Sequences in Plants. *Plant Cell*. **31**, 563–578 (2019).
39. Paterson, A. H. *et al.* Comparative genome analysis of monocots and dicots, toward characterization of angiosperm diversity. *Curr. Opin. Biotechnol.* **15**, 120–125 (2004).
40. Chen, F. *et al.* Genome sequences of horticultural plants: past, present, and future. *Hortic. Res.* **6**, 1–23 (2019).
41. Onstein, R. E. Darwin's second 'abominable mystery': trait flexibility as the innovation leading to angiosperm diversity. *New Phytol.* **228**, 1741–1747 (2020).
42. Igloi, G. L. Evidence for fungal sequence contamination in plant transcriptome databases. *Plant Syst. Evol.* **305**, 563–568 (2019).
43. Zhu, J., Wang, G. & Pelosi, P. Plant transcriptomes reveal hidden guests. *Biochem. Biophys. Res. Commun.* **474**, 497–502 (2016).
44. Simion, P. *et al.* A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. *BMC Biol.* **16**, 28 (2018).
45. Chang, Z., Wang, Z. & Li, G. The Impacts of Read Length and Transcriptome Complexity for De Novo Assembly: A Simulation Study. *PLOS ONE*. **9**, e94825 (2014).
46. Sundell, D. *et al.* High-Spatial-Resolution Transcriptome Profiles Reveal Uncharacterized Modularity of Wood Formation in *Populus tremula*. *Plant Cell*. **29**, 1585–1604 (2017). AspWood
47. Takahagi, K., Inoue, K. & Mochida, K. Gene Co-expression Network Analysis Suggests the Existence of Transcriptional Modules Containing a High Proportion of Transcriptionally Differentiated Homoeologs in Hexaploid Wheat. *Front. Plant Sci.* **9**, (2018).
48. Copley, S. D. Evolution of new enzymes by gene duplication and divergence. *FEBS J.* **287**, 1262–1283 (2020).
49. Shi, T. *et al.* Distinct Expression and Methylation Patterns for Genes with Different Fates following a Single Whole-Genome Duplication in Flowering Plants. *Mol. Biol. Evol.* **37**, 2394–2413 (2020).
50. Des Marais, D. L. & Rausher, M. D. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*. **454**, 762–765 (2008).
51. Xu, Z. *et al.* Comparative Genome Analysis of *Scutellaria baicalensis* and *Scutellaria barbata* Reveals the Evolution of Active Flavonoid Biosynthesis. *Genomics Proteomics Bioinformatics*. **18**, 230–240 (2020).
52. Pothiraj, R., Ravikumar, M. J., Suthanthiram, B., Subbaraya, U. & Krishnamurthy, P. Genome-scale analyses of polyketide synthases in banana: Phylogenetics and expression profiling forecast their candidacy in specialized metabolism. *Gene*. **778**, 145472 (2021).
53. Koduri, P. K. H. *et al.* Genome-wide analysis of the chalcone synthase superfamily genes of *Physcomitrella patens*. *Plant Mol. Biol.* **72**, 247–263 (2010).
54. Kuo, Y. T., Chao, Y. T., Chen, W. C., Shih, M. C. & Chang, S. B. Segmental and tandem chromosome duplications led to divergent evolution of the chalcone synthase gene family in *Phalaenopsis* orchids. *Ann. Bot.* **123**, 69–77 (2019).
55. Koch, M. A., Haubold, B. & Mitchell-Olds, T. Comparative Evolutionary Analysis of Chalcone Synthase and Alcohol Dehydrogenase Loci in *Arabidopsis*, *Arabis*, and Related Genera (Brassicaceae). *Mol. Biol. Evol.* **17**, 1483–1498 (2000).
56. Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K. & Shiu, S. H. Importance of Lineage-Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental Stimuli. *Plant Physiol.* **148**, 993 (2008).
57. Rizzon, C., Ponger, L. & Gaut, B. S. Striking Similarities in the Genomic Distribution of Tandemly Arrayed Genes in *Arabidopsis* and Rice. *PLOS Comput. Biol.* **2**, e115 (2006).

Figures



Figure 1
 Photographs of study species used. *B. conchifolia* (A) and *B. plebeja* (B).

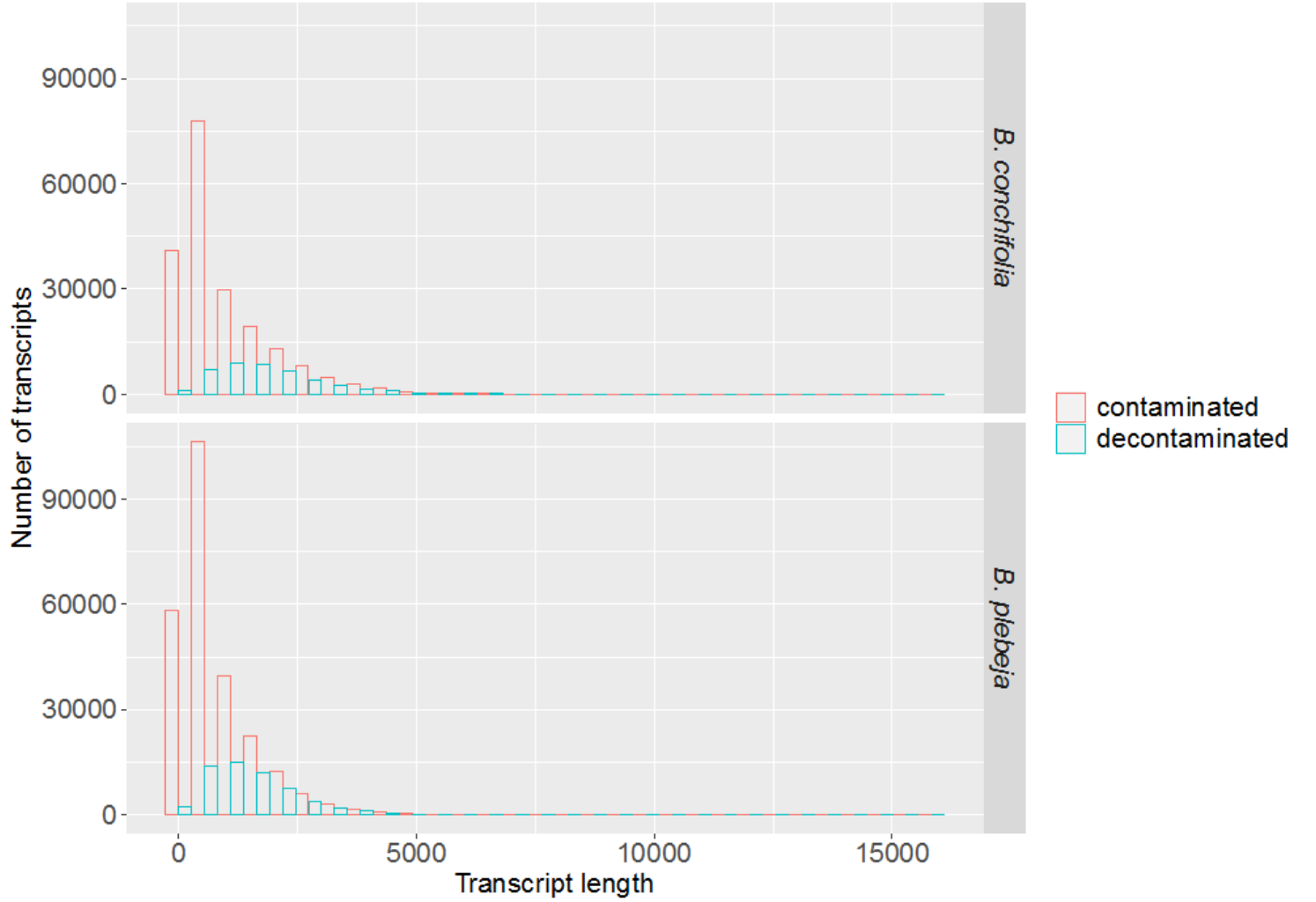


Figure 2
 sequence length histograms showing *B. conchifolia* and *B. plebeja* sequence length distributions in decontaminated and contaminated assemblies

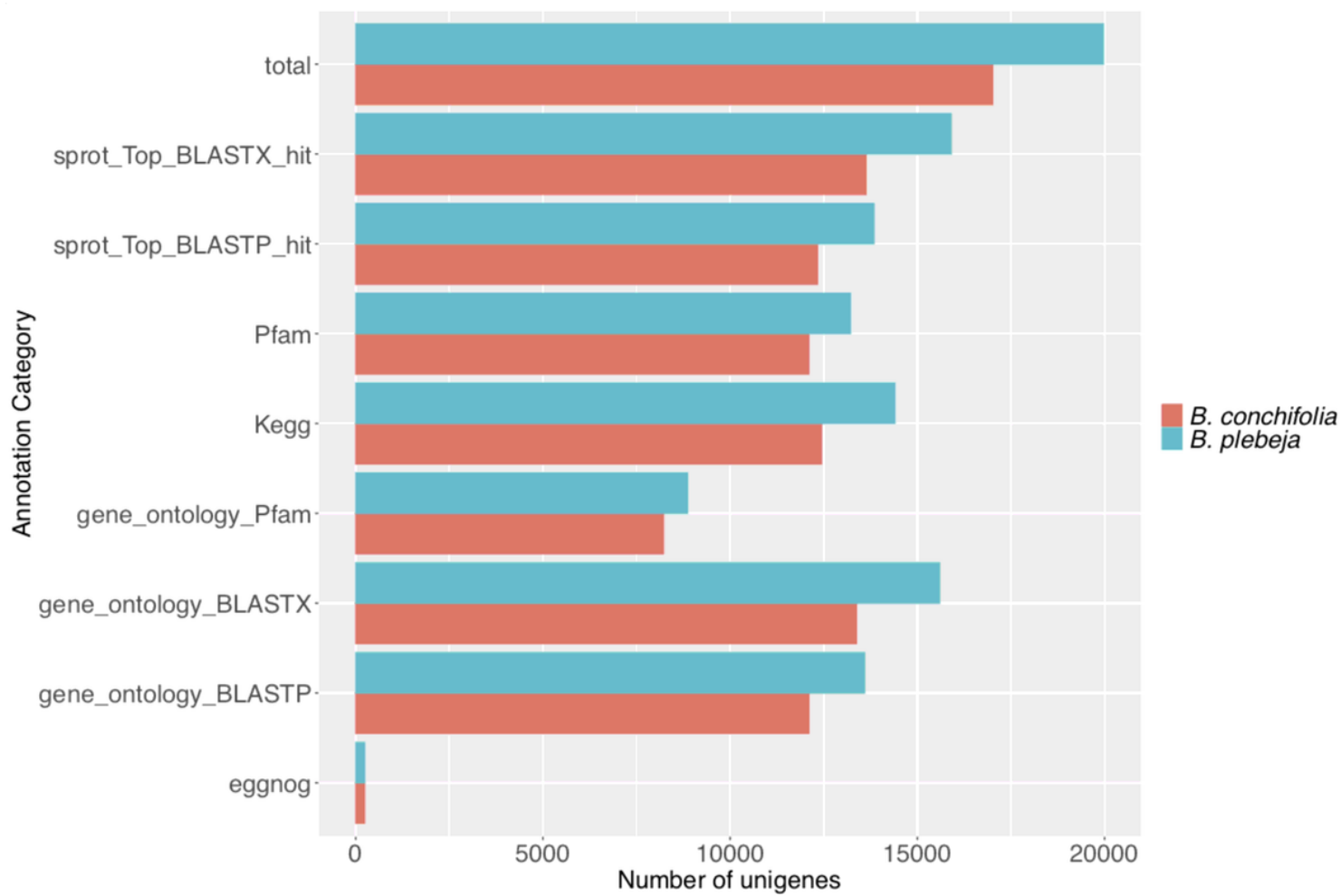


Figure 3

Number of unigenes annotated using different annotation categories in *B. conchifolia* and *B. plebeja*

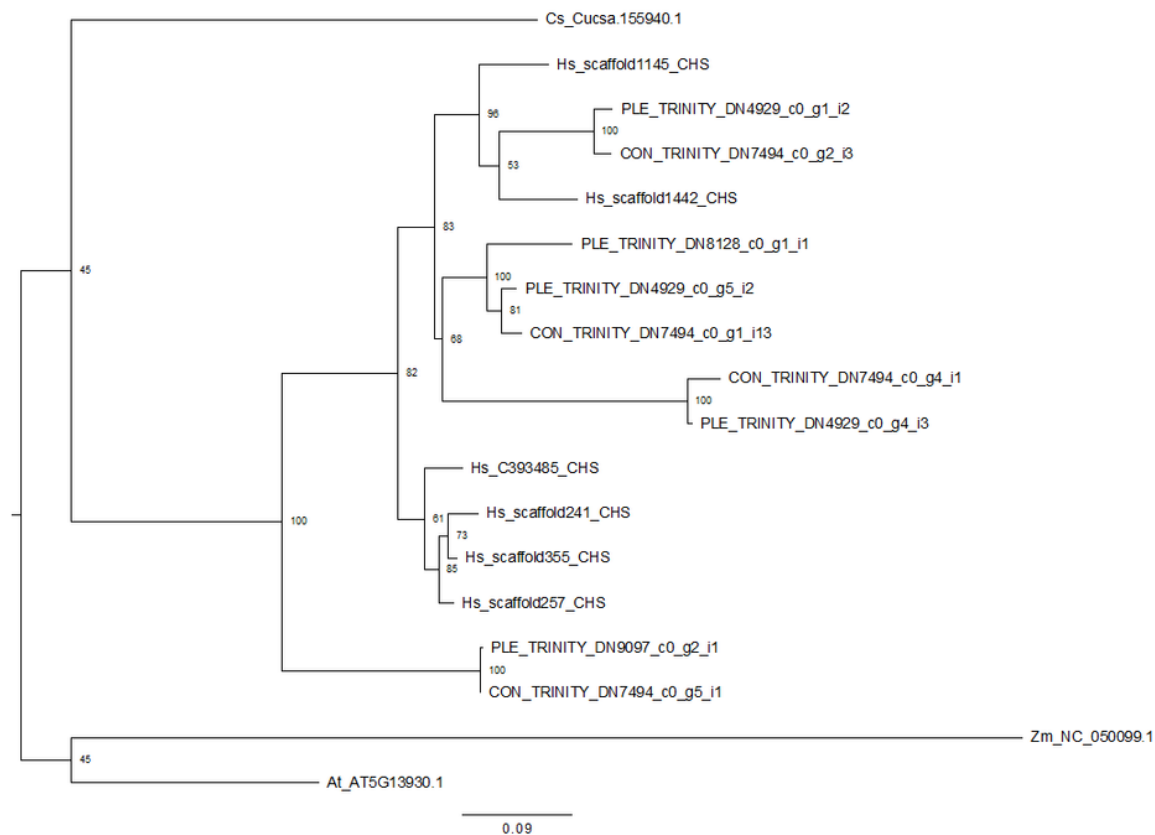


Figure 4

ML tree of CHS sequences from *B. conchifolia* (CON), *B. plebeja* (PLE), *H. sandwicensis* (Hs), *C. sativus* (Cs), *A. thaliana* (At) and *Z. mays* (Zm). Pairs of orthologs are highlighted as follows: blue: group 1, pink; group 2, yellow: group 3, green: group 4, red: group 5.

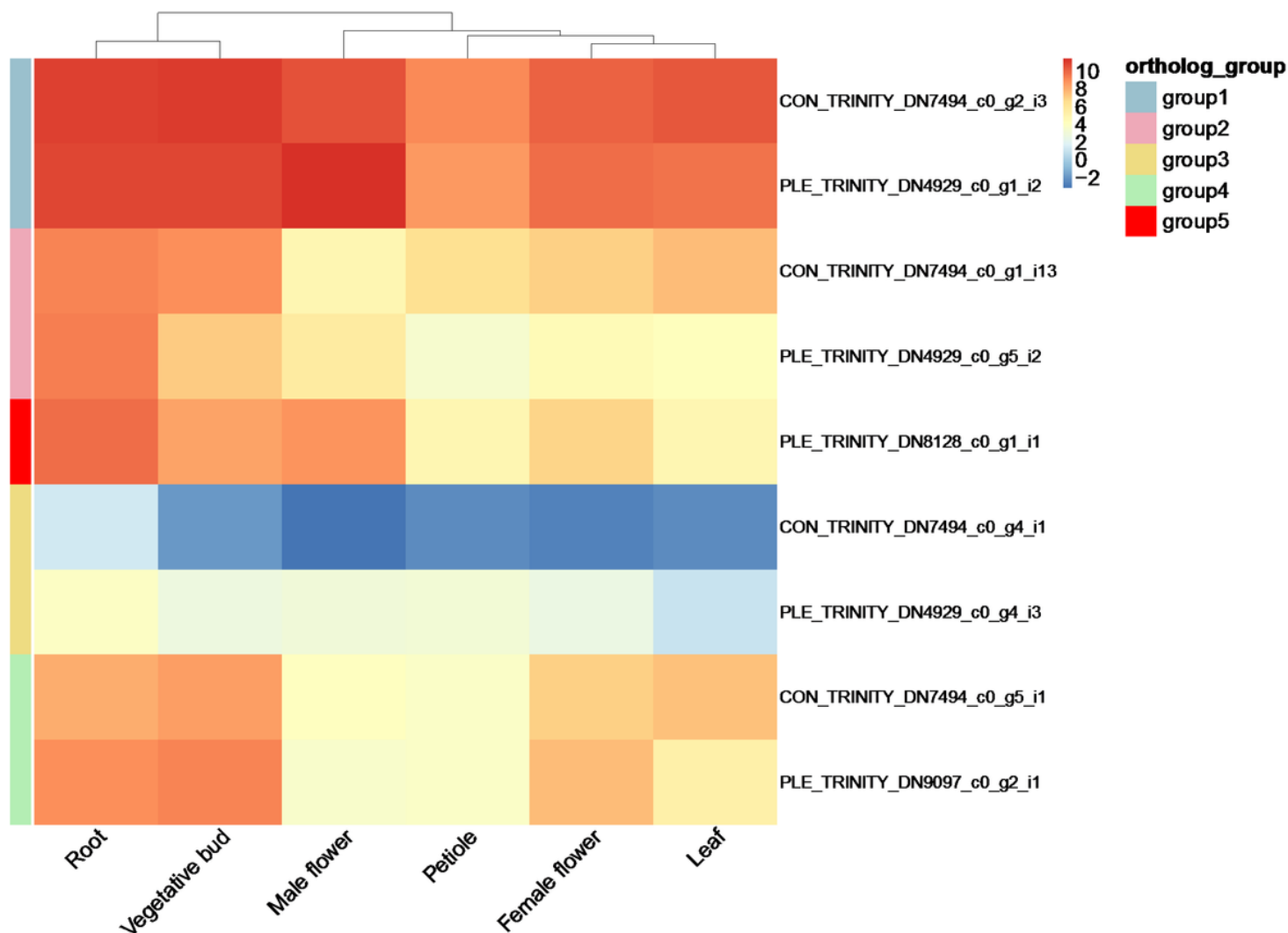


Figure 5

Heatmap of log2 FPKM of CHS homologs in *B. conchifolia* and *B. plebeja*. Ortholog groups refer to highlighted homologous CHS pairs in the ML gene tree.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementaryfigures.pdf](#)
- [supplementarytable1readcounts.docx](#)
- [supplementarytable2blastresults.docx](#)
- [supplementaryfile1conplehilATZMCUalignmentlongestonlyextraction16102901manuallycorrectedwithaddedplebejaseq.fasta](#)
- [supplementaryfile2RAxMLbipartitions.t15](#)
- [supplementaryfile3contrinotateannotationreport.xls](#)
- [supplementaryfile4pletrinotateannotationreport.xls](#)